

University of Southampton Research Repository

Copyright © and Moral Rights for this thesis and, where applicable, any accompanying data are retained by the author and/or other copyright owners. A copy can be downloaded for personal non-commercial research or study, without prior permission or charge. This thesis and the accompanying data cannot be reproduced or quoted extensively from without first obtaining permission in writing from the copyright holder/s. The content of the thesis and accompanying research data (where applicable) must not be changed in any way or sold commercially in any format or medium without the formal permission of the copyright holder/s.

When referring to this thesis and any accompanying data, full bibliographic details must be given, e.g.

Thesis: Weiqi Liao (2020) "The use of sequence analysis to study primary care pathways: an exploratory study of people at high risk of lung cancer in England", University of Southampton, School of Health Sciences, PhD Thesis, pagination.

University of Southampton

Faculty of Environmental and Life Sciences

School of Health Sciences

The use of sequence analysis to study primary care pathways: an exploratory study of people at high risk of lung cancer in England

DOI: [enter DOI] hyperlink, optional

by

廖炜圻 Weiqi Liao BMgt, MMed, MRes

ORCID ID: 0000-0002-8605-3749

Thesis for the degree of Doctor of Philosophy in Health Sciences

November 2020

University of Southampton

Abstract

Faculty of Environmental and Life Sciences

School of Health Sciences

Thesis for the degree of Doctor of Philosophy

The use of sequence analysis to study primary care pathways: an exploratory study of people at high risk of lung cancer in England

by

廖炜圻 Weiqi Liao

Research background, gaps, and aim: Lung cancer (LC) is a research priority in the UK, due to its high incidence and mortality, and poor survival. Current population-based studies are focused on investigating ‘route to diagnosis’, factors associated with late diagnosis and poor survival, and the implications of different intervals (e.g. primary care interval, diagnostic interval, treatment interval) in the cancer care pathway. However, the longitudinal sequence of interdependent patient-GP events over time (patient’s help-seeking behaviours and general practitioner (GP) management) preceding cancer diagnosis is a research area less investigated. Therefore, this PhD study proposes a new perspective of studying primary care sequences for early diagnosis research, using a novel statistical method – sequence analysis (SA), to identify meaningful typologies in a less investigated population – patients at high risk but not yet diagnosed with LC.

Methodology: A systematic scoping review was conducted to understand how SA has been applied to study disease trajectories and care pathways in health services research, to learn the lessons from published studies and inform the application of SA in the main study.

- **Study design, setting, and participants:** 899 community patients at high risk of developing LC (based on patient's smoking history) but not yet diagnosed with LC from eight general practices in the South coast of England consented to participate in this study. Their primary care records from June 2010 to October 2012 (29 months) were reviewed. Information was extracted from GP notes in free text and transcribed manually.
- **Research process:** Two study phases, methodological exploration and empirical analysis, were involved to address three research objectives: how to construct primary care sequences from discrete health events, how to use different features of SA to obtain meaningful cluster patterns (the outcome of SA), and how patients’ sociodemographic and clinical

characteristics can help explain the variation in the cluster patterns and help-seeking behaviours.

- **The primary outcome, covariates, and statistical methods:** SA and cluster analysis were used to obtain the primary outcome – typology of clusters. Descriptive statistics were used to characterise patient profiles for each cluster, followed by traditional statistical tests (ANOVA, chi-square test, Kruskal-Wallis test) to compare patients' sociodemographic and clinical features among clusters. Generalised linear models were used to explore the association between patient characteristics and clusters, and to quantify the relative risk ratios.

Key findings: The study sample was classified into seven clusters. The subgroups of patients who presented with potential LC symptoms were categorised into four clusters with different GP management: GP ordered tests or prescribed medications for transient symptoms (Cluster 1, n=133/899, 14.8%); GP ordered chest X-ray or referred patients to specialists (Cluster 2, n=65, 7.2%); GP offered health advice to patients (Cluster 3, n=60, 6.7%); and patients presented symptoms multiple times and received repeated prescriptions from GP (Cluster 4, n=37, 4.1%). Cluster 5 was patients without potential LC symptoms but presented in general practice with cardiorespiratory comorbidities and/or other alarm symptoms (n=326, 36.3%). Patients in Cluster 6 only had minor care needs (n=237, 26.4%). Patients in Cluster 7 did not visit GP at all during the whole study period (n=41, 4.6%). For patients who had ≥ 2 visits with potential LC symptoms, the median interval was 61.5 days, interquartile range [16, 217] days. Age and the number of comorbidities were the two significant variables in different models. Variation of patient management among practices (practice effect) was observed.

Conclusions, clinical relevance, and implications: This PhD thesis has made an original contribution by establishing the feasibility and analytical framework of using SA to study complex primary care sequences in the field of early diagnosis research. This study demonstrates the potential of applying SA in a larger scale study with a more representative population, to investigate the complex and heterogeneous primary care sequences leading to LC diagnosis, which has clinical implications for patient care and management, promote early cancer diagnosis from primary care, and eventually improve LC survival in the UK.

Table of Contents

Table of Contents	i
List of tables	xiii
List of figures	xv
Research Thesis: Declaration of Authorship	xix
Acknowledgements	xxi
Definitions and Abbreviations	xxiii
Chapter 1 Introduction	1
1.1 Research context and the motivation to conduct this PhD study	1
1.1.1 Synopsis of this thesis	1
1.1.2 Why lung cancer	1
1.1.3 Research gap and a new research angle	2
1.1.4 A novel statistical method for the new research angle	2
1.2 Research aims and the key research questions (RQs)	3
1.3 The journey of the PhD study and this thesis	5
1.4 The structure of this thesis.....	6
Chapter 2 Literature review	9
2.1 Chapter introduction	9
2.1.1 Search strategy and database	9
2.2 LC survival in England	10
2.2.1 LC survival in England	10
2.2.2 English LC survival in the European context	11
2.2.3 English LC survival in the international context	12
2.2.4 Section summary and discussion: early diagnosis is the key to improve LC survival in England.....	13
2.3 Policies and initiatives to promote early LC diagnosis in England	14
2.3.1 National initiatives and campaigns for LC	15
2.3.2 LC screening in England.....	16
2.3.3 NICE recommended LC care pathways, diagnosis and staging procedures	17

Table of Contents

2.3.4	Waiting time targets in the cancer care pathway in England.....	19
2.3.5	Section summary	20
2.4	Current research models and paradigms of early diagnosis research	20
2.4.1	The conceptual model of the cancer care pathway	20
2.4.1.1	The Aarhus statement.....	20
2.4.1.2	Intervals and delays relevant to primary care investigation	22
2.4.1.3	Current research evidence between diagnostic interval and outcomes (stage at diagnosis and survival)	23
2.4.1.4	The limitation of the conceptual model of the cancer care pathway ..	24
2.4.2	The route to diagnosis of LC	24
2.4.2.1	Two-week Wait	24
2.4.2.2	Emergency presentation.....	25
2.4.2.3	The limitation of ‘route to diagnosis’	26
2.4.3	Proposing a new research paradigm – primary care sequence/pathway.....	26
2.5	Lung cancer diagnosis from primary care in England	28
2.5.1	The role of primary care in England, and the opportunities and challenges of early diagnosis of LC from primary care	28
2.5.2	The difficulty of early LC diagnosis – lack of specific symptoms	29
2.5.3	Comorbidities and LC.....	31
2.5.3.1	The rationale on how comorbidities could influence cancer diagnosis and survival	31
2.5.3.2	COPD and LC.....	32
2.5.4	GP’s perspectives on early diagnosis of LC.....	34
2.5.5	Patients’ help-seeking behaviours and influencing factors.....	35
2.5.6	Patient’s SES and LC	38
2.5.6.1	The rationale on how patients’ SES could influence cancer diagnosis, treatment, and survival.....	38
2.5.6.2	How SES is operationalised in English health research	39
2.5.6.3	SES and LC diagnosis and intervals	39
2.5.6.4	SES and LC treatment.....	40

2.5.6.5	SES and LC survival	41
2.5.7	Missed opportunities for early LC diagnosis and the lessons learnt from audit.....	42
2.5.8	An empirical study investigating LC diagnostic pathways from primary care .	42
2.5.9	Section summary and the rationale of this PhD: why studying primary care sequences/pathways for LC is needed in England at this moment	43
2.6	Chapter summary.....	44
Chapter 3	Sequence analysis (SA).....	45
3.1	An overview and the basic concepts of SA	45
3.1.1	The evolution and application of SA in different disciplines	45
3.1.2	Basic concepts of SA.....	45
3.1.3	An overview of the analytical process of SA and relevant terminology	47
3.2	Measuring (dis)similarity between sequences.....	48
3.2.1	The importance of measuring similarity between sequences.....	48
3.2.2	The commonly used distance measure – Optimal Matching (OM).....	49
3.2.3	The other commonly used dissimilarity measures	50
3.2.4	Cost setting.....	50
3.2.5	Typical sequences.....	51
3.2.6	Event sequence analysis (ESA)	53
3.3	Grouping sequences – cluster analysis	54
3.3.1	Introduction.....	54
3.3.2	How the agglomerative hierarchical clustering algorithm (Ward’s method) works	54
3.3.3	Assessing the clustering quality	55
3.3.4	Heterogeneity and outliers	56
3.3.5	Deciding the optimal number of clusters.....	56
3.4	Multiple interdependent dimensions in sequences	57
3.4.1	Examples of multiple interdependent dimensions.....	57
3.4.2	Possible approaches to analyse multiple interdependent dimensions	57
3.4.3	Multi-Channel Sequence Analysis (MCSA).....	58

Table of Contents

3.4.4	An exploration of the suitability of MCSA and SA in this PhD study	58
3.4.5	Brief discussion: the two approaches to analyse interdependent patient and GP events and the cluster patterns	60
3.5	Mainstream statistical software and packages to conduct SA.....	61
3.6	Chapter discussion: the technical uncertainties of SA	62
3.6.1	The nature of SA and the technical uncertainties	62
3.6.2	Appropriate dissimilarity measure and cost setting.....	62
3.6.3	The optimal number of clusters and the validity of results.....	63
3.6.4	The order, timing, and complex interdependencies of states in sequences over time.....	63
3.7	Chapter summary: how SA fits in this PhD study	64
Chapter 4	The application of sequence analysis in health services research: a systematic scoping review.....	65
4.1	Introduction	65
4.1.1	The objectives of this chapter.....	65
4.1.2	Position of this review	65
4.1.3	Formulate the search question for the review	67
4.2	Methods.....	67
4.2.1	Search strategy and databases	67
4.2.2	Inclusion and exclusion criteria	68
4.2.3	Selection of critical appraisal tool and critique of individual paper	70
4.2.4	Structured data extraction from the included papers.....	72
4.3	Results and interpretations	73
4.3.1	Results of critical appraisal of the included papers.....	73
4.3.2	Characteristics of the included studies, research design and context, data source and sample size	73
4.3.3	The added value of applying SA in respective studies and some comments for further improvement.....	74
4.4	Summary and discussion of the methodological/statistical decisions related to SA	78

4.4.1	The interval and sequence length	78
4.4.2	Dissimilarity measures and cost setting	79
4.4.3	Sequence clustering/partition	79
4.4.4	Approaches to identify significant covariates among the clusters	80
4.4.5	Dealing with missing data	81
4.4.6	A final brief note	82
4.4.7	Strengths and limitations of this review	82
4.5	Conclusion of the review and the relevance to this PhD study	93
4.5.1	Conclusion: the current application of SA in health services research	93
4.5.2	Research gaps and the opportunity for this PhD study	93
4.5.3	How the findings of this review can be used to inform the decision and application of SA in this PhD study	94
Chapter 5	Data source and methodology	95
5.1	Introduction of the original study and research ethics	95
5.1.1	Ethical approval for this study	95
5.1.2	Original and independent work by using the NAEDI data for this thesis	95
5.2	Available data and variables	96
5.2.1	Participants, observation period, and timeline	96
5.2.2	Primary care events	96
5.2.3	Coding framework for the primary care events	97
5.2.4	Available information and variables for patient characteristics	101
5.3	Statistical analysis plan and methods	101
5.3.1	Patient characteristics and pairwise correlation between variables	102
5.3.2	Comparison of patient characteristics in different clusters after SA	102
5.3.3	Patient characteristics associated with primary care attendance and potential LC symptom consultations	103
5.3.3.1	Modelling count data	103
5.3.3.2	Modelling count data with excessive zero	104
5.3.3.3	Patient characteristics as independent variables to model the count data	104

Table of Contents

5.3.4	Exploration of practice effect	104
5.3.5	Statistical software and packages.....	105
5.4	Discussion of the study data	105
5.4.1	The study sample size	105
5.4.2	Sample representativeness.....	105
5.4.3	Strengths and limitations of the study sample.....	106
5.4.4	Strengths and limitations of the primary care data	107
5.4.5	Assessment of the data quality.....	108
5.4.6	Potential missing data in health records	109
5.4.7	Assumptions and handling missing data	110
5.5	Chapter summary	110
Chapter 6	Methodological exploration on the use of sequence analysis to identify typologies of primary care sequences among community-based patients at high risk of developing lung cancer.....	111
6.1	Introduction: research objectives of methodological exploration.....	111
6.2	Addressing the methodological issues	111
6.2.1	State specification for SA (1 st methodological issue)	111
6.2.1.1	The rationale of state specification	111
6.2.1.2	The states for patient's reason to visit the general practice	112
6.2.1.3	The states for GP actions	113
6.2.1.4	The frequency of cross-tabulating patient and GP states	113
6.2.1.5	The combined states.....	115
6.2.1.6	A brief summary of the whole process of state specification	119
6.2.2	Constructing primary care sequences (1 st methodological issue).....	119
6.2.2.1	The first approach: sequences constructed by visits.....	119
6.2.2.2	The second approach: sequences constructed in a timeline with equal intervals.....	120
6.2.2.3	Brief discussion: applications and implications of two sequence construction approaches	121
6.2.3	Dissimilarity measure and cost setting (2 nd methodological issue)	122

6.2.4	Criteria to determine the optimal number of clusters (3 rd methodological issue).....	122
6.3	Presentation and interpretations of the whole primary care sequences from all study subjects.....	123
6.3.1	Sequence profile 1: state distribution of primary care sequences constructed by visits.....	123
6.3.2	Sequence profile 2: state distribution of primary care sequences constructed in a timeline.....	125
6.3.3	Brief summary and discussion of the whole primary care sequences.....	126
6.4	Subgroup analysis 1 – sequences of patients presented with potential LC symptoms.....	127
6.4.1	Introduction and rationale of the subgroup analysis.....	127
6.4.2	Sequence profile 3: LC symptom sequences in subgroup analysis.....	128
6.4.3	Dendrograms and regression trees for LC symptom sequences.....	130
6.4.3.1	Substitution cost matrix.....	130
6.4.3.2	Dendrograms – the clustering structure of sequences.....	131
6.4.3.3	Regression tree – how patterns in clusters changed by splitting the nodes in the dendrogram.....	131
6.4.3.4	Deciding the optimal number of clusters for the typologies.....	132
6.4.4	Comparison and interpretation of the cluster patterns in the typologies of LC symptom sequences.....	134
6.4.4.1	Interpretation of the cluster patterns of $OM_{[1,2]}$ in the LC symptom sequences.....	135
6.4.4.2	Further explanations of the mechanism of grouping sequences.....	135
6.4.5	The most frequent LC symptom sequences in each cluster of the typology by $OM_{[1,2]}$	136
6.4.6	Event SA for LC symptom sequences in each cluster of the typology by $OM_{[1,2]}$	137
6.4.6.1	Explanations and examples of the results from event sequence analysis.....	137

Table of Contents

6.4.6.2	Similarities and differences between the most frequent sequences and event sequence analysis	139
6.4.7	Brief summary and learning points: the analytical process of LC symptom sequences	139
6.5	Subgroup analysis 2 – potential LC symptoms situated in smoking-related comorbidities and other alarm symptoms	140
6.5.1	Sequence profile 4: high-risk sequences	140
6.5.1.1	The whole high-risk sequences	140
6.5.1.2	Introducing a common reference point in the sequences – the first observed presentation of potential LC symptoms	141
6.5.2	Dendrograms and regression trees for the high-risk sequences.....	142
6.5.2.1	Dendrograms and the outlier sequence	142
6.5.2.2	Regression tree and deciding the number of clusters	145
6.5.3	Comparison and interpretation of the cluster pattern in the typologies of high-risk sequences.....	145
6.5.4	Representative high-risk sequences in each cluster of the typology by $OM_{[1,2]}$	147
6.5.5	Event SA of high-risk sequences in each cluster by $OM_{[1,2]}$	148
6.5.6	The left-truncated part – events happened before the first observed presentation of potential LC symptoms	151
6.5.7	Comparison of the typologies between the two sets of subgroup analysis .	152
6.5.8	A brief summary and learning points: subgroup analysis of high-risk sequences	154
6.6	High-risk primary care consultations among patients without potential LC symptoms during the observation period	155
6.7	General discussion of the methodological exploration and the empirical findings of the primary care sequences in the NAEDI study	156
6.7.1	Summary of the methodological exploration in this chapter.....	156
6.7.2	Simplification of the primary care events and categorisation into states	156
6.7.3	Dissimilarity measures, cost settings, and the implications.....	157

6.7.4	Determine the number of clusters and the implications for further statistical analysis	159
6.7.5	The empirical findings and the implications	160
6.7.6	Relevance and implications in the broader field of early diagnosis research	161
6.7.7	Strengths and limitations of the methodological exploration	161
6.8	Conclusion of the methodological exploration phase	162
Chapter 7 Results of the empirical analysis.....		163
7.1	Introduction.....	163
7.2	Results and interpretations.....	163
7.2.1	Sociodemographic and clinical characteristics of the study sample	163
7.2.2	Results of the association between sociodemographic and health variables	166
7.2.3	Patient characteristics among the seven clusters.....	166
7.2.3.1	Patient profile and statistical tests among the seven clusters	166
7.2.4	Results of multinomial logistic regression model	173
7.2.4.1	The reference outcome cluster and the independent variables	173
7.2.4.2	Significant results in the univariable multinomial logistic regression model and the interpretations.....	173
7.2.4.3	Differentiate clusters by significant patient characteristics	176
7.2.4.4	Multinomial logistic regression model with multiple predictors.....	176
7.2.5	Results of modelling count data.....	178
7.2.5.1	Model of primary care attendance and the interpretation of results	178
7.2.5.2	Model of potential LC symptoms consultations and the interpretation of results.....	180
7.2.6	Practice effect.....	181
7.2.6.1	Patient characteristics among practices	181
7.2.6.2	Practice effect in patient's help-seeking behaviours	183
7.2.6.3	Cluster membership among eight practices	184
7.3	Discussion.....	186
7.3.1	Statistical methods in the empirical analysis.....	186

Table of Contents

7.3.1.1	Summary of the whole analysis process.....	186
7.3.1.2	Different statistical techniques provide different angles to find useful information from the data	186
7.3.2	Summary and discussion of the findings from empirical analysis.....	187
7.3.3	Boarder connection with other literature	189
7.3.3.1	Symptom recognition and the help-seeking behaviours.....	189
7.3.3.2	Patient’s characteristics and the patterns of primary care consultations.....	190
7.3.4	Clinical relevance and implications.....	191
7.3.4.1	Potential opportunity to encourage patients at high risk to seek help more promptly in primary care.....	191
7.3.4.2	New campaign to improve smoker’s lung health.....	191
7.3.4.3	The practice effect and clinical implications.....	192
7.3.5	Strengths and limitations of the empirical analysis.....	193
7.4	Conclusion of the empirical analysis phase	194
Chapter 8	General discussion and conclusion	195
8.1	The central argument of this thesis.....	195
8.2	Key messages and findings from this PhD study	195
8.2.1	The research gap this study addresses	195
8.2.2	Characteristics of patients at high-risk and their help-seeking behaviours ..	195
8.2.3	Cluster patterns of primary care sequences in the NAEDI study	196
8.2.4	The association between patient characteristics and cluster patterns.....	196
8.2.5	The gap between patients experiencing symptoms and help-seeking behaviours	197
8.2.6	Practice effect	197
8.3	Reflection and further discussion	198
8.3.1	The face validity of the study findings	198
8.3.2	Reflection on the strengths and limitations of the whole study.....	199
8.3.3	Rethinking the conceptual models of the cancer care pathway	199
8.4	Original contributions of this PhD study.....	200

8.5	Recommendations for other researchers on how to use SA in health research...	201
8.5.1	Data preparation	201
8.5.2	State specification	202
8.5.3	Construct and analyse sequences	202
8.5.4	The choices of dissimilarity measures and cost setting	202
8.5.5	Clustering structure and deciding the optimal number of clusters	203
8.5.6	Presentation of the cluster patterns	203
8.5.7	A final reminder	204
8.6	Recommendations for future research and potential implications for policy and practice	204
8.6.1	Recommendations for future empirical studies for early diagnosis research	204
8.6.2	Research is needed to assess the impact of COVID-19 on early cancer diagnosis	205
8.6.3	Recommendations for future methodological work	206
8.7	Connect this thesis to the wider research fields and disciplines	206
8.7.1	Data source: big data, real-world data, and EHRs	206
8.7.2	Unsupervised and supervised learning	208
8.7.3	Phenotypes and stratified medicine	209
8.8	Conclusion and final message	210
	Appendix A Search strategy for the systematic scoping review (Chapter 4).....	211
A.1	Search strategy in the EBSCO platform	211
A.2	Search screenshots from the EBSCOhost platform.....	214
A.3	Search strategy in Ovid.....	215
A.4	Search screenshots from Ovid	217
	Appendix B Joanna Briggs Institute (JBI) Critical Appraisal Checklist for Case Series	219
	Appendix C National Heart, Lung, and Blood Institute Quality Assessment Tool for Case Series Studies.....	223
	Appendix D Substitution cost matrix of $OM_{[1, TR]}$ for high-risk sequences in subgroup analysis-2.....	225

Table of Contents

List of References227

List of tables

Table 3.1 – Results of the ten most frequent sub-sequences from event SA in an earlier phase of a study using data from the HHRAD.....	53
Table 4.1 – Summary of the indications to conduct a systematic review or a scoping review ...	66
Table 4.2 – Summary of study characteristics and how SA was used in each study	84
Table 5.1 – Patients’ reasons for primary care consultations.....	97
Table 5.2 – HCP actions (the outcomes of consultation).....	99
Table 6.1 – Cross-tabulation of patient and GP states in primary care consultations, n(%)	114
Table 6.2 – The meaning, frequency, and the colours of the combined states (combined patient and GP events) in R.....	117
Table 6.3 – Frequency of the combined states in the subgroup analysis, n(%)	128
Table 6.4 – The number of consultations related to potential LC symptoms and the intervals between two visits (days)	129
Table 6.5 – Substitution cost matrix of $OM_{[1, TR]}$ for LC symptom sequences (subgroup analysis-1)	130
Table 6.6 – The common sub-sequences in the four clusters of LC symptom sequences by $OM_{[1,2]}$	138
Table 6.7 – The common sub-sequences in the five clusters of the high-risk sequences by $OM_{[1,2]}$	149
Table 7.1 – The sociodemographic and clinical characteristics of the study sample (N=899) ..	164
Table 7.2 – Descriptive statistics and tests comparing patient characteristics among the seven clusters.....	169
Table 7.3 – Significant results in univariable multi-nominal logistic regression (unadjusted relative risk ratio and 95% CI).....	174
Table 7.4 – Multinomial logistic regression model with multiple variables comparing the four clusters of GP management of patients presented with potential LC symptoms	177

List of tables

Table 7.5 – Multivariable multinomial logistic regression comparing the three patient groups	178
Table 7.6 – Results of the negative binomial regression model for the number of primary care attendances (n=829)	179
Table 7.7 – The number of eligible patients, responders, non-responders, response rate, and patients included in this study from each participating practice	181
Table 7.8 – Descriptive statistics of the key patient characteristics by practices	182
Table 7.9 – Distribution of cluster membership among patients from the eight general practices, n (column %)	185

List of figures

Figure 1.1 – Conceptual model of pathways to cancer treatment (Walter et al., 2012).....	4
Figure 2.1 – Relative survival of LC up to 10 years after diagnosis by sex and year (Rachet et al., 2008).....	11
Figure 2.2 – The latest NICE lung cancer care pathway (October 2020)	18
Figure 2.3 – The latest NICE pathway for diagnosis and staging LC (October 2020)	18
Figure 2.4 – Waiting time targets for referral and treatment in cancer within NHS (Forrest et al., 2014a).....	19
Figure 2.5 – A conceptual model of cancer care pathway illustrating the milestones and intervals from the first symptom to the start of treatment (Olesen et al., 2009)	22
Figure 3.1 – An illustration of basic concepts (sequence, states, and alphabet) of SA	46
Figure 3.2 – The outcome of SA presented in state distribution plot (a very simple example) ..	48
Figure 3.3 – The outcome of SA (a typology of clusters in state distribution plot), from Gabadinho et al. (2011a).....	48
Figure 3.4 – An illustration of three basic operations (insertion, deletion, and substitution) in the optimal matching (OM) algorithm.....	49
Figure 3.5 – Dendrogram of agglomerative hierarchical clustering (Ward’s method) based on OM distances	55
Figure 3.6 – Typology of two clusters from single sequences with states combining interdependent patient-GP events two years before LC diagnosis (month as interval)	59
Figure 3.7 – Typology of two clusters by MCSA, patient (left) and GP (right) channels two years before LC diagnosis (month as interval)	60
Figure 4.1 – The PRISMA flow diagram of the whole screening process for study selection and exclusion	70
Figure 4.2 – States containing information from multiple dimensions and the typology of sequences from Darak et al. (2015).....	77
Figure 4.3 – An illustration of regression tree, from Le Meur et al. (2019).....	80

List of figures

Figure 6.1 – The legend of the combined states for the primary care sequences in the NAEDI study	116
Figure 6.2 – Representation of primary care sequences for individual patients (sequences constructed by visits)	120
Figure 6.3 – Representation of primary care sequences for individual patients (sequences constructed in a timeline)	121
Figure 6.4 – Constructing a timeline and setting intervals to represent health sequences in different calendar years	122
Figure 6.5 – State distribution plot of all 858 primary care sequences (percentage relative to the total number of sequences)	124
Figure 6.6 – Alternative presentation of state distribution plot of all 858 primary care sequences (percentage at each visit, legend as the same in the above figure)	124
Figure 6.7 – State distribution plot of primary care sequences in the calendar month	125
Figure 6.8 – Sequence index plot (left, sorted by the number of visits) and state distribution plot (right) for LC symptom sequences	129
Figure 6.9 – Dendrograms of $OM_{[1, TR]}$ (left) and $OM_{[1, 2]}$ (right) for the LC symptom sequences (subgroup analysis-1)	131
Figure 6.10 – Regression tree demonstrating how the cluster patterns changed step by step in $OM_{[1, TR]}$ (left) and $OM_{[1, 2]}$ (right) for the LC symptom sequences	133
Figure 6.11 – The typologies (4 clusters) of $OM_{[1, TR]}$ (left) and $OM_{[1, 2]}$ (right) for the LC symptom sequences.....	134
Figure 6.12 – The five most frequent LC symptom sequences in each cluster by $OM_{[1, 2]}$	137
Figure 6.13 – The state distribution plot for the high-risk sequences (subgroup analysis-2) ...	141
Figure 6.14 – State distribution plot for the high-risk sequences after the first presentation of potential LC symptoms.....	142
Figure 6.15 – An outlier high-risk sequence in subgroup analysis 2	142
Figure 6.16 – Dendrograms of the high-risk sequences (top: $OM_{[1, TR]}$; bottom: $OM_{[1, 2]}$; left: all 295 sequences; right: after excluding the outlier sequence).....	143

Figure 6.17 – Regression tree of the high-risk sequences by $OM_{[1, TR]}$ (left) and $OM_{[1, 2]}$ (right) in subgroup analysis 2 (after excluding the outlier sequence, $n=294$).....	145
Figure 6.18 – Typologies of high-risk sequences by $OM_{[1, TR]}$ (left) and $OM_{[1, 2]}$ (right) in subgroup analysis 2.....	146
Figure 6.19 – The five most frequent high-risk sequences in each cluster by $OM_{[1, 2]}$	148
Figure 6.20 – The state distribution plots before (left) and after (right) the first observed presentation of potential LC symptoms in the five clusters by $OM_{[1, 2]}$	152
Figure 6.21 – The typologies of LC symptom sequences (left) and high-risk sequences (right) by $OM_{[1, 2]}$	153
Figure 6.22 – State distribution plot for patients without potential LC symptoms but consulted cardiorespiratory diseases and other alarm symptoms ($n=326$).....	155
Figure 7.1 – Distribution of patient's comorbidities and IMD quintile by practices.....	183

Research Thesis: Declaration of Authorship

Print name: 廖炜圻 Weiqi Liao

Title of thesis: The use of sequence analysis to study primary care pathways: an exploratory study of people at high risk of lung cancer in England

I declare that this thesis and the work presented in it are my own and has been generated by me as the result of my own original research.

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. None of this work has been published before submission.

Signature: Weiqi Liao Date: 1 December 2020

Acknowledgements

Pursuing a PhD is a very challenging academic journey. First of all, I would like to thank my lead supervisor, Dr Lucy Brindle, who came up with the initial idea of this study and helped me secure the prestigious studentship, and her contributions of numerous insightful comments that shaped and improved my thesis. I very much appreciate the academic guidance and supervision from Dr Lucy Brindle and Dr Bronagh Walsh. Thank you for sharing your knowledge and research experiences with me, discussing the directions of my project, reading many rounds of draft chapters, and providing me with detailed and thorough written feedback. I benefited a lot from this process, which helped me challenge myself, think deeper of the whole study, and write clearer of my thesis. You both set excellent examples as academics with expertise, acumen, and rigour. I also thank Prof Dankmar Böhning and Prof Peter W. F. Smith (Social Statistics) for their helpful discussion, Prof Paul Roderick, Dr Olga Maslovskaya (internal examiners), and Prof Angela Tod (external examiner, University of Sheffield) for their constructive feedback in my PhD confirmation and final viva.

I would also like to thank the following funders and sponsors for providing financial support and training opportunities during my study in the UK.

- I am very grateful that the UK Economic and Social Research Council (ESRC) awarded me a full international studentship (Grant Number: ES/J500161/1) for four years to study two programmes – Master of Research (Clinical and Health Research) and PhD (Health Sciences) in Southampton, and the associated Research Training Support Grant to meet my learning needs and develop my transferable skills.
- The South Coast Doctoral Training Partnership (SCDTP) provided extensive training opportunities (in-house workshops, residential training, and writing boot camps in the beautiful Cumberland Lodge), and granted me two international visits to the University of Western Australia (Perth) and Iceland in 2017. I appreciate the support from the Directorate of the SCDTP, Prof. Pauline Leonard, Prof. Athina Vlachantoni, Dr Amos Channon, and the management team, Mr Glenn Miller and Ms Gemma Harris.
- The Alan Turing Institute fully sponsored my attendance of their Data Study Group (2019) in London.
- The Southampton Opportunity Scholarship provided me funding for a summer school course in the Netherlands on lecturing in the higher education sector (2019).

Acknowledgements

I would like to thank the NHS South, Central and West Commissioning Support Unit and the Hampshire Health Record Information Governance Group for granting me access to the HHRA data, Mr Matthew Johnson for his help in extracting the data, and the NIHR CLARHC Wessex for generously covering the data extraction cost. Although the HHRA dataset was not used and reported in this thesis, it was still valuable, and extensively used for methodological exploration and adaptation of sequence analysis for a very long period in both my MRes and PhD programmes. I appreciate the NAEDI research team (School of Health Sciences, Southampton) for allowing me access to their data for analysis, which is the main data source for this thesis. I also thank the university research engagement librarian, Ms Vicky Fenerty, for her consultations about the search strategy for the reviews.

Peer support is essential to overcome the difficulties encountered within my PhD study. I have met many brilliant fellow PhD students and early career researchers. I appreciate their trust, sharing their experiences and useful tips, successes and struggles, ups and downs with me. I very much enjoy many meaningful and inspirational personal conversations.

Finally, I would like to thank my beloved family, relatives, and friends, especially my parents, for their unconditional love, support, and encouragement all these years.

Thousands of miles as I travelled, from the southern coast of China (Guangdong Province) to England, I think I have been practising a Chinese proverb “Read 10,000 volumes of books and travel 10,000 miles” in person. In retrospect, this has been a long and challenging academic journey for in-depth knowledge acquisition, also an interesting and unforgettable life experience with lots of joyful memories in a foreign country with a unique culture. I have almost come to the end of my formal education; yet I believe, it is the commencement of a new chapter in my life.

Definitions and Abbreviations

AF: Atrial Fibrillation

ASW: Average Silhouette Width

BMI: Body Mass Index

CHD: Coronary Heart Disease

CI: Confidence Interval

CINAHL: Cumulative Index to Nursing and Allied Health Literature

COPD: Chronic Obstructive Pulmonary Disease

CPRD: Clinical Practice Research Datalink, called GPRD before 29 March 2012, also see GPRD

CRUK: Cancer Research UK, a major leading non-profit organisation focusing on cancer research and awareness in the UK

CT: Computed Tomography

CVD: Cardiovascular Disease (6)

CXR: Chest X-Ray

DHD: Dynamic Hamming Distance

EHR: Electronic Health Record

ESA: Event Sequence Analysis

FEV1: Forced Expiratory Volume in 1 second

FVC: Forced Vital Capacity

GOLD: Global initiative for chronic Obstructive Lung Disease

GP: General Practitioner

GPRD: General Practice Research Database, a primary care record database called during 1994-2012, now known as CPRD

HAM: Hamming Distance

Chapter 1

HCP: Healthcare Professional

HES: Hospital Episode Statistics, secondary care database in England

HHRAD: Hampshire Health Record Analytical Database

ICBP: The International Cancer Benchmarking Partnership Study

ICD10: International Classification of Diseases (10th revision)

IHD: Ischaemic Heart Disease

IMD: The English Indices of Multiple Deprivation

IQR: Inter Quartile Range

IRR: Incidence Rate Ratio

LC: Lung Cancer

LDCT: Low Dose Computed Tomography

LLP: Liverpool Lung Project

LSOA: Lower-Layer Super Output Areas

MCSA: Multi-Channel Sequence Analysis

MDT: Multi-Disciplinary Team

MI: Myocardial Infarction

NAEDI: The National Awareness and Early Diagnosis Initiative

NHS: National Health Service

NICE: The National Institute for Health and Clinical Excellence

NSCLC: Non-Small Cell Lung Cancer

OM: Optimal Matching

ONS: Office for National Statistics

OR: Odds Ratio

OTC: over-the-counter medications

PLCO_{M2012}: a validated prediction model from the Prostate, Lung, Colorectal, and Ovarian (PLCO) Cancer Screening Trial for lung cancer risk

PPV: Positive Predictive Value

QOF: Quality and Outcomes Framework

RCTs: Randomised Control Trials

RQ: Research Question

RR: Relative Risk

RRR: Relative Rate Ratio

RS: Relative Survival

SA: Sequence Analysis

SCLC: Small Cell Lung Cancer

SD: Standard Deviation

SEER: the US Surveillance, Epidemiology and End Results Program, classified tumours as localised, regional, and distant

SES: Socioeconomic Status

THIN: The Health Improvement Network, an English primary care database

TNM: a globally recognised cancer staging system to classify the extent of tumour (T), the degree of nodal involvement (N), and metastases (M)

UK: United Kingdom

WHO: World Health Organisation

Chapter 1 Introduction

1.1 Research context and the motivation to conduct this PhD study

1.1.1 Synopsis of this thesis

This thesis is about exploring the application of a statistical method – sequence analysis (SA), to study primary care sequences/pathways using events from health records involving patients and healthcare professionals (HCPs, mainly general practitioners, GPs) among a group of clinically relevant patients at high risk of developing lung cancer (LC) in the southern coast of England. The motivation to conduct this study is to address the poor LC survival in the UK, from a less investigated research angle (primary care sequences) and population (high-risk patients before diagnosis).

1.1.2 Why lung cancer

LC was identified as one of the research priorities by the leading cancer charity, Cancer Research UK (CRUK) in 2014, due to its high incidence, mortality, difficulty to early detect and diagnose, 'hard to treat', and poor survival. According to the most recent cancer statistics published by Cancer Research UK (2021), LC was the third most common cancer in incidence (after breast and prostate cancers) and the most common cause of cancer death in the UK. Incident LC cases took up to 13% of all new cancer cases in 2017. LC deaths accounted for 21% of all cancer deaths in 2017, more than twice of the second highest cancer mortality (bowel cancer, 10%). The highest incidence and mortality rates of LC were in the age group of 85-89 (2015-2017). LC incidence and mortality rates decreased in males, but increased in females in the last decade, most likely related to the change of **smoking pattern** by sex. LC incidence and deaths in England were more common in people living in the most deprived areas (lower socioeconomic status, SES), where smoking and other risk factors were common in this population. Compared with other cancers, LC survival is poor. Only 40.6% of patients survived one year or longer (2013-2017), 16.2% survived for five years, and 9.5% for ten-year survival. Statistics showed that around three quarters of LC cases were diagnosed at late stages in England (2014), Scotland (2014-2015) and Northern Ireland (2010-2014). Nevertheless, if patients were diagnosed at the earliest stage, about 57% of patients could survive five years or more, compared with around 3% when diagnosed at the latest stage (Cancer Research UK, 2021). These cancer statistics demonstrated the burden of LC and the necessity to improve late diagnosis and the associated poor survival in the UK, making LC one of the research priorities.

1.1.3 Research gap and a new research angle

Current population-based cancer studies are mainly focused on identifying risk factors for late diagnosis (Macleod et al., 2009, Maclean et al., 2015) and poor survival of LC (O'Dowd et al., 2015), investigating route to diagnosis (Elliss-Brookes et al., 2012), and diagnostic intervals in the cancer care pathway (Lyrtzopoulos et al., 2013, Redaniel et al., 2015, Neal et al., 2015b). However, there is still very little knowledge about **the primary care pathways** relevant and leading to LC diagnosis, which is a research gap to fill. Primary care sequence/pathway involves patient's help-seeking and GP's response to the patient's presentation. Studying sequences of primary care events preceding diagnosis may help us understand the reasons for delayed diagnosis. In addition, primary care is the foundation of the British health system, the setting where most health conditions are assessed and managed by GPs. Patients have different levels of health literacy, comorbidities, and performance status, which may affect their help-seeking behaviours and health services utilisation. Different frequencies of patient-GP interactions may affect the stage of cancer diagnosis. If patients seek help in a timely manner, and the abnormal symptoms are picked up by vigilant GPs, patients may have a higher chance of being timely referred for further investigation in secondary care. However, if patients neglect abnormal symptoms and do not see their GP at all, the chance of emergency presentation would increase, which is proven to be associated with worse clinical outcomes – late stage diagnosis and very poor survival (McPhail et al., 2013). Currently, emergency presentation is the most common route to LC diagnosis (Cancer Research UK, 2021). A shift of the diagnostic route from emergency presentation to rapid GP referral (two-week wait) is vital for early diagnosis and improving survival. **Primary care involvement** is essential to achieve this goal. Research evidence on the patterns of primary care pathways may help shift the diagnostic route from emergency presentation to two-week wait.

1.1.4 A novel statistical method for the new research angle

Primary care events are now well documented and stored in electronic health records (EHRs), which can be easily retrieved. Patients' help-seeking events and the subsequent GP management over time can be constructed as a sequence for each patient. Regression model is a family of variable-based statistical methods, widely used in biomedical and health services research to investigate the association between dependent and independent variables and to predict outcomes. But if the research interest is studying the primary care pathway as a whole, regression is not able to identify common patterns, nor typologies of events. Sequence analysis (SA) has the potential to achieve the goal.

SA is an exploratory, process-based statistical method. It has been used to study life course and career trajectories in social sciences since the 1980s (Abbott, 1983, Abbott and Forrest, 1986), but is less applied in health research until recent years. It can examine the whole sequence holistically as an entity, summarise the patterns of sequences from individual patient to group level as clusters, and visualise the cluster patterns as typology (the outcome of SA). The typology then could be used in regression models, either as a dependent (categorical) variable, exploring its association with subject characteristics at the individual level; or served as an independent variable, exploring its relationship with other important outcomes (e.g. quality of life, survival outcomes). The initial idea for this PhD study came up from the inspiration of these features of SA. Although primary care sequences are not entirely the same as life course and career trajectories, some lessons can be learnt from the application of SA in social sciences, when trying to use it in health services research. However, when applying methods from one field to another, some adaptations for different research contexts are needed.

1.2 Research aims and the key research questions (RQs)

This PhD study aims to re-contextualise and adapt the application of SA from social sciences to early diagnosis research, to explore the use of SA to study complex primary care sequences, to provide new research perspectives to promote early diagnosis of LC, and ultimately, to improve LC survival in England/the UK. Figure 1.1 illustrates where this PhD study is situated in the field of early diagnosis research. **This PhD study is positioned as an exploratory study** in general, to lay down the groundwork for future empirical studies of this kind. Therefore, the methodological explorations were driven by the need to answer important empirical research questions (RQs) that could provide research evidence. The exploration process was more focused on making good use and adapting the existing methodological framework of SA, rather than purposing a new algorithm or modifying the current algorithms.

There are three interlinked research stages for the whole study. The key research questions for each stage are:

The first stage: systematic scoping review for the use of SA in health services research

- a. How has SA been used in health services research?
- b. What are the added values of using SA in health services research in the published studies?
- c. What lessons can be learnt from the previous research and applied in this study?

The second stage: methodological exploration of using SA in primary care sequences

Chapter 1

- A. How can SA be contextualised and used to represent primary care sequences from discrete health events involving patients and GPs?
- B. How can methodological decisions be made to obtain meaningful typologies that make sense in empirical context?

The third stage: empirical analysis of the cluster patterns, patient characteristics, and the practice effect

- i. What were the cluster patterns of primary care sequences among community-based patients at high risk of developing LC?
- ii. What were the patient profiles for each cluster?
- iii. What patient characteristics can help explain the variation in the cluster patterns, especially among patients presenting with potential LC symptoms?
- iv. Was there a practice effect on patients' primary care attendance, consultations of potential LC symptoms, or the cluster patterns of sequences?

These RQs imply the exploratory nature of this PhD study. Therefore, the analysis and results will be reported step-by-step in each chapter.

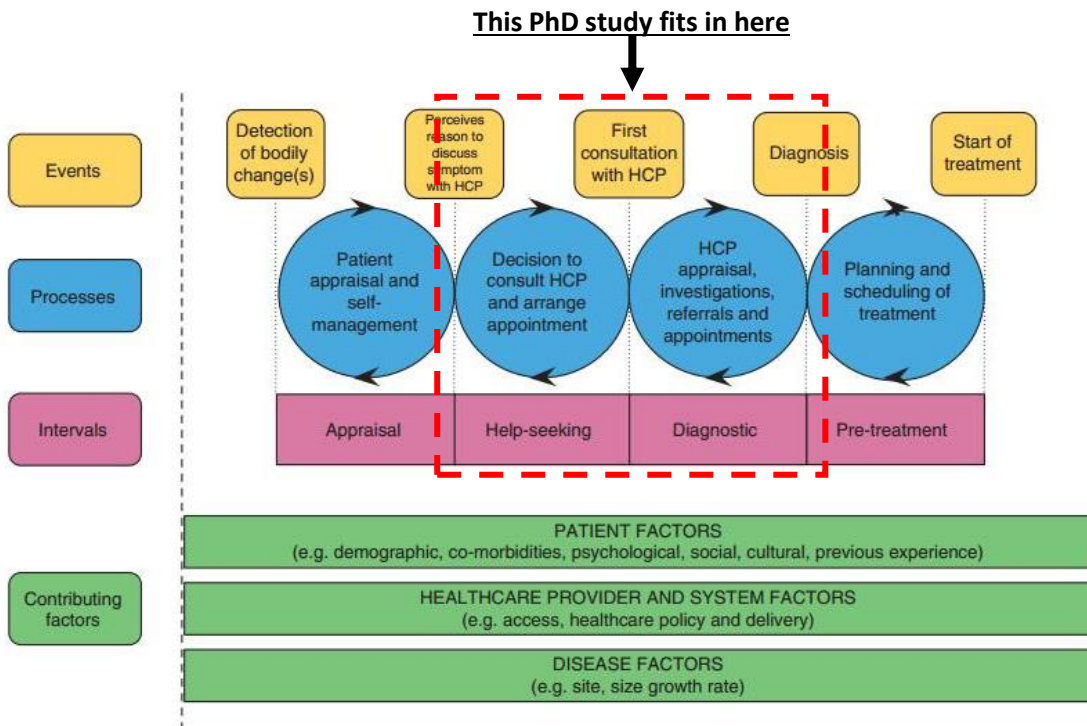


Figure 1.1 – Conceptual model of pathways to cancer treatment (Walter et al., 2012)

1.3 The journey of the PhD study and this thesis

The journey of this PhD study was like travelling through a rugged mountain road, with lots of twists and turns, ups and downs. The final product of this thesis was adjusted to the unforeseeable circumstances during my PhD candidature.

Originally, EHRs of primary care events from the Hampshire Health Record Analytical Database (HHRAD) was planned as the main data source of this PhD study. HHRAD is a clinical database containing about 1.4 million patients registered in 145 general practices in Hampshire and the neighbouring towns in southern England (Hampshire Health Record, 2017), which was an ideal data source for population research. The original research plan was to use HHRAD to study primary care sequences among patients diagnosed with primary LC between 2000 and 2015, with an observation period of two years before diagnosis.

During the data cleaning process (the first and second year of my PhD), it was found that some deceased patients were selectively deleted from the HHRAD, as this group of patients did not need clinical care anymore, but it had implications of using this data source for research purposes. The mechanism of missingness (who was deleted) was unknown, and further information on the deleted subjects was unavailable. This problem resulted in a reduced study sample having a higher survival rate and a longer survival period than the general LC population in the UK, which meant the study sample was biased and not representative. After communication with the relevant personnel, it was confirmed that it was not possible to rectify this problem within the period of my candidature. Therefore, due to the data quality problem, an empirical study was no longer an option. The extracted dataset was used for methodological exploration in two studies – sequences of patients' help-seeking behaviours (one domain), and how GP managed patients with COPD two years before LC diagnosis (two dimensions, interdependent patient-GP events). Although addressing methodological issues, it was found that the decisions on methodological issues were very much tied to and dependent on the empirical context. Without drawing empirical conclusions, the methodological conclusions were context-dependent and unlikely to transfer and generalise to a different study. Despite tremendous time and efforts being put into the research using the HHRAD dataset, regrettably, the results of these two studies were not reported in this thesis.

A silver lining in this very difficult situation, another dataset in the English primary care setting was available, which allowed me to conduct methodological work and draw empirical conclusions. The project was previously funded by the National Awareness and Early Diagnosis Initiative (NAEDI, called the NAEDI study in this thesis hereafter). The population of the NAEDI study was

Chapter 1

community patients with a long smoking history and at high risk of developing LC, but not yet diagnosed with LC at the point of data collection (Wagland et al., 2016). As McCutchan et al. (2019) pointed out, it is still a lack of evidence on how this group of population attributes to potential LC symptoms and decide to seek medical help, and the optimal methods of promoting earlier presentation through interventions targeted at high-risk, highly deprived groups. This thesis addresses this research gap, using the data from the NAEDI study to explore how SA can be used to identify meaningful patterns of primary care sequences for early diagnosis of LC. The previous experiences of using HHRAD made the research process in the NAEDI dataset more smoothly. Examples and lessons learnt from the exploration using HHRAD are used in this thesis to illustrate, discuss, and support academic arguments where appropriate.

1.4 The structure of this thesis

This introductory chapter sets out the research context and background for the whole thesis, followed by a detailed literature review in Chapter 2, to elaborate the research gap in this field, to justify the rationale and the necessity to conduct this study, also to explain the novelty of this PhD. The interest of studying primary care sequences relevant and leading to LC diagnosis, and the two important elements in the sequences (patient's help-seeking behaviours and GP's response) will be fully discussed in the literature review chapter. Chapter 3 will introduce and cover the technical details of SA, situating this method in health research, explain and discuss why SA is a potentially helpful statistical method to study primary care sequences. Since SA is still a relatively new method in health research, a systematic scoping review was conducted to understand how SA has been used to study disease trajectories, care pathways, and health services research, and to learn the experiences from other studies, which is reported in Chapter 4. Chapters 2-4 together provide comprehensive contextual and technical background for this study.

Community-based patients at high risk of developing LC but not yet diagnosed with LC are the target population for early diagnosis. The NAEDI dataset was used to investigate the primary care sequences among this population. The data source is introduced in Chapter 5 (Methodology), followed by a statistical analysis plan, including all the methods used in the whole study. After that, Chapter 6 reports the whole methodological exploration process, which aims to address the methodological issues and explore the best analytical approach that can yield meaningful cluster patterns that makes sense in the empirical context. This is the first phase of the main study. The second phase is to understand patients' sociodemographic and clinical characteristics in each cluster, to compare the patient profiles across clusters, to investigate what patient characteristics

can enhance the understanding of the variation in the cluster patterns of primary care sequences, and the practice effect. All findings of the empirical analyses are reported in Chapter 7.

In the final chapter of this thesis (Chapter 8), I revisit the central argument of this thesis, summarise the key findings and messages, share the lessons learnt from the whole study, discuss the original contributions of this PhD to the research field, connect this study to wider research areas, and make recommendations on how to use EHRs and SA to design future studies that can better understand primary care pathways to achieve the goal of early diagnosis and improving LC survival in the UK.

Chapter 2 Literature review

2.1 Chapter introduction

This chapter starts to review relevant literature about LC survival in England, and compare the LC survival statistics in England with other European and developed countries, with the aim to understand why LC survival in England was poorer than other western developed countries (section 2.2) and how the British government, health system, charities, and research organisations planned to address this problem (section 2.3). Early diagnosis is the key to improving cancer survival. Therefore, the current national guidelines, research paradigms on the cancer care pathway and route to diagnosis were reviewed and appraised. Their respective limitations are discussed in section 2.4. The role of primary care in the British health system and its importance in improving early diagnosis and cancer survival is discussed in section 2.5. Through the literature review, the current research paradigms are not ideal to provide research evidence that can fully understand the primary care process leading to LC diagnosis. Therefore, a new research angle using a novel statistical method is proposed to address this research gap in this PhD study, which is studying primary care pathways using sequence analysis. Further explanation and discussion about why this research angle and statistical method have the potential to provide new knowledge to address the late diagnosis and poor survival of LC in England are at the end of this chapter.

2.1.1 Search strategy and database

Relevant medical subject headings (MeSH) and free text keywords were searched in PubMed to identify literature about the above topics of LC diagnosis and survival in England. Lung Neoplasms (MeSH) and the relevant sub-headings (diagnosis, diagnostic imaging, epidemiology, mortality, organisation and administration, prevention and control, radiotherapy, statistics and numerical data, surgery, therapy) were searched in PubMed, together with United Kingdom, help-seeking behaviour, primary health care (general practice, general practitioner), and social class (all MeSH terms). Free text keywords were searched in the field of “title and abstract” in PubMed, including lung cancer, England (the UK/United Kingdom), diagnosis, survival, high risk, screening, prevention, symptom, comorbidity, chronic obstructive pulmonary disease (COPD), primary care (general practice/general practitioner/GP), help-seeking, interval, delay, route, treatment (therapy/surgery/radiotherapy/chemotherapy), and socioeconomic status (SES, socioeconomic position/social class). The initial search and review of literature started in 2016, when I started my

PhD, and the search was updated continuously for new publications during my candidature. The last search was conducted in March/April 2020 for this thesis.

2.2 LC survival in England

This section aims to review the research evidence on LC survival in England, compared with that in other European countries (the EURO CARE-5 study), and some developed western countries with similar health systems (the International Cancer Benchmarking Partnership, the ICBP study). Most common cancers were included in both EURO CARE and ICBP studies, but only LC is discussed here, as other cancer types are not relevant to the research context of this study. Two cancer staging systems, TNM and SEER, are used in different cancer registries. TNM is the acronym for the extent of tumour (T), the degree of nodal involvement (N), and metastases (M). SEER stands for the US Surveillance, Epidemiology and End Results Program, classifying tumours as localised, regional, and distant.

Relative survival (RS) is a standard approach to compare population-based cancer survival across countries. It is the ratio of the observed survival in patients with cancer and the expected survival in the general population for all causes of death (i.e. background mortality) in the same region/country, stratified by age groups, sex, and calendar year. RS can be understood as survival from cancer after adjusting for other causes of death. There are three common indicators of RS, 1-year RS, 5-year RS, and 5-year survival conditional on 1-year survival (conditional 5-year survival, denoted as 5|1-year RS). The reason to use 5|1-year RS is that short-term survival is an important driver for longer-term survival in some cancers (Coleman et al., 2011, De Angelis et al., 2014).

2.2.1 LC survival in England

Based on the data from 392,000 adult patients (aged 15-99 years) diagnosed with primary LC in England and Wales during 1986-1999 (followed up to 2001), Rachet et al. (2008) estimated the trend of LC survival up to 10 years by sex, as shown in Figure 2.1. RS dropped steeply in the first year, especially in the first several months after diagnosis, then continued to drop, but much slower and more stable afterwards, until the tenth year.

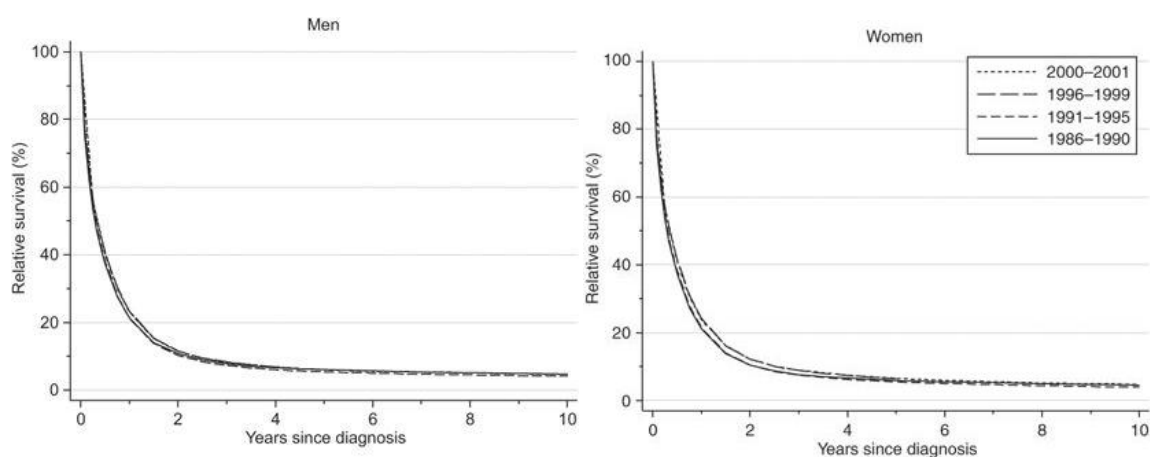


Figure 2.1 – Relative survival of LC up to 10 years after diagnosis by sex and year (Rachet et al., 2008)

McPhail et al. (2015) used data from the National Cancer Registration Service to estimate 1-year RS. Patients diagnosed in 2012 ($n=34,997$, the whole English population) were followed up to the end of 2013. The completeness of cancer stage was unusually high in this study (close to 90%). The results showed that most LC cases were still diagnosed at advanced stages. The distribution of LC stages in TNM from Stage 1 to 4 were 13.2%, 7.5%, 20% and 49%, respectively, and 10.2% for unknown stage. The percentages for men and women in each stage were comparable. Age-standardised 1-year RS was 37.2% (95% CI, 36.7-37.8), ranging from Stage 1 for 85.1% (84.0%-86.2%) to stage 4 for only 17.6% (17.0%-18.2%). Another study reported similar stage-specific survival statistics in LC (Blandin Knight et al., 2017). **Such a big difference in LC survival demonstrated the importance of early diagnosis. Stage and age** were the two most important contributors for the 1-year RS, while other significant factors included sex, income deprivation, and geographic area. The 1-year RS substantially decreased with each level of increase in stage at diagnosis of LC, which demonstrated **the need to shift LC diagnosis to earlier stages**, and the efforts to improve stage-specific survival for all stages of LC.

2.2.2 English LC survival in the European context

EUROCARE was the largest cooperative study on cancer survival in Europe. The most recent study (EUROCARE-5) assessed cancer survival from 29 European countries with cases diagnosed during 2000-2007. The results showed that the 5-year age-standardised RS of LC in the four UK countries (8.8% in England, 8.7% Scotland, 8.6% Wales, and 11.0% Northern Ireland) were much lower than the mean European level (about 13.0%)(De Angelis et al., 2014). Among 87 cancer registries included in the EUROCARE-5 study, 52 registries provided stage information for LC, but only ten registries (166,554 patients) provided data up to quality standards to estimate stage-specific survival, which indicated the data quality of cancer stage was still a big issue. This study found

that older patients (aged ≥ 70 years) were diagnosed at more advanced stages, and had worse stage-specific survival than those aged < 70 years. For early stage (localised, $< 20\%$ of LC cases), 5-year RS for patients aged < 70 years (55%) was 20 percentage points higher than those aged ≥ 70 years (35%). For metastatic patients (41% cases), 1-year RS was around 20% and declined to 3%-4% for 5-year RS. **Stage at diagnosis** and **age** were the two crucial factors for LC survival (Minicozzi et al., 2017). Older people having worse survival outcomes were perhaps due to existing comorbidities and performance status, which made them less suitable for potentially curative treatment when they were diagnosed at advanced stages. Early diagnosis may increase the chance of survival for older people, not to mention the life years gained for younger patients.

Abdel-Rahman et al. (2009) estimated that the number (and percentage) of **avoidable deaths** for British patients diagnosed with LC was 3548 (2%) during 1985-1989 (EUROCARE 2), 3735 (2%) during 1990-1994 (EUROCARE 3), and 4923 (3%) during 1995-1999 (EUROCARE 4), if RS of the three UK nations (England, Wales, and Scotland) could catch up with the mean of the European 5-year RS. Much more premature deaths could be avoided, if compared with the highest level of European survival. Such estimates showed significant **public health implications of cancer survival**, and supported the argument that **even small improvements in survival from common cancers (like LC) can prevent large numbers of premature deaths in the whole population** (Coleman et al., 2011, De Angelis et al., 2014).

2.2.3 English LC survival in the international context

The International Cancer Benchmarking Partnership (ICBP) study aimed to estimate the up-to-date survival of the selected cancers, to understand whether international differences in cancer survival have changed, to investigate the causes of survival disparities, and to provide high-quality research evidence for policymakers to reduce inequalities in cancer survival (Coleman et al., 2011). Six countries (Australia, Canada, Denmark, Norway, Sweden, and the UK) participated in the ICBP study. These countries had similar economic development, comparable nation wealth, tax-financed universal health systems, and high-quality cancer registries. 715,330 patients from 12 jurisdictions (6 countries) during 1995-2007 were included to estimate age-specific and age-standardised 1-year, 5-year, and conditional 5-year RS in LC. The results showed that all three indicators of RS in LC in the UK continuously improved in the three periods (1995-99, 2000-02, 2005-07). Almost all the improvement could be attributed to the increase of 1-year RS. However, 1-year and 5-year RS in the UK were at the bottom of the six countries. For 2005-07, 1-year RS was around 30% in the UK nations, 35% in Denmark, and 39-44% in Australia, Canada, Norway, and Sweden. The 5-year RS was low in the UK, at around 15% in 2010-14, compared with Australia and Canada at about 22% (Arnold et al., 2019). The ICBP studies found that patients in the UK

nations had the lowest age-standardised 1-year RS in both non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC), and the lowest proportion of adenocarcinoma (25.2%), a histological type that had a better prognosis. Stage at diagnosis could explain some of the international differences in LC survival, and wide disparities in stage-specific survival still existed. Late diagnosis was the main problem for the poor survival in the UK, particularly for patients aged 65 years and older (Coleman et al., 2011, Walters et al., 2013, Eden et al., 2019).

2.2.4 Section summary and discussion: early diagnosis is the key to improve LC survival in England

For population-based studies estimating and comparing cancer survival across countries, the authors usually just reported age-standardised RS in different calendar periods, which was possibly due to the nature and the scale of such studies. Studies like EURO CARE or ICBP often reported RS for 20-30 common cancers in different countries. The length of a research paper does not allow authors to report more details. These studies can provide a good overview of the incidence, mortality, or survival of a wide range of cancer types across different countries over time, which is very informative to compare the performance of health systems in providing care and addressing cancer survival. The main limitation of this type of study is that it usually does not provide sufficient details for a particular type of cancer in a country, like LC in England.

Two major issues that appeared in these studies were comparability and data harmonisation (Walters et al., 2013) and the **lack of robust cancer stage data**. Firstly, the data were from different countries. Although efforts had been continuously made to harmonise the data collection procedures, there were still variations in data quality. Different cancer staging systems (TNM or SEER) were used in different countries. Secondly, despite the critical role in cancer survival, cancer stage was not used until the most recent EURO CARE study (Minicozzi et al., 2017). The quality and completeness of staging information in cancer registries were still a big concern. Missing stage was more frequent in older patients, including those in Clinical Commissioning Groups (CCGs) with high completeness in England (Di Girolamo et al., 2018). In addition, it was difficult to link cancer registry data to large electronic health databases, like the Clinical Practice Research Datalink (CPRD) in primary care, or the Hospital Episode Statistics (HES) in secondary care. **The lack of robust and complete cancer staging data greatly limits broader exploration using routine EHRs in empirical studies in the UK.**

There were some possible explanations for the international variations in LC survival among countries, including the differences in cancer screening opportunities and the coverage of population, diagnostic intensity, stage at diagnosis, cancer biology and histological types, the

prevalence of comorbidities in different populations, and provision and accessibility to quality treatments and cancer care (Richards, 2009b, De Angelis et al., 2014). Late diagnosis has been established as the main reason for the poor LC survival in England, especially for the short-term (1-year) survival (Richards, 2009b, Thomson and Forman, 2009). If the analyses were restricted to include patients who survived at least one year from diagnosis, the difference in the 5-year survival between England and other European countries was smaller (Thomson and Forman, 2009; Holmberg et al, 2010). McPhail et al. (2015) and Exarchakou et al. (2018) argued that the short-term survival reflected the speed and efficiency of patient management in the health system, including diagnosis, staging, first definitive treatment, the quality of surgery, and postoperative care. As such, Richards (2009b) suggested that 1-year RS could be used as a proxy to separate early versus late diagnosis. For the long-term (5-year) survival, it reflected more on the effects of treatment and patient management.

The low LC resection rate was likely to be another factor for the poor LC survival in England. For patients diagnosed with NSCLC during 1998-2006, the surgery rate was 9.9% (19,153/192,657), which was lower than that in other developed European countries, e.g. 24% in Italy and 18% in Sweden. LC surgery is more invasive and technically more complex than breast cancer. A lower number of thoracic surgeons, and a higher prevalence of comorbidities among the English patients, were the two possible reasons for the low surgery rate for LC (Richards, 2009b, Nur et al., 2015b). The proportion of patients receiving LC surgery increased from around 10% in 2008 to 17% in 2015. Such improvement may partly be due to a higher number of specialised surgeons (Exarchakou et al., 2018).

The UK government recognised the importance of early diagnosis, which was a pivotal avenue to improve LC survival. The current government policies and initiatives to promote early LC diagnosis in England are summarised in the next section.

2.3 Policies and initiatives to promote early LC diagnosis in England

The NHS Cancer Plan for England (Department of Health, 2000) aimed to improve cancer services and raise the 5-year survival to the levels of the best performing countries in Europe by 2010. LC survival in the UK did improve in the last 20 years. Age-standardised 5-year RS increased from 7.2% (95% CI 7.0%-7.3%) during 1995-1999, to 14.7% (14.5%-15.0%) in 2010-2014, with an absolute increase of 7.5 percentage points (Arnold et al., 2019). However, the UK still lagged behind the best performing countries, as LC survival improved in those countries as well, even faster. Denmark used to be in the same tier of cancer survival as the UK, but now it had the largest absolute increase (10.7 percentage points, from 8.2% in 1995-99 to 18.9% in 2010-14)

(Arnold et al., 2019). The Danish three-legged strategy (Vedsted and Olesen, 2015), a **differentiated approach for referrals from general practice to support early cancer diagnosis**, seemed to work well.

If the UK wanted to achieve the best European LC survival, it required a shift of diagnosis to earlier stages. Early diagnosis may lead to prompt treatment and prolonged survival, but also can result in overdiagnosis and lead-time bias. A health economic analysis was conducted to assess the likely impact of early LC diagnosis on the costs and benefits to the NHS (Department of Health, 2011a). The estimated additional costs for diagnostic services would be around £95 million, and the additional treatment costs £9m. Treatment costs would be higher if patients were diagnosed earlier, as more treatments would be needed during the prolonged survival. However, the model suggested a population benefit of 42,083 life-years gained. The average cost to save a life was £2,376, which was very cost-effective.

2.3.1 National initiatives and campaigns for LC

The National Awareness and Early Diagnosis Initiative (NAEDI) was launched in November 2008 in England, as part of the Government's strategy to improve cancer outcomes. The main aim of NAEDI was to address the poor cancer survival by reducing the number of patients diagnosed and treated in late stages. LC was prioritised in this initiative. NAEDI also tried to improve and optimise the diagnostic pathway within the health system. Actually, the data for this PhD study was from a study funded by NAEDI. The Cancer Reform Strategy (2007) emphasised the importance of public awareness of symptoms recognition and signs indicative for LC to early diagnosis. NAEDI echoed and acted on this suggestion. National campaigns were carried out to raise public awareness of symptom presentation and timely help-seeking behaviours (Richards, 2009a, Thomson and Forman, 2009, Hiom, 2015). The national campaign Be Clear on Cancer (BCOC) was launched in England in 2012, to raise public awareness of the key symptoms of bowel and lung cancers, targeting people aged 50 years and over, especially for the lower socioeconomic groups, as this group of patients were associated with lower health literacy and less aware of cancer symptoms (Rubin et al., 2014, Whitaker et al., 2015). The key message for the LC campaign was "If you have been coughing for three weeks or more, tell your doctor" (Moffat et al., 2015). The future campaign should raise public awareness of other symptoms indicative of LC, and increase engagement with the lower socioeconomic groups.

2.3.2 LC screening in England

Screening is one possible route to achieve early cancer diagnosis. The US Prostate, Lung, Colorectal and Ovarian (PLCO) screening trial concluded that screening asymptomatic populations with chest X-ray (CXR) did not reduce LC mortality (Oken et al., 2011). The US National Lung Screening Trial Research Team et al. (2011) reported a 20% reduction in LC mortality with an annual low dose computed tomography (LDCT) in an **asymptomatic high-risk population**, which was defined as participants aged 55-74 years at the time of randomisation in that study, who had a cigarette smoking history for at least 30 pack-years or quit smoking within the previous 15 years for ex-smokers. LDCT would be less cost-effective if applied to individuals at lower risk of LC, because more people need to be screened, in addition to the cost of LDCT. Therefore, the common practice is to use a validated model/questionnaire, e.g. PLCO_{M2012} (developed in the US population) (Tammemägi et al., 2013), or the Liverpool Lung Project (LLP) risk model (the UK population) (Cassidy et al., 2008), to select subjects at high-risk who are more likely to develop LC.

The UK Lung Cancer Screening trial was the first population LC screening trial using LDCT. A large population sample (n=88,897) aged 50-75 years were approached, with a questionnaire to determine the risk. Those with an estimated risk of at least 5% of developing LC in the next five years (based on the LLP risk model) were invited to participate in the trial. Higher socioeconomic status positively correlated with the response rate, but inversely correlated with risk (McRonald et al., 2014). The UK Lung Cancer Screening trial randomised 4,055 subjects, 1,994 underwent LDCT, 42 (2.1%) patients confirmed LC, 85.7% (36/42) were in stage I or II, and 83.3% (35/42) had surgical resection. Health economic analysis suggested that LDCT be cost-effective, £8,466 per quality-adjusted life-year gained (95% CI, £5,542-£12,569)(Field et al., 2016).

The community-based nurse-led 'Lung Health Check' LDCT screening in Manchester used the PLCO_{M2012} risk model to select participants in deprived areas based on the risk of developing LC. This project started in 2017. Initial findings suggested high screening adherence (90%, 1,194/1,323), although most participants were from the lowest decile of deprivation in England. The incidence rate was 1.6% (n=19) and 79% (15/19) patients were diagnosed in stage I (Crosbie et al., 2019), which indicated an early diagnosis of LC through LDCT screening in a deprived population seemed achievable.

LDCT screening can cause harm. According to a systematic review, a relatively high proportion of subjects (about 20%) were identified with nodules, but most nodules were benign (Bach et al., 2012). Additional radiological investigation of these nodules can trigger increased radiation exposure. The value of LDCT screening should be weighted between the risk of LC and the potential harms. Overdiagnosis and patient's psychological burden should not be neglected.

Patients may experience distress and anxiety when waiting for the scan results, and fear even with a slight suspicion of LC.

Smoking could easily cancel out the potential gains from LC screening. LDCT screening offered a teachable moment for participants to consider giving up smoking, especially for those who received a positive scan result. The smoking cessation rate among participants in the UK Lung Cancer Screening trial was 11% in the short term and 22% at two years follow-up, higher than 4% in the general UK population (Brain et al., 2017). But another study reported that pulmonary nodule detection during LC screening had little impact on smoking behaviours (Clark et al., 2018). Further behavioural research is needed to find the optimal strategies to integrate evidence-based smoking cessation interventions with stratified LC screening, especially for smokers from socioeconomically deprived backgrounds (Brain et al., 2017).

The UK National Screening Committee makes recommendations for all screening programmes. The debate of potential benefits and harms, diagnostic statistical indicators (sensitivity, specificity, positive predictive value, negative predictive value), and cost-effectiveness may be the reasons why LDCT is not currently recommended for LC screening in the UK. The Committee may review this decision following the final published results from the largest European (Dutch-Belgian NELSON) trial (Bradley et al., 2019b). **The majority of LC diagnosis is still made through symptomatic presentations by patients** and appropriate investigation by vigilant GP. This is why this PhD study focuses on understanding the interdependent patient-GP events in the primary care setting. **Smoking cessation** remains a valuable and cost-effective preventive measure for chronic cardiorespiratory diseases and LC.

2.3.3 NICE recommended LC care pathways, diagnosis and staging procedures

Accurate diagnosis and staging LC are the premises to offer patients the best possible treatment. Optimising the steps in cancer diagnosis and staging, and completing the necessary procedures as quickly as possible can reduce delays in diagnosis and treatment. The National Institute for Health and Care Excellence (NICE) is a British organisation that provides guidance and advice for clinical practice to improve patient care. The LC care pathway was available online <https://pathways.nice.org.uk/pathways/lung-cancer> (last accessed on 20 October 2020). The latest recommendation on the LC care pathway and the diagnosis and staging pathway are in Figure 2.2 and Figure 2.3, respectively.

Lung cancer overview

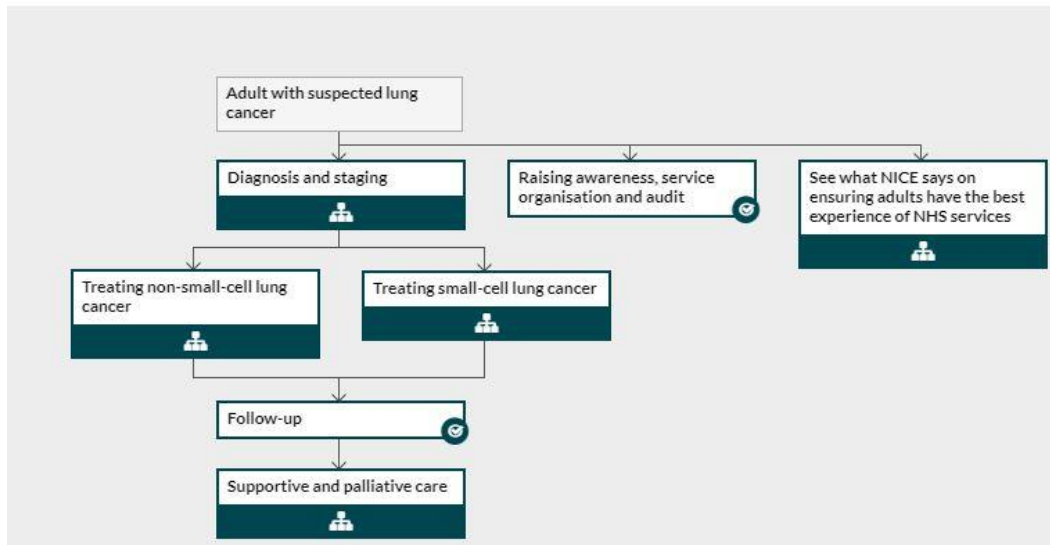


Figure 2.2 – The latest NICE lung cancer care pathway (October 2020)

Diagnosis and staging of lung cancer

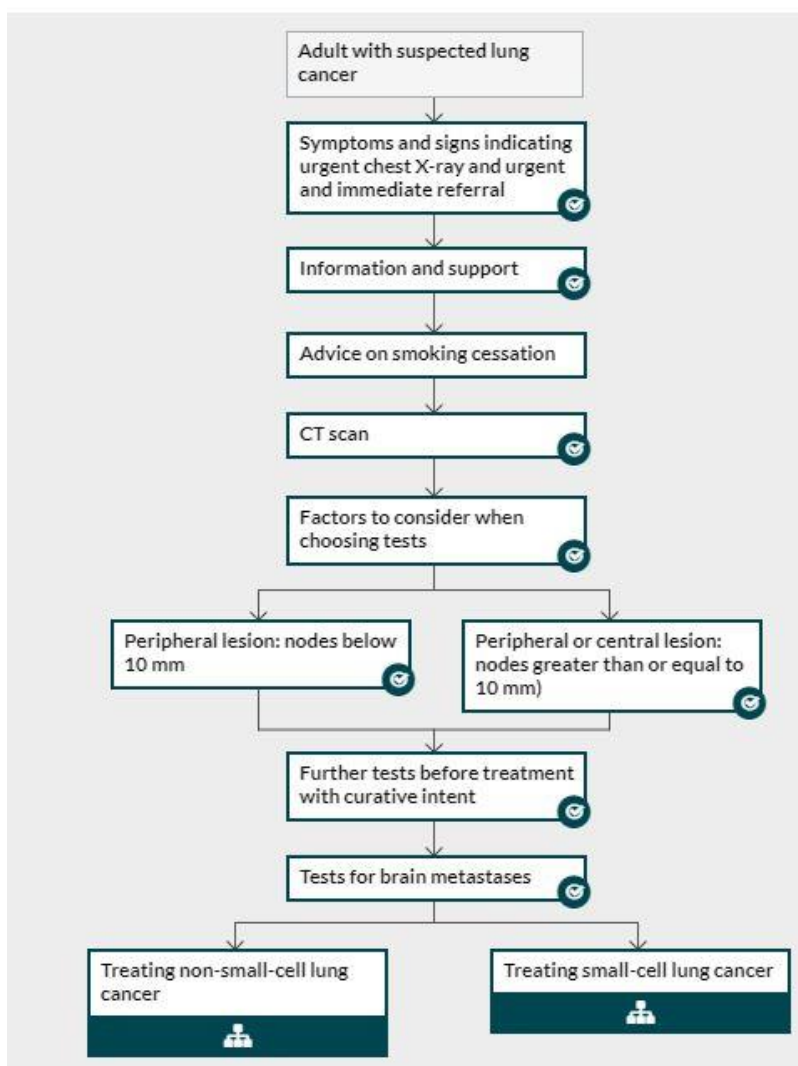


Figure 2.3 – The latest NICE pathway for diagnosis and staging LC (October 2020)

GP is advised to offer a chest X-ray for patients aged 40 and over if they have two or more of the following unexplained symptoms, or have one or more of the following symptoms but patients have smoked before: cough, fatigue, shortness of breath, chest pain, weight loss, and appetite loss. The GP should consider offering an urgent CXR (to be performed within two weeks) for patients presenting with haemoptysis, or patients aged 40+ years with any of the following unexplained or persistent (last for more than three weeks) symptoms or signs: cough, chest/shoulder pain, dyspnoea (breathlessness), weight loss, chest infection, hoarseness, finger clubbing, features suggestive of metastasis from an LC (for example, in brain, bone, liver, or skin), cervical/supraclavicular lymphadenopathy.

For patients whose CXR findings suggest LC or aged 40 and over with unexplained haemoptysis, they qualify for a suspected LC rapid referral – an appointment with a chest physician within two weeks. If the CXR is normal but there is a high suspicion of LC, and the symptoms continue being present in the patient, or new symptoms develop, then further investigation is warranted. A repeat CXR could be ordered, or offer the patients an urgent referral to a member of the LC multidisciplinary team (MDT), usually a chest physician (The National Institute for Health and Care Excellence (NICE), 2019).

2.3.4 Waiting time targets in the cancer care pathway in England

The Department of Health (2011b) set three waiting time targets in the cancer care pathway, including the two-week wait referral, the 31-day and 62-day wait for treatment (Figure 2.4). The two-week wait is to support early cancer diagnosis, meaning that patients with potential cancer symptoms should be seen by a specialist within two weeks of an urgent GP referral for suspected cancer. Patients should start their first definitive cancer treatment within 31 days of a decision to treat, or within 62 days of a GP referral for suspected cancer. The underlying rationale is that the sooner the patients receive their treatment, the better chance they can have a favourable clinical outcome.

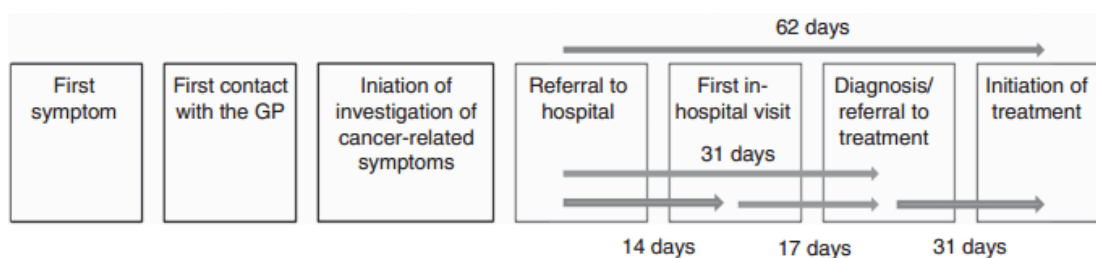


Figure 2.4 – Waiting time targets for referral and treatment in cancer within NHS (Forrest et al., 2014a)

Chapter 2

An English study found that late stage, SCLC, and poor performance status were more likely to meet the recommended target of referral interval (14 days), diagnostic interval (31 days), and treatment interval (62 days). Older patients were significantly less likely to receive treatment within the 31 days (and 62 days) targets. Sicker patients were more likely to attract physicians' attention, they were referred, investigated, diagnosed, and treated more quickly (less system delay), but had worse survival outcomes (Forrest et al., 2014a). This phenomenon was called '**waiting time paradox**' (Torrington et al., 2013) or the '**sicker quicker effect**' (Forrest et al., 2015). A possible explanation was that these patients were in critical conditions or terminally ill, and the presented severe symptoms were more likely to draw physician's attention and discover the underlying aggressive disease, which reflected as a quicker referral, shorter interval, prompt diagnosis, but poorer outcome of shorter survival (Torrington et al., 2013). Studying the primary care pathways to LC diagnosis among this group of patients may help us understand how patients utilised primary care services and how GP responded to their care needs, which may shed some light on this phenomenon.

2.3.5 Section summary

This section reviews how the UK government and research charities addressed the poor LC survival. Health policies are in place. The NICE referral guideline, LC diagnosis and the care pathway for the best practice are summarised. Fast track referral (two-week wait) was introduced and national initiatives were launched to engage the public and raise public awareness of LC symptoms. All of these are the strategies to promote early LC diagnosis. Smoking cessation plays a vital role in prevention, especially for heavy smokers in socioeconomically deprived backgrounds. Due to a lack of an effective screening program in England, early diagnosis of LC still relies on symptomatic presentation in primary care. The next section will be the current research paradigms of the cancer diagnosis and care pathway and discuss their respective limitations. A new research angle – studying primary care sequences/pathways, is proposed at the end of the next section, which is complementary to the current research paradigms.

2.4 Current research models and paradigms of early diagnosis research

2.4.1 The conceptual model of the cancer care pathway

2.4.1.1 The Aarhus statement

The cancer care pathway is a trending research topic, as researchers are interested in knowing the problems and the gaps in the current pathways, and then finding solutions to fix the problems,

addressing the gaps, and optimising the pathways to improve patient outcomes. Studies investigating cancer care pathways often use different ways to define and measure intervals or delays. A scoping review including 65 articles from 21 countries published in English during 2007-2016 found **96 variations of intervals** related to LC diagnosis and treatment (Jacobsen et al., 2017). The lack of consensus on methodology and measurement make it difficult to compare the results across studies from different populations and countries. A panel of experts came up with the Aarhus statement (Weller et al., 2012) to guide researchers in the field of early cancer diagnosis. The Aarhus checklist defined some milestone dates and intervals in the cancer care pathway, including:

- **The date of the first symptom:** defined as ‘the time point when the patient first noticed bodily changes and/or symptoms’. However, it may be difficult for patients to appreciate the onset of the first symptom and recall relevant dates (prone to recall bias), also difficult for physicians or researchers to track back this information from EHR database.
- **The date of the first presentation:** the date that the patient presented in general practice with signs or symptoms probably due to cancer;
- **The date of referral:** the date that the referral letter was sent, which was a transfer of responsibility from a GP to a specialist;
- **The date of the first attendance in secondary care:** the date that the patient was assessed and investigated in an outpatient clinic or hospital admission;
- **The date of diagnosis:** there were four possible ways to determine the date of cancer diagnosis, listed below. Researchers should clearly state which date they use in their studies.
 - 1) when the pathological specimen was reported as malignancy;
 - 2) when the MDT met to make the diagnosis;
 - 3) when the patient was told the results;
 - 4) when the diagnosis from the confirmation letter was coded in the EHRs.

The dates of referral, first attendance in secondary care, and diagnosis can be easily accessed from EHRs, and are relatively accurate. The Aarhus Statement provides general recommendations to study all types of cancer. Considering the symptoms of LC are non-specific and hard to detect, whether it is possible to know the first symptom of LC and the date of the first presentation will be discussed in the final chapter of this thesis, with examples from this study.

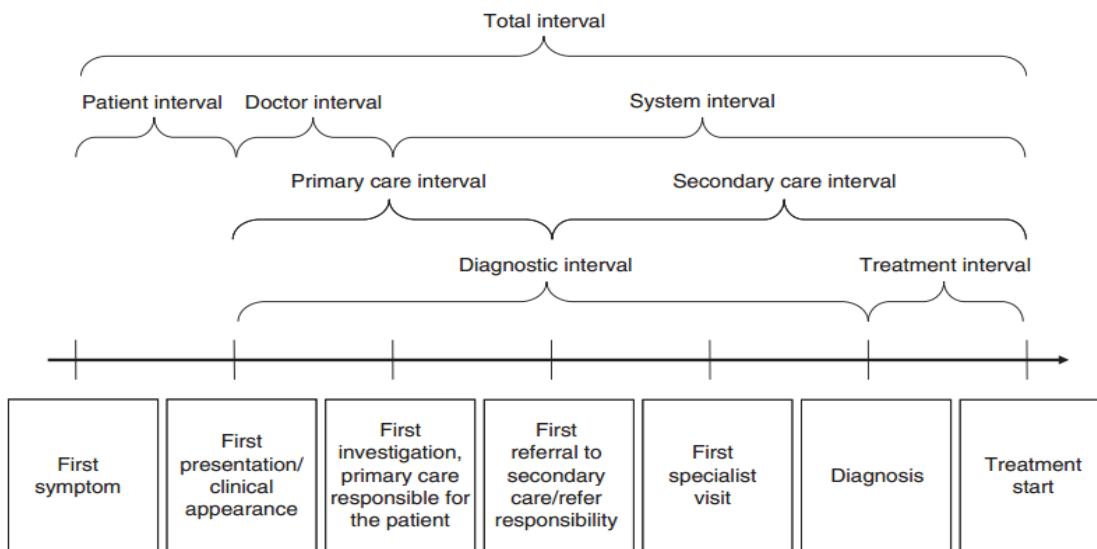


Figure 2.5 – A conceptual model of cancer care pathway illustrating the milestones and intervals from the first symptom to the start of treatment (Olesen et al., 2009)

2.4.1.2 Intervals and delays relevant to primary care investigation

The Aarhus statement recommends using the term ‘interval’ rather than ‘delay’ to describe the duration between any of the two milestone dates. The term ‘delay’ was criticised as judgemental and potentially stigmatising, due to its implications of intent (Dobson et al., 2014). Intervals are the periods between two milestone events, including:

- **Patient interval:** the period from the patient’s first experience of a potential cancer symptom to the first visit in primary care. It is difficult to accurately measure patient interval, as it is hard for patients to recognise the symptoms of LC, thus often not able to remember the date of non-specific symptoms;
- **Primary care interval:** the time between the first symptomatic presentation and the first specialist referral;
- **Referral interval:** the period between referral and first attendance in secondary care;
- **Diagnostic interval:** the duration from patient’s first presentation in primary care to the date of a confirmed cancer diagnosis

Neal (2009) summarised some possible approaches to reduce different intervals:

- **Reducing patient interval:** by increasing public awareness of symptoms and signs of cancer, and knowing how and when to act;

- **Reducing primary care interval:** by increasing the awareness of potential cancer symptoms among GP, and lowering the thresholds for urgent referral and GP-initiated investigations;
- **Reducing system delay:** by improving healthcare infrastructure to shorten the time that the health system needs to process investigations, revising and implementing urgent cancer referral guidelines to shorten patients' waiting time.

2.4.1.3 Current research evidence between diagnostic interval and outcomes (stage at diagnosis and survival)

The rationale for studying pre-diagnostic indicators and primary care characteristics before LC diagnosis (e.g. various intervals, the number of consultations) was that shorter diagnostic intervals could lead to earlier stages at diagnosis and better cancer outcomes (Rubin et al., 2014). The English National Audit of Cancer Diagnosis in Primary Care (2009/10) found that 80% of LC cases (1,200/1,494) had GP-initiated investigation (Lyrtzopoulos et al., 2013). The mean primary care interval was 34.5 days (median 13, IQR 3, 39 days). Compared with patients without primary care investigation, 23.6 more days (95% CI 16.8-30.0) were needed for primary care investigation ($P < 0.0001$), 4.5 more days (1.6-10.1) for referral interval, 28.6 more days (19.4-36.8) for the pre-hospital interval. The number of consultations and primary care investigation had a positive correlation with the median primary care interval. This study only measured the promptness of cancer diagnosis in primary care. It was unclear whether the prolonged interval led to a delay in LC diagnosis.

In another English study, patients with alert symptoms for LC (9.3%) had a much shorter median diagnostic interval (35 days, IQR 17-78 days) than those with non-alert symptoms (99 days, IQR 38-222 days). The 5-year RS for LC patients with alert symptoms (12.0%, 95% CI 9.4%-15.1%) was also higher than patients with non-alert symptoms (7.4%, 6.6%-8.3%), no matter how long the diagnostic interval was (varied from <1 month to >6 months) (Redaniel et al., 2015). For patients with non-alert symptoms, longer diagnostic intervals were associated with lower mortality, as it was likely that this group of patients were still in the early stages of the disease. The longer diagnostic interval was probably because GP applied the 'watchful waiting' strategy to allow the evolution of non-alert symptoms to rule in or rule out cancer.

According to a systematic review, the association between different intervals and the two outcomes (stage at diagnosis and survival) in LC was still inconclusive (Neal et al., 2015b). There were similar numbers of studies reporting positive, negative, and no associations (mixed findings, 20 studies in total). Study design, population, variation in the definitions of different intervals, covariates included, and different statistical models used across studies, may be the possible

explanations why the findings were inconsistent. Future studies following the recommendations of the Aarhus statement may produce a more robust estimate of the intervals, which can help us better understand how different intervals would influence the clinical outcomes of LC.

2.4.1.4 The limitation of the conceptual model of the cancer care pathway

The major limitation of this research paradigm is that **the rich and complex information of primary care consultations over time is simplified as a primary care interval**, which is a number, usually using day as a unit. **What makes it worse is that this interval is not even accurately measured in current research.** A possible new research paradigm is to use concrete primary care events to construct primary care sequences (see the following subsection 2.4.3), and investigate the association between primary care sequences and the clinical outcomes (stage at diagnosis and cancer survival). Primary care sequences can provide richer information than intervals, for instance, to explain the waiting time paradox (subsection 2.3.4). The pre-diagnostic activities in primary care may help explain why patients who had shorter diagnostic intervals had worse clinical outcomes.

2.4.2 The route to diagnosis of LC

'**Route to Diagnosis**' is used to categorise a patient's journey to cancer diagnosis, based on the **setting** and the **referral route** to secondary care. Elliss-Brookes et al. (2012) categorised eight routes to diagnosis for all cancer types using multiple English routine datasets, namely, screen-detected, GP referral, two-week wait, emergency presentation, inpatient admission, outpatient appointment, death certificate, and unknown. The three most frequent routes for LC were emergency presentation (39%), two-week wait (24%), and GP referral (17%), accounting for 80% of all LC cases included in that study (n=96,735). One-year survival for two-week wait and GP referral was around 40%, while emergency presentation was only 12%, much lower than the average of all eight routes combined (29%). Emergency presentation is often related to a poorer prognosis, as it indicates that cancer develops to a late stage.

2.4.2.1 Two-week Wait

The two-week wait referral was introduced in 2000 to improve early diagnosis, which can be seen as a fast-track GP referral. Patients urgently referred for suspected cancer investigation by their GPs can expect to be seen by a specialist within two weeks. The implementation of the NICE urgent referral guidelines for suspected cancer in 2005 resulted in an average reduction of diagnostic interval for 2.4 days in LC (nonsignificant change though, P=0.47). NICE qualifying symptoms for urgent referral had shorter diagnostic intervals than non-NICE qualifying symptoms

(Rubin et al., 2014). Moller et al. (2015) assessed the overall effect of the English urgent referral pathway on cancer mortality in England. After four years of follow-up (215,284 patients from 8049 general practices, 2009-2013), the authors concluded that using the urgent referral pathway was associated with reduced mortality in colorectal, lung, breast, prostate, and other cancers. However, the author did not report each type of cancer individually. The impact of urgent referral on LC survival is still unknown.

2.4.2.2 Emergency presentation

Cancer diagnosis through emergency presentation is that the diagnosis occurs shortly after an emergency/unscheduled/unplanned hospital admission or emergency treatment. The admission could be initiated by patient seeking help through an emergency portal, the acute and emergency (A&E) department, or by a GP referring the patient to the A&E (including out of hours), emergency transfer, or by a secondary physician admitting the patient directly from an outpatient clinic (Elliss-Brookes et al., 2012, Mitchell et al., 2015). If emergency presentation is the result of a direct referral by the GP, technically, the route is 'emergency presentation via GP'. For LC, the percentage of patients diagnosed through emergency presentation was the third highest (39%) among all cancer types, after brain and pancreatic cancers (Elliss-Brookes et al., 2012). Other English studies reported similar percentage in emergency presentation, 36% (n=9601, male 35%, female 38%)(McPhail et al., 2013), 38.5% (62,498/162,543, 2006-2010)(Abel et al., 2015), and 39.4% (145/368) in a Scottish study (Murchie et al., 2017).

Diagnosis through emergency presentation is associated with poorer outcomes, including advanced stages (III, especially IV), less suitable for surgery, shorter survival, and higher mortality (McPhail et al., 2013, Tataru et al., 2015). The proportion of LC patients diagnosed through emergency presentation in stage IV (59%) was substantially higher than that in other common cancers (e.g. breast, colorectal, and kidney, around 30%)(Zhou et al., 2017). Patients emergently admitted were significantly less likely to receive LC surgery (0.1%), compared with those electively admitted (9.7%)(Raine et al., 2010). Emergency presentation was highly predictive of short-term cancer mortality (1 year), especially in the first three months after diagnosis (McPhail et al., 2013). LC cases in emergency presentation could be missed opportunities for possible earlier diagnosis. Emergency presentation was more likely for patients registered at GP practices with poor quality and outcomes framework (QoF) performance, more non-UK qualified GPs, and fewer 48h appointments (Murchie et al., 2017). Patients' sociodemographic characteristics were also associated with emergency presentation. The oldest-old group (85+) had the highest percentage (54%) of emergency presentations (McPhail et al., 2013). Female and patients in the most deprived quintile were at greater risk of emergency presentation in LC (Raine et al., 2010, Abel et

al., 2015). Non-respiratory, atypical, non-red flag symptoms, more comorbidities, patients lack a regular source of primary care, and lower primary care use (no visits in the previous 12 months before diagnosis), and higher secondary care utilisation were identified as risk factors for emergency presentation in reviews (Mitchell et al., 2015, Zhou et al., 2017). The effectiveness of screening programs to avoid LC diagnosis through emergency presentation is unknown, as there is no LC screening programme available for the whole population in the UK at the moment.

2.4.2.3 The limitation of 'route to diagnosis'

'Route to diagnosis' is only concerned with the referral mechanism and how patients arrive in secondary care, which is more like the **destination**, i.e. **the final step of the process, rather than the whole process**. Two-week wait, GP referral, and emergency presentation are the three main routes to LC diagnosis. Taking regular GP referral as a reference, two-week wait is a fast track route, usually having better outcomes, while emergency presentation results in the opposite direction (worse outcomes). Maringe et al. (2018) reported that the proportions of GP-led emergency presentation dropped from 28.3% in 2006 to 16.3% in 2013, whilst patient-led emergency presentation increased from 62.1 to 66.7%. High proportions of emergency presentation in LC were mainly due to more patient-initiated attendances in A&E, rather than the problems within specific bad performance practices. The finding of this study demonstrated the importance of studying primary care pathways to gain a deeper understanding of why patients bypassed primary care and presented in the emergency department. Most patients in these routes had primary care contacts to different extents. Studying the primary care sequences may have the potential of providing new knowledge on why patients end up in different routes to diagnosis (two-week wait, emergency presentation) and why the "missed opportunity" cases happen.

2.4.3 Proposing a new research paradigm – primary care sequence/pathway

Due to the limitations of the current research paradigms (discussed in subsections 2.4.1.4 and 2.4.2.3, respectively), a new research paradigm for early diagnosis research, which is studying primary care sequences/pathways, is proposed and explored in this thesis. **Primary care sequence/pathway** is defined as a series of consultations between patients and HCPs (mainly GPs) in general practice, which happens in chronological order. It includes two elements in each consultation – the reason the patient attends to the surgery and the response of GP gives to the patient. Patient and GP events are interdependent. Such interdependence is relevant in primary care, as help-seeking is generally patient-initiated, compared with more scheduled appointments and elective procedures in secondary care in the English health system (more discussion in

subsection 2.5.1). GP's response is dependent on the patient's presentation and health conditions. Interdependence also means that the patient can revisit the practice if the previous GP action does not work for him/her.

A primary care sequence includes all relevant patient and GP events, rather than milestone events (snapshots) in the pathway. Therefore, the sequence can include more details in the primary care pathways that could have been missed in the conceptual model. Studying primary care sequences preceding diagnosis is particularly important for "harder to suspect" cancers, which was defined using the proportion of patients with multiple consultations, as a measure of diagnostic difficulty (Lyrtzopoulos et al., 2014). LC was classified as "harder to suspect" cancer. 32.8% of patients with potential LC symptoms consulted their GP three times or more before referral (Lyrtzopoulos et al., 2013). Essentially, **non-specific and atypical symptoms for LC** are the main reason for multiple consultations and difficult to suspect LC in primary care. More consultations significantly increase the median referral interval for LC. Multiple consultations and prolonged diagnostic intervals may affect clinical outcomes and patient experiences. Longer diagnostic interval is associated with increased mortality (Rubin et al., 2014). **Studying primary care sequences/pathways retrospectively may provide us new insight on how GP can better manage patients in general practice, how to reduce the number of consultations, shorten primary care interval, and optimise the primary care pathway.** As to the conceptual model, it may be more suitable to be used to study "easier to suspect" cancers (e.g. breast, melanoma, testicular, and others), with fairly specific symptoms (e.g. a lump or a visible mole), and simpler pathways. Furthermore, the conceptual model more focuses on the lengths of different intervals providing numeric information, while primary care sequences can provide information about what exactly happened within these intervals (concrete events). Therefore, studying primary care sequences may elucidate the reasons and the variation in different lengths of the diagnostic interval among patients.

The central argument of this thesis is that **primary care sequence/pathway should be a significant component** in early diagnosis research, **complementary to** (NOT substitute) the current **conceptual model of the cancer care pathway** (Figure 2.5) and 'route to diagnosis'. The next section will explain why primary care plays an important role to achieve early diagnosis of LC in England, and further justify the need to study primary care sequences. Comorbidities and SES are the two constructs that can influence the whole LC care pathway. One research aim in this study is to investigate the relationship between patients' comorbidity burden, SES, and the patterns of primary care sequences. Research evidence between comorbidities, SES, and LC is summarised in the following section.

2.5 Lung cancer diagnosis from primary care in England

2.5.1 The role of primary care in England, and the opportunities and challenges of early diagnosis of LC from primary care

The WHO defines primary care as “first-contact, accessible, continued, comprehensive, and coordinated care” (World Health Organization, 1998). General practice can provide a range of services for common health problems in the community. It is the setting where symptoms are first assessed and managed. Encouraging patients to seek help in general practice in the first instance is more cost-effective, which can reserve hospital resources for patients in more severe conditions. Through continuous care, GP can know the patient’s long-term health problems, build a relationship with the patient, and provide a whole-person, rather than disease-centred care. GP usually acts as the role of coordinator between patients and specialists. However, in countries with a gatekeeper system, the referral process itself might generate inequalities. For example, it is more challenging for socioeconomically deprived patients to navigate the health system. Deprived patients usually have more chronic health conditions and psychological problems than wealthy patients. They wait longer to see a GP, have a shorter time in clinical consultations, and are less satisfied with the experiences (Rubin et al., 2015).

Most residents in the UK have a registered GP. Over 90% of patient contacts occur in primary care, with less than 10% of the NHS budget. It was estimated that about 300 million consultations occurred in general practice in England every year (Rubin et al., 2011). **The broad coverage of the general population, the role of GP as a gatekeeper, and the intensity of contacts, make primary care an ideal setting to facilitate early diagnosis of cancer.** Usually, a primary care consultation tackles ‘one problem at a time’. According to a report by the Royal College of General Practitioners (RCGP) (2019), the duration of GP appointments in the UK was 9.2 minutes on average, considerably shorter than that in other Western European nations (Round et al., 2013). Such a short time does not allow GP to thoroughly obtain a patient’s health history or elicit clinical signs and symptoms in one consultation, which may increase the risk of multiple consultations, prolonged diagnostic interval (subsection 2.4.1.2), delayed investigation and diagnosis. RCGP (2019) suggested that all primary care appointments should be at least 15 minutes long. However, a shortage of qualified GP, heavy workload and time pressure on GP make this recommendation infeasible at the current English general practice.

Some organisational characteristics of the primary care system in England were associated with stage at diagnosis in LC. Easier access to primary care (indicator: fewer patients per GP), younger (aged <50 years) and female GP, efficient use of the referral system (higher percentage of two-

week wait) and faster investigation by GP reduced the proportion of patients diagnosed in advanced stages (III/IV) of LC (protective factors). In contrast, being at a practice with a higher emergency admission rate was associated with a higher percentage of advanced stages at diagnosis (Maclean et al., 2015).

GP needs to deal with a wide range of complaints and symptoms in daily consultations. A major challenge for GP is to determine whether the non-specific symptoms are due to severe health problems (e.g. cancer) or common self-limiting diseases. Clinical decision making in primary care is mainly based on risk estimation. In the presence or absence of symptoms and physical signs, GPs can make a judgment on how likely the assumed disease would be and the severity. If the risk of severe disease is low, then no further action is needed, and the patient could afford a 'watch and wait' approach (Rubin et al., 2015). Time will allow GP to observe whether the undifferentiated symptoms develop to more specific ones or resolve spontaneously. In this sense, time is a diagnostic tool. But if the risk is high, prompt intervention is warranted. Bjerager et al. (2006) summarised some **possible GP actions after a patient presenting with respiratory symptoms**: treatment for lung infection, intensified treatment for chronic diseases, treatment for a new suspected disease, acute admission, referral to CXR, referral to an outpatient clinic, referral to physiotherapy, blood tests, other investigations, wait and see.

The main investigation for suspected LC in primary care is **CXR**, which is cheap and accessible to GP. But the biggest problem is in its high false-negative rate to detect LC. Stapley et al. (2006) reported 23% (38/164) of patients diagnosed with LC, whose CXRs were negative. A systematic review found there was still a paucity of evidence on examining the sensitivity of CXR in detecting LC among symptomatic patients. The highest-quality studies (only three) suggested that the sensitivity of CXR for symptomatic LC was 77%-80% (Bradley et al., 2019a). No particular symptom was significantly associated with negative CXRs. A CXR reported as normal could be truly normal. It was also possible that the lesion was too small for the radiologist to visually identify, or the tumour hidden behind the intra-thoracic structures or the skeleton. GP considered CXR as a "blunt instrument" to investigate potential LC. Further investigation may be needed in high-risk patients who have a negative result of CXR. They expressed a need to direct access to more sensitive diagnostic tools in primary care, like CT scans (Wagland et al., 2017).

2.5.2 The difficulty of early LC diagnosis – lack of specific symptoms

Possible LC symptoms include persistent or unexplained cough, haemoptysis, dyspnoea, chest pain, chest/lung infection, hoarseness, unexplained weight loss, anorexia (appetite loss), fatigue, finger clubbing, shoulder pain, upper back pain, stridor, superior vena cava obstruction (Del

Giudice et al., 2014). Cough is the most common symptom seen in primary care, very often with repeated attendances. As the first symptom, cough is present in nearly a quarter of patients with LC. The positive predictive value (PPV, the probability that subjects with a symptom or a positive screening test truly have the disease) of cough for LC increases with each attendance, but remain <1% after three presentations. If there are no other suspicious symptoms along with cough, both patient and physician can adopt a 'watchful wait' strategy to allow the disease to become clearer (Hamilton et al., 2005). Respiratory symptoms with a low risk of LC increase time to diagnosis, and are more likely to be benign respiratory diseases rather than LC (Walter et al., 2015). Due to the non-specific symptoms, GP may not necessarily link the common symptoms with LC. For chest and shoulder pain, GP may initially consider cardiac or musculoskeletal diseases. For hoarseness, laryngeal cancer is more likely than LC. Examination of the vocal cords would be the priority (Hamilton and Sharp, 2004).

Patients being referred due to suspicion of LC often have complex symptomatology. Synchronous first symptoms are common (Walter et al., 2015). Cough, dyspnoea, and haemoptysis (coughing up blood) are the three most common symptoms, either reported by patients or recorded by GPs (Shim et al., 2014). Symptoms and signs with a PPV $\geq 5\%$ are regarded as highly predictive of cancer. A systematic review (Shapley et al., 2010) found that **only haemoptysis** have a PPV $>5\%$ for LC, which is the symptom consistently indicated as a predictor of LC. Other symptoms independently associated with LC diagnosis include appetite loss, weight loss, fatigue, and fever/flu (Hippisley-Cox and Coupland, 2011, Shim et al., 2014). Having a second symptom with haemoptysis markedly increases the PPV for LC, especially recurrent haemoptysis (Hamilton et al., 2005). However, only 4.6% of patients diagnosed with LC presented haemoptysis as the first symptom in primary care (Walter et al., 2015). A systematic review found that **the diagnostic values of most symptoms for LC are inconclusive**, and there is insufficient evidence to suggest a symptom profile for LC across different stages (Shim et al., 2014). **Diagnosis of LC based on symptoms is difficult**. Vedsted and Olesen (2015) argued that it is important to consider the symptom epidemiology throughout the pathway, because it has implications on how GPs interpret the presentation of symptoms and the decisions around patient management and investigation. Their opinion echoes **the emphasis of interdependent patient-GP events in primary care sequences in this thesis**.

2.5.3 Comorbidities and LC

2.5.3.1 The rationale on how comorbidities could influence cancer diagnosis and survival

Comorbidity was defined as the coexistence of health conditions in addition to the primary disease of interest (Feinstein, 1970). There are several reasons why other diseases coexist with cancer. First, many comorbidities and cancers share common risk factors, like older age, smoking, heavy alcohol drinking, poor diet, lack of physical exercise, inactivity, obesity. Second, the biological mechanisms associated with comorbidity may predispose to cancer. Chronic infections, diseases in the immune system, and diabetes mellitus are associated with an increased risk of cancer (Sarfati et al., 2016). Comorbidity could influence the morphology, histology, differentiation, proliferation status, the growth of cancer, and affect the prognosis (Islam et al., 2015). Comorbidity can influence the timing of cancer diagnosis in two ways. Patients may have more contact with HCPs, thus increasing the chance of getting an early diagnosis. Conversely, comorbidity may distract either the patient or HCPs (or both) from the primary disease (cancer), resulting in delayed diagnosis (Sarfati et al., 2016).

Comorbidity is an important moderator for cancer outcomes. Evaluating the impact of comorbidity on cancer outcomes in a causal relationship is difficult, as different comorbidities may interact with each other through various mechanisms in a very intricate and complex human body (Islam et al., 2015). There is no gold standard (Sarfati, 2012, Sharabiani et al., 2012), nor validated tool available (Grose et al., 2011) to assess the comorbidity burden specific to LC, although the Charlson comorbidity index is the most widely used instrument to measure the disease burden of patients. The common research practice is to cut the total comorbidity score into two or three categories, compare with the reference category of “no comorbidity”. However, the Charlson comorbidity index does not measure comorbidity as precisely as clinical data, as it only includes less than 20 health conditions, while there are more possibilities in the real world. Furthermore, the Charlson comorbidity index may be outdated now. For example, the weighting for HIV/AIDS is probably too high, as treatments for HIV/AIDS have been significantly improved, compared with the time when the Charlson comorbidity index was created (the 1980s).

Comorbidity is found to have an adverse impact on cancer survival, although the magnitude of the impact was different among studies. It depends on how comorbidity is operationalised, study population, cancer type, cancer stage, treatment modalities, and the outcome measure (Sarfati et al., 2016). The number and severity of comorbidities at the time of cancer diagnosis strongly influence the probability of dying from non-cancer causes, and also may influence cancer-specific survival (Edwards et al., 2014). For potentially curable NSCLC, patients who are offered surgery have fewer and less severe comorbidities, while patients who receive the least active palliative

treatments have the most severe comorbidities (Grose et al., 2014). It means that comorbidities, patient's age, and performance status have a significant impact on the treatment modality, which would influence patient's survival outcome. The Liverpool Lung Project also found that the severity of comorbidity (based on the Charlson comorbidity index) was associated with higher LC specific mortality (Marcus et al., 2015).

2.5.3.2 COPD and LC

COPD is characterised by irreversible airflow obstruction in the lungs and symptoms related to decreased expiratory volume. Breathlessness, cough, sputum, and wheeze are common symptoms for both COPD and LC. Such symptoms are more likely due to chronic respiratory diseases than LC in the community setting. Spirometry is essential to confirm the diagnosis of COPD, using the Global initiative for chronic Obstructive Lung Disease (GOLD) criteria to determine the grade of clinical severity (Ytterstad et al., 2016). Chronic bronchitis and emphysema are two phenotypes of COPD. Chest CT can be used to diagnose emphysema and the existence of LC (Mets et al., 2011). COPD in advanced stages tends to be treated as chronic rather than terminal conditions. When a patient is diagnosed with a localised LC, other comorbidities, like COPD, is often considered less important (Jeppesen et al., 2016).

Some researchers argued that COPD and LC might be different stages or manifestations of the same disease, shared with common aetiology and mechanisms, such as ageing in the lungs, smoking, and genetic predispositions (Durham and Adcock, 2015). The normal decline in lung function in ageing is accelerated in patients with COPD leading to premature loss of lung function. Inflammatory diseases can predispose to cancer. Almost all cancerous tissues show inflammation. It is possible that local pulmonary and systemic chronic inflammation in COPD is a potent driver for LC development, as chronic inflammation causes repeated tissue damage, stimulating cell division to restore homeostasis. Inhaled corticosteroids are anti-inflammatory drugs for COPD treatment. It also decreases the risk of LC in a dose-response relationship (Huang et al., 2015). Linkage studies have implicated regions in chromosome 6 linked to both diseases. Genome-wide association study (GWAS) in large COPD and LC cohorts have found the same risk loci (specific physical locations of a gene or DNA sequence on a chromosome)(Durham and Adcock, 2015).

A systematic review reported that the prevalence of COPD in patients with LC ranged from 28%-40% and emphysema ranged 47-76%, in studies specifically designed to investigate the relationship between COPD and LC. COPD is an independent risk factor for LC. Patients with COPD have approximately 4-6 times of risk to develop LC, particularly for squamous cell carcinoma (Raviv et al., 2011), compared with smokers with normal lung function. The risk increases with the decline of FEV1 (Forced Expiratory Volume in 1 second, a proxy of airway obstruction),

independent of age and smoking history (Mouronte-Roibas et al., 2016). An English study found that a recent diagnosis of COPD (within six months) was strongly associated with a subsequent diagnosis of LC. However, the odds of getting an LC diagnosis dropped rapidly after six months of COPD diagnosis, and remained relatively stable for 5-10 years after the diagnosis of COPD (Powell et al., 2013). **Ascertainment bias** might overestimate the magnitude of the association, as patients with suspected LC symptoms were likely to undergo more clinical assessment and intensive investigations than those without LC symptoms. This may lead to a diagnosis of COPD a few weeks or months before the diagnosis of LC. In addition, patients with COPD were more likely to have more contact with health professionals and monitored more regularly, thus more likely to have a subsequent LC diagnosis (Guldbrandt et al., 2017). Therefore, studying primary care sequences of **the pre-diagnostic activities** (like lung function tests, prescriptions of antibiotics and/or COPD medications), **we can know how GP manages these patients in primary care, which may further elucidate the relationship between patients' comorbidity status (especially COPD), primary care sequences, and stage at diagnosis.**

Patients staged in T₁₋₂N₀M₀ of NSCLC may be eligible for surgery to remove the tumour (Jeppesen et al., 2016). The existence of COPD can worsen the outcomes of the 5-year overall survival and progression-free survival for patients diagnosed with NSCLC in early stages (IA-IIB) and treated with surgical resection, particularly in men and squamous cell carcinoma (Zhai et al., 2014). In addition, older patients with severe COPD are unlikely to tolerate pneumonectomy for LC, as they may not have good cardiopulmonary functions and enough ventilation reserve (Sarfati et al., 2016). Given the relationship between smoking and respiratory complications in the postoperative period, smoking cessation before surgery has been proposed to minimise the risks of postoperative complications of LC surgery (Raviv et al., 2011).

Smoking cessation is a crucial intervention available in primary care to manage patients with COPD and prevent disease progression, irrespective of the disease stage of COPD. Socioeconomic deprivation is a key factor in the prevalence of smoking, COPD and its severity. The prevalence of COPD is twice among people living in socioeconomically deprived areas, but the smoking cessation rate is much lower than those living in less deprived areas (Simpson et al., 2010). Besides smoking cessation, self-management programmes and vaccination against influenza have been proven cost-effective and can improve patients' quality of life with COPD. Inhaled bronchodilators and corticosteroids can improve respiratory symptoms and reduce the risk of COPD exacerbation (Broekhuizen et al., 2012). These are the GP actions for mediating the adverse effect of smoking, managing COPD and LC prevention, and are common events in the primary care pathways for this patient group. This part was included and investigated in the main study.

2.5.4 GP's perspectives on early diagnosis of LC

Not all patients presented with respiratory alarm symptoms lead to a referral for further examination, prescriptions, or malignant diagnoses (Sele et al., 2016). A vignette study found that **GPs did not investigate everyone with the same symptoms equally**, nor more likely to initiate cancer investigations for patients with higher risk symptoms. **The decisions to investigate LC were more influenced by whether GPs sought out relevant clinical information about the presence of symptoms.** Even when GPs elicited sufficient information about symptoms, there remained inequalities in the decisions of LC investigation in patients' age and ethnicity. Older patients and black ethnicity were less likely to be investigated than younger and white patients in an English study (Sheringham et al., 2017).

Qualitative studies were conducted to gain GPs' perspectives and insights in early detection and LC diagnosis from primary care. GPs mentioned that the burden of early cancer detection in general practice was intensified by a perceived fragmentation of services and care within the NHS, due to the increased number of part-time and salaried GPs, and locum cover, which made it more difficult for patients to see the same doctor (less continuity of care) (Green et al., 2015). It is difficult to elicit the symptoms from patients. Some patients do not recognise symptoms, or perceive symptoms as normal, or do not report symptoms in consultations, whilst some may "change the stories" of symptoms between consultations. All these situations make it difficult for the GP to judge the severity of symptoms that the patient experience. GPs perceived the three most relevant symptoms for diagnosing possible LC: significant recent weight loss, persistent cough for longer than six weeks, and haemoptysis. However, patients who rarely attended the practice, but suddenly presented with symptoms, would trigger GP's concern (Wagland et al., 2017). Patients who presented with vague symptoms and turned out to be cancer eventually, were more likely to receive a late diagnosis because their symptoms did not meet the NICE urgent referral criteria. In such cases, the referral criteria became a barrier for early diagnosis (Green et al., 2015). In the absence of red flag symptoms, GP's hunch/gut instinct developed through clinical experience, played an important role in their ability to identify patients who should be referred for further investigation of LC (Green et al., 2015, Wagland et al., 2017). GP practicing in more affluent areas recognised an increased number of consultations after the public health initiatives, whereas those working in more deprived areas perceived that many of their patients were less affected by the initiatives (Green et al., 2015).

There are some possible solutions to address these issues. GP can allocate additional consultation time for infrequent attendees presented with symptoms and patients with multimorbidities.

Safety netting is a diagnostic strategy used in the UK primary care to ensure patients are

monitored until their symptoms or signs are explained, regarded as the best practice to protect against inaccurate working diagnoses (Evans et al., 2018). GP should communicate with the patient the uncertainty of any diagnosis, and discuss with patients what to do and when to seek further help if symptoms come back after a negative investigation (Round et al., 2013, Rubin et al., 2015). GP may feel more responsible to ensure patient's complaints are followed up, if familiar with the patient and his/her health problems (Ridd et al., 2015).

2.5.5 Patients' help-seeking behaviours and influencing factors

Patients' help-seeking behaviours are of research interest, as it is usually patient initiates the visit to general practice in the UK. Seeking help for symptoms is a complex cognitive, psychological, and behavioural process, involving perception, interpretation, and appraisal of symptoms, decision-making, and the motivation to transform the decision into action by visiting a healthcare professional (HCP) (Scott and Walter, 2010). Symptoms could be a subjective interpretation of bodily changes, not necessarily indications of an underlying disease (Elnegaard et al., 2015). Not every detected or experienced symptom is interpreted as an illness by patients. Even interpreted as illness, it does not mean the patient believes it warrants medical attention. There is a gap between the intention and the actual behaviour of visiting an HCP. Help-seeking behaviour not only concerns the decision of whether to seek help or not, but also the timing of making the decision and transforming that decision into action.

Campbell and Roland (1996) conducted a literature review to understand patients' pathways to care and the factors associated with low and high consultation rates. They found a wide range of sociodemographic, psychological, and health systems factors that could influence patients' help-seeking behaviours, specifically broken down as, demographic (age, sex, ethnicity, education), socioeconomic (social class, employment status, income, housing tenure, distance from house to a health facility), family and social network (marital status, social support), cognitive and psychological factors (medical knowledge, perceived susceptibility and severity of symptom/illness, benefits and costs of seeking medical care, the effectiveness of self-care, stressful life events), healthcare provider and system attributes (accessibility, universal health system or otherwise, coverage by health insurance, affordability of medical bills, availability of appointments, queueing and waiting time, trust of health providers, communication between patient and HCP). At least 70 different factors have been shown to play a role in help-seeking behaviour (Scott and Walter, 2010), which would be impossible to study all these factors and quantify the influence of each factor in a single study. In this PhD study, some important sociodemographic variables are available from the NAEDI dataset that could be used to investigate their association with patients' help-seeking behaviours.

Chapter 2

Besides doing nothing at all, there are two main ways for patients to manage emerging symptoms. One is seeking information and leading to self-management, including looking for information about the experienced symptoms, discussing with friends and family, and taking the over-the-counter (OTC) medications. The concept of 'symptom iceberg' describes the situation that people manage most symptoms in the community without seeking professional care (Last, 1963, Hannay, 1980). Another approach is seeking professional advice via different channels, e.g. phone NHS 111¹ (a 24/7 telephone line and online service in the UK) for health advice, consulting with a GP or a nurse in general practice, a local pharmacist, or a complementary therapist (could be private)(Elliott et al., 2011). Researchers can only study the proportion of symptoms presented in the care settings, often documented in the EHR databases nowadays. This is the visible part of the iceberg based on the 'symptom iceberg' theory. It is not possible to know the bigger proportion of the 'submerged' iceberg, i.e. patients manage the symptoms by themselves, without consulting an HCP.

Patients considered episodic, non-specific, non-progressive LC symptoms as part of 'everyday fluctuations' of body functioning (Corner et al., 2006), or normalised as the ageing process or lifestyle (Brindle et al., 2012). It was more difficult for patients who had comorbidities to become aware of a new and different illness. Patients may only seek help when the symptoms were escalated and affected their daily life, or they fail to resolve the symptoms by themselves and could not tolerate any more (Corner et al., 2006, Elliott et al., 2011, Whitaker et al., 2016). Recognition of warning signs is associated with faster help-seeking for potential cancer symptoms (Quaife et al., 2014). Long-term heavy male smokers, older people with more comorbidities, those from lower SES backgrounds and living alone, often have lower expectations of good health, lower health literacy and knowledge of cancer symptoms. Thus, it often takes them longer to consult potential LC symptoms. Even realising something is wrong in their bodies, they may choose to avoid help-seeking if they fear the adverse outcome of consultations (that they have a serious underlying disease)(Whitaker et al., 2015). Approaches to promote early presentation should aim to increase the awareness of important cancer symptoms (Forbes et al., 2014), and social campaigns should be more targeted at the population with the aforementioned sociodemographic characteristics. Family members (especially partner), friends, and other social networks may notice patient's symptoms and have an influence on the sanctioning of symptom seriousness and help-seeking, directly or indirectly forcing the individual to consult with HCPs who may otherwise be reluctant (Smith et al., 2005, Smith et al., 2009, Chatwin and Sanders, 2013).

¹ As 'NHS 24/NHS Direct' published in the article (2011). NHS 24 is a similar service as NHS 111 in Scotland, while NHS Direct was established in March 1998, discontinued on 31 March 2014, and replaced by NHS 111.

Patients with haemoptysis, breathlessness, cough, loss of appetite, having a history of chest infection and renal failure, or previous diagnoses that required hospital treatments were more alert to symptoms and more likely to seek help quickly (Smith et al., 2009, Hannaford et al., 2020). Quaife et al. (2014) reported that people with higher education (degree or above) had a greater delay (>2 weeks) in help-seeking for persistent cough, because they were too busy.

Perceived blame, lecturing, or reprimand prevents patients with potential LC symptoms from seeing their GPs promptly (Corner et al., 2006, Walton et al., 2013). Smokers might feel ashamed or fear being stigmatised, judged, or ignored (Chapple et al., 2004), and choose not to contact GP. Due to the smoking habits and/or perceived SES, some patients feel unworthy of health care. For the most socioeconomically deprived patients in the UK, they need to overcome difficult life circumstances and the challenges of living with no or minimal income. They may need to work for long hours, have family issues, care responsibilities, or other burdens. When competing with other life demands, patients may put health in a lower priority (Dixon-Woods et al., 2006, Smits et al., 2018).

The worry of wasting doctors' time is common among British patients (Rubin et al., 2015). Thus, patients may not fully use the consultation time to communicate the symptoms they experience with their GPs. Therefore, some missed opportunities for earlier diagnosis may be due to insufficient symptom elicitation and ineffective doctor-patient communication. Patient engagement campaigns could boost patients' confidence to fully communicate their health problems with their GPs and not feel guilty of wasting doctors' time. Using plain language (like aches, out of breath) is more likely to elicit symptoms than disease terminology (e.g. breathlessness, or even, dyspnoea)(Brindle et al., 2012). Providing translation services for ethnic minority patients who are not competent in using English in consultation would enhance GP-patient communication (Lyratzopoulos et al., 2015).

In summary, patient mediated factors that may influence the delayed presentation and prolonged help-seeking in LC include not recognising the seriousness of symptoms or signs (cognitive), fear of receiving a cancer diagnosis (emotional), and reluctance to interact with the health system (psychological and behavioural). Symptom characteristics (severity, duration, and interference with daily life) that have a negative impact on daily life are *protective* factors for *decreasing* delay. How patient characteristics (age, sex, SES, education level, marital status, comorbidities) would help explain the different patterns of primary care sequences/pathways was investigated in the empirical analysis of this PhD study, reported in Chapter 7.

2.5.6 Patient's SES and LC

2.5.6.1 The rationale on how patients' SES could influence cancer diagnosis, treatment, and survival

Socioeconomic status (SES), or socioeconomic position (SEP), is a latent construct related to the social and economic factors that influence the positions an individual holds within a society, which may vary in different stages across one's life (Galobardes et al., 2006a). SES is relevant in the context of cancer because it is associated with and influences various aspects of cancer. Health behaviours, such as smoking, alcohol drinking, physical exercise, and diet, are socially patterned (Singh-Manoux and Marmot, 2005). SES may interact with comorbidity that could influence cancer development. Patients in lower SES usually have a lower level of health literacy and symptom awareness, less able to navigate the health system, which could influence their help-seeking behaviours and health services utilisation. Patient-GP communication and symptom elicitation may indirectly influence GP's responses to patients' presentation, which may have an impact on delayed referral and diagnosis. Comorbidity burden, performance status, and stage at diagnosis would influence treatment modalities, also associated with survival outcome. Socioeconomic inequality in cancer survival and prognosis has been observed and well-documented (more in subsection 2.5.6.5). Patients in lower SES have a higher percentage of premature death from cancer than those in higher SES. Each step of the cancer care pathway is closely linked, and often unfavourable to patients in lower SES. Therefore, the national campaigns should target and engage people from disadvantaged socioeconomic backgrounds, increase their awareness of cancer symptoms, and encourage them to seek help more promptly. **Socioeconomic equality is important in the UK because the NHS is a universal health system. It should be accessible and equal to anyone who needs it, irrespective of one's SES.** Furthermore, socioeconomic equality is not only a concept in health services, but also an important social construct.

SES is generally measured by a composite index of proxy indicators, such as education, occupation, employment status, income, family wealth, residential value. Education level substantially influences other socioeconomic measures (occupation, employment status, income, and wealth). Income directly measures the component of material rewards, and strongly influences one's purchasing power in housing and health expenditure. Housing is a key component of material resources in most people's wealth. Housing tenure, location, conditions, and household amenities are extensively used as measures of SES, which are comparatively easy to collect for research. Neighbourhood location and quality (e.g. access to facilities for physical

exercises and healthcare providers within the proximity of neighbourhood) also influences one's health behaviours and outcomes (Galobardes et al., 2006a, Galobardes et al., 2006b).

2.5.6.2 How SES is operationalised in English health research

Individual socioeconomic indicators are usually unavailable in health databases. Area level indicator (ecological) is often used as a proxy of patients' SES. The Index of Multiple Deprivation (IMD) is the official measure of relative deprivation in England, updated periodically, based on the postcodes of residence within small administrative geographical areas (Lower Super Output Area, LSOA). The whole England was divided into 32,482 LSOAs in IMD (2015), with an average of approximately 1,500 residents or 650 households per LSOA. The IMD is a composite measure using census and administrative data to calculate the score from 7 domains with different weights, including income deprivation (22.5%), employment deprivation (22.5%), education, skills and training deprivation (13.5%), health deprivation and disability (13.5%), crime (9.3%), barriers to housing and services (9.3%), and living environment deprivation (9.3%). The small areas are ranked in a continuum from the most deprived to the least deprived by scores, and then mapped into the corresponding geographical areas to facilitate public resource allocation (Department for Communities and Local Government, 2015). Area level measure holds the assumption that small areas are comparatively socioeconomically homogeneous. However, even for people living in the same area, their SES may be different. It may misclassify individual SES at area level. The larger the area is, the greater chance of misclassification tends to be. In research, the IMD is operationalised as an ordinal variable, equally divided in quintiles by the rank of LSOAs. Some English studies used Townsend or Carstairs index before 2010. Only a few studies used professions by the Registrar General's classification to reflect social class (Hart et al., 2001, Neal and Allgar, 2005). Patients' education level, employment status, and IMD (rank and further categorised as quintiles) were collected in the NAEDI study and used in this thesis.

2.5.6.3 SES and LC diagnosis and intervals

Patient's SES was thought to be an important factor in the differences of the stage at diagnosis and diagnostic interval of LC. However, according to a systematic review and meta-analysis, there was no evidence of socioeconomic inequalities in late-stage at diagnosis, compared the most with the least deprived group (OR=1.04, 95%CI 0.92-1.19)(Forrest et al., 2017). No evidence of socioeconomic inequalities in the patient interval (from symptom recognition to presentation) was found, nor the treatment interval (from diagnosis to treatment). The 'sicker quicker' effect (subsection 2.3.4) may cancel out socioeconomic related delays, which might otherwise result in longer intervals among more deprived patients. There was no consistent pattern of referral and diagnostic intervals (still inconclusive).

Chapter 2

The findings from Forrest et al. (2017) may not be directly applicable to the UK because of the following methodological concerns. Firstly, although half of the studies (20 out of 39) included in the systematic review were from the UK, there were still substantial studies from countries with a non-universal health system (like the US). Considerable heterogeneity existed ($I^2=60\%$) when pooling the results from universal (5 studies) and non-universal health systems (2 studies) in the meta-analysis. I^2 statistic describes the percentage of variation across studies due to heterogeneity rather than chance (Higgins and Thompson, 2002). Secondly, many studies included in the review were not high quality, and may suffer from recall bias and ascertainment bias in observational studies, and potential publication bias. Thirdly, very few included studies had good quality data to examine intervals of the LC diagnostic pathway, which made it difficult to compare the results across studies. The authors of the original studies only investigated some intervals of the whole trajectories (fragmented information). It would be more informative to investigate the whole disease trajectory, which is what I propose to do in this PhD study. Fourthly, **the lack of complete cancer staging data** is the biggest methodological concern. Fifthly, only about 20% of patients were diagnosed at an early stage in the UK. The meta-analysis may be underpowered to detect the differences between early and late stages by SES. Finally, there was no study examining primary care interval (from first symptomatic presentation to first specialist referral), which is a research gap, partly because LC symptoms are non-specific, and thus, it is difficult to determine “the first symptom”.

2.5.6.4 SES and LC treatment

Socioeconomic deprivation was significantly associated with diagnosis following an emergency hospital admission and a lower surgical rate for LC (Raine et al., 2010). It was consistently reported that the most deprived patients (the bottom quintile of IMD) were less likely to receive LC surgery in the English population, although the inequality narrowed over time (McMahon et al., 2011, Berglund et al., 2012, Forrest et al., 2013, Forrest et al., 2014a, Forrest et al., 2014b, Nur et al., 2015b). Compared with LC surgery, the evidence of the receipt of chemotherapy or radiotherapy among patients in different SES was less consistent across studies.

Patients in the lowest SES were associated with a reduced likelihood of receiving any kind of treatment for LC (OR=0.79, 95% CI 0.73-0.86, $P<0.001$), surgery (31 papers included, OR=0.68, 0.63-0.75, $P<0.001$), and chemotherapy (23 papers included, OR=0.82, 0.72-0.93, $P=0.003$), but not radiotherapy (18 papers included, OR= 0.99, 0.86-1.14, $P=0.89$), reported in a systematic review and meta-analysis (Forrest et al., 2013). These inequalities could not be explained by the differences in cancer stage, or health system (universal or not), controlling for comorbidity. Inequalities in receipt of treatment may contribute to inequalities in cancer survival.

2.5.6.5 SES and LC survival

The relationship between SES and survival outcomes is complicated, very often mediated by other factors. Socioeconomic inequalities in LC survival were continuously observed. The most recent study reported that the deprivation gap of 1-year net survival in LC between patients in the most and the least deprived quintile in males did not narrow from 1996 to 2013 in England (remained the same level, -4%), but notably widened in female, from -3.7% in 1996 to -4.8% in 2013 (widened 1.1 percentage points), although the overall 1-year net survival was gradually improving (Exarchakou et al., 2018). Females had a larger observed improvement in the adjusted 1-year survival, from 27.2% in 1996 to 45.8% in 2013; and from 25.9% to 38.6% in males. The improved LC survival could be due to a higher number of specialised surgeons available to perform LC surgery, which increased the proportion of patients receiving surgery (17% in 2015). In addition, more patients were managed in specialised centres, which reduced postoperative mortality.

The deprivation gap in the 1-year survival was even worse among younger patients (over 10% in patients aged 15-44 years) in both sexes (Nur et al., 2015a). Socioeconomic inequalities in LC survival could be statistically explained by the receipt of treatment, but not by the timeliness of referral (two-week wait target) or treatment (31 days from diagnosis target)(Forrest et al., 2015). The number of comorbidities and performance status varied among patients in different SES, which might explain the inequalities in receipt of treatment and survival.

Loss in life expectancy due to cancer is another way to measure cancer burden and the negative impact of cancer on the rest of life in a person, which is defined as the difference of life expectancy between the general population free of cancer and the patients diagnosed with cancer (Syriopoulou et al., 2017). Compared with other cancers, LC had the largest overall loss in absolute life years and the proportion of expected life remaining. Over 30,000 total life years were lost among patients diagnosed with LC in England in 2013. Male LC patients in the least and most deprived quintile lost 12.8 and 11.8 years on average, respectively, equivalent to 87.5% and 88.1% of their average remaining expected life years; and female LC patients lost 14.4 (86.1%) and 13.8 (88.2%) years respectively. Generally, patients in the least deprived group were expected to have a higher life expectancy than those in the most deprived group, which meant the least deprived patients had more life-years to lose due to cancer (Syriopoulou et al., 2017). This study provided a new perspective to measure the cancer burden by the loss of life expectancy. However, it was unclear how the cancer stage and other important variables would impact on the loss of life expectancy.

In summary, the associations between SES and diagnosis (intervals and stage), treatment and survival of LC have been widely explored in the UK. Some associations are more consistent than

others. **The complexity of socioeconomic inequalities in LC diagnosis and survival justifies further health services research to understand how patients in different socioeconomic groups utilise primary care services for their symptoms and comorbidities, and the relationship to the 'route to diagnosis', the lengths of primary care and diagnostic intervals.** The empirical part of this PhD study explored the differences in patients' help-seeking behaviours and the patterns of primary care sequences among patients in different socioeconomic groups.

2.5.7 Missed opportunities for early LC diagnosis and the lessons learnt from audit

Missed opportunities refer to the cases that alternative medical decisions or actions could have been made for more timely diagnosis. They may occur in any part of the diagnostic pathway and may involve patients, physicians and the care team, and health system factors. However, not all delayed diagnoses are missed opportunities (Lyrtatzopoulos et al., 2015).

Significant event analysis and performance review through audit are useful methods for quality improvement to enhance patient safety and care in British primary care (Rubin et al., 2015). Analysis of 132 significant events in LC diagnosis from 92 general practices in the North of England Cancer Network was conducted to gain insights of the LC diagnostic process (Mitchell et al., 2013). The lessons from these significant events included GPs need to be vigilant of patients with atypical symptoms and comorbidities, the limitations of CXR as a diagnostic tool for complicated situations, the importance of safety netting, and patient education about smoking cessation and cancer symptoms. Mitchell et al. (2013) summarised what a fairly typical pattern of presentation and referral pathway of LC diagnosis would look like: presentation of chest-related symptom → initial treatment → GP review if no improvement → CXR and report → referral. A reasonable interval for the whole process would be one month. However, diagnostic pathways in real life rarely follow this pattern.

2.5.8 An empirical study investigating LC diagnostic pathways from primary care

There is only one study looking at the current LC diagnostic pathway. Case reports of 118 patients with LC diagnosis from 96 Welsh general practices (one or two cases per practice) were collected by incentivised GP using a standard template (Neal et al., 2015a). Ninety-six patients (81.4%) presented with respiratory symptoms; 79 patients (66.9%) had GP-initiated CXR before diagnosis; 23 CXRs did not initially show suspicion of LC (false negative of CXR); 25 patients (21.2%) were diagnosed after a GP-initiated acute admission; 14 patients (11.8%) with CXR reported as normal were subsequently diagnosed with LC (misleading CXR). Half of the patients had three or more

consultations before referral or investigation. The authors **manually** categorised 11 mutually exclusive diagnostic pathways for LC. The three most frequent pathways were:

1. Symptoms → (patient visited) GP → CXR → referral (n=72, 61.0%);
2. Symptoms → GP → admission (n=22, 18.6%);
3. Symptoms → GP → no CXR → referral (n=7, 5.9%);

This study provided some useful information about primary care activities related to LC diagnosis, and the authors concluded that the **pathways to LC diagnosis were often not straightforward and rarely linear**. Small sample size and selection bias were the two major limitations of this study. Cases in this study were less likely to be representative, as the researchers asked GPs to report 1-2 recent cases of diagnosis per practice. In addition, the pathways were more like a snapshot, selecting key events, rather than presenting a holistic perspective of the whole diagnostic pathway from primary care. **The complexity of LC warrants more research in the primary care pathways/sequences.**

2.5.9 Section summary and the rationale of this PhD: why studying primary care sequences/pathways for LC is needed in England at this moment

This section provides a comprehensive overview of the current situation of LC diagnosis in England, and established the opportunity for early diagnosis from primary care, in the context of a universal health system in the UK (great coverage of the population, GP as a gatekeeper, the intensity of primary care contacts, and the setting where most symptoms are first assessed and managed). Due to the lack of an effective screening programme for LC in England, LC diagnosis is still based on symptomatic presentation in primary care. However, because of the non-specific symptoms for LC, it is difficult for patients and GP to recognise LC at the early stages. Patient and GP mediated factors for delayed LC diagnosis and how to avoid missed opportunities are summarised from the literature. Current research in this field mainly focuses on investigating the length of different intervals in the diagnostic pathway, route to diagnosis, and risk factors associated with delayed LC diagnosis and poor survival. Their respective limitations are discussed in the previous subsections. **Primary care process (patient's help-seeking patterns, the recorded reason of patient's attendance to the general practice, and how GP responded to patient's visits over time) is less investigated**, probably due to **the complexity of interdependent patient-GP events in the primary care pathways** related to LC and **the lack of proper statistical methods to cope with such complexity**. It offers an opportunity for this PhD study to explore a statistical method (sequence analysis) to address this research gap.

Chapter 2

Considering the disease development of LC is insidious and most patients diagnosed with LC are smokers or ex-smokers, it is crucial to understand the patterns of primary care pathways among a high-risk population **without** a diagnosis of LC, as this is the target population for early diagnosis, also a population less investigated. Most of the published studies used patients diagnosed with LC as the study population. For those studying high-risk patients, the focus was on how to effectively select eligible patients at high risk for LC screening using LDCT. The primary care pathways in the population of the NAEDI study are representative of the situation in daily general practice (more discussion in subsection 5.4.2). Smoking, patients' sociodemographic variables (e.g. age, SES), comorbidities, COPD and LC, are correlated. When studying the primary care pathways, they are all indispensable components. Given the current evidence on the association of age and SES with the route to diagnosis, it is sensible to investigate the association between patients' sociodemographic characteristics and primary care sequences in this study. The findings of this PhD study can provide new knowledge on the patterns of patients' help-seeking behaviours and how GPs manage patients at high risks over time, and may provide some clues on early intervention and diagnosis in this population. Identifying typical primary care pathways and the variables significantly associated with the pathways might enable us to further explore the opportunities for earlier cancer diagnosis.

2.6 Chapter summary

In this chapter, I argue the importance of studying primary care sequences (involving interdependent patient-GP events over time) in early diagnosis research. Through literature review, **studying primary care sequences is a new research direction**, different from studying risk factors, diagnostic intervals, and route to diagnosis, which is **the originality of this PhD study**. A novel statistical method, sequence analysis (SA), is proposed to study primary care sequences among a group of community patients at high risk of developing LC, who are also the target population for early diagnosis. The **key research question (RQ)** is **how SA can be used to study complex primary care sequences**, to enhance our understanding of the longitudinal primary care process to improve earlier cancer diagnosis. **The attempt of using SA to identify cluster patterns of primary care sequences would be an opportunity for this study to generate new knowledge and to make an original contribution to the field of early diagnosis research**. The next chapter provides an overview of SA and discusses relevant methodological issues about how SA can be contextualised in health research and applied in this study. Methodological RQs are summarised at the end of the next chapter.

Chapter 3 Sequence analysis (SA)

The literature review chapter identified a potential opportunity for SA to investigate primary care pathways. This chapter introduces this method, provides an overview of the background, development, rationale, and technicalities of SA, followed by a discussion of its strengths, limitations, controversies, and finally explains how SA can be contextualised and used to explore primary care sequences involving interdependent patient and GP events in this study.

3.1 An overview and the basic concepts of SA

3.1.1 The evolution and application of SA in different disciplines

SA was initially designed to analyse DNA, RNA, and peptide sequences in Bioinformatics (Needleman and Wunsch, 1970). In the 1980s, Andrew Abbott made adaptations of SA and introduced it in social sciences to study life course trajectories (Abbott, 1983, Abbott and Forrest, 1986, Abbott and Hrycak, 1990). Through the advocacy of Abbott and colleagues over the years (Abbott, 1990, Abbott, 1992, Abbott and Tsay, 2000), SA becomes increasingly popular in sociology and demography to understand important transitions in an individual's life course trajectory, from education in adolescence, to employment, marriage, forming family, raising children, and housing tenure in adulthood. SA is often used to classify individual sequences into distinct groups of trajectories. The application of SA in social sciences has expanded to political sciences (Blanchard, 2011) and survey methodology (Durrant et al., 2018) in recent years. Compared with its application in social sciences, SA in health research is still in its infancy. Based on the recent publications, this method appears promising and worth further exploration in health research. Therefore, to inform the analysis plan of this PhD study, a systematic scoping review on how SA has been used to study disease trajectories, care pathways, and health services research was conducted and is reported in the next chapter.

3.1.2 Basic concepts of SA

SA is data-driven and exploratory in nature and provides descriptive and graphic results. It takes the whole sequence from all subjects in the analysis. It does not require any modelling or assumptions of the distribution. **Sequence** is sometimes used interchangeably with other terms, like **trajectory** or **pathway**, to describe the longitudinal process, representing the stability or change of categorical states over time, e.g. sequences of the health status of older people, or pathways to disease diagnosis. There are two types of SA, i.e. **state SA** and **event SA** (further

explained in subsection 3.2.6). If without specification, **SA is usually referred to as state SA** in literature, same as in this thesis.

Sequence has an intrinsically time-ordered structure. Longitudinal processes, coded as **mutually exclusive states** in successive time periods along a timeline, constitute a **sequence**. Therefore, patients' visits to general practice over time could be reconstructed as sequences. **How to construct sequences from discrete health events is the first methodological issue** that needs to be addressed in this study. The position in a sequence could be a relative, rather than an absolute time point. For example, subjects may be born in different calendar years. We can still take age as a relative time point, investigating their life course trajectories from 18-40 years old. Alternatively, we can recruit a cohort of 18 years old, and use calendar years as a timeline, e.g. from 1998 to 2020, to delineate individual trajectory.

A **state** refers to a period of relative structural or functional stability, which is a combination of two elements, **event** and **timing**. An example is a subject in a "healthy" state in one observed period, such as hour, day, week, month, year. A **transition** means a change between two states, e.g. a subject transits from "healthy" status to "ill" or "hospitalised". A simple sequence, or say a stable sequence, has fewer transitions, like a subject in "healthy" status for the whole observation period; while a complex sequence has more transitions and perhaps involves many different states. For example, a patient had multiple chronic diseases and visited physicians frequently, hospitalised and discharged multiple times. An **alphabet** is a pool of all the states, e.g. "healthy", "ill", "hospitalised", "terminal", "deceased", and "missing". Usually, different states are assigned different colours for the purpose of presentation in figures, to facilitate visualisation and interpretation of the result. Figure 3.1 illustrates the key concepts in this paragraph. Other features and technicalities of SA relevant to this PhD study are introduced and discussed in the following sections, while the **contents of SA irrelevant to the study are not included in this thesis**.

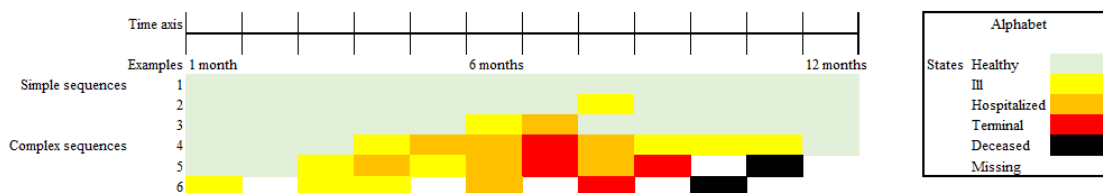


Figure 3.1 – An illustration of basic concepts (sequence, states, and alphabet) of SA

3.1.3 An overview of the analytical process of SA and relevant terminology

SA involves sorting, comparing, and grouping sequences by calculating distances based on the similarities between sequences by algorithms. A typical SA includes the following five steps. **The first step** is to specify states in an alphabet, construct and align sequences (a timeline may be helpful in some situations), as shown in Figure 3.1.

The second step is to define the costs for operations between states (more in subsection 3.2.4). **The similarity of the two sequences is quantified as “distance”**. The distance between two sequences is measured by the number of edits (**edit distance**), mainly insertions, deletions (called **indel** together), and substitutions, to change one sequence identically to another. **Costs** are numeric values for the three operations. For example, the cost for insertion and deletion could be set as 1, and substitution as 2. **Distance** is the total cost of a series of operations between any two sequences. The only difference between sequences 1 and 2 in Figure 3.1 is the state in Month 8, i.e. “healthy” in green in sequence 1 but “ill” in yellow in sequence 2. To make these two sequences identical, we can delete the different states first (yellow in sequence 2, cost 1) and then insert the same one (green, cost 1), or use substitution directly (substitute yellow to green, cost 2). The total cost for either approach is 2. Therefore, the distance between sequences 1 and 2 is 2. We can use the same process to calculate the distances between sequences 1 and 3, 1 and 4, ..., 5 and 6, called **pairwise distances**.

The third step is using an algorithm to calculate pairwise distances. It is labour intensive to calculate pairwise distances manually. We can use algorithms to do that for us, which would yield a symmetric matrix $\frac{n \times (n-1)}{2}$, where n is the number of sequences. There are six sequences in Figure 3.1, and we can get a matrix of 15 pairwise distances ($\frac{6 \times 5}{2}$). **Optimal Matching (OM)** (Abbott and Forrest, 1986, Abbott and Tsay, 2000) is the most commonly used algorithm to calculate the distances, but it is possible to use other algorithms as well.

The fourth step is grouping similar sequences in the same cluster. Cluster analysis (Fowlkes and Mallows, 1983, Estivill-Castro, 2002), a data reduction technique, is often used to achieve this goal. Cluster analysis is a connecting (intermediate) step of the whole analytical process of SA. The distances indicate the similarity among sequences. The more similar the two sequences are, the shorter the distance is. For example, sequences 1 and 2 in Figure 3.1 are more similar (both healthy) than sequences 1 and 6 (one healthy, the other one ill and died). Sequences in the same cluster are more similar to each other than those in other clusters (intra-cluster homogeneity and inter-cluster heterogeneity). Sequences 1-3 may be grouped in one cluster (healthy patients), while sequences 4-6 in another cluster (ill patients), as shown in Figure 3.2.

Chapter 3

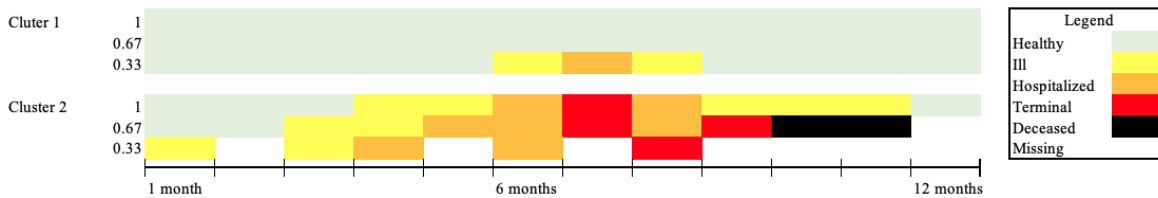


Figure 3.2 – The outcome of SA presented in state distribution plot (a very simple example)

The fifth step is result visualisation. The pattern of clusters can be presented as figures for visualisation and facilitate the interpretation of meanings in each cluster. The result of SA is a typology of several clusters, as the examples illustrated in Figure 3.2 and Figure 3.3. Due to its descriptive nature, SA cannot establish any form of causality, nor conduct statistical testing. Some consider it as a weakness. However, the typology (clusters) could be used as either an independent or dependent variable to investigate the association with other variables in regression models.

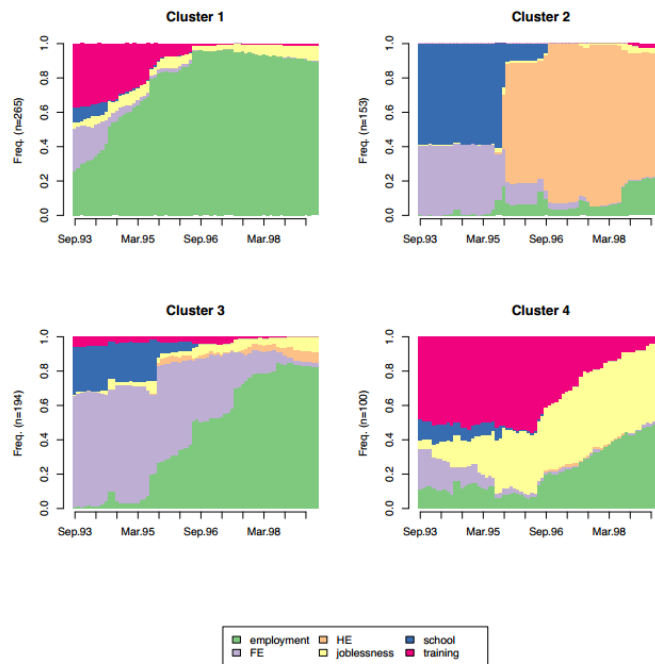


Figure 3.3 – The outcome of SA (a typology of clusters in state distribution plot), from Gabadinho et al. (2011a)

3.2 Measuring (dis)similarity between sequences

3.2.1 The importance of measuring similarity between sequences

Measuring the distances between sequences in pairs is a crucial step of SA, which quantifies how similar (or different) the two sequences are. Dissimilarity measures usually account for 1) the order of states and transitions, 2) the temporality of transitions, and 3) the duration of stay in

each state (Gabadinho et al., 2011a). Dissimilarity measures can be generally classified into two types:

1. Dissimilarities based on the counts of common attributes: it includes the Longest Common Prefix (LCP), Longest Common Suffix (RLCP), and Longest Common Subsequence (LCS) (Gabadinho et al., 2011a). It means that states are in the same order and for the same duration. It is relatively strict and does not allow moving any part of a sequence.
2. Edit distance. For example, without shifting, the evident similarity between $x = ABAB$ and $y = BABA$ can be very distant, but they can become quite similar by shifting just one position in y , shifting the first "B" to the last. Such algorithm allows sliding states in the sequence from one place to another, if it improves matching. The most popular algorithm of this type is Optimal Matching (OM).

3.2.2 The commonly used distance measure – Optimal Matching (OM)

The use of OM, followed by cluster analysis, has become the predominant approach to analyse life course trajectories in social sciences since its initial application in the late 1980s (Abbott and Forrest, 1986, Abbott and Tsay, 2000). OM is also known as Levenshtein distance, or edit distance, using three operations: insertion, deletion (indel), and substitution of states, illustrated in Figure 3.4. Each operation is assigned a cost. The distance of the paired sequences is calculated as the total cost of edits to transform one sequence identical to another. One common practice is to set constant costs. The cost of insertion and deletion is the same (e.g. 1), which could be half of the cost of substitution (e.g. 2), as substitution is equivalent to sequential operations of deleting the different state and then inserting the same one (illustrated as the second approach of sequences 4-b and 4-c in Figure 3.4).

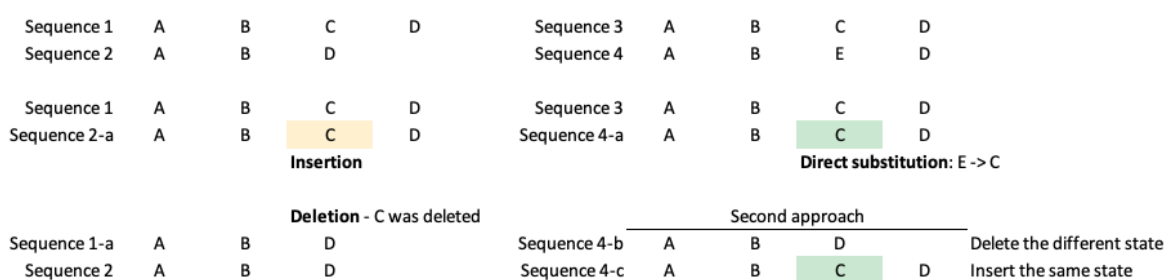


Figure 3.4 – An illustration of three basic operations (insertion, deletion, and substitution) in the optimal matching (OM) algorithm

OM is flexible and allows a time-shift penalty (through indel cost). There may be many ways to change one sequence identical to another. The reason why this algorithm is named as 'optimal matching' is that the algorithm is 'smart' enough to find the best way to get the **shortest distance**

between two sequences **with minimal cost**. Therefore, with the same OM algorithm, **the more operations are required to make the two sequences identical, the more costs they need, the larger the distance is, the more dissimilar the two sequences are**. Usual distance measures, such as the Euclidean distance, are ineffective for sequential data when the lengths of sequences are different (Kruskal, 1983). Based on OM, researchers have proposed modified algorithms, for example, context-dependent indel costs (OMloc)(Hollister, 2009) and costs weighted by spell length (OMslen)(Halpin, 2010).

3.2.3 The other commonly used dissimilarity measures

Hamming Distance (HAM) and Dynamic Hamming Distance (DHD)(Lesnard, 2010) are the two dissimilarity measures with substitution only, without indel. **They can only apply to sequences with equal length**. The distance between sequences is the sum of the period-by-period difference between states. The difference between HAM and DHD is that DHD applies position-wise state-dependent weights, while HAM only counts the number of mismatches. In simpler words, HAM could be understood as substitution cost set as constant (e.g. 1, 2). What 'dynamic' in DHD means is substitution cost varies in different positions of the sequence (time-varying costs), and the cost is dependent on the transition rates among states at each time point.

Different measures pick up different sequence characteristics. They offer a variety of options to analyse sequence similarity from different aspects. Studer and Ritschard (2016) nicely compared and summarised different dissimilarity measures in their paper. There is **no golden rule to select a dissimilarity measure**. It is necessary to consider the research questions at hand and refer to other relevant published studies in the research field. A study compared some dissimilarity measures using real datasets and statistical simulation, and concluded that no one measure was dominant or out-performing than others in all situations based on the χ^2 statistic. However, OM and Hamming Distance (HAM) were ranked in the top tier. DHD was close to HAM, but in a slightly lower tier (Halpin, 2012). Since SA is exploratory in nature, it is worth **trying different dissimilarity measures with the same dataset and comparing the typologies to find the best one that makes sense of the data**, which is **one of the research objectives and the analysis plan of this study**.

3.2.4 Cost setting

Setting costs for operations (indel and substitution) is one major methodological issue for algorithms like OM in SA, especially for the substitution cost. Different cost setting schemes may significantly influence the results. Some may argue that cost setting is arbitrary and consider it as a limitation (Wu, 2000). Therefore, it needs careful consideration. Gauthier et al. (2009)

summarised some possible approaches to set costs and discussed the possible options. The first option is setting all the costs as constants, e.g. 1 for indel and 2 for substitution, denoted as $OM_{[1,2]}$ hereafter. This option can be used when there is no theoretical basis or background information available to support the cost setting. A second approach is based on theory to set costs and give weights to the costs. A third choice is applying empirical costs based on common sense or face value, considering the problem at hand. The results from prior analyses or published studies may be used as a reference. The fourth strategy to set costs is a combination of common sense (the third strategy) and the likelihood of transitions between states in empirical data. For example, in a study investigating the cohort differences in a career at Lloyds Bank (Stovel et al., 1996), substitution costs were set proportional to the transition rates (TR) among states, denoted as $OM_{[1, TR]}$ hereafter. **Transitions between two states that happen more often cost less in substitution.** Compared with setting costs as constant, this approach is **data-driven**, more able to reflect the differences among states over time.

The characteristics and differences of the common dissimilarity measure and cost settings are: $OM_{[1, TR]}$ and DHD are **data-driven** approaches for substitution cost (dependent on transition rates among states). The difference is that DHD is time-varying (position-wise), while $OM_{[1, TR]}$ is time-invariant, considering all the time points in all the sequences. $OM_{[1,2]}$ and HAM use **constant** cost. As to the relationship between the indel and substitution cost, when the substitution cost is set as more than two times of the indel cost, substitution will not be used, as it costs less to delete the unwanted state and insert the desired one than to substitute states. This means that if one set substitution cost 2.5 times of the indel cost in OM, the algorithm will always choose indel because indel cost less. Conversely, if substitution costs are very low compared to the indel costs, then indel is no longer used, as the algorithm always chooses the lowest cost.

Considering the importance of cost setting, how other studies set costs when using SA in health services research were reviewed, summarised, and reported in the next chapter. The findings are used to inform the cost setting in this study. In addition, as part of the methodological exploration, results from different cost setting schemes in the main study will be compared, to find to what extent cost setting influences the results.

3.2.5 Typical sequences

SA can identify representative and the most frequent sequences for each cluster. One possible way of doing SA is pre-defining several empirical or theoretical ideal types of 'reference sequences' and then calculating the distances of all the sequences against those 'reference sequences' (Abbott and Hrycak, 1990). For instance, in a study investigating the trajectories of

work, partnerships, and housing, related to the quality of life in early old age (Wiggins et al., 2007), the typical sequences were constructed by the research group to capture the 'ideal trajectories'. They were artificial/synthetic sequences, rather than from the data. The distance of each sequence was calculated by comparing with the 'ideal type sequences'. Individual sequences were then assigned to one of the 'ideal type' categories based on the shortest distance to that 'ideal trajectory'. In some cases, it may be difficult to determine which 'ideal types' the sequences should belong to, especially when the distances to several 'ideal sequences' are close. In addition, the number of 'ideal sequences' is the number of clusters. Without theoretical knowledge, it would be difficult to select the 'ideal sequences' and justify their validity. Another way to create a synthesised sequence is to pick up the most frequent state at each time t and get the **modal** state sequences (Gabadinho et al., 2011a), which cannot ensure the plausibility of such a synthesised sequence. For example, it may end up in a representative sequence with a state 'married', followed by a state 'single', which does not make sense in reality.

Using a **medoid** sequence is perhaps a better solution to describe and represent the clusters, rather than an artificial one. The **medoid** is one real sequence and the most central sequence, which has the shortest (or the weighted sum of) distance from the other individual sequences in one cluster (Gabadinho et al., 2011b). An important benefit of describing a cluster by medoid is that it can easily compute the dispersion of sequences around the medoid sequence (the minimum, the maximum, and the average distance within a cluster), which cannot be done in a modal sequence. The dispersion feature can tell us the extent of homogeneity (or heterogeneity) in any given cluster. A cluster containing homogenous sequences tends to be very similar to the medoid sequence, whereas a high dispersion suggests that sequences within the same cluster are heterogeneous, even though they are grouped together (Aassve et al., 2007).

The possibility of identifying typical sequences for each cluster was explored in this study. Care pathways could be highly heterogeneous, as every patient has his/her health situation and care needs, which reflects the complexity of clinical practice and health research. It is easier to identify typical sequences to represent clusters in social sciences, because life course trajectories are simpler than care pathways, in terms of a smaller number of states and fewer permutations of states in the sequences. Individuals may have variations in education, employment, marriage, having children, and housing tenure in their life course, but it is possible to summarise the group patterns and identify typical trajectories in a population.

3.2.6 Event sequence analysis (ESA)

Up to this point of this chapter, the previous contents are about state SA. If ignoring the durations of states (no timing information), they are just a string of characters (i.e. events). **Event sequence analysis (ESA)** finds common **sub-sequences**, which are one or more events occurring in the chronological order of the original sequences. Sub-sequences could be understood as part of the whole sequence by deleting some elements. For example, A, CE, ABD, and ABDE are sub-sequences of ABCDE, which are fragments of a whole sequence. The most frequent sub-sequences are usually from the prefix or suffix of a whole sequence.

The limitation of the ESA is that it only provides fragmented information of the whole sequence. Table 3.1 illustrates the results of the 10 most frequent sub-sequences (with two events) from ESA in an earlier phase of the PhD study using data from HHRAD. We can know the frequency of events before LC diagnosis, but the previous common events identified by ESA **could be anywhere in the diagnostic pathway**, rather than the one right before the LC diagnosis. From this aspect, event SA is less useful, and sometimes could be misleading or cause confusion for someone who does not know this method, as they may think the events before diagnosis are the ones right before cancer diagnosis, but the sum of the percentages of the sub-sequences is over 100%. Therefore, ESA could be used as a supplementary method to state SA in this study, but not the focus, as **analysing sequence holistically is the research interest**. Despite this, **ESA** is still helpful to **provide sequential patterns in text within each cluster** (complementary to SSA), to allow a better understanding of the typology presented in the graph.

Table 3.1 – Results of the ten most frequent sub-sequences from event SA in an earlier phase of a study using data from the HHRAD

The 10 most frequent event categories before lung cancer diagnosis		
247	82.89%	(Antibiotics & vaccination)-(Cancer diagnosis made)
176	59.06%	(Chest X ray)-(Cancer diagnosis made)
163	54.70%	(Respiratory symptoms)-(Cancer diagnosis made)
152	51.01%	(Smoking cessation)-(Cancer diagnosis made)
111	37.25%	(Chest related symptoms)-(Cancer diagnosis made)
105	35.23%	(Lung cancer related referral)-(Cancer diagnosis made)
102	34.23%	(Other referral)-(Cancer diagnosis made)
88	29.53%	(Systematic symptoms)-(Cancer diagnosis made)
83	27.85%	(Spirometry & respiratory function tests)-(Cancer diagnosis made)
60	20.13%	(Other tests)-(Cancer diagnosis made)

3.3 Grouping sequences – cluster analysis

3.3.1 Introduction

After calculating the pairwise sequence distance, the next step is to group similar sequences together. Agglomerative hierarchical clustering (Ward's method)(Ward, 1963), an unsupervised clustering technique, is commonly used to group sequences after OM. Dlouhy and Biemann (2015) compared eight clustering algorithms to find out which one can yield the best results, including Ward's method, single linkage clustering (nearest neighbour method), complete linkage clustering (furthest neighbour method), average linkage clustering (between-groups linkage), centroid clustering, median clustering, McQuitty's method, and k-means². The results showed that Ward's method delivered the best results among the eight clustering algorithms, consistently having the lowest **misclassification rates** in sequence lengths ranging from 5 to 100, where the misclassification rate was defined as the percentage of incorrect assignments of sequences to wrong clusters in that study.

3.3.2 How the agglomerative hierarchical clustering algorithm (Ward's method) works

The agglomerative algorithm starts to find similar sequences locally. Sequences with shorter distances indicate greater similarity, which is grouped first. This process continues, until all the sequences are grouped together, under one roof. The basic idea of the algorithm is to minimise the intra-group variance and maximise the inter-group variance. The outcome of the whole process is a tree-structured dendrogram, as illustrated in Figure 3.5. The shape of the dendrogram provides information about how the sequences are grouped, and the structure of clusters. Decisions on the optimum number of clusters (discussed in the following subsection 3.3.5) will determine where the dendrogram is cut, as the red line is shown in Figure 3.5, so there would be four clusters in this example. The typology of the four clusters has been illustrated in Figure 3.3. The cluster membership for each sequence can be stored in a variable, as the outcome of cluster analysis.

² The first seven algorithms are in the family of hierarchical clustering, while k-means is a partitioning clustering algorithm. One important limitation of k-means is that the number of clusters needs to be prespecified. In this PhD study, we could not know the number of clusters beforehand, as we do not have previous knowledge on this.

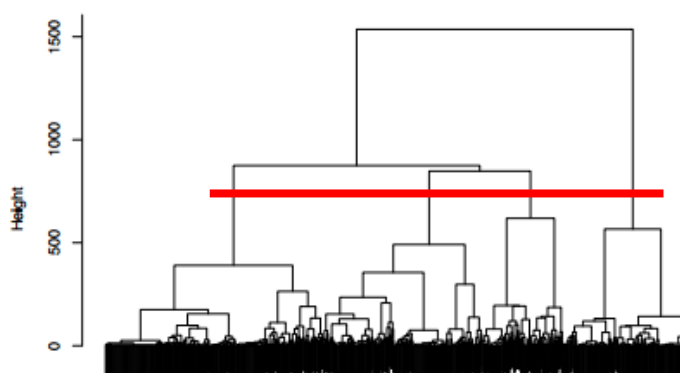


Figure 3.5 – Dendrogram of agglomerative hierarchical clustering (Ward's method) based on OM distances

3.3.3 Assessing the clustering quality

Due to the hierarchical structure, a bigger number of clusters is always possible by splitting the nodes at a higher level to get more clusters at a lower level. As shown in Figure 3.5, the red line could be anywhere in the dendrogram, to get the number of clusters from 1 to n (sample size, the number of sequences). Therefore, additional indicators would help decide the number of clusters.

Over ten statistical indicators are available to assess the quality of clustering/partition in R packages, to name a few, Point Biserial Correlation, Hubert's Gamma, Hubert's Somers'D, Hubert's C, average silhouette width (ASW), Calinski-Harabasz index (Pseudo F), Pseudo R^2 . Understandably, these indicators have different algorithms, and they may not come to the same conclusion. It is better to choose one indicator suitable for SA (edit distance, not Euclidean or Manhattan distance, subsection 3.1.3), and use it consistently to compare the results from different sets of analysis in this study.

Silhouette is a technique providing a graphical representation of how well each object is classified. The ASW value ranges from -1 to 1, measuring how similar an object (sequence in this study) is to its cluster compared with other clusters (Rousseeuw, 1987), which can be calculated for any distance metric (edit distance here). A high value (close to 1) indicates high between-group distances and substantial within-group homogeneity (Kaufman and Rousseeuw, 1990). The magnitude of ASW value is generally divided as 0.71–1.00, indicating strong and appropriate clustering structure, 0.51–0.70 reasonable structure, 0.26–0.50 weak structure and probably artificial, where ≤ 0.25 indicates the homogeneity of the groups is low (Studer, 2013).

Other commonly used indicators are less suitable to assess the clustering quality after SA, due to their respective limitations. For example, the Calinski-Harabasz index ranged $[0, +\infty]$, is based on the F statistic of ANOVA (Pseudo F computed from the distances). Its extension to non-Euclidean

distance (such as OM) is subject to debate. Pseudo R^2 , which ranges from 0 to 1, can only compare partitions with the same number of groups, as it does not penalise complexity (Studer et al., 2011).

3.3.4 Heterogeneity and outliers

Even within the same cluster, considerable heterogeneity could still exist among sequences, which could reflect on the indicator of clustering quality. The distances between the deviants and the majority are bigger. The deviants could be chaotic or complicated sequences and considered as outliers. For health sequences, they are worth being further investigated to understand the reasons behind, e.g. why some patients had more frequent visits within a certain period. On the other hand, defining some criteria for outliers (e.g. a large number of GP visits) and selecting them out beforehand would help increase the efficiency of clustering and make the interpretation of cluster patterns easier.

3.3.5 Deciding the optimal number of clusters

Given the exploratory nature of SA, the number of clusters was usually decided by the researchers in a qualitative way in most published studies, looking at the extent to which the clusters could explain the research problem in the most informative way. The authors often presented the result (typology of clusters) without elaborating how they came to that solution and what criteria they used to decide the optimal number of clusters in their studies.

As the outcome of the analysis, the rationale for deciding the number of clusters should be clear to readers and justifiable. One of the objectives to conduct a systematic scoping review (in the next chapter) is to understand how researchers decided the number of clusters after using SA. If the criteria are available from the published studies, their applicability in this PhD study could be assessed. If not available, then one of the objectives in the methodological exploration phase of this study is trying to establish some criteria to decide the optional number of clusters and then test the applicability.

When deciding the optimal number of clusters, the empirical research context should be considered. Otherwise, if only relying on an indicator, the research is no different from a statistical exercise. Researchers could explore the results from a different number of clusters, presented in state distribution plots, to see whether the pattern in each cluster makes sense or not (interpretability). The best solution (typology) should represent the patterns of the data (not oversimplified or overcomplicated), with a reasonable number of sequences in each cluster (sample size), and the meaning of the clusters should not strongly deviate from the existing

theory and evidence (at least not counterintuitive or against the reality). These general principles could be used to decide the optimal number of clusters.

3.4 Multiple interdependent dimensions in sequences

3.4.1 Examples of multiple interdependent dimensions

In some situations, sequences could be complex and contain multiple correlated dimensions (like education, employment, marriage, raising children). Another example is the primary care events in this PhD study, which are the reasons patients visited the general practice, and how HCPs (GP and practice nurse) managed patients after the presentation. Patient and HCP events are the two interdependent dimensions. HCP manages the patient's health conditions; the patient follows HCP's advice, takes the prescribed medications, and may come back if the previous action does not work.

3.4.2 Possible approaches to analyse multiple interdependent dimensions

There are four possible ways to analyse sequences with multiple correlated domains. The first one is constructing a typology for each domain individually and then using the most important dimension (e.g. employment) as the dependent variable in multinomial logistic regression, and other dimensions (e.g. marital status, housing tenure) as independent variables in the model. Instead of regression, the second approach is combining the result of distinct types of trajectories from each domain, like cross-tabulating typologies. These two approaches are not suitable to study interdependent patient-GP events. The main problem of these two approaches is that they do not consider the local or the temporal interdependence of the correlated dimensions, as each dimension is analysed and clustered independently. The timing information is not fully used, and all the correlated sequences are condensed and simplified as categorical variables. The results in each dimension may not be equally reliable or informative, as it is potentially sensitive to noisy data (missing data, poorly recorded information, or heterogeneous contents). In addition, cross-tabulating typologies of categorical outcomes of multiple domains from SA may overestimate the number of clusters. Some combinations possibly have a very small sample size of sequences and thus are not informative.

The third approach is combining patient and GP events together in states and using traditional SA to analyse sequences. The fourth approach is separating patient and GP states as two correlated channels, and using multi-channel sequence analysis (MCSA) in the analysis. These two approaches are more likely to work and further discussed their suitability below.

3.4.3 Multi-Channel Sequence Analysis (MCSA)

Pollock (2007) and Gauthier et al. (2010) did pioneering work to extend traditional SA in one dimension to multiple dimensions, and called it multi-channel sequence analysis (MCSA) or Multidimensional SA. This approach can fully use the longitudinal information and analyse multiple correlated domains simultaneously from a holistic perspective. Compared with using SA multiple times and calculating distances of each dimension separately, Gauthier et al. (2010) argued that the strength of MCSA is its ability to take into account of all dimensions together by creating an extended alphabet, i.e. combining alphabets from individual dimensions. Calculating the distances and considering all the dimensions can produce more robust and informative results than analysing sequences in each domain separately. MCSA is thought to better account for the local interdependence and the interconnectedness of states at each time point of the alignment among channels. However, as the number of states becomes larger in the extended alphabet and the combinations become more heterogeneous, it may be challenging to set up and justify a cost scheme. In addition, the local contribution from each dimension to the overall distance is unknown.

3.4.4 An exploration of the suitability of MCSA and SA in this PhD study

An early exploration between combining patient and GP events together as the combined states and analysed using traditional SA (the third approach) and separating patient and GP events as the two interdependent channels and analysed by MCSA (the fourth approach) was conducted to decide which approach would be more suitable in this study. The background of this exploration is investigating how GP managed patients with COPD two years before LC diagnosis using the HHRA dataset. The outcomes of the analysis are present in Figure 3.6 (the typology of sequences with combined patient and GP states and analysed by SA) and Figure 3.7 (the typology by MCSA). Further discussion of these two approaches is in the next subsection.

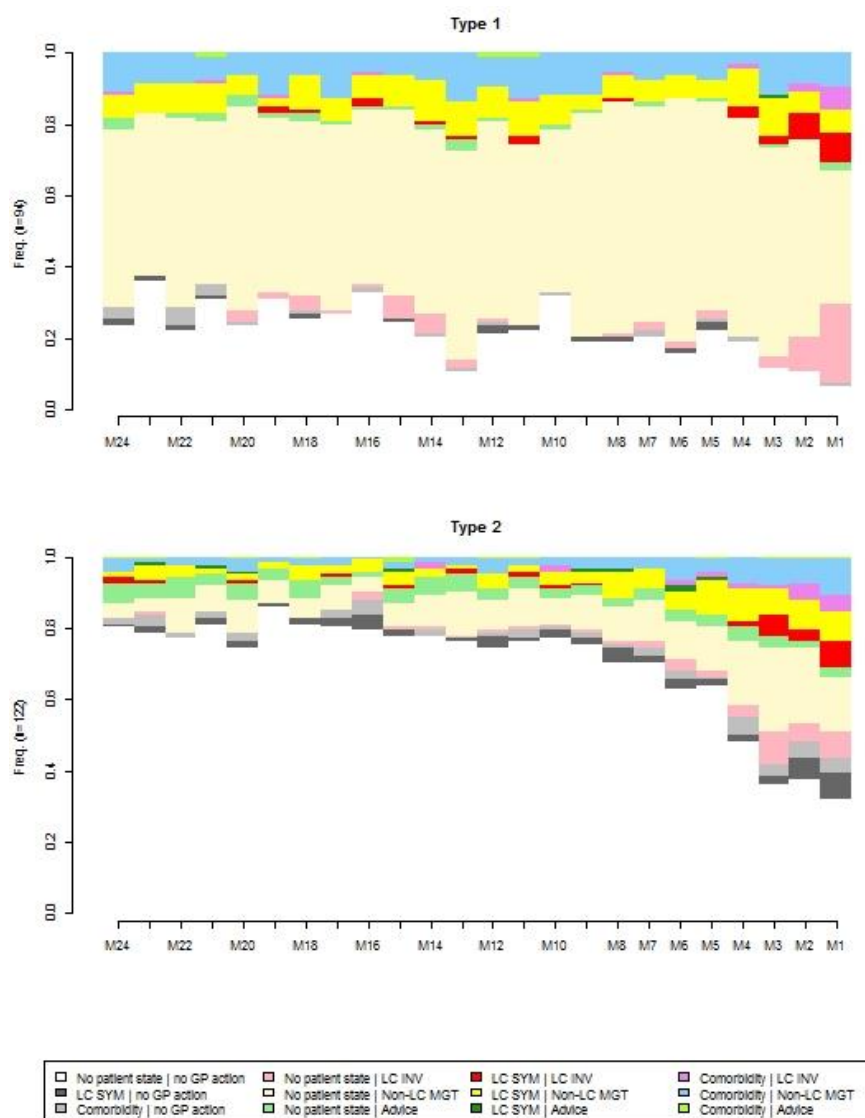


Figure 3.6 – Typology of two clusters from single sequences with states combining interdependent patient-GP events two years before LC diagnosis (month as interval)

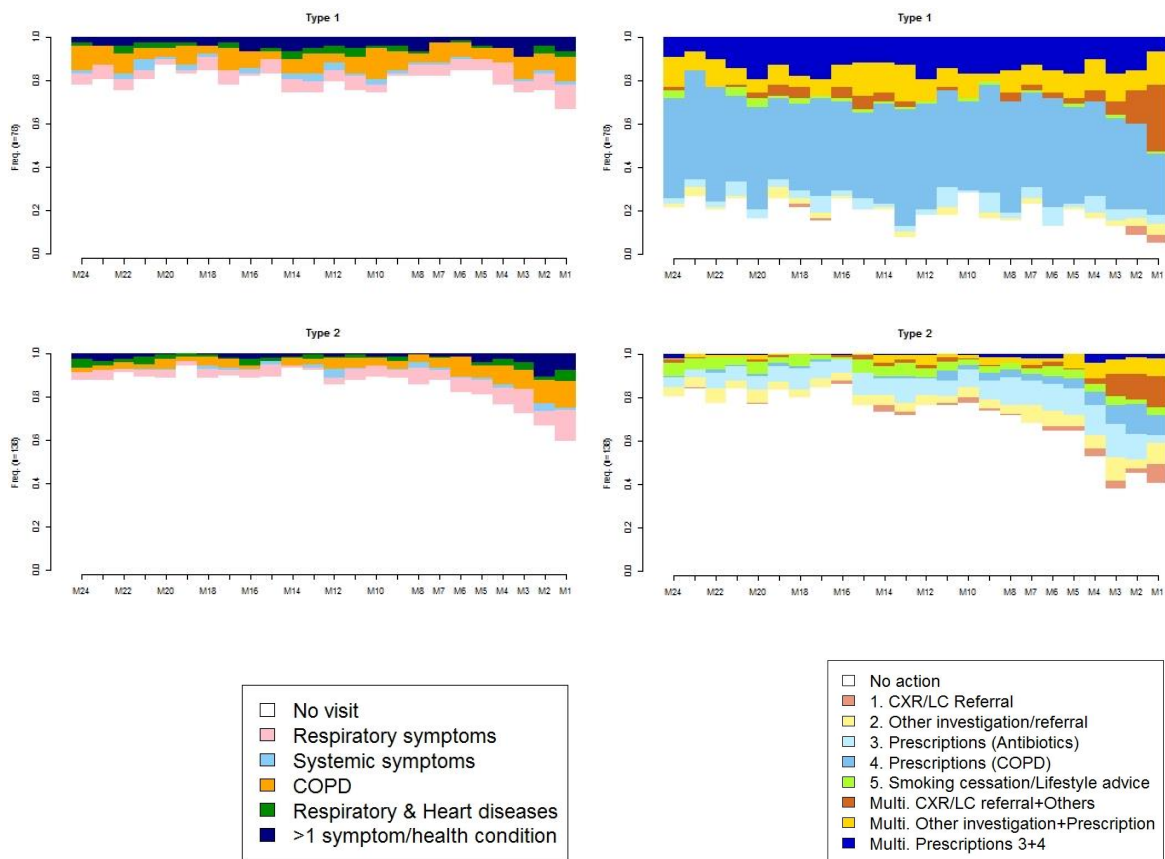


Figure 3.7 – Typology of two clusters by MCSA, patient (left) and GP (right) channels two years before LC diagnosis (month as interval)

3.4.5 Brief discussion: the two approaches to analyse interdependent patient and GP events and the cluster patterns

Separating patient and GP events can include more states in the alphabet and reduce the complexity in MCSA, while in traditional SA, patient and GP parts need to combine together for each state. The biggest challenge is how to categorise the combination of patient and GP events, and ensure the number of states is appropriate for analysis, graphic presentation, and result interpretation. If the number of combined states is too large (e.g. 20), it would be difficult to use different colours to represent all the states and detect the differences in the patterns in distinctive colours by human eyes.

The separate presentation of the patient and GP channels in the typology of MCSA (Figure 3.7) allows the readers to know the patient and GP states at each time point (x-axis), which is the strength of MCSA. When calculating the distance between sequences, the patient and GP states are considered as a whole in the analysis of MCSA. However, the patient and GP sequences are presented separately in the figure. Therefore, we could not know how GP responded to patient's presentation from the figures, which is the biggest limitation of MCSA. In addition, the proportion

of events classified as GP management was much larger than those classified as patient events in the HHRAD dataset. Therefore, a greater weighting was given to the GP channel than to the patient channel in MCSA, which had implications for the clustering structure. Using traditional SA avoids such a situation. Understanding how GP managed patient's presentation at each consultation is of research interest. Therefore, **combining patient and GP events together, representing the reason and outcome of the consultations** in one single sequence, and analysed by traditional SA is a better approach in this study. MCSA is still a valid approach to analyse interdependent/correlated dimensions in other research contexts.

3.5 Mainstream statistical software and packages to conduct SA

The 'TraMineR' package (Gabadinho et al., 2011a) in R is the most popular package for both state and event SA. It can analyse and visualise sequential data, also able to conduct MCSA. Two packages, 'SQ' (Brzinsky-Fay et al., 2006) and 'SADI' (Halpin, 2017), and a plug-in 'seqcomp' (Lesnard and Kan, 2011), are available in Stata to conduct SA. SAS can produce sequence index plots but cannot perform OM, while SPSS cannot implement SA at the moment. In this study, the 'TraMineR' package was used to conduct SA, followed by the 'cluster' package in R to do hierarchical clustering (Ward's method).

The computing power required to run SA should not be underestimated. For n sequences with a total observation period t , it needs to calculate pairwise distances between sequences to yield a symmetric distance matrix. If both n and t are very large numbers, the computing power will be very demanding. In addition, the length of a vector/object in R is fixed, i.e. $2^{31}-1$ (2,147,483,647) elements, which is unchanged regardless of the increase of computing power. Computer power is related to the speed of computation, while the limit of a vector is the maximal storage space for the distance matrix. The dissimilarity matrix needs a storage space of n^2 , which should be less than $2^{31}-1$. Therefore, the maximal possible number of sequences is 46,340, theoretically. The 'TraMineR' package may not be able to conduct OM with a large number of long sequences. It was estimated that OM could be applied in R for up to 35,000 sequences, depending on the length of sequences (Durrant et al., 2018). It has been tested to use OM to analyse around 2,000 sequences with a maximal length of 100 positions in the 'SQ' package in Stata (Brzinsky-Fay et al., 2006). If the computing power is limited, researchers can randomly select a proportion of samples from the original dataset to conduct SA. An example can be referred to Mattijssen and Pavlopoulos (2017). This process can be repeated several times and compare the stability of the results.

3.6 Chapter discussion: the technical uncertainties of SA

3.6.1 The nature of SA and the technical uncertainties

The sociological rationale of SA is “life patterns are structured by variations in the timing, duration, and order of events” (Elder, 1985). SA represents a trend in social sciences toward thinking about “events in context”, rather than “entities with variable attributes” (Aisenbrey and Fasang, 2010). “Event in context” more emphasises having a holistic view of events in the sequence (to see parts in the entirety). In such a case, events are not fragmented or isolated from each other in a sequence. Instead, they are analysed and understood as a whole within specific contexts, while “entities with variable attributes” more reflect the relationship between independent and dependent variables in regression models.

Unlike regression based on modelling, SA is algorithmic, non-parametric, exploratory, and descriptive without a dependent variable. Regression aims to understand the association between independent and dependent variables, and establish models for prediction, while SA aims to extract simplified information from the data, uncover the patterns among the sequences, and categorise the patterns into a limited number of representative clusters. Such an approach allows an overview of the whole trajectory at both individual and group levels (Halpin, 2012). SA provides a new perspective to understand longitudinal data.

SA remains a marginal analytical tool. It sits somewhere between purely narrative and traditional variable analysis (Pollock, 2007). As an algorithmic method without an explicit probabilistic base, a key question is whether using SA can obtain meaningful results. Compared with other statistical methods, subjective judgement is involved in almost every step of the analytical process, let alone the technical uncertainties of SA (Aisenbrey and Fasang, 2010), including how to choose an appropriate dissimilarity measure and determine the costs of operations, how to decide the optimal number of cluster and justify the results are valid, how to treat the missing values and censored observations, and how to deal with the complex interdependencies over time.

3.6.2 Appropriate dissimilarity measure and cost setting

Measuring similarity between sequences is central in SA. Choosing the dissimilarity measure and setting the costs are the base for calculating the distance. Therefore, it is fair to say they are the most important parts of SA, which determines the results. Dlouhy and Biemann (2015) conducted a Monte Carlo simulation of career sequences to test how different sequence lengths, the sample size of sequences, and missing items in OM would influence the results. That study concluded that sequence length was the most crucial factor for results, and the authors recommended a

minimum sequence length of 25 should be met to ensure high-quality results, where 'high-quality result' was defined as <10% misclassification of sequences as a threshold by the authors. Sample size does not substantially affect result quality. The simulation found that OM performed quite well, even with a small sample size ($n > 50$), although a larger sample size with better representativeness of the population was favoured for greater generalisability of the results. Another important finding was that OM could tolerate sequences with up to 30% of missing elements for a low misclassification rate.

The four ways of cost setting were introduced and discussed in subsection 3.2.4. Wu (2000) commented that the choice of costs was a major concern in using OM in life course analysis because of the arbitrariness and the weak link to the theory. The distances could be meaningless in a particular field of sociology. Halpin (2003) counter-argued that indel and substitution are simply computationally efficient to calculate the distances between sequences, and the criticism of no clear sociological interpretation was mainly a red herring. Halpin made an analogy that we did not need to worry about the lack of a sociological interpretation of the Newton-Raphson method in the maximum-likelihood estimation. Theories in the health and social sciences are rarely precise enough to answer questions related to cost setting. Without theoretical support or previous empirical evidence, it could be challenging to justify the decision of cost setting. Two general approaches of cost setting are data-driven and constant costs. Both approaches are used in this study, and the results are compared in Chapter 5.

3.6.3 The optimal number of clusters and the validity of results

Ward's method was proved as the best algorithm after OM with the lowest misclassification rates (Dlouhy and Biemann, 2015). Such a conclusion solves the problem of which clustering algorithm should be used, and boosts researchers' confidence to use Ward's method. However, the optimal number of clusters is still an unsolved problem. This study aims to establish some criteria to help decide the optimal number of clusters, and to improve the validity of results.

3.6.4 The order, timing, and complex interdependencies of states in sequences over time

States in life course sequences are time-referenced. Timing is relevant to patient's help-seeking behaviours, and is especially important before the stage of cancer diagnosis. Since indel operations can move states forward or backward, when analysing patient and GP events in a single combined sequence, the states could combine both parts, like "patient presented with cough, GP ordered CXR". Putting the GP part after the patient part could be considered as the outcome of the consultation. If separating them, when matching sequences, it may result in

several patient states, followed by several GP states, as OM may consider this way as the 'optimal' solution with the lowest cost; but this situation does not make sense. Researchers should consider the possible adverse effects and the implications when using the algorithms to avoid counterintuitive results.

3.7 Chapter summary: how SA fits in this PhD study

There are still many unresolved problems in SA, especially in the field of health research. These unsolved problems provide new research opportunities and academic debates, which could lead to improvement in the theory and application of SA in various disciplines. This PhD study tries to contextualise SA to study complex primary care sequences involving patient and GP events. **SA can take the whole sequence from all the patients to identify cluster patterns. Such a unique strength is not achievable by regression.** Considering the exploratory nature of SA, it needs to compare results from different methodological/technical options to get meaningful empirical cluster patterns. In summary, **the key methodological issues of SA relevant to this study** include:

1. How can sequences be constructed using events extracted from discrete health records? How the variation of visit intervals in the sequence can be accommodated among different patients? Constructing sequences is the premise of conducting SA. Without sequences, there is no way to perform the analysis.
2. How would different dissimilarity measures (e.g. OM, HAM, DHD) and cost setting schemes (constant and data-driven) influence the results of SA?
3. What criteria are helpful to decide the optimal number of clusters and improve the validity of the result?
4. Is it possible to identify some 'typical sequences' for each cluster of the typology? Are they helpful and meaningful?

A systematic scoping review was conducted to understand how these key methodological issues have been addressed in the published health studies, and to further evaluate the value of SA for each study. This is reported in the next chapter. The findings of the review were used to guide the analysis plan and address the above methodological issues in the main study.

Chapter 4 The application of sequence analysis in health services research: a systematic scoping review

4.1 Introduction

4.1.1 The objectives of this chapter

The previous chapter provided an overview of SA and discussed some methodological issues in the application of SA in this study. SA has been applied in genetics, biology, and social sciences, but is relatively new in health research. This chapter is a review of studies that used SA in health research. Generally, there are three objectives to conduct this review:

1. To understand how SA has been applied in health research;
2. To document how the methodological decisions related to SA were made in each study (the issues identified at the end of Chapter 3) and to critically appraise the strengths and limitations of those methodological decisions;
3. To evaluate whether there was any added value of using SA, or better use alternative methods to answer the RQs in respective studies, considering the research context.

At the end of this chapter, the findings of this review, the general problems in the current application of SA in health services research, and the research gaps are summarised and discussed.

4.1.2 Position of this review

This review could be positioned either as a methodological systematic review or a systematic scoping review. Methodological systematic review is one of the ten types of systematic reviews in medical and health sciences (Munn et al., 2018b). It aims to examine a research method and its potential impact on research quality. An example of this type of review is investigating the effect of editorial peer review processes on improving the quality of reports in biomedical studies (Jefferson et al., 2007). Alternatively, scoping review is a common approach to review evidence in an emerging field or topic, and to identify knowledge and research gaps (Munn et al., 2018a). The first framework for scoping review was published in 2005. Scoping review has great utility in synthesising research evidence in a field regarding the nature, features, and volume of existing literature (Arksey and O'Malley, 2005). Munn et al. (2018a) provided a guide for authors to

choose between conducting a systematic review or a scoping review, and the indications to conduct them, summarised in Table 4.1.

Table 4.1 – Summary of the indications to conduct a systematic review or a scoping review

Systematic review	Scoping review
1) Uncover international evidence;	a) To identify the types of available evidence in a given field;
2) Confirm current practice/address any variation/identify new practices;	b) To clarify key concepts/definitions in the literature;
3) Identify and inform areas for future research;	c) To examine how research is conducted on a specific topic or field;
4) Identify and investigate conflicting results;	d) To identify key characteristics or factors related to a concept;
5) Produce statements to guide decision-making.	e) As a precursor for a systematic review;
	f) To identify and analyse knowledge gaps.

Compare the research objectives of this review against the above indications:

- To identify available studies using SA in health research – falls in scoping review indication (a)
- To investigate the methodological issues related to the design, conduct, and analysis of studies using SA – scoping review (c)
- To synthesise evidence and identify knowledge gaps – scoping review (f)
- To guide decision-making and inform possible research approaches for this PhD study – systematic review (3, 5)

Based on the above comparison, this review is more inclined to be a scoping review, despite one objective fit in the indication of systematic review. Scoping reviews can be carried out to investigate the way how research has been conducted. For example, a scoping review was to understand how scoping reviews have been conducted (Pham et al., 2014). A variety of study designs are usually included to support a greater breadth of scoping review. A difference between scoping review and systematic review is that scoping review aims to provide an overview of the existing evidence base **regardless of the quality of the included studies**. A formal assessment of methodological quality is generally not performed (Peters et al., 2015). However, methodological quality is central in this review, and it should be appraised. Studies were not excluded even if the methodological decisions were challenged, as they provided some examples of pitfalls and learning points for this study.

Therefore, this review is positioned as a **systematic scoping review**, focused on a particular statistical method, i.e. SA, applied in health research, and follows the guidance of conducting and reporting reviews of this type (Peters et al., 2015). In order to improve the quality of this review and the utility of the results, this review was conducted systematically, in terms of transparent search strategy, systematically search in multiple key medical and health databases, explicit inclusion and exclusion criteria, critically appraising the quality of studies using recognised and recommended checklists, a structured format for data extraction of the included studies, synthesis of research evidence, and reporting the findings following the guidelines of the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) statement (Moher et al., 2009).

4.1.3 Formulate the search question for the review

The search question was, “**how has SA been used to study disease trajectories, care pathways, or longitudinal health process (health services) in medical and health research?**” Two acronyms, PCC for Population, Concept and Context (Peters, 2016), and SDMO for the types of Studies, Data, Methods, and Outcomes (Clarke et al., 2011) were used to deconstruct the search question, and made it more concrete and specific, which also involved specification of the inclusion and exclusion criteria to select studies for this review.

- **P**opulation: any type of population in human species, regardless of demographic characteristics, but not animals;
- **C**oncept/**M**ethods: sequence analysis;
- **C**ontext/type of **S**tudies: health research related to disease trajectories, care pathways, or health services, not limited to any type of disease, but NOT in genetics or biology;
- **D**ata: longitudinal **quantitative** data, either observational or interventional, with or without randomisation or control group;
- **O**utcome: clusters or typology of sequence analysis, likely to be multiple outcomes (clusters), but the outcomes **must** be in the health domain, NOT social sciences (e.g. transition in employment, career trajectories)

4.2 Methods

4.2.1 Search strategy and databases

The search strategy was discussed and refined multiple times with the support from a research librarian, who provided consultation services for the School of Health Sciences, University of

Southampton. The librarian kindly provided tips to improve the search strategy. Medical Subject Headings (MeSH) and free text keywords were used in the search strategy wherever possible. If MeSH terms were not available or applicable, free text keywords were used instead. For example, the six MeSH terms of “sequence analysis” were all related to a technique to analyse protein, DNA, or RNA in genetics, which was not the intended research field in this review. Therefore, only keywords were used in such a circumstance. The key words included pathways, trajectories, and health services in different care settings (primary, secondary, tertiary care, and other similar and relevant expressions), as well as care pathways, disease trajectories, care trajectories for any type of disease. When conceptualising and developing the search strategy, keywords used to describe the technical aspects of SA were considered (e.g. cost, distance matrix), but not included in the final search strategy, because it was possible that some papers focused on a specific healthcare issue mentioned SA but none of the methodological terms – e.g. the basic operations (indel and substitution), the most common dissimilarity measures (e.g. OM, Hamming, or others), statistical packages and software used to conduct SA (e.g. ‘TraMineR’).

In order to be more inclusive, the keywords were searched in all text (TX) field, not limited to the title (TI) and/or abstract (AB), as some keywords were not necessarily described in the title or abstract. The search strategy was searched in four main medical and health databases: MEDLINE, PsycINFO, CINAHL (Cumulative Index to Nursing and Allied Health Literature) Plus with Full Text, and Embase (including Embase Classic, since 1947). The first three databases were integrated into the EBSCOhost research database, purchased by the University of Southampton, while the Embase was on the Ovid platform. The whole searching strategy and the number of records in the EBSCOhost (MEDLINE, PsycINFO, and CINAHL) and Ovid platforms (Embase) are in Appendix A. The records were restricted to publications in English, abstract available for screening, and published up till 31 December 2018. Some accepted papers (pre-print, in process) were retrieved by the search engine and included in this review, although they were formally published in 2019. The reference lists of the included papers were reviewed. Potentially relevant papers were hand-searched and further assessed their eligibility to be included in this review.

4.2.2 Inclusion and exclusion criteria

Since SA was still emerging in health research, it was expected that studies using SA had a great variety in different health-related contexts. A broader inclusion criterion may increase the chance to identify a broad range of potential uses of SA in health research. Publications using SA to study any type of disease trajectories, care pathway, or health services utilisation were of interest, and potentially eligible to be included in this review. However, the following types of studies were excluded:

1. Studies of genetic pathways in biology or genetics;
2. Work/family life-course trajectories in sociology or demography. Notably, some social epidemiological studies investigated the association between work/family life-course trajectories and health-related measures or outcomes, like Pedersen et al. (2016) and Benson et al. (2017), and others. The reason to exclude this type of study was that SA was used to study trajectories or pathways **in the social sciences domain** (employment history, career development, forming family), rather than health trajectories, which was out of the scope of this review;
3. As discussed in subsection 3.2.6, event sequence analysis can be used only to identify **fragments** of health events. This review aims to understand how SA has been used to analyse and identify patterns of the whole trajectories. Therefore, studies only used event SA like Rao et al. (2018a) were also excluded;
4. Review and pure methodological studies (rather than empirical studies), qualitative studies, conference abstracts, and unavailable full-text articles were all excluded.

Based on the search strategy, the EBSCOhost and Ovid databases returned 706 and 86 records, respectively (screenshots from each platform are in Appendix A). It was worth mentioning that there were only 659 records available after downloading from the EBSCOhost platform, with a gap of 47 (706-659) records. This could be that the EBSCOhost platform removed the exact duplicates among the MEDLINE, PsycINFO, and CINAHL databases for the users. The extracted references were managed by EndNote X9. The 33 duplicates in the PRISMA flow diagram (Figure 4.1) were those in both EBSCOhost and Ovid databases. After removing the duplicates, 655 records were excluded based on title and abstract screening, 97% of which (633/655) were in genetics and biology, and the remaining were in social sciences or conference abstract. The next step was reading the full-text articles and assessing their eligibility to be included in this review. Another 53 references were excluded at this stage for two main reasons: the majority of papers (89%, 47/53) were health studies but not involving SA. The other six were social epidemiological studies, and SA was used in the social domain (work-family life course trajectories). Finally, 13 studies were included in this systematic scoping review. The whole review process was present in Figure 4.1, using the PRISMA flow diagram (Moher et al., 2009). I was the main reviewer. The two supervisors, Dr Lucy Brindle and Dr Bronagh Walsh, each independently reviewed a random sample of 5.5% (n=40) from the excluded references at the stage of title and abstract screening (n=721) and 9.4% (5/53) at the stage of full-text screening, as a quality assurance measure. Any uncertainty was discussed and finally agreed.

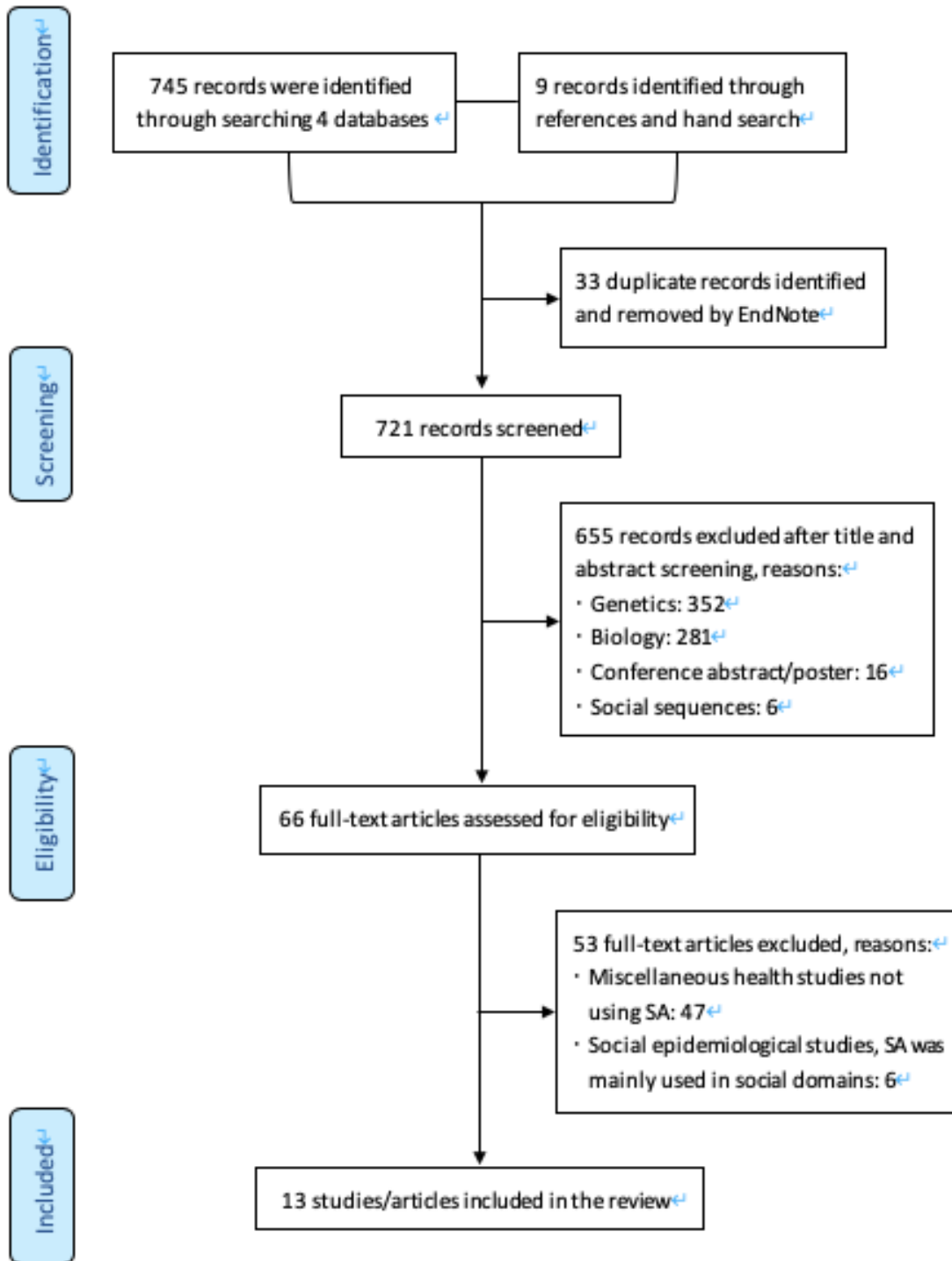


Figure 4.1 – The PRISMA flow diagram of the whole screening process for study selection and exclusion

4.2.3 Selection of critical appraisal tool and critique of individual paper

Critical appraisal tools would be very helpful to evaluate the included studies objectively and consistently with the same standards. Bучcheri and Sharifi (2017) summarised the commonly used critical appraisal tools and reporting guidelines. Critical Appraisal Skills Programme (CASP, the UK) and Joanna Briggs Institute (JBI, based in the University of Adelaide, Australia) provide the

most comprehensive checklists for different types of studies, to name a few, systematic review, meta-analysis, RCT, observational studies (case-control/cohort study), qualitative study, diagnostics, economic evaluation, clinical prediction rules.

The first step to choose the right appraisal tool was to know what types of studies were selected in the review. The common characteristics of the 13 included studies were **uncontrolled, longitudinal, quantitative, observational studies**. These characteristics greatly limited the choice of choosing an appropriate appraisal checklist, as they were:

- Longitudinal: not cross-sectional, and it was not necessary that the studies had an exposure, a control group, or confounders. Therefore, checklists developed for cohort studies, case-control studies, and prevalence studies may not be directly applicable to appraise these studies;
- Observational: not interventional/experimental, checklists for RCT or quasi-experimental studies were also not suitable;
- Quantitative: checklists for qualitative, text or opinion were irrelevant.

The NICE algorithm for classifying quantitative study designs (The National Institute for Health and Care Excellence, 2018) was referred to solve this dilemma and guide the decisions on which checklist should be used. The algorithm indicated the study type as case series (uncontrolled longitudinal study) and recommended three checklists for studies of this type, i.e. Institute of Health Economics (2016), the JBI checklist (Moola et al., 2017), The National Heart Lung and Blood Institute (NHLBI) (2013) checklist for case series studies. In the context of evidence-based medicine, case series was defined as “a report on a series of patients with an outcome of interest. No control group is involved” (Glossary of EBM terms). The Oxford Handbook of Medical Statistics (Peacock and Peacock, 2010) defined case series as “a descriptive study involving a group of patients who all have the same disease or condition. The aim is to describe common and differing characteristics of a particular group of individuals” (P34). These two definitions reflected the research design of the included studies the best. Of course, there are other definitions of case series in clinical settings.

The three checklists recommended by NICE are recognised and accepted in the scientific community. The checklists have different lengths and wording in each item, but cover general and vital elements to appraise a case series study, including RQ or research objective, study design, population, study outcome, statistical methods, results, and conclusions. After comparing the three checklists, it was decided to create a customised framework to assess the quality of the included studies by combining the strengths of the JBI and NHLBI checklists for the following reasons: firstly, none of the checklists was 100% suitable for the research context of the included

studies. For example, items related to intervention in case series did not apply to observational studies in this review. Secondly, SA focuses on studying sequential patterns of states over time, meaning that the outcomes in these studies are likely to be multiple and transitional. Thirdly, because the checklists were designed to appraise case series, substantial items were focused on the 'cases' (participants) – clear criteria for inclusion, valid methods to identify participants, consecutive and complete inclusion of participants, and comparable subjects. These criteria were relevant and essential to a clinical case series study, but somewhat repetitive in this review. Furthermore, none of the checklists advised a clear cutoff point to include or exclude the paper, which could probably lead to a subjective decision despite an objective assessment process. In conclusion, both JBI (in Appendix B) and NHLBI (Appendix C) checklists are easy to implement and user friendly, especially JCI, as it provides further explanation for each item with some examples, which makes a difference from those checklists without clear and explicit explanations. Ambiguous wording in some standalone checklists could confuse the users.

The customised critical appraisal checklist includes the following items. The options for these items are: yes, no, cannot determine, not reported, or not applicable.

1. Was the study question or objective clearly stated?
2. Was the study population clearly and fully described?
3. Were the states clearly defined, valid, reliable, and implemented consistently across all study participants?
4. Were the statistical methods (not just SA) well described and appropriate?
5. Were the results well described?

Besides the checklist, the application of SA in respective studies was critiqued about its strengths and limitations – whether there was any added value or novelty of using SA to reveal the key findings related to the RQ or fit in the research context in each study, and discussed other methodological issues (e.g. dealing with missing data, any potential selection bias).

4.2.4 Structured data extraction from the included papers

Key information from the included papers was extracted in a structured manner, based on the recommendations by Peters et al. (2015), including the following fields:

- Author(s) and the year of publication (together formatted as Harvard referencing style as required in this thesis);

- Country of origin;
- Aims/purpose of the study, with a particular focus on why SA was used to answer the RQ in each study;
- Research context, study population, and sample size

Additional information/fields regarding the methodological decisions of SA were also extracted for this review, including:

- Duration of the observation – intervals and timeline of sequences;
- How states in the sequences were defined and measured;
- Dissimilarity measures (e.g. OM, Hamming, DHD, or others) and cost settings in SA;
- Clustering algorithm and the decision on the number of clusters

The extracted information in respective publications is organised and reported in the Result section (Table 4.2).

4.3 Results and interpretations

4.3.1 Results of critical appraisal of the included papers

For all the 13 included papers, the study objectives were clearly stated, and the study population was fully described in each study. The states for sequence construction in each study were clearly defined and implemented consistently across all study participants. The statistical methods used in each study (not just SA) and the results were critiqued, reported in the following subsections. Important information for each study is summarised and presented in Table 4.2.

4.3.2 Characteristics of the included studies, research design and context, data source and sample size

The included studies indicate that SA is still quite a new method in applied health research. The number of publications started to increase in 2015. The method was more used by researchers from the western developed countries (the United States 5 papers, France 3 articles, England and Germany 2 studies for each, and the Netherlands 1 publication). Despite the research context and the study population of two studies being from developing countries (India and Malawi, both studied HIV), the first and last authors in these publications were from developed countries. In almost all manuscripts reviewed, the authors claimed the originality of their studies by stating that their attempt of applying SA to solve a problem was a novel approach, and the first one in their respective research field.

The application of SA in health research among the 13 included studies varied considerably. Three studied mental health – two in chronic mental illness among homeless persons (Wuerker, 1996, Lim et al., 2018), and one explored long-term clinical course (the psychopathological status and syndromes) of schizophrenia (An der Heiden and Hafner, 2015). Four studied physical illnesses, two in heart failure (Rao et al., 2018b, Vogt et al., 2018), one about community-acquired pneumonia (Hougham et al., 2014), the other one about end-stage renal disease (Le Meur et al., 2019). Two studied HIV in developing countries, one about the reproductive trajectories (marriage and childbearing) and the awareness of HIV infected status among women in Western India (Darak et al., 2015), while the other aimed to understand participants' opinions about the prioritisation of receiving antiretroviral therapy (ART) in different populations in Malawi (Yeatman and Trinitapoli, 2017). Two French studies used SA to explore the level of care consumption in health services, among pregnant women in their trimesters (Le Meur et al., 2015), and among patients with multiple sclerosis (Roux et al., 2018). The remaining two studies explored the trajectories in non-patient populations, one in vision changes among English elder people (Whillans et al., 2016), the other in BMI changes among American pupils (Moreno-Black et al., 2016).

As to the source of longitudinal data, nine studies used electronic health records (EHRs), administrative data, health insurance claim data, or existing data (the English Longitudinal Study of Ageing, ELSA) for secondary analysis. The remaining four were studies with primary data collection – two studies of HIV (Darak et al., 2015, Yeatman and Trinitapoli, 2017) collected data by interview. Sample sizes of the studies were ranged from around 50 to 10,000. Studies using existing databases (EHRs and administrative data) tended to have larger sample sizes, and the authors were more likely to claim their samples representative of the intended study population.

4.3.3 The added value of applying SA in respective studies and some comments for further improvement

SA has been applied in different ways to solve problems in health research. In most studies, sequences were consecutive states in the timeline. SA was used together with cluster analysis in most studies to explore trajectories, identify typical patterns, and create typologies.

SA is a useful descriptive tool to present and visualise complex sequences over a long period in various figures (e.g. state distribution plot, state frequency plot, sequence index plot, regression tree), which is the biggest strength of SA. Using the right graphic presentation can ease the communication of the findings. The included studies provide some examples of researchers using figures exclusive to SA to effectively communicate their findings. Sequence frequency plot was

used to present the most frequent sequences in each cluster of vision trajectories among older people in ELSA (Whillans et al., 2016). It has been popular to use the most frequent sequences to represent the cluster patterns. In most situations, multiple sequences are needed to reach a certain level of coverage (e.g. covering 25% of all the sequences in a cluster). Vogt et al. (2018) also reported the ten most frequent sequences in each cluster in figures, which allowed readers to understand how patients diagnosed with incident heart failure consumed health services in Germany through the typical procedures, specialities, and medication sequences. But if the sequences have considerable heterogeneity among each other, it could be challenging to find a few sequences to represent the clusters, as they do not reach the defined threshold (e.g. 25%).

Sequences were creatively used to represent the order of events in health services in two studies. Hougham et al. (2014) constructed the sequences based on the order of stability in seven clinical indicators (blood pressure, return to baseline mental status, ability to feed by oral intake, respiratory rate, temperature, heart rate, blood oxygen saturation) among adult patients (age 18+ years, n=1,461) after hospital admission because of community-acquired pneumonia from five academic medical centres in the US. That study aimed to understand the variation of sequences, and to assess the association between the patterns of stabilisation and patient-level outcomes (30-day mortality, length of stay, and hospitalisation costs). Although the approach of constructing sequences was interesting, the technical issues were not addressed appropriately. For example, the stability of the clinical indicators was dynamic. Some stable indicators might become unstable at a later time point of hospitalisation. The sequences in that study did not represent the possible changes in the clinical indicators, which was a limitation.

In another study, sequences were constructed by the order of the perceived prioritisation to receive antiretroviral therapy for HIV among six groups of people with combinations of the three characteristics (healthy-looking or sick pregnant women, healthy-looking or sick non-pregnant women, and healthy-looking or sick men). That study aimed to understand the awareness and perceived fairness of a health policy among participants in Malawi (Yeatman and Trinitapoli, 2017). Despite its novel use, SA was simply used as a descriptive tool without cluster analysis. Simple descriptive statistics (e.g. a frequency table) could have generated the same result, as the order of priority did not have any sequential or temporal implication. It was just the perception of the participants.

Chapter 4

Le Meur et al. (2015) used an EHR database to investigate the consumption of prenatal care in trimesters among French pregnant women. Among 2518 women, 27 distinct sequences were identified, which meant the sequences were relatively homogeneous, as the most frequent sequence could represent 34.7% (n=873) women, while the least frequent one still covered 3.3% (n=84) women. The reason for such homogeneous results was probably because only three states were specified – ‘no use of care’, ‘intermediate level’, and ‘high level of care consumption’, which were also the three themes identified by cluster analysis, representing 21%, 66%, and 11% of the population, respectively. The authors could have improved that study by taking a further step, comparing the results between the typology and the descriptive statistics, and concluded whether it was possible to use a simpler method (e.g. a frequency table summarising the patterns of prenatal care consumption in three gestational periods, with the frequency and percentage for each pattern) to substitute the more complicated and troublesome approach (SA and cluster analysis). The authors may conclude whether using simpler descriptive statistics could answer the RQs in that study or not. This study was a special case, as it only involved three states in three periods and highly homogeneous sequences. For longer sequences with a larger number of states, the sequences would become more heterogeneous. In such a case, using simple descriptive statistics could not identify meaningful patterns of sequential changes over time, nor effectively communicate the information in a frequency table.

How to specify states was dependent on the RQs in respective studies. The number of states of the included studies ranged from 3 to 15. Applying prior knowledge of the research context in state specification would be more likely to yield meaningful typology, especially for the sequences containing complex information from multiple dimensions. One good example was the study by Darak et al. (2015). Each state comprised information from three dimensions (marital status, awareness of HIV status, and childbearing status), as shown in Figure 4.2. The graphic typology allowed the readers to appreciate the distribution and the change of states over time across participants’ life course in each cluster, alleviating readers’ cognitive burden to digest complex information. Another good way to make sense of the cluster pattern was to use medoid sequences or summarise the pattern in text. For example, in the same study, the most central (typical) trajectory for the cluster “HIV diagnosis concurrent with childbearing” (Cluster 1, Figure 4.2) was (women) getting married at an average age of 21.5 years -> pregnant within a year of marriage -> tested HIV positive during the first pregnancy -> partner tested HIV positive immediately thereafter -> living concordant with one child. The explanation in text helped readers to understand what the graphic presentation of the cluster meant. However, it was not always possible for all the studies to do this, as some clusters may have heterogeneous sequences

leading to less apparent patterns. Demographic variables (woman's age at marriage, education, urban/rural residence, and the period of HIV diagnosis) were identified as risk factors using multinomial logistic regression (one cluster as reference). This study not only provided a holistic perspective of the reproductive trajectories and the HIV infected status among Indian women, but also identified demographic risk factors for each cluster. All of these were new knowledge generated by the unique features of SA, which could not have been provided by other methods.

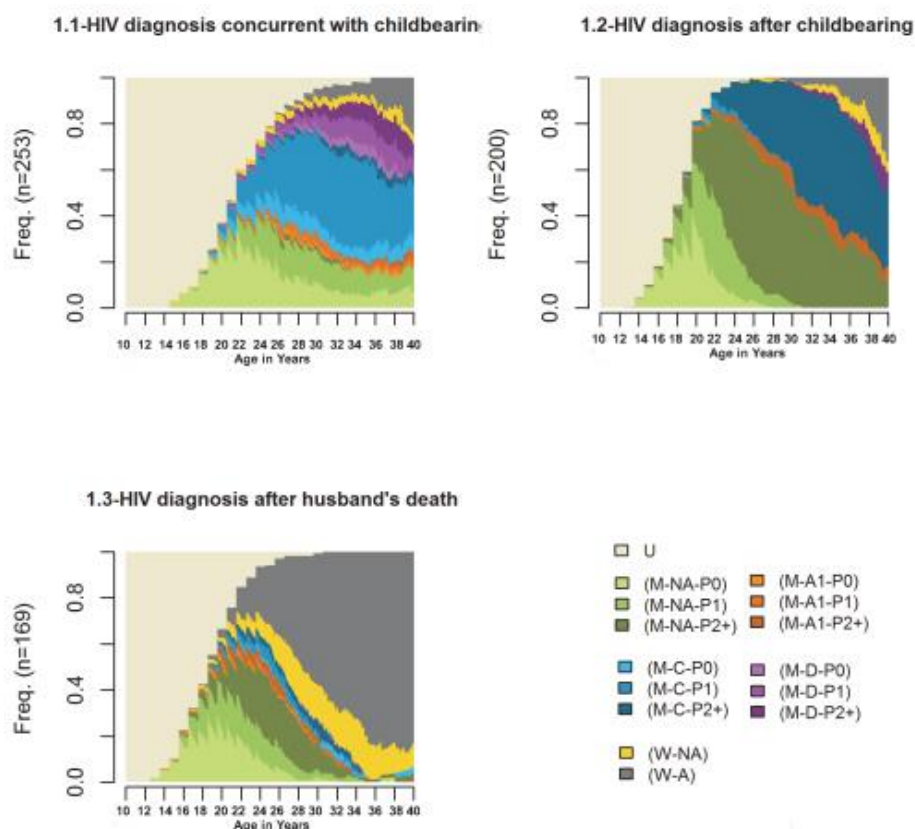


Fig 1. Clustered typologies of reproductive career trajectories of ever married HIV infected Indian women (N = 622). U- unmarried; M-married; W-widowed; NA-both the woman and her husband are not aware of their HIV status; A1-only one partner is aware of his/her HIV status; C-concordant (the woman and her husband are known HIV infected); D-discordant (only woman is known HIV infected and husband is HIV uninfected); P0- parity zero; P1- parity one; P2+-parity two and over. See [Table 1](#) for a detailed description of the different states.

Figure 4.2 – States containing information from multiple dimensions and the typology of sequences from Darak et al. (2015)

Another way to present and analyse sequences with complex information was to construct sequences in multiple channels, each channel for one specific dimension. Vogt et al. (2018) aimed to identify typical treatment sequences in ambulatory care for German patients with an incident diagnosis of heart failure. Instead of integrating information into states, the authors constructed three sequences for each patient – speciality sequence (GP, internist, cardiologist, missing), procedure sequence (electrocardiogram, echocardiography, lab test, missing), and medication sequence (ACE hemmer, angiotensin receptor blockers, beta-blockers, missing), and analysed each sequence separately. The findings of cluster patterns were informative for patient care and management, as well as health services planning. However, the authors did not synthesise and

triangulate the information from the three sequences. The authors could have at least tabulated the patterns from the three sequences. Alternatively, they could have used multichannel sequence analysis to analyse the three channels together, rather than analysing each channel separately.

Rao et al. (2018b) used event sequence analysis to identify six significant sub-sequences of the common causes of two emergency readmissions (e.g. cardiorespiratory symptoms/signs -> chest infection; external injury -> chest infection) among five groups of patients with heart failure. It was useful knowledge for clinicians to act proactively and avoid such incidents. But the limitation of event sequence analysis was that we could not know when these events happened in the timeline or the gap between the two events (it could be one month or several years), or whether there were any other medical events that happened between the two emergency readmissions. All of these were of clinical interest, but event SA could not adequately address them. In addition, there were substantial overlapping health events in the sub-sequences. Such knowledge may have already been reported in other studies in that field. Group-based trajectory modelling based on zero-inflated Poisson regression was used to classify patients into five subgroups. State SA was only used as a descriptive tool in that study. Both group-based trajectory modelling and SA can identify subgroups as clusters. The main difference is that the former method is often used to investigate the change of a single outcome (often a continuous variable, e.g. physical development in children, cognitive decline in the elderly) over time using a likelihood function, while state SA is used to investigate the stability or transition of different states over time and based on algorithmic edit distance.

4.4 Summary and discussion of the methodological/statistical decisions related to SA

4.4.1 The interval and sequence length

The intervals of the timeline included month, quarter, and year. The longest timelines of the sequences were 134 months (by month) (An der Heiden and Hafner, 2015) and 30 years (by year) (Darak et al., 2015), respectively; while the shortest length of sequences was only five (5 years or five waves of data) in 3 studies. Sequences had the same length in each study. Although sequences could have unequal length, this situation did not appear in the included studies. Dlouhy and Biemann (2015) recommended a minimum of 25 elements in sequences (subsection 3.6.2), but it is not always possible.

4.4.2 Dissimilarity measures and cost setting

OM was the most commonly used algorithm, while in three studies, researchers chose alternative algorithms. Whillans et al. (2016) used dynamic Hamming distance (DHD) but did not further explain why they chose DHD in that study, nor discussed its implication in the results. Vogt et al. (2018) and Le Meur et al. (2019) chose the longest common subsequence (LCS), as they were interested in the most common attributes occurring in the same order among sequences. This approach is similar to event sequence analysis, selecting the same events from sequences, as the word 'subsequence' in LCS indicates, with an additional calculation of the number of common sub-sequences between two sequences. As to cost setting for the three basic operations in OM, it was very common to set indel cost as one and substitution cost proportional to the transition rates among states, i.e. $OM_{[1, TR]}$. This data-driven approach to set substitution cost seemed reasonable, when there was no theoretical knowledge available. An der Heiden and Hafner (2015) used a customised substitution cost matrix to represent the severity of psychopathological syndromes of schizophrenia. Although the author made a footnote saying that it yielded an almost identical cluster pattern as the traditional cost setting (indel cost=1, substitution cost=2), it was not clear whether only substitution was used, or together with indel in the customised substitution cost matrix. Because if the substitution cost is greater than two times of the indel cost, substitution would not be used in OM, as they are not computationally efficient. Indel operations are used instead (discussed in subsection 3.2.4).

4.4.3 Sequence clustering/partition

Ward's method (agglomerative hierarchical clustering based on the dissimilarity matrix) was the clustering technique used most often. However, very few studies reported the criteria to decide the number of clusters. Hougham et al. (2014), Le Meur et al. (2015) and Moreno-Black et al. (2016) used **average silhouette width (ASW)** to assess the cluster quality, as introduced in subsection 3.3.3. The last two studies reported a reasonably good clustering structure in general, with the overall ASW value of 0.52 and 0.72 in respective studies. Hougham et al. (2014) examined the ASW values from 2 to 20 clusters of sequences, and the ASW values were ranged from 0.11 to 0.26. The authors finally chose eight clusters (ASW=0.14) as the optimal solution, considering the cluster quality (although not the highest ASW), discrimination between the types of sequences and patients, and cluster size (all >30 patients), which was a sensible decision.

Regression tree (also known as **discrepancy analysis**) was used in some studies to split the nodes and identify the key determinants to explain the variation among clusters. For example, it was

used to split groups of homogeneous care trajectories by a number of variables (e.g. age group, sites, diabetes, nephropathy, disability) among patients with end-stage renal disease (Le Meur et al., 2019), as shown in Figure 4.3. Only a small proportion (12%) of the total variation in the care trajectories could be explained by the covariates in regression tree, as other variables or information that may be more helpful to explain the variation among the sequences might not be available.

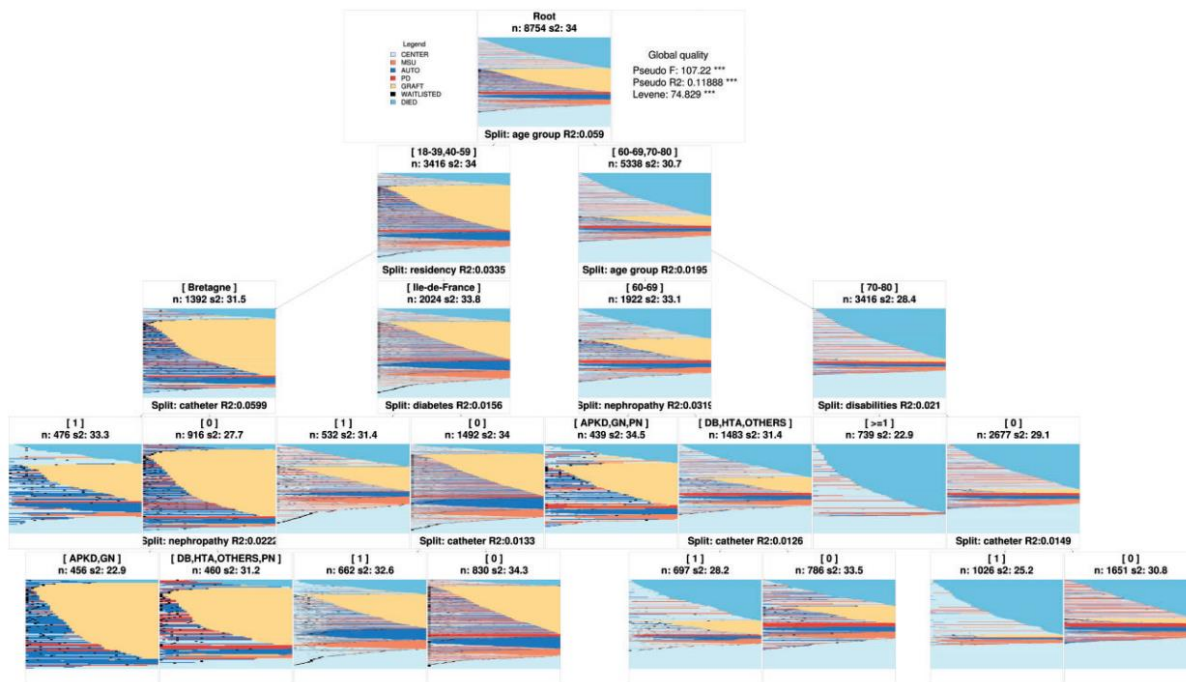


Figure 4.3 – An illustration of regression tree, from Le Meur et al. (2019)

4.4.4 Approaches to identify significant covariates among the clusters

Three ways were used in the included studies to identify significant variables among clusters: regression tree (discrepancy analysis) (Le Meur et al., 2019), multinomial logistic regression (Darak et al., 2015), and simple parametric and non-parametric tests (e.g. ANOVA, Pearson's chi-square test, Kruskal-Wallis test) (Roux et al., 2018).

Regression tree is essentially different from the other two approaches, which is done without cluster analysis. It starts from one big cluster including all the sequences (the initial parent node), and then recursively partition the sequences by splitting a binary node with the biggest R^2 at each step. The result (two child clusters) would differ as much as possible from one another, while the sequences within the same cluster as homogenous as possible. Regression tree is a powerful graphic tool to present the result, as shown in Figure 4.3. It allows the readers to understand the entire process of how the covariates partition the sequences from the beginning more intuitively, and how the covariates explain the variance of sequences at different steps. However, the biggest

limitation of regression tree is that it can only split a **binary** node at each time. Continuous and categorical variables need converting to several binary variables if researchers want to use regression tree, which could be burdensome to find the best cutoff values for each variable to partition the clusters.

Multinomial logistic regression is another possible analytical approach to identify significant variables for different clusters. The clusters are the dependent variable (categorical), with one cluster as the reference category, and the independent variables could be any data type (continuous, binary, categorical, or ordinal). Researchers can use multinomial logistic regression to do uni-variable and multivariable analyses. But if the number of clusters is big, with some categorical variables in the multivariable multinomial logistic regression model, more advanced statistical knowledge is needed to interpret the results. A simpler solution would be using parametric and non-parametric statistical tests first, to compare the characteristics among the clusters and identify significant variables. Such analysis could be done before running multinomial logistic regression to select significant variables for the multinomial logistic regression model.

When using regression tree, it would be better to have some background knowledge (from theory or previous empirical evidence) on the relationship between the covariates and the sequences, as the covariates directly partition the sequences and determine the clusters. Without previous knowledge, one may not select the right covariates to explain the variance of sequences, or set the right cutoff points to convert the covariates to binary variables, which has an impact on the final typology and the interpretation of the cluster patterns. Whereas in multinomial logistic regression model, the exploration of the association between the covariates and the clusters is done after cluster analysis. This approach is perhaps more pragmatic during the early stage of research, to obtain preliminary empirical knowledge. This approach was used in this PhD study.

4.4.5 Dealing with missing data

In most situations, SA uses longitudinal data in the analysis. Dropouts and missing data are common problems in longitudinal data. It is possible to specify a “missing data” state in the sequences, as Vogt et al. (2018) did in their study, and use a full dataset in the analysis so that more participants could be included, rather than only using a subset with complete data only. In a simulation study with different levels of missing values, Dlouhy and Biemann (2015) concluded that OM could perform well even with up to 30% of elements missing in the sequences. More missing data would increase the misclassification rate, i.e. sequences could be mistakenly classified in a wrong cluster. Discarding sequences with missing data completely would be a waste of data, and probably would lead other researchers to question the reliability and generalisability

of the conclusions. Moreno-Black et al. (2016) used a complete dataset only including 22.4% of the original subjects (414 in the analysis out of 1,847 recruited at baseline). A similar issue was in the study by Whillans et al. (2016), where the authors only included participants who responded to the first five waves of ELSA (complete dataset). The authors could have tried to do the analysis using a full dataset, and additional sensitivity analysis with complete data only, and compare the findings between the two sets of analysis. Alternatively, the authors could also compare the demographic characteristics between participants with complete data and missing data, and evaluate whether it was appropriate to use complete data only in analysis, whether there was any potential selection bias, whether the group with complete data had more favourable outcome than those with missing data, as this issue was relevant to the representativeness of the study population and the generalisability of the study conclusion. The readers may find such an analytical approach more informative.

Certainly, other statistical methods may be more able to cope with missing data than SA, but they are probably not able to present the sequential patterns. The authors should explain the possible reasons for missing data wherever possible, assess the impact of missing data (e.g. whether the missing data were in a considerable amount to compromise the whole sequence or not), make a sensible decision on what statistical method is appropriate to answer the RQs in a particular study, and discuss the implications of missing data to the study findings.

4.4.6 A final brief note

Two studies (Yeatman and Trinitapoli, 2017, Rao et al., 2018b) used SA only as a descriptive tool to present sequences at the individual level by sequence index plot. Therefore, dissimilarity measures, cost setting, and clustering techniques were not relevant in those two studies.

4.4.7 Strengths and limitations of this review

This review followed the guidelines and the best practice to conduct and report a systematic scoping review. Scoping review is useful when a body of literature has not been comprehensively reviewed, or the studies in a particular field exhibits a complex or heterogeneous nature not amenable to a more precise systematic review of the evidence (Peters et al., 2015). The focus on the methodological issues related to SA and the analysis of the added value of using SA in each included study are the unique strengths of this review. The great variety of the included studies posed a great challenge to adopt a critical appraisal tool applicable to each study. Despite this, efforts were made to choose appropriate, recognised, and widely accepted tools and modify the items to appraise the quality of each study in this review. Some may criticise the customised items

for critical appraisal and the appropriateness of combining items from the two tools, and argued that this process was liberal, without going through the whole process of pretesting and establishing item validity, usefulness, reliability before using it. This could be a limitation, but developing a critical appraisal tool to assess the quality of studies using SA was not in the scope nor the purpose of this review. This research gap could be an opportunity for future studies to address. However, there were precedents of creating a customised checklist based on the developed tools in the field of early diagnosis research. For example, being unable to find a suitable tool to assess the quality of cohort studies, Forrest et al. (2017) produced their bias assessment checklist by adapting a previously validated tool. They claimed their customised tool was unique and highly specific to detect bias in the type of studies included in their systematic review and meta-analysis. Grey literature was not included, which was another limitation of this review.

Table 4.2 – Summary of study characteristics and how SA was used in each study

No	Study and reference	How SA was used, research context and aim of the study (country of origin)	How patients were selected and sample size	States in the sequence	Interval	Distance measure	Cost setting
1	Wuerker (1996)	To describe the patterns of service use over time by a group of homeless with chronic mental illness (US)	49 subjects, each had 25 or more admissions to any of the Los Angeles County Department of Mental Health Services before 1 July 1993	Six service settings: inpatient, outpatient, emergency, jail, day treatment, residential, and others	The interval between two admissions was not represented in the sequence.	OM	Not mentioned
2	Hougham et al. (2014)	To describe the variation in the sequences – the order of stability in 7 clinical indicators among patients hospitalised with community-acquired pneumonia, and to assess the associations between the patterns of stabilisation and patient-level outcomes (US)	1,461 patients age 18+ years hospitalised due to community-acquired pneumonia from 2000 to 2003 across five academic medical centres in the US; Seven clinical indicators: blood pressure, return to baseline mental status, ability to feed by oral intake, respiratory rate, temperature, heart rate, and blood oxygen saturation	Ten states in total: the rank/order of stability in clinical indicators (from 1 to 7) during the hospital stay, 0 if stable at admission, 8 - indicators not stabilised at discharge and alive, and 9 - dead	The order of the stability among seven clinical indicators, sequence length = 7	OM	Not reported
3	An der Heiden and Hafner (2015)	To explore the long-term clinical course (psychopathological status/syndromes) of schizophrenia (Germany)	107 patients participating in the Age-Beginning-Course (ABC) Schizophrenia Study (launched in 1987)	Seven psychopathological states of schizophrenia: inconspicuous, unspecific, depressive, negative, positive, negative + depressive, positive + negative + depressive	Month, 134 months in total	OM	Indel cost=1, substitution=2, also with a self-defined substitution cost matrix, ranged from 1 to 6

4	Darak et al. (2015)	To investigate the reproductive trajectories and the awareness of HIV infected status among Indian women (the lead author was based in the Netherlands)	Retrospective data of 622 ever married HIV infected women aged 15-45 attending an HIV clinic in Pune, Maharashtra, Western India, were collected through interviews.	15 states, the states were combined with three dimensions (marital status, awareness of HIV status, and childbearing)	Year, the life course from 10 to 40 years in the participated women's lives	OM	Not mentioned
5	Le Meur et al. (2015)	To mine care trajectories and assess the disparities in prenatal care consumption (France)	2,518 women who gave birth without complications in 2009, using data from electronic health databases	Three levels of care consumption: absence, intermediate, and high	Month, nine gestational months in total	OM	Not mention indel cost, substitution cost matrix used transition rates between states
6	Moreno-Black et al. (2016)	To explore the changes in obesity status (BMI transformed into four categories) in elementary school children (grades K-5, US)	Data were collected annually from kindergarten and first-grade students participating in the Community and Schools Together (CAST) project (n=414).	Four BMI categories: normal, overweight, underweight, and obese	Year, data were collected for five years from 2008 to 2013	OM	Not explicitly stated, but a table of transition probabilities was provided
7	Whillans et al. (2016)	To identify typical trajectories of self-reported vision, and factors associated with different vision trajectories among older people (UK)	2,956 respondents, aged 60+ years at wave 1 of the English Longitudinal Study of Ageing (ELSA)	Four vision categories: poor vision or blindness, fair vision, good vision, excellent or very good vision	By waves (5 waves of data across eight years)	DHD	No indel cost, time-varying substitution costs inversely proportional to the observed transition frequencies

Abbreviations: DHD – dynamic hamming distance; LCS – longest common subsequence; N.A. – not applicable; OM – optimal matching

No	Study and reference	How SA was used, research context and aim of the study (country of origin)	How patients were selected and sample size	States in the sequence	Interval	Distance measure	Cost setting
8	Yeatman and Trinitapoli (2017)	To depict the perceived and the ideal order of prioritisation to receive antiretroviral therapy (ART) among different populations in Malawi (Lead author was based in the US).	Young women (n=1,440) and their partners (n=574) in southern Malawi	Six types of populations: sick man, healthy-looking man, sick non-pregnant woman, healthy-looking non-pregnant woman, sick pregnant woman, and healthy-looking pregnant woman	N.A.	N.A.	N.A.
9	Rao et al. (2018b)	1. State sequence analysis was used to characterise healthcare visits (health services use) among different subgroups of heart failure patients. 2. Event sequence analysis was used to identify common distinct causes of emergency readmissions (sub-sequences) for each patient subgroup (UK).	Patients with a primary diagnosis of heart failure during 2008-2009 were identified and followed up for five years (n=9,466) CPRD linked to HES and ONS (UK)	Six states of healthcare visits: no visit, elective GP visit, elective hospital admission, emergency GP visit, emergency hospital admission, and death	Year (5 years in total)	N.A.	N.A.

10	Vogt et al. (2018)	To explore whether sequence clustering techniques can be used to identify typical treatment sequences in ambulatory care for heart failure patients (Germany)	Anonymised regional statutory health insurance claims data from 2009 to 2011, among patients with an incident diagnosis of heart failure (n = 1,577)	Three separate sequences: Procedure sequences (electrocardiogram, echocardiography, lab test, missing); Speciality sequences (GP, internist, cardiologist, missing); Medication sequences (ACE hemmer, angiotensin receptor blockers, beta-blockers, missing)	Two years (8 quarters), as health insurance payment was billed by quarters in Germany.	LCS	N.A.
11	Lim et al. (2018)	To assess the association between supportive housing tenancy program and the Medicaid savings in the New York City (US)	New York City housing program applicants (long-time homelessness), with serious mental illness (N=2,827; 737 placed, 2,090 not placed)	Monthly Medicaid cost (health expenditure): 0. no Medicaid coverage; 1. \$0; 2. \$1-\$292; 3. \$293-\$690; 4. \$691-\$1,379; 5. \$1,380-\$2,898; 6. ≥\$2,899	2-year pre-housing period, by month (24 months in total)	OM	Substitution costs were estimated based on transition probabilities.
12	Roux et al. (2018)	To investigate the care pathway among patients with multiple sclerosis (France)	Multiple sclerosis patients were identified from a French health insurance database (n=1,000).	6 levels of care consumption: 0 (no consumption at all), (0-Q1], (Q1-Q2], (Q2-Q3], >Q3, and missing, where Q represented annual quartiles.	Year (7 years in total, from 2007 to 2013)	OM	Substitution costs were estimated empirically using the observed transition rates.
13	Le Meur et al. (2019)	To study the care trajectories of patients with end-stage renal disease, and identify homogeneous care trajectories at group level explained by different covariates (France)	5,568 incident patients aged 18-80 years old (2006-2009), using data from the French Renal Epidemiology and Information Network registry	Six renal replacement therapy modalities: in-centre haemodialysis, medical unit, autonomous haemodialysis, peritoneal dialysis, death, and transplantation.	Month (48 months in total), the timeline was reconstructed, as patients had different start points of dialysis	LCS	N.A.

continued (part 2)

No.	Study and reference	Clustering and the decision on the number of clusters	Strengths and limitations of using the SA in this study	Other methodological issues and comments
1	Wuerker (1996)	Based on peaks in the cubic clustering criterion, pseudo F, and pseudo t^2 of cluster analysis in SAS	This was a very early attempt of using SA in health services research. However, this study had a relatively small sample size (n=49) and may have a potential selection bias, as the author chose subjects who had at least 25 admissions and picked the first 25 episodes of care. Eight clusters were identified, with only a few sequences in each cluster. The authors acknowledged that the sample was not representative.	Due to the availability of the software and package, the presentation of sequences was not as attractive as it is nowadays. Despite this, the authors communicated the sequences clearly.
2	Hougham et al. (2014)	Ward's clustering method	Strengths: eight clusters of the order of stabilisation among clinical indicators were identified for patients admitted to the hospital with community-acquired pneumonia. The authors compared the patient characteristics among eight clusters, also explored the associations between the eight clusters and the three important clinical outcomes (30-day mortality, length of stay, and hospitalisation costs), which had implications for patient care and management, health services and cost. Limitations: the stability of clinical indicators was dynamic. Some indicators could be stabilised simultaneously. The authors assigned them at the same rank. Alternatively, the authors could have given them different ranks based on the importance of indicators relative to survival/mortality. It was possible that some stable indicators may subsequently destabilise. The authors did not account for this possibility and reflect it in the sequences.	Parametric and non-parametric statistical tests were used to compare patient characteristics among five different medical centres and eight clusters of stability in clinical indicators. Generalised linear regression was used to investigate the associations between eight clusters and the 30-day mortality, length of stay, and hospitalisation costs.

3	An der Heiden and Hafner (2015)	Agglomerative hierarchical clustering (Ward's)	<p>The data in this study were collected from a clinical setting, specifically designed for schizophrenia, with a relatively long follow-up period (134 months, 11+ years)</p> <p>The authors used a customised substitution cost matrix, representing the severity of psychopathological symptoms. Although the author made a footnote saying that it yielded an almost identical cluster pattern as the traditional cost setting (indel cost=1, substitution cost=2), it was not clear whether only substitution was used, or together with indel in the customised cost matrix.</p>	<p>Further comment: if the substitution cost was greater than two times of the indel cost, it would not be used in OM, as it is not efficient in computation. Indel operations would be used instead.</p>
4	Darak et al. (2015)	Not clearly described, but mentioned using dendrogram to examine the clustering of sequences	<p>The states were a combination of three dimensions (applying prior knowledge of the research context). In such a way, it was possible to use traditional sequence analysis (rather than multi-channel sequence analysis).</p> <p>The authors also used medoid sequences to represent at least 25% of sequences in each cluster. Typical trajectories were described.</p>	<p>Multinomial logistic regression was used to identify significant demographic variables (age at marriage, women's education, urban/rural residence, and year of HIV diagnosis) belonging to the identified clusters (one cluster was used as reference)</p>
5	Le Meur et al. (2015)	Agglomerative hierarchical clustering (Ward's method)	<p>The trajectories of care consumption among pregnant women could have been more informative by classifying the states in more detail. There were only three states in this study.</p> <p>Traditional descriptive statistics (frequency table) may achieve the same study objectives, which would be a much easier approach, without going through all the complicated statistical analysis (SA and cluster analysis).</p>	<p>Logistic regression was used to analyse the association between the variables of interest and the identified clusters in both univariable and multivariable analyses. However, the authors ran logistic regression three times by comparing each cluster against the other two (e.g. Cluster 1 VS Clusters 2&3). It would be more appropriate to use multinomial logistic regression.</p>

No.	Study and reference	Clustering and the decision on the number of clusters	Strengths and limitations of using the SA in this study	Other methodological issues and comments
6	Moreno-Black et al. (2016)	Agglomerative hierarchical clustering (Ward criterion)	Due to the dropout at each wave, the authors only included 414 students with complete data from all five waves, while there were 1847 subjects at baseline (only 22.4% of all subjects were included in SA). Regression tree was used to split binary nodes. Two covariates, socioeconomic status (using subsidised meal as a proxy) and ethnicity (Hispanic VS white and others), were significant predictors at the significance level of 0.01, accounting for the greatest discrepancy among sequences.	Although it was an interesting attempt to use SA and regression tree in this study, another possible analytical approach was to take BMI as a continuous outcome and use multilevel modelling or latent growth model, which may be more able to cope with missing data and repeated measures than SA.
7	Whillans et al. (2016)	Agglomerative hierarchical clustering (Ward's method)	SA was used to identify eight clusters of vision trajectories among older people (aged over 60) from ELSA, which helped authors to understand how vision changed in this population over time. Based on this, the authors were able to predict vision trajectories by social position and age groups. Sequence frequency plot was used to present the most frequent sequences (as representative sequences) in each cluster. The authors only included participants with complete information on the first five waves of the study. The sample size was further reduced by the age limit.	Multinomial logistic regression was used to examine the sociodemographic characteristics associated with different trajectories.
8	Yeatman and Trinitapoli (2017)	N.A.	Sequences in this study were the order of priority to receive antiretroviral therapy among different sub-populations from participants' perception. There was no interval and group classification (cluster analysis). SA was used as a descriptive tool. It was an interesting attempt of using SA to explore how participants perceived health policy.	No further comment for this study

9	Rao et al. (2018b)	N.A.	<p>1. State sequence analysis was used as a descriptive tool to delineate the nature of healthcare visits among different subgroups of patients with heart failure during follow-up. It could argue that some states overlapped with the outcome (readmission rate) of the group-based trajectory modelling (presenting similar information in different ways). 2. Common symptoms/signs of two consecutive emergency admissions were identified by event sequence analysis for each group. It was useful knowledge for clinicians to act proactively and avoid such incidents. However, the limitation of event sequence analysis was that we could not know when these events happened, whether it was one month or close to five years during the observation period.</p>	<p>State sequence analysis was not used to identify trajectories. Instead, group-based trajectory modelling based on zero-inflated Poisson analysis was used to classify patients into five groups and to predict the development of their trajectories. The outcome was readmission rate. However, sequence analysis could be used as an alternative method for trajectory classification. It would be interesting to compare the group memberships of the patients between the two methods.</p>
10	Vogt et al. (2018)	k-medoids clustering	<p>The authors constructed three sequences – speciality, procedure, and medication sequences for each patient with incident heart failure, which was informative from the perspective of health services use. However, the authors did not synthesise the information together from three individual sequences. They could have tried multichannel sequence analysis to run three channels together, rather than doing it separately. Alternatively, they could have tabulated the cluster patterns from the three dimensions.</p>	<p>Patient characteristics were compared among clusters using traditional statistical tests (e.g. ANOVA, χ^2 test, and others) Logistic regression was used to explore the association between cluster membership and hospitalisation, adjusting for age, sex, and the Charlson comorbidity index.</p>
11	Lim et al. (2018)	Hierarchical cluster analysis (Ward method)	<p>The authors used SA to identify six distinct patterns of Medicaid users (from very low coverage to high users) by month for two years. The authors also broke down the Medicaid expenditure by service categories (e.g. outpatient, inpatient, emergency, prescription, others) and compared across clusters, which allowed readers to understand how health expenditure was spent on different health services.</p>	<p>Propensity score matching was used to minimise the baseline differences between the placed (applicants into the housing programme) and the unplaced group. Traditional statistical tests were used to compare the characteristics of subjects in different clusters of Medicaid expenditure.</p>

No.	Study and reference	Clustering and the decision on the number of clusters	Strengths and limitations of using the SA in this study	Other methodological issues and comments
12	Roux et al. (2018)	Agglomerative hierarchical clustering analysis with Ward's criterion on the dissimilarity matrix	The authors illustrated the whole process of SA and discussed relevant methodological aspects of SA. The authors' purpose to conduct this study was achieved. However, the authors did not clearly explain how different care consumption pathways could be used to improve the services of multiple sclerosis.	General statistical tests (Pearson's chi-square test, Fisher's exact test, and Kruskal-Wallis) were used to identify significant characteristics among patients in different clusters.
13	Le Meur et al. (2019)	Regression tree (discrepancy analysis) was used to split the nodes and identify the key determinants for clusters	Regression tree was able to estimate the association and interaction of multiple factors with the sequences of renal replacement therapy modality. However, only 12% of the variation of the care trajectories could be explained by regression tree. The limitation of regression tree was that the split at each step was binary. It may be challenging to determine the right cutoff point for continuous variables (e.g. age) and excessive use of the same variable (e.g. age) with different cutoff points at different levels of clusters.	The authors argued that the commonly used OM algorithm was less justified in their research context since short time deletion or substitution transformations had less meaning. An alternative analytical approach for regression tree was to use multi-nominal logistic regression after identifying the typology by cluster analysis.

4.5 Conclusion of the review and the relevance to this PhD study

4.5.1 Conclusion: the current application of SA in health services research

This systematic scoping review was conducted to inform the analysis of the main study. This review identified a range of applications of SA in applied health research, such as health services and care settings, medications and therapies, the level, frequency, and expenditure of care consumption. Some studies used SA only for descriptive purposes, which could have been done through less complicated statistical methods such as using a frequency table. Other studies used SA to generate new knowledge, or facilitated the understanding of a research problem, which could not have been obtained from other commonly used statistical methods. **The added values of using SA in the included studies and the relevant methodological issues have been summarised and discussed in subsections 4.3.3 and section 4.4.** Generally, **the unique strength of SA is its ability to identify meaningful patterns of sequential changes over time, and present such patterns in different forms of figures to facilitate the understanding of complex information within sequences and cluster patterns.** A common limitation of the included studies in this review was that very few studies clearly reported how the key methodological decisions were made. The authors were more focused on using SA to obtain empirical results. The reporting of using SA still needs strengthening. A guideline for clear reporting of the technical details related to the analytical process and a critical appraisal tool to assess the quality of studies using SA would be helpful for both authors and readers. These research gaps could be the research directions for future studies.

4.5.2 Research gaps and the opportunity for this PhD study

SA has not been used to study **complex primary care sequences** involving events between patients and GPs yet, **nor the pathway to the cancer diagnosis.** There is no precedent knowledge on how to distinguish whether the health records are patient events or care provider events. Such distinction is particularly important in constructing primary care sequences to improve early diagnosis, because it may help us to know whether there is any missed opportunity for earlier diagnosis and identify the sources of the problem – whether it is due to the help-seeking behaviour of patients (e.g. late presentation), or because of the vague symptoms (despite patients present symptoms multiple times), or the main problem is from GP (e.g. not recognise the symptoms or inappropriate management). After identifying the problem, we can then make more specific recommendations to improve the current situation. In addition, sequences had the same length in the studies of this review. But patients have different care needs and the number

of visits to general practice may vary significantly among patients, even in different periods of the same patient. All these uncharted territories are the research gaps, which provide an opportunity for this PhD study to explore, but also pose a considerable challenge, as it is less known about how to specify the states that could best represent the health sequences involving patient-initiated and GP events and to make sense of the sequential patterns. It is challenging to summarise and simplify a wide range of complex primary care events over a relatively long period and categorise all the events into a handful of mutually exclusive states that are meaningful to clinical practice. Therefore, to get meaningful patterns in typology, this PhD study explores the key methodological issues around the research questions step by step, to find the best way to make sense of the patterns from primary care sequences. The purpose of methodological exploration is to lead to meaningful empirical findings.

4.5.3 How the findings of this review can be used to inform the decision and application of SA in this PhD study

From the results of this review, OM with indel cost set as 1 and substitution cost set as proportional to the transition rates among states ($OM_{[1, TR]}$), and agglomerative hierarchical clustering (Ward's method) were commonly used in empirical health studies. Among the 13 studies, none used multichannel sequence analysis. Although there was a study that correlated channels were analysed separately, no attempt was made to integrate the findings from different channels. Subsection 3.3.5 discussed some criteria to decide the optimal number of clusters, which were consistent with the three criteria used by Hougham et al. (2014). The criteria included cluster quality assessed by the ASW value, discrimination between the clusters of sequences and patients, interpretability of cluster patterns, and reasonable sample size of sequences in each cluster. These criteria were tested in the NAEDI study.

The next chapter will introduce the data source for the main study, followed by a statistical analysis plan. The data quality is also discussed.

Chapter 5 Data source and methodology

5.1 Introduction of the original study and research ethics

Smoking is a known risk factor for many chronic cardiorespiratory diseases and LC. It increases the risk of developing LC (relative risk=9.3, 95% CI [8.3-10.4])(Blakely et al., 2013) and in a dose-response relationship (Jemal et al., 2008). Current or ex-smokers aged 40 years and above, living in deprived areas (the bottom quintile of IMD), with severe lung comorbidity like COPD, are at increased risk of developing LC, and more likely to delay symptomatic presentation. This is the target population for early diagnosis, but less investigated than populations diagnosed with LC. A study titled “Symptom prevalence and help-seeking amongst patients at risk of lung cancer”, funded by the UK National Awareness and Early Diagnosis Initiative (NAEDI) programme (grant number C3801/A14137) was designed to fill this research gap. Two papers were published using the data collected for this study. One was a mixed-method study to understand patients’ help-seeking behaviours in primary care in response to symptoms (Wagland et al., 2016). The other was a qualitative study, exploring GP’s views regarding the potential for early diagnosis of LC in primary care (Wagland et al., 2017). This PhD study uses the NAEDI study data, with a new perspective to analyse patients’ health records to explore the patterns of primary care sequences.

5.1.1 Ethical approval for this study

The data used for the main study in this thesis, including methodological exploration (Chapter 6) and empirical analysis (Chapter 7), were previously collected by a research team based in the School of Health Sciences, University of Southampton. This study was governed by the NHS ethics agreement and protocol, with additional sponsor (University of Southampton) approval for the inclusion of the analyses using unlinked anonymised primary care data for a PhD thesis (REC: 12/SC/0049; RGO REF: 8388). Participating GP practices and patients were anonymised using pseudo-anonymised practice/participant identifiers (e.g. 01/001, 09/068) to protect participants’ anonymity and privacy.

5.1.2 Original and independent work by using the NAEDI data for this thesis

The original study team developed early drafts of the coding framework for the primary care events. I was granted access to the manually transcribed records on paper. I reviewed all the transcribed notes and made additional revisions and corrections of the codes and the categories to better suit the purpose of this study. I came up with the analysis plan, conducted all **the**

primary analyses reported in this thesis, and interpreted the results. Although it is a **secondary use of the data source, the work presented in this thesis remains original and by myself**. None of the contents in this thesis has been previously published.

5.2 Available data and variables

5.2.1 Participants, observation period, and timeline

The eligibility, inclusion and exclusion criteria of participants were reported in a previous publication (Wagland et al., 2016). Individuals who were older than 50 years and had a smoking history (either current smokers or quit smoking within the previous ten years) were considered at high risk of developing LC. These patients registered in eight general practices across three counties in south England were invited to participate in the NAEDI study. They received a letter from their respective general practice and a participant information sheet. They were invited to complete the IPCARD (Identifying Symptom Predictors of Chest and Respiratory Disease) questionnaire and post it back to the research team. The research team also asked for individual patient's consent to review their primary care records. For those who consented (n=912), the research team reviewed the records of **each consultation in GP notes** (not EHRs) from respective general practice, dated back from two years on the date patients gave consent. The observation period was mainly between June 2010 and October 2012 (29 months in total). Among 912 patients, 13 patients (1.4%) were excluded due to unknown/uncertain smoking history in patient characteristics.

5.2.2 Primary care events

During the observation period, each time the patient visited the general practice, the date of visit, the reason(s) why the patient visited, and the outcome(s) of visit (HCP actions) were manually transcribed from GP notes in free text (not Read codes) in standardised data extraction forms. In addition, the mode of consultation (face-to-face or through telephone) and the staff who provided the services (GP or practice nurse) were included. Generally, practice nurses were more in a capacity of managing minor health issues (e.g. measure blood pressure, drawing blood for tests, administering flu vaccine, wound dressing, syringe ear wax). Thus, the HCP actions were mainly GP actions, while nurses may be involved in some minor clinical work.

After data cleaning, a total number of 8,896 episodes of primary care consultations were included for SA. A small number of patients (n=41) with a smoking history but did not have any visits during the observation period. It was not possible to conduct SA in patients without health records, but

they could form one cluster (no use of primary care services). The sociodemographic characteristics of this group of patients could be compared with other groups of patients in the empirical analysis.

5.2.3 Coding framework for the primary care events

After several rounds of iterations and with contextual inputs from the supervisory team and clinical input from a chest physician, the reasons and outcomes of each primary care visit were thematically grouped into different categories, presented in Table 5.1 and Table 5.2, respectively. For each consultation, it could be more than one reason and/or outcome coded either in the patient or GP part. For example, the patient could present with a symptom (e.g. cough) and a long-term health condition (e.g. COPD, two codes); the GP reviewed the treatment plan, prescribed new medication, and ordered tests (three codes). Similarly, it was possible that only the patient or the GP part, i.e. either the reason or the outcome of the visit, was recorded. For example, a patient went for a regular health check (monitoring chronic disease), or a practice nurse administered a flu vaccine to the patient. The clinical situations were very complex and heterogeneous. The complexity significantly increases if investigate care sequences in a large population for a long period (e.g. several years).

Table 5.1 – Patients' reasons for primary care consultations

Codes	Categories (Patient)	Examples
1	Chest symptoms indicative of LC ³	Cough, chest pain, breathing changes, chest infections, haemoptysis
2	Indicative systemic symptoms	Weight loss, voice changes, sweats, fatigue
3	Monitoring/review (smoking-related) chronic cardiorespiratory comorbidities	Chronic obstructive pulmonary disease (COPD), ischemic heart disease (IHD), asthma, atrial fibrillation (AF), hypertension
4	Monitor/review non-respiratory chronic illnesses	Diabetes, hypothyroidism, chronic kidney disease (CKD), transient ischaemic attack (TIA), cerebral

³ Some patients presented symptoms indicative of LC, but none received a diagnosis of LC during the observation period in this PhD study. Three participants were diagnosed with LC/mesothelioma, within a range of 4 weeks to 11 months **after** completing the questionnaire (Wagland et al., 2016).

		vascular accident (CVA), Crohn's disease, prostate cancer
5	Acute problems of chronic respiratory conditions	COPD exacerbation
6	Acute problems of non-respiratory chronic conditions	Rheumatoid arthritis, colitis, sciatica
7	Other alarming symptoms potentially indicative of serious problems	<ul style="list-style-type: none"> • Epigastric/abdominal pain later found to be ovarian cysts/peptic ulcer • Palpitations/fainting • Dysphagia/vomiting • Rectal bleeding/altered bowel habit/constipation
8	Other health problems commonly seen in general practice	<ul style="list-style-type: none"> • Bunions/perianal itching/thrush/minor injuries/dog bites/insect stings/bursitis • Skin problems: warts/skin tags/ulcers/abscesses/rash/varicose veins • Musculoskeletal pain: lower back/leg/knee/ankle • Ear/eye problems • Menopause problems/cystitis • Mild side effects of prescribed medications
9	Health checks, health information, and advice	<ul style="list-style-type: none"> • NHS health checks/operation (surgery) follow-up • Cervical smear test • Foreign travel advice
10	Mental health issues	<ul style="list-style-type: none"> • Anxiety/depression/psychological distress • Bipolar affective disorder/schizophrenia

		<ul style="list-style-type: none"> • Panic attacks/stress at home or work • Alcoholism/drug abuse
11	Social/caring problems/help with benefits	Discuss problems in general (usually social or financial problems – often linked to alcoholism/drugs and mental health issues)
12	Others	<ul style="list-style-type: none"> • Fear of flying – request diazepam • Requesting a medical certificate/sick note

Table 5.2 – HCP actions (the outcomes of consultation)

Codes	Categories (HCP actions)	Examples/explanations
1	Watch & wait/pro re nata (PRN) review/continue current treatment	No definitive treatment was provided. Patients were often given advice and asked to return if the problems persisted/worsened, or other related symptoms occurred
2	General advice	This was sometimes given in combination with PRN R/V (above), i.e. to take fluids when having viral infections, stretches for back pain, sleep hygiene for insomnia, exercise, weight control
3	A Prescribed antibiotics	
	B Prescribed other medications	Steroids, analgesia
	C Prescribed antibiotics and other medications	
4	Review medications and no change	Medications were often reviewed when new problems arose, but not always clear whether changes have taken place from the transcribed notes.

Chapter 5

5	Review medication and change	If patients were prescribed medications, then another code of 3A/B/C was recorded.
6	A Request full blood count (FBC)	
	B Requested chest X-ray (CXR)	
	C Requested spirometry	
	D Requested FBC & CXR +/- sputum	
	E Request FBC & CXR & acute cardiac/respiratory referral (2WW)	
	F Referral to cardiac/respiratory consultant	Chest pain may be due to angina, then the patient was usually referred to a cardiologist.
7	Request test – others	X-rays for patients with back/shoulder/limb problems, blood tests, urine tests, ECGs, INR (International Normalised Ratio) monitoring
8	Acute non-cardiac/respiratory referral (2WW)	e.g. colorectal consultants for altered bowel habit
9	Routine referral to secondary care or community services	ENT/ophthalmology/audiology/physiotherapy /orthopaedics/podiatry
10	Problem resolved/identified and treated	
11	Immediate referral to MAU/A&E	
12	Smoking cessation service	
13	Minor interventions or vaccinations	<ul style="list-style-type: none"> • Cryotherapy/vitamin 12 injections/minor surgery/syringe ear wax/wound dressings • Flu vaccine

5.2.4 Available information and variables for patient characteristics

The following patient sociodemographic characteristics were collected, including date of birth, sex, ethnicity, marital status, the highest qualification earned, employment status, and the English index of multiple deprivation (IMD, 2007 version) based on participant's postcode of residence, as a proxy of individual SES. Participant's age was calculated as the consent date subtracted from the date of birth and then divided by 365.25. The IMD quintile was equally divided by the rank of Lower-Layer Super Output Areas (LSOA, 32,482 LSOAs in IMD 2007), from the most deprived quintile (Q1) to the least deprived (Q5). Smoking behaviour is usually characterised by several variables in research, including current smoking status (current or ex-smokers), smoking intensity per day and pack-years, start age of smoking, cumulative smoking duration, and the time since quitting smoking for former smokers (Huang et al., 2015). This study collected participant-reported data on these variables to understand their smoking status, intensity, and lifetime exposure.

Comorbidities were collected from the GP notes, and categorised as no comorbidity, 1, 2, and ≥ 3 comorbidities by the original team. The researchers reviewed patients' comorbidities for a period longer than the observation period. Therefore, even for those who did not attend general practice ($n=41$), the comorbidity burden among some of them was still available. I extracted four common cardiorespiratory comorbidities (either relevant to LC or smoking-related chronic diseases) from a string variable. These four comorbidities were COPD, asthma, hypertension (recorded as HBP – high blood pressure), and a series of heart diseases, recorded as multiple abbreviations across practices, including CVD (cardiovascular disease), IHD (ischaemic heart disease), CHD (coronary heart disease), AF (atrial fibrillation), and MI (myocardial infarction) in free text. These four variables were operationalised as binary variables, i.e. whether the patients were diagnosed with respective diseases or not.

5.3 Statistical analysis plan and methods

There were two phases in this study. The first phase was methodological exploration, in Chapter 6. SA and cluster analysis were used to classify primary care sequences into different groups. Ward's method is sensitive to outliers. Therefore, it is necessary to identify outlier sequences first (if any). A quick run of SA and cluster analysis could get the preliminary dendrogram to know whether there are any outlier sequences or not. The distances between the outliers and other sequences are larger, which could be easily spotted from the dendrogram. After excluding the outlier sequences, SA and cluster analysis are run again to explore and compare the solutions with different numbers of clusters. The final typology should make sense in the empirical context.

The second phase is a series of planned empirical statistical analyses. The first analysis was to characterise and compare the patient profile in different clusters, and to understand how patient characteristics can help explain the variations in the cluster patterns of primary care sequences. The second analysis was to investigate the variables significantly associated with patient's primary care attendance and the number of consultations relevant to potential LC symptoms (help-seeking behaviours). The third analysis was to explore whether there was a practice effect on patients' use of primary care services and the cluster patterns. Methods involved in the two phases are introduced below.

5.3.1 Patient characteristics and pairwise correlation between variables

Descriptive statistics were conducted to understand the patient characteristics of the study sample. After that, the relationship among variables was explored, as understanding the association among variables could help choose the right variables in the regression models to increase model fit and minimise collinearity. These two parts will be reported in subsections 7.2.1 and 7.2.2. In the literature review (subsection 2.5.6.1), it was established that education and employment would influence patient's SES (IMD quintile), which could affect patient's lifestyle (e.g. smoking), health status (e.g. the number of comorbidities), and help-seeking behaviours (e.g. the number of primary care visits). The years of smoking may associate with age. Older people may have more comorbidities and care needs. Age, the years of smoking, and the total number of visits were continuous variables, while IMD quintile, qualification, and the number of comorbidities were ordinal variables. Pearson correlation was used to explore the pairwise correlation between continuous variables. Spearman correlation (Spearman, 1904) was used between continuous and ordinal variables, and two ordinal variables.

5.3.2 Comparison of patient characteristics in different clusters after SA

Descriptive statistics were used to report the characteristics of the study sample and patients in different clusters. Statistical tests were used to compare whether patient characteristics were significantly different among different clusters or not. For continuous variables, Shapiro-Wilk test (Shapiro and Wilk, 1965) was used to test the normality of the data, and Bartlett's test (Bartlett, 1937) for equal variances among groups. Parametric test (analysis of variance, ANOVA) was used to compare continuous variables (e.g. age, years of smoking) among multiple groups, where the hypotheses of normality and homogeneous variance were fulfilled. Šidák's method (Šidák, 1967) was used for post hoc multiple comparisons between two groups after ANOVA. If the hypothesis of normality or equal variance was rejected, non-parametric approach, Kruskal-Wallis rank test (Kruskal and Wallis, 1952) was used instead, but did not perform post hoc pairwise comparison, as

the software did not provide such function. Chi-square test was used to compare the proportions of binary, categorical, and ordinal variables. Finally, multinomial logistic regression was used to explore the association between patient characteristics and the cluster membership and report the relative risk ratio (RRR). Multinomial logistic regression is the extension of logistic regression. One cluster is served as the reference category, and the other clusters are compared against the reference category (Agresti, 2002, Long and Freese, 2014). **The equation for multinomial logistic regression is $\log \frac{\pi_j(x)}{\pi_1(x)} = \alpha_j + \beta_j'x, j = 1, \dots, J - 1$** (Agresti, 2002)(P268). **Interaction between sex, IMD quintile, and the number of comorbidities were tested, as comorbidity burden may differ in sex and patients in different SES.**

5.3.3 Patient characteristics associated with primary care attendance and potential LC symptom consultations

Count data, such as the number of visits, the interval between two two milestone events, are often used in the field of early diagnosis research to characterise cancer diagnostic pathways. The number of primary care attendances and the number of potential LC symptom consultations are the two variables of count data in this study. The number of primary care visits and potential LC symptom consultations are indicators of service use, reflecting patient's care needs and their awareness of well-being and potential LC symptoms. It is reasonable to assume that patients who had more comorbidities need more care, and those who were more concerned about their health conditions would be more likely to visit their GP. Understanding what patient characteristics are more (or less) likely to use primary care services and consult potential LC symptoms would help to inform social marketing campaigns for the target patient subgroups (Niksic et al., 2015).

5.3.3.1 Modelling count data

Poisson regression is usually used to model the outcome variable of count data. **The equation for Poisson regression is $\log \frac{\mu_i}{t_i} = \alpha + \beta x_i$** (Agresti, 2002)(P385). The Chi-square test for the likelihood ratio of **the dispersion parameter, alpha (α)**, determines whether **Poisson regression** or **negative binomial regression** should be used. Alpha is negative two times the difference of the log likelihood between Poisson regression and negative binomial regression, i.e. $-2 \times (\log \text{likelihood}_{\text{Poisson regression}} - \log \text{likelihood}_{\text{negative binomial regression}})$. The null hypothesis is $\alpha=0$, i.e. the count data are epi-dispersed, then Poisson regression is suitable for such a situation. If the null hypothesis is rejected, which means the data are over-dispersed (the variance is greater than the mean), then negative binomial regression should be used instead (Hilbe, 2011, Long and Freese, 2014).

5.3.3.2 Modelling count data with excessive zero

If the count data have excessive zero, then the zero-inflated model should be used. Whether using zero-inflated Poisson regression or zero-inflated negative binomial regression still depends on whether the count data are over-dispersed or not, i.e. the likelihood ratio test for the dispersion parameter – alpha (Long and Freese, 2014). There are two parts in the zero-inflated model – a count model to predict the response variable, and the inflate part to predict the excessive zero.

5.3.3.3 Patient characteristics as independent variables to model the count data

Based on the published studies, the following variables are known factors influencing patient's help-seeking behaviours, including age, sex (female as a reference category in the analysis), IMD quintile (the most deprived quintile, Q1 as reference), the number of comorbidities (no comorbidity as reference), current smoking status (no smoking as reference), and the years of smoking. [Point 3](#) The interaction terms were the same as those stated in subsection 5.3.2. Two respiratory comorbidities, COPD and asthma, were specifically tested to see whether they were associated with the number of potential LC symptom consultations. The Backward approach was used to eliminate nonsignificant variables step by step. Incidence rate ratios (IRR), obtained by exponentiating the coefficients in the regression, and the 95% CI, were reported for significant predictors.

5.3.4 Exploration of practice effect

Practice effect has implications in health services research and primary care audit (to evaluate the performance of general practices). General practice is a natural cluster, as patients are registered with practices. Different practices may have different ways of managing their patients. Therefore, practice effect was explored to see whether there was any difference in patient's willingness to participate in this study (response rate), patient's characteristics and help-seeking behaviours, and the cluster patterns of primary care sequences, using the statistical methods described above. When practice was introduced in the model as a categorical variable, Practice 1 was the reference category. Multilevel multinomial logit model (Skron dal and Rabe-Hesketh, 2003) was not used to explore the practice effect in this study for two main reasons. First, the practice effect was exploratory, which aimed to provide some initial evidence about practice effects of patients' help-seeking behaviours and how GPs managed their patients. Traditional statistical methods are sufficient to achieve this goal. Secondly, the outcome of SA and cluster analysis is a nominal variable. The number of clusters of primary care sequences was unknown before analysis, which could be between two and nine clusters. Compared with other practices, one practice had a very

small sample size (Practice 7, n=24). Using multilevel multinomial logit model would increase the complexity of the analysis and may not achieve the best result. However, this method may be useful in a population-based study with a large sample size and a large number of practices, further discussed in subsection 8.6.1.

5.3.5 Statistical software and packages

Sequence analysis, cluster analysis, and graph visualisation were conducted in R (Version 3.6.0)/R Studio (Version 1.2.1335), using the packages 'TraMineR' (Version 2.0-12), 'cluster', and 'WeightedCluster' (Version 1.4). 'Graphviz', an open source software for graph visualisation, was used to produce the figures of regression tree. Besides SA, data cleaning, management, and all other statistical analyses were conducted in Stata 16.0.

5.4 Discussion of the study data

5.4.1 The study sample size

The sample size of the original study was calculated based on the estimated response rate of the IPCARD questionnaire, the number of participating practices, and the number of estimated eligible patients in each participating practice. The second round of posting questionnaires to the eligible participants who did not send back their questionnaires at the first mailshot was an attempt to increase the response rate. The study sample in this PhD study was those who gave consent to the researchers to review their medical records, which was a subset of the original study sample. Nothing could be done to increase the number of primary care sequences after data collection.

5.4.2 Sample representativeness

All patients who met the inclusion and exclusion criteria in the eight general practices were invited to participate in this study. Of 4,621 patients invited, 1,172 (25.3%) completed and returned the questionnaire. Patients who gave consent to the research team to review their health records were part of the questionnaire respondents (77.5%, 908 out of 1172 patients). The sampling approach (Bower et al., 2017) of this study could be deconstructed as:

- Target population: people at high risk of developing LC in England;
- Source population: patients in the target population registered in the eight general practices that agreed to contribute to this study;
- Sampling frame: all eligible patients in the source population;

- Study sample: patients in the sampling frame who consented to participate in the study and health record review.

A representative sample means that the study sample matches **some** characteristics of the target population (Bower et al., 2017). There are many ways to evaluate representativeness; for example, based on age, sex, SES, education, health status, and so on. Non-response could be a threat to sample representativeness. The previous publication (Wagland et al., 2016) reported the patient characteristics between the responders and non-responders. Females had a slightly higher (but nonsignificant) response rate (26.2%) than males (24.7%). Patients in two age groups (60-69 and 70-79 years) had higher response rates (27.8%) than younger (50-59, 21.9%) and older patient groups (80+, 24.3%). There was a descending trend of responses rate from the least deprived quintile (27.2%) to the most deprived quintile (24.2%). All of these results make sense, as women and people in higher SES are generally more concerned about their health (Whitaker et al., 2015). Younger patients (aged 50-59) may not worry about this, while older patients (80+) may have additional barriers (e.g. poor eyesight, frailty) to finish the questionnaire even if they wished to. The research team did not compare the other sociodemographic characteristics between the responders and non-responders. Such data were not available to me to conduct further comparisons. The good sides of the study sample were the responders were in different age groups and had relative balanced proportions in sex and SES. If comparing with the criteria for the English targeted lung health check (NHS England, 2019) for early diagnosis of LC (ever smokers aged between 55 and 75 years old), this study sample was in the “high-risk” group and eligible for the lung health check. The United States Preventive Services Task Force recommends LC screening (Jonas et al., 2021), also based on age (50 to 80, age band 10 years wider than the UK) and smoking intensity as eligibility criteria. Therefore, **this study sample was representative of a specific population of interest, i.e. smokers or ex-smokers age ≥ 50 years old in England, who are the candidates for screening and early intervention of LC.** At the event level, primary care consultations should be complete and representative, as the research team systematically review and transcribe patient’s records, further discussed in the “data quality” subsection (5.4.5) below.

5.4.3 Strengths and limitations of the study sample

The study participants were very unique in the study of this type, as they were the target population for early diagnosis but less investigated, let alone to understand their primary care pathways. Therefore, the choice of the study sample was part of the strength of this study.

Using health survey as a research design needs careful consideration, including sampling methods, coverage of geographical areas, local contact and logistical support, funding, manpower

(especially the number of dedicated researchers), and timeline. The study design (postal questionnaire survey with additional review of health records) and the available resources influenced the scale of this study and the sample size. The geographical coverage was mainly around the south coast of England. Patient's willingness to fill in the questionnaire and post it back (the response rate) and informed consent were out of researchers' control. Non-response bias is a well-recognised limitation of studies using postal questionnaire survey to collect data for research. Patients who are interested in this study are more likely to participate than those who are less interested. The length of the questionnaire (10 pages) may be a barrier and cause a psychological burden for some potential participants, but the response rate was comparable with other primary care postal surveys (Wagland et al., 2016). The invitation letter was sent to each eligible patient by the practices. If the practices did not elect to take part in the study, there was no way that researchers could reach the patients in those practices. This reflected the importance of local support. The inclusion of practice and participating patients was much more complex than selecting patients from EHRs.

This study sample was mostly white British (93.1%), and the percentage of known ethnic minority patients was only 2.5%. The small study sample size and patient characteristics (ethnicity and the distribution of SES) were due to the nature of the study design discussed above, but another English study with a similar study design (Walabyeki et al., 2017) also reported similar problems. This signifies the importance of including and engaging ethnic minority patients in health research. A better understanding of their thoughts and health conditions can inform health policy to provide health services to meet their care needs.

5.4.4 Strengths and limitations of the primary care data

The data of primary care visits were manually collected. Using EHRs can greatly increase the sample size and the coverage of the population. But the advantage of using GP notes over primary care events indexed by Read codes in EHRs is that free text provides richer information to answer the RQs in this study. For the visits related to potential LC symptoms, additional clinical information (symptoms, tests, treatments) was recorded in detail, which was important information to study primary care sequences. The symptom terminology used by GP is more accurate than those reported by patients. These are the strengths.

Manual transcription of data was more prone to error. For example, unclear handwriting may be difficult for other people to recognise, which could cause barriers to double-checking and revision. Typos may occur when typing into the computer. In addition, coding events in free text is time-consuming. Researchers undertaking this task need some qualitative coding experience to summarise all the texts into categories. It would be helpful to develop a rough coding framework

with prior knowledge before coding. Researchers can make further adjustments during the coding process. The coding framework in this study was less systematic and structured than the well-established ones like the ICD-10.

5.4.5 Assessment of the data quality

Health informatics literature provides some dimensions to assess the data quality of health records for research (Weiskopf and Weng, 2013, Kahn et al., 2016, Weiskopf et al., 2017, Feder, 2018). Here discuss the relevant dimensions of data quality to this study (the terminology may vary in different literature).

Completeness is referred to whether patients' health conditions are completely recorded in the EHRs or not (Weiskopf and Weng, 2013). Researchers could determine whether the EHR data are complete enough for a specific research purpose, or sufficient in quantity for the task at hand. This criterion is subjective. If based on this, the data extraction of health records for this study was quite complete, because small events such as wound dressing, vaccinations, patients consulted with family issues were recorded by the GPs and transcribed by the researchers in paper form.

Data plausibility: synonyms like correctness, accuracy, credibility, reliability are used in different literature for the same meaning. It means whether the EHR data could be trusted or not, or whether the data are suspected of quality issues. EHR data are plausible if they are in agreement with general medical knowledge and user-perceived reality, which means that the variable values should "make sense" based on external knowledge and clinical context (e.g. height and weight should be positive values and within a reasonable range) and free from error (Feder, 2018). I reviewed the transcribed paper notes and had several rounds of discussion and queries with the original research team for contextual inputs. The face validity of primary care events and variable values are good. The data are relatively credible, as they are in alignment with medical knowledge and common sense. The demographic dataset has been cleaned for analysis for the previous publication, and there were no outlier values for the variables.

Concordance and data availability: this means that the desired data elements are available to check the agreement between different data elements, for example, blood glucose (HbA1c) level to corroborate the diagnosis of diabetes (Weiskopf et al., 2017). Due to human resources and time, it was not possible to transcribe all the information (e.g. test results) from GP records on paper, and infeasible to ask another researcher to cross-check between the transcribed and the original GP notes. It was not possible to verify every diagnosis (e.g. COPD).

Data timeliness/recency: this means health events should be “time stamped”, as some data may be collected using techniques or laboratory procedures no longer in practice due to the advance in clinical knowledge. Aged data may reflect patient characteristics not necessarily generalisable to the contemporary populations (Feder, 2018). The events were dated, but this dimension did not have much impact on the methodological exploration. However, clinical guidelines for treating patients and referral criteria may change over the years.

Uniqueness: patient characteristics (high risk but not yet diagnosed with LC) and data source (GP notes rather than Read codes) are the two aspects of data uniqueness of this study.

In conclusion, the available primary care health records are quite complete, plausible, and unique. The data quality is good enough for the research objectives of this study.

5.4.6 Potential missing data in health records

Missing data are pervasive in longitudinal data. Health record is a special type of longitudinal data. Traditional longitudinal data for research are usually collected around the same interval (e.g. week, month, year). But individual patient’s health services utilisation over time is unique, as visiting the care providers largely depends on individual patient’s health conditions and care needs. Therefore, the frequency of visiting a GP and the intervals between two visits may vary greatly among patients, even within the same patient at different stages of life.

Healthcare and administrative events are recorded using specific coding systems, e.g. Read code in primary care, ICD-10 in secondary care in NHS England. When extracting the data but no record returned, it could be one of the following three possibilities. The first one is the patient did not attend the care facilities at all in that period, or the patient declined to provide the health data for research. Patients may selectively report some symptoms/health problems, but omit others. The second possibility is that physicians or other staff forgot to record the events (Feder, 2018), or GPs recorded some key problems indexed by Read codes, but put others in free text, which could not be extracted. It is not possible to verify how complete the care providers document health events. Researchers have little to do in these two situations. The third possibility is uncommon codes were used to record the events but those codes were not included for data extraction. To avoid this possibility, when preparing the codes for data extraction, researchers can search the coding database systematically, consult with experienced colleagues, and include relevant codes as extensively as possible. But in most cases, the absence of events means the event did not occur (the first possibility), rather than missing data. This is a commonly accepted assumption of using health records in research. Therefore, a state of ‘no visit’ can accommodate such a situation in SA,

if needed. I have also discussed some strategies for missing data in the systematic scoping review in subsection 4.4.5.

5.4.7 Assumptions and handling missing data

For patients without potential LC symptoms recorded in the GPs' notes, the assumption was patients did not have those symptoms. For those with fewer visits to the general practices, it was assumed patients had different levels of care needs and help-seeking behaviours. As reported in Table 7.1 (Result chapter), the completeness of patients' sociodemographic information was good. Missing values in demographic characteristics did not affect methodological exploration. By comparing the patient profile among different clusters in empirical analysis, missing values of some variables were coded as unknown/unreported, and reported as a separate category. The unreported proportions in each variable are 3.1% in the number of comorbidities, 4.4% in ethnicity, 5.2% in marital status, and 8.6% in qualification. It is assumed that the unknown/unrecorded data in this study are missing at random (MAR), which means that missingness could be explained by the observed data, rather than by the unobserved data. It is particularly challenging for the imputation model to converge when involving several categorical variables using multinomial logistic regression in multiple imputation by chained equations (MICE)(White et al., 2011). In addition, it was not possible to construct a rich multiple imputation model based on the limited available study data. [Considering this thesis is positioned as an exploratory study and does not intend to generalise findings to the wider population, more detailed analyses for missing data were not conducted. Complete case analysis was used as the primary analysis, as the proportion of missing data was <5% \(Jakobsen et al., 2017\).](#) The unknown/unreported status in ethnicity, marital status, and qualification had a minimal impact, as these characteristics were not significantly different among the patient clusters, and not included in the regression models (discussed in subsection 7.3.2).

5.5 Chapter summary

This chapter introduces the data source and statistical methods applied in the main study. The data quality, strengths and limitations of the data source are discussed. The study data are complete, plausible, and unique; and the data quality is good enough for the research objectives of this study. The next chapter reports the process of addressing relevant methodological issues, which would lead to empirical findings of the cluster patterns.

Chapter 6 Methodological exploration on the use of sequence analysis to identify typologies of primary care sequences among community-based patients at high risk of developing lung cancer

6.1 Introduction: research objectives of methodological exploration

This first phase of this study focuses on methodological exploration. It aims to explore how SA could be used to represent and analyse primary care sequences and **identify meaningful cluster patterns**, to evaluate and discuss how SA could be used to provide new knowledge in the field of early diagnosis research. Around this aim, there are four research objectives in this phrase (cross-reference to section 3.7):

- 1 To construct primary care sequences from discrete health records (methodological issue: sequence construction, which is the premise of analysis, involves state specification and setting interval);
- 2 To investigate how dissimilarity measures (e.g. OM) and cost setting (constant and data-driven) would influence the cluster patterns;
- 3 To establish and test the criteria that would help decide the optimal number of clusters;
- 4 To identify the most frequent sequences of potential LC symptoms

To make the whole analytical process more logical for the readers, the methodological exploration is reported step by step in this chapter, with brief discussions in place to explain why decisions are made in particular ways.

6.2 Addressing the methodological issues

6.2.1 State specification for SA (1st methodological issue)

6.2.1.1 The rationale of state specification

The states need to be **mutually exclusive** to each other. The numbers of categories for the patient and GP parts in Table 5.1 and Table 5.2 were too large. They needed to be reorganised in a way to make the number of states manageable and appropriate for the final presentation of the typology

in figures, while the states were still informative and distinguishable from each other. Therefore, based on the similarity and importance, the categories were further grouped as five patient states and six GP states, respectively. The importance of states was in descending order, concerning the possibility of LC. The hierarchy of states was decided by the research interest of this study. If multiple categories were recorded and in different states in one consultation, a state at a lower position (e.g. Pt-3. non-respiratory somatic illnesses) would render the position for the state at a higher position (e.g. Pt-1. potential LC symptoms). This was **a way to give more weights for more important states of greater interest and relevance for investigating potential LC.**

6.2.1.2 The states for patient's reason to visit the general practice

The grouping of patient events and the hierarchy of five patient states below was based on whether the patient events might signify LC to GP, the relevance of comorbidities to this study, and the severity of health conditions. The numbers in the parenthesis were the codes in Table 5.1.

- Pt-1. Potential LC symptoms: chest (1) and systemic (2);
- Pt-2. Smoking-related cardiorespiratory comorbidities: review/monitoring progression (3), acute exacerbation of COPD (5);
- Pt-3. Other non-respiratory somatic illnesses: review/monitoring progression (4), acute problems (6), other alarming symptoms indicative of potentially serious problems (7), other health problems commonly seen in general practice (8);
- Pt-4. Other primary care services (less relevant to COPD/LC): health checks, seeking for health information and advice (9), mental health issues (10), social care problems/help with benefits (11), and others (12);
- Pt-0. The patient part was not recorded.

LC shares some common symptoms with other respiratory comorbidities, like COPD. Even in the same state, the symptoms have a different level of severity. Take the Pt-1 state as an example, cough is a very common symptom in primary care, which would be caused by different reasons. Chest infection is a more severe problem, while haemoptysis should be an alarming symptom for both patient and GP. The GP action, whether managing a long-term respiratory health condition, or starting LC investigation, becomes vital in primary care sequences towards LC diagnosis, which should be reflected in the states.

6.2.1.3 The states for GP actions

There were six states for the outcome of primary care consultations. The numbers in the parenthesis below were the codes in Table 5.2.

- GP-1. Requested CXR (6B) and/or (rapid) referral: GP requested full blood count (FBC) & CXR +/- sputum (6D), request FBC & CXR & acute cardiac/respiratory referral (2WW, 6E), referral to cardiac/respiratory consultant (6F). **This state indicated that GP started to investigate potential LC symptoms;**
- GP-2. Other tests, investigation, and/or referral: request FBC (6A), requested spirometry (6C), request other tests (7), acute non-cardiac/respiratory 2WW referral (8), routine referral to secondary care or community services (9), immediate referral to medical acute units (MAU)/accident & emergency (A&E) (11);
- GP-3. Prescribe medications: antibiotics (3A), others (3B), antibiotics & other medications (3C);
- GP-4. Smoking cessation service (12);
- GP-5. Review treatment plans for chronic conditions and health advice: watch & wait/pro re nata (PRN) review/continue current treatment (1), provide advice (2), review medications and no change (4), review medications and change (5), problem resolved/identified and treated (10);
- GP-0. No GP action was recorded.

6.2.1.4 The frequency of cross-tabulating patient and GP states

As discussed in subsection 3.4.4, interdependent patient and GP events should be combined, called **combined states**, and use traditional SA to analyse the primary care sequences. The GP part could be after the patient part, representing the combination of the reason and the outcome of a primary care consultation. For five patient states and six GP states, theoretically, it could have 30 combinations. Table 6.1 presented the frequency (and percentage) of the 24 combinations of the patient and GP states, with a massive difference in frequency, which varied from 17(0.2%) to 1,963 (22.1%).

Patients' help-seeking behaviours related to the potential LC symptoms (629 out of 8,896 episodes, 7.1%) were the key research interest in this study, and the GP response – requested a CXR or referred the patient to a chest physician (CXR/Referral), although only in a small

Chapter 6

percentage, was of top interest (n=118, 1.3%). Cardiorespiratory presentations excluding potential LC symptoms were about 10% (n=879, 9.9%), 2.8% higher than those due to potential LC symptoms. This was partly because when the patients presented with both potential LC symptoms and cardiorespiratory diseases, the latter gave up the position for the former, as potential LC symptoms were defined as a more important state in this study. Almost half of all the consultation episodes were related to somatic illnesses (n=4,284, 48.2%). More than a quarter of all consultations (n=2,415, 27.2%) did not have a code in patient's part. When reviewing the transcribed notes, it was found that most of these situations were GP reviewed or ordered tests (e.g. blood test) to monitor patients' chronic conditions. Smoking cessation service was only coded in GP part, without a patient code accompanied (n=237, 2.7%). Four combinations with a percentage >10% were underlined in Table 6.1, which were GP ordered tests for patients (likely to be monitoring chronic diseases, n=1,963, 22.1%), patients presented with somatic illnesses and GPs reviewed patients' condition (n=1,318, 14.8%), or GPs did nothing about it (n=1,293, 14.5%), or GPs ordered tests (n=901, 10.1%). These four combinations more reflected patients' comorbidity burden but were less relevant to the research interest of this study (LC).

Table 6.1 – Cross-tabulation of patient and GP states in primary care consultations, n(%)

N=8,896	Pt-1/Symptom	Pt-2/Respiratory	Pt-3/Somatic	Pt-4/Others	Pt-0
GP-1/CXR	118 (1.33)	17 (0.19)	47 (0.53)	NA	24 (0.27)
GP-2/Tests	115 (1.29)	77 (0.87)	<u>901 (10.13)</u>	51 (0.57)	<u>1,963 (22.07)</u>
GP-3/Meds	242 (2.72)	68 (0.76)	725 (8.15)	44 (0.49)	5 (0.06)
GP-4/Smoking	NA	NA	NA	NA	237 (2.66)
GP-5/Review	115 (1.29)	534 (6.00)	<u>1,318 (14.82)</u>	176 (1.98)	186 (2.09)
GP-0	39 (0.44)	183 (2.06)	<u>1,293 (14.53)</u>	418 (4.70)	
Column total	629 (7.07)	879 (9.88)	4,284 (48.16)	689 (7.75)	2,415 (27.15)

Note: NA meant no such combinations in the dataset. The denominator in this table was the total number of events (N).

Some combinations may look odd, but actually possible in reality. For example, a patient presented with cough (symptom indicative of LC), which could be caused by benign respiratory problems. The GP may just provide some advice (watchful wait) and see how the symptom would

develop. The GP action may not be recorded, and thus resulting in the combination of Pt-1+GP-0. The patient had a subsequent attendance several days later, as the symptom did not resolve but worsened. The GP made a rapid referral. The patient part was not recorded this time and resulted in the combination of Pt-0+GP-1. Such coding could be due to GP's recording habits or during the transcription process. It was possible that some notes were not entirely written down in the data extraction form by the researchers, and thus affecting the coding. A similar situation may exist in EHRs. The use of Read codes is not standardised national wide. GPs may have their own habits and preferences to record clinical events.

6.2.1.5 The combined states

The 24 combinations should be further grouped conceptually, to decrease the number of combined states. The ranking of the combined states was again in a hierarchy, based on the research interest in this study. Again, the combined states need to reflect the complex clinical situation, but still be able to distinguish from each other. In addition, the percentage of the combinations was considered. Otherwise, if the percentages were too small, even for important states, they would be drowned by other states in bigger percentages and become invisible in the typology (as Figure 6.5 and Figure 6.7 demonstrated in the following subsections to support this point). Therefore, the process of grouping the combined states carefully considered the importance of the combined states and maintaining a manageable number of combined states for visualisation in figures.

The grouping and ranking of the combined states were established based on the hierarchies of the patient and GP states (in subsections 6.2.1.2 and 6.2.1.3, respectively). Generally, we were more interested in potential LC symptoms (Pt-1) than cardiorespiratory comorbidities (Pt-2), then somatic illnesses (Pt-3), and finally, non-specific care needs (Pt-4 and Pt-0) in patient's part. For the GP part, investigation of potential LC symptoms (GP-1) was more important than requesting other tests or investigations for non-LC problems (GP-2) or prescriptions (GP-3). Review (GP-4) was at a lower level, as it indicated no particular medical intervention from GP (either investigation or prescription), but slightly better than doing nothing at all (GP-0). The meaning, frequency, and colours in R (to facilitate communication of the results) for the combined states are in Table 6.2. The legend of the combined states is illustrated in Figure 6.1.

■	1. LC symptoms CXR/referral
■	2. All 'non-LC symptom' presentation CXR/referral
■	3. LC symptoms Tests/investigation/prescription
■	4. LC symptoms Review/advice
■	5. Cardiorespiratory presentation Tests/investigation/prescription
■	6. Other somatic illnesses Tests/investigation/prescription
■	7. GP offered smoking cessation service
■	8. All 'non-LC symptom' presentation Review/advice
■	9. Pt:non-specific care needs Tests/investigation/prescription
■	10. Pt:any presentation GP:no action

Figure 6.1 – The legend of the combined states for the primary care sequences in the NAEDI study

Here are some further explanations for the combined states. Patients presented with potential LC symptoms that triggered GP initiating LC investigation by requesting CXR or making a referral to a chest physician was on top of the list (combined state 1, denoted as Combined-1). Prioritising this state over the others was because it indicated that GP considered the possibility of LC in patients and started to investigate LC. Due to the non-specific symptoms of LC, patients may present in other forms, or the event was not recorded as potential LC symptoms, but GP still investigated potential LC. Such a situation became the second in the hierarchy (Combined-2). The hierarchy of the states from Combined-3 to Combined-6 is self-explanatory, based on the rules stated above. Smoking cessation service (Combined-7) is an important intervention in general practice, to prevent disease progression to COPD and/or LC. Thus, it was considered less important than other valid interventions (tests, investigation, or prescription), but more important than medical review and/or advice from GP (Combined-8). Combined state 9 was mostly monitoring patient's chronic comorbidities, while no GP action regardless of any reason of patients' presentation was placed at the bottom of the hierarchy of the combined states.

Table 6.2 – The meaning, frequency, and the colours of the combined states (combined patient and GP events) in R

	The meaning of the combined states (with brief explanation)	The original combined codes	n (%)	Colour in R
1	Patient presented with potential LC symptoms GP: potential LC investigation (CXR/referral to chest physician)	Pt-1 (potential LC symptoms) GP-1 (CXR/Referral)	118 (1.33)	red
2	Patient: other 'non-LC symptoms' presentations GP: potential LC investigation	Pt-2 (Cardiorespiratory disease) GP-1; Pt-3 (Somatic disease) GP-1; Pt-0 (Patient part not recorded) GP-1	88 (0.99)	mediumpurple1
3	Patient: potential LC symptoms GP: other tests, investigation, or prescriptions (managed as benign conditions)	Pt-1 GP-2 (other tests, investigation) Pt-1 GP-3 (prescribe medications)	357 (4.01)	lightgoldenrod1
4	Patient: potential LC symptoms GP: review and/or health advice	Pt-1 GP-5 (review and/or health advice)	115 (1.29)	chocolate
5	Patient: cardiorespiratory presentations GP: other tests, investigation, or prescriptions	Pt-2 GP-2 (other tests, investigation) Pt-2 GP-3 (prescribe medications)	145 (1.63)	aquamarine
6	Patient: other non-cardiorespiratory somatic diseases GP: other tests, investigation, or prescriptions	Pt-3 GP-2 (other tests, investigation) Pt-3 GP-3 (prescribe medications)	1,626 (18.28)	lemonchiffon

7	GP offered smoking cessation service to patients	Pt-0 (Patient part not recorded) GP-4	237 (2.66)	lightgreen
8	Patient: all non-LC symptoms presentation GP: review and/or advice	Pt-2 (Cardiorespiratory disease) GP-5; Pt-3 (Somatic disease) GP-5; Pt-4 (Other services needed) GP-5; Pt-0 (Patient part not recorded) GP-5.	2,214 (24.89)	lightpink
9	Patient: non-specific care needs GP: other tests, investigation, or prescriptions (likely to be repeated prescriptions and/or monitoring chronic diseases)	Pt-4 (Other services needed) GP-2; Pt-0 (Patient part not recorded) GP-2; Pt-4 (Other services needed) GP-3; Pt-0 (Patient part not recorded) GP-3	2,063 (23.19)	skyblue
10	Patient: any presentation GP: no action	Pt-1 (Potential LC symptoms) GP-0; Pt-2 (Cardiorespiratory disease) GP-0; Pt-3 (Somatic disease) GP-0; Pt-4 (Other services needed) GP-0;	1,933 (21.73)	grey75

6.2.1.6 A brief summary of the whole process of state specification

State specification is the first step of constructing sequences. The whole process went through three major steps to come up with these ten **mutually exclusive** combined patient-GP states:

1. A thematic categorisation of the reasons and outcome of visits from transcribed GP notes in free text;
2. Decreased the number of categories and created patient states and GP states individually;
3. Based on Step 2, the patient and GP states were combined to further reduce the number of states based on the clinical importance and the research interest of this study. However, some states still had very small frequency.

6.2.2 Constructing primary care sequences (1st methodological issue)

There are two possible ways to construct primary care sequences from the combined patient-GP states. This section introduces the two approaches, and discusses the strengths and limitations of each approach. The presentation and comparison of the figures between the two approaches are in section 6.3.

6.2.2.1 The first approach: sequences constructed by visits

The first approach to construct primary care sequences is to align each consultation successively in the timeline. For each patient, all primary care consultations, coded as combined states, can be aligned one by one in the same row, parallel to the x-axis (visits), and constitutes a primary care sequence. Figure 6.2 illustrates how the primary care sequences for all the patients look like in this approach. The sequences were sorted by the total number of visits, where the x-axis is the number of visits in chronological order. Each row is a whole sequence, consisting of short lines in different colours, representing the corresponding combined states (the reason and outcome of that visit). The two indicators widely used in early diagnosis research, the total number of visits (in the x-axis) and the interval between visits (only applicable for sequences having ≥ 2 visits during the observation period), could be calculated and used to describe the sequence characteristics. The **strength** of this approach is that the sequences are conceptually straightforward, easy to understand, while the two major **limitations** are the loss of timing information and bringing redundant information into the sequences. From Figure 6.2, we could not know when the consultation happened in the timeline. When reviewing the GP notes, it was found that some patients had very frequent visits to the practices after surgery (e.g. hip replacement) for wound dressing or post-operative monitoring. Although these were actual care needs of the patients, they were introduced and treated equally in the sequence as the more important events in this

study (e.g. visits due to potential LC symptoms). Such uncensored information could introduce excessive noise in the sequences, which may NOT result in satisfactory and meaningful patterns in the typology, because the number of consultations related to ordinary care needs was much bigger than that related to potential LC symptoms (Table 6.1). They were ‘more powerful’ in determining the clustering structure than potential LC symptoms.

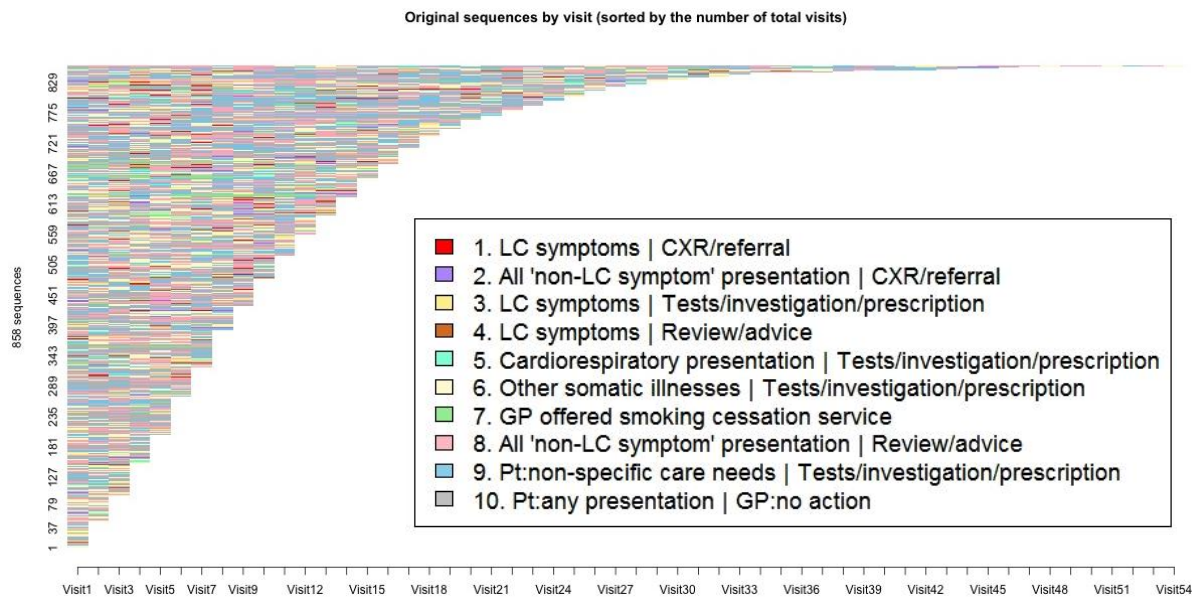


Figure 6.2 – Representation of primary care sequences for individual patients (sequences constructed by visits)

6.2.2.2 The second approach: sequences constructed in a timeline with equal intervals

The second approach to construct sequence is to create small equal intervals (e.g. week, fortnight, month) in the timeline (x-axis). Based on the date information, health records can be located at one of the corresponding intervals (calendar month here in the example). Some patients may visit their GPs more than once in a month. In such a situation, we can keep the combined (only one) state at the highest rank, as we are more interested to know the most important reason and outcome of the patient’s visit in that month. Figure 6.3 presents the primary care sequences of individual patients constructed in a timeline. The length of the sequences is different when constructing sequences by visits (in Figure 6.2), as there are variations in the frequency of patients’ help-seeking behaviours. But sequences are in the same length in this approach, with the same start and end point of the timeline. Most of the patients did not attend general practice every month. An additional state of “no primary care visit” was added to accommodate for such a situation. This is a special state, and not missing data (discussed in subsection 5.4.6).

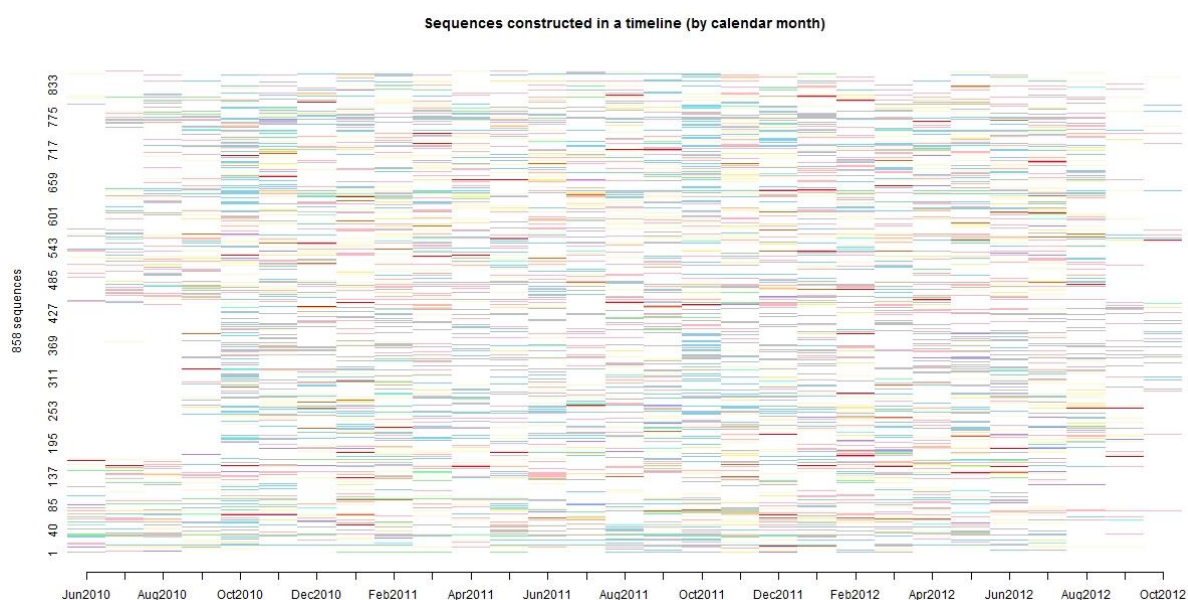


Figure 6.3 – Representation of primary care sequences for individual patients (sequences constructed in a timeline)

6.2.2.3 Brief discussion: applications and implications of two sequence construction approaches

The second approach (timeline with interval) compensates for the two limitations of the first approach (by visit). We could know when the consultations happened (timing) from the timeline, and the redundant information is controlled by only keeping the most important combined state in one interval (month in this example). Although this study sample was at increased risk, we could not know where the individual patients were in the disease trajectory of LC, as they did not have a clinical diagnosis of LC. Some of them may develop LC a few years later; some may not in their whole lives. Therefore, **constructing sequences by visit was fine for this population**. But it would be problematic if studying a population with a definitive clinical outcome (e.g. LC diagnosis), because timing is important and relevant in such context. We would be interested to know when the important consultations (e.g. consultations related to potential LC symptoms) happened relative to the date of LC diagnosis. If two patients had only one consultation and the sequences were constructed by visits, they were in the same position (Visit 1) in the figure. We cannot know the timing of consultation in this approach. It could be two years or one month before the diagnosis. It has huge differences in clinical implications between these two timings for symptoms like haemoptysis.

The observation period in this study was from June 2010 to August 2012. It is sensible to use the original calendar month in the x-axis for sequence construction. But if we study primary care sequences before LC diagnosis (e.g. 2 years before diagnosis) in a population-based study, it is

likely to include patients diagnosed with LC in different calendar years with a big time span (e.g. from 2001 to 2020). If using the original dates, it would be difficult to align all the sequences diagnosed in different calendar years in the same timeline. Making a relative timeline, rather than using the original dates, is an important step to align all the sequences in such a situation. Using the date of diagnosis as the endpoint of the sequence, health events can be reorganised backwards for 2 years, with month as an interval, as Figure 6.4 shows. In this way, all the sequences can be in the same timeline, with the same start and end points. It is easier to understand the timing of health events relative to the date of cancer diagnosis.

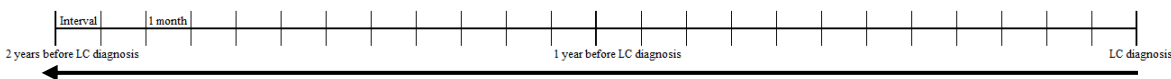


Figure 6.4 – Constructing a timeline and setting intervals to represent health sequences in different calendar years

6.2.3 Dissimilarity measure and cost setting (2nd methodological issue)

SA can cope with sequences of either equal length (second approach, in a timeline) or unequal lengths (first approach, by visits). Dissimilarity measures and cost setting are the second methodological issue, which could be considered as sensitivity analysis of applying different methods on the same data. Some technical aspects are already discussed in subsections 3.2.2 to 3.2.4. **To use Hamming Distance (HAM) and Dynamic Hamming Distance (DHD), it requires sequences of equal length.** Therefore, **they are NOT suitable for sequences constructed by visit** (unequal length), but it is possible to use optimal matching and setting cost as constant ($OM_{[1,2]}$) or proportional to transition rates among states ($OM_{[1,TR]}$). Sequences constructed in a timeline have equal length, which has more choices of dissimilarity measures and cost setting, including $OM_{[1,TR]}$, $OM_{[1,2]}$, HAM, and DHD.

6.2.4 Criteria to determine the optimal number of clusters (3rd methodological issue)

The ASW value was used to assess the clustering quality in three studies (Hougham et al., 2014, Le Meur et al., 2015, Moreno-Black et al., 2016) in the systematic scoping review. Whether the ASW is helpful to decide the optimal number of clusters has been tested in this study. In addition, subjective criteria such as interpretability of the cluster patterns in the research context, and pragmatic criteria like the number of clusters in typology and the number of sequences in each cluster, were considered when making the decision. These criteria were mentioned in the previous subsections (3.3.5 and 4.4.3) and are further discussed in subsection 6.7.4.

6.3 Presentation and interpretations of the whole primary care sequences from all study subjects

6.3.1 Sequence profile 1: state distribution of primary care sequences constructed by visits

There are 767 distinct sequences out of all 858 sequences. The median number of GP consultations is 8 times during the observation period, IQR [5, 14], and a maximum of 54 times. For patients having ≥ 2 visits, the median interval between two visits is 28 days, IQR [11, 70] days. Figure 6.5 presents how the ten states are distributed by visits (x-axis). The y-axis is the percentage of states at each visit, relative to all the sequences. As the number of visits increases, more sequences end, shown as the curve drops and more area in white in Figure 6.5. The whole figure looks similar to a survival curve. If not making a relative percentage, at the end of the x-axis, there are only one or two states at each visit, and the single state could take up to 100%, as shown from Visit 47 to Visit 54 in Figure 6.6, which is very misleading and confusing for readers who do not know this method.

In the state distribution plot, the states are presented the same order as they are specified in the legend, from the bottom (the first state) to the top (the tenth state), not by the percentage of states. States with higher percentages are easier to recognise in the figure. ‘Lightpink’ (8th state, 24.9%), ‘sky-blue’ (9th state, 23.2%), ‘grey75’ (10th state, 21.7%), and ‘lemon chiffon’ (6th state, 18.3%) are the four dominant colours in Figure 6.5. Despite only taking up a very small percentage (1.3%), the first state “potential LC symptoms | CXR/Referral” is the most important state and the research interest in this study. A red colour may help readers easier to find it at the bottom of the figure. The illustration of these two figures echoes the point mentioned in subsection 6.2.1.5 that important states in small percentages would be drowned by states in bigger percentages, and become invisible in figures.

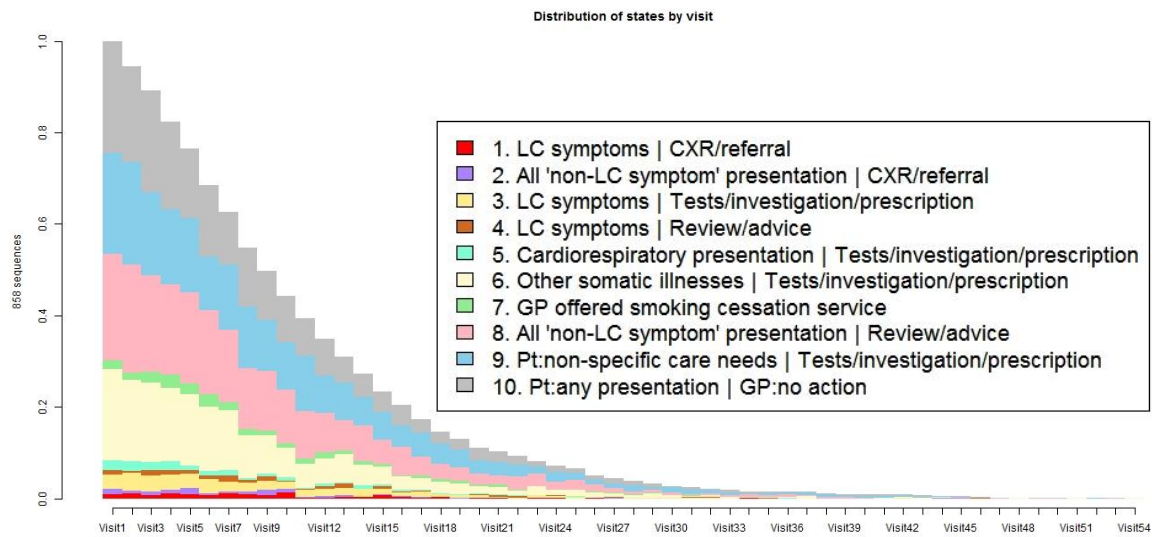


Figure 6.5 – State distribution plot of all 858 primary care sequences (percentage relative to the total number of sequences)

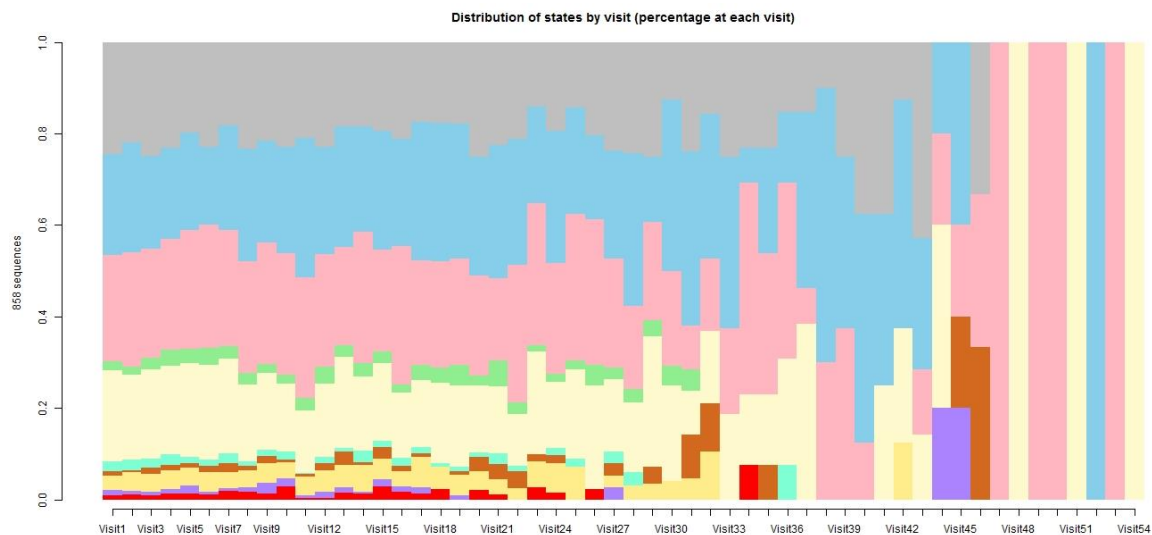


Figure 6.6 – Alternative presentation of state distribution plot of all 858 primary care sequences (percentage at each visit, legend as the same in the above figure)

Note: after Visit 47, there is only one sequence, and that only one state takes up 100%, which is not comparable to what 100% represented at the first several visits (with a much larger sample size).

6.3.2 Sequence profile 2: state distribution of primary care sequences constructed in a timeline

The state distribution plot of primary care sequences constructed in a timeline is presented in Figure 6.7, with the interval of month in chronological order. The sequences are in equal length (29 months). There are 846 distinct sequences out of 858, 79 more distinct sequences (846-767) compared with those constructed by visits, which is due to the timing and a new state – “no primary care visit” (coloured in white in the figure). The same state located in different months were considered as different sequences. Besides the beginning (June 2010) and the end (July and August 2012) of the timeline, the states are distributed relatively evenly, rather than heavily stacked at the beginning of the first several visits in Figure 6.5. The frequency of states in each month was not the original frequency, but a condensed result, because only the state most relevant to the RQ was kept in the sequence if a patient visited more than twice in a month. This is a way to reduce the noise of less interested states. Even in this situation, there was an increase of frequency every October. GPs explained the peak was because patients caught viral infections in flights during holidays. GPs requested tests and/or prescribed medications for these patients. What is reflected in the figure is that the 9th state in ‘sky blue’ has a higher percentage every October than any other month of the year.

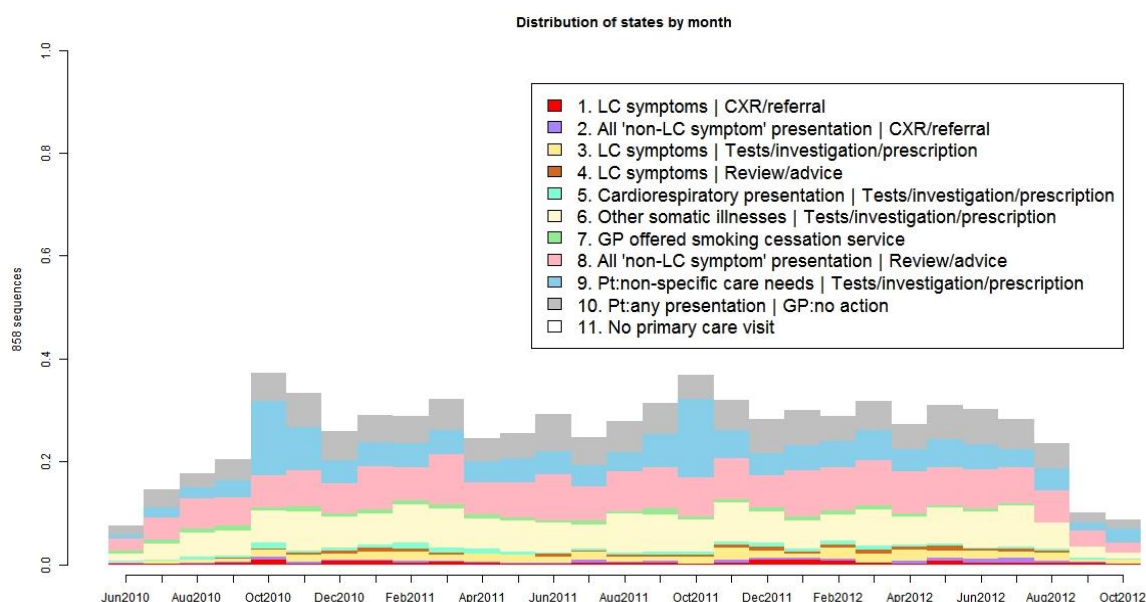


Figure 6.7 – State distribution plot of primary care sequences in the calendar month

6.3.3 Brief summary and discussion of the whole primary care sequences

Constructing sequences in different ways using the same original dataset gets different sequence profiles. The whole sequences for all patients were presented to **provide background information** for the study and explore the implications. Cluster analysis was not conducted at this stage because the whole sequence reflected the health problems, routine care needs, and the visit patterns in general practice among patients. **Using a full dataset to create typologies would provide limited useful patterns of the consultations related to potential LC symptoms or COPD.** Not all of the patients had consulted potential LC symptoms. Most of their care needs were still for systemic comorbidities. Understanding different levels of care needs among patients from the typology may be interesting, but it is not the research interest of this PhD study. During the review of the transcribed notes, quite a substantial proportion of primary care events were about repeated attendances of wound dressing, preventive measures like flu vaccination or immunisation, vitamin 12 injections, or even counselling patient's family issues with GP. **The consultations related to potential LC symptoms only took up a very small percentage** in the community, even among a high-risk population. The state of most research interest "patients presented with potential LC symptoms and GP requested CXR or made referral" was scattered at the bottom of the figure and barely visible. It needed a different strategy to analyse the sequences.

In order to get more meaningful patterns relevant to potential LC symptoms, to investigate the risks among different patient subgroups, and to discuss the implications of patient management to achieve early diagnosis of LC, it is better to separate the study sample into two groups – patients who presented with potential LC symptoms and patients who did not, and conduct analysis separately for each group. Hence, investigating the patterns of consultations among patients who presented with potential LC symptoms is reported in the following two sections (6.4 and 6.5). Patients who did not present with potential LC symptoms are still of research interest and have clinical implications, as they are still a high-risk population and the target population for early diagnosis. They may present and consult potential LC symptoms before or after the observation period, just not in the timeframe of this study. The analysis of patients in this group is reported in section 6.6.

6.4 Subgroup analysis 1 – sequences of patients presented with potential LC symptoms

6.4.1 Introduction and rationale of the subgroup analysis

The previous section established the need to separate the study sample into two groups – patients presented with and without potential LC symptoms in general practice. This section reports the exploration of how SA was used to identify meaningful patterns of sequences among patients who presented with potential LC symptoms and how GP managed these patients (e.g. ordered CXR, prescriptions, offered health advice). Only around one-third of patients (34.4%, 295/858) presented with potential LC symptoms at least once during the observation period. These subjects were included in this part of subgroup analysis. The sequences were constructed and analysed in two steps:

1. Sequences with potential LC symptoms only, called **LC symptom sequences** for short hereafter;
2. More complex sequences, including consultations with potential LC symptoms, smoking-related cardiorespiratory comorbidities (e.g. COPD, respiratory and heart diseases), and other non-respiratory alarm symptoms potentially indicate serious problems (category 7 in Table 5.1). The reason to include this was that these symptoms might compete for GP's attention against respiratory problems. They were also likely to form part of the GP's clinical judgment and diagnostic reasoning. Because these events were high risk related to LC, the sequences were called **high-risk sequences** for short hereafter.

The first step is a simplified approach of the second step, isolating potential LC symptoms from other healthcare events, which is a good start to understand the cluster patterns from the simplest situation. However, for each patient, health events are connected. GP may take a quick review of the patient's previous attendances and personal medical history before meeting the patient. When making clinical decisions, GP may take the previous consultations into account. Therefore, in the second step, patients' potential LC symptoms were situated in a broader and relevant clinical context. Other consultations in the whole sequence, i.e. GP and practice nurse attended to patients' general care needs (e.g. wound dressing, flu vaccination), were excluded from both 'steps', as they were less relevant to the research interest of this study (LC). Excluding them in the subgroup analysis was to reduce unnecessary noise in the sequences, and enable us to find meaningful patterns. The reasons for patients' presentation were in three states, and the GP actions in five states, leading to 15 combined states in total. Table 6.3 presents the frequency of the combined states in the subgroup analysis, taking up 13.8% (1,232/8,896) of all

consultations. Step 1 included the five states in column 2 of Table 6.3 (Pt-1), while step 2 included the states in all three columns. The results of step 2 are reported in the next section.

Table 6.3 – Frequency of the combined states in the subgroup analysis, n(%)

N=1,232	Pt-1: Potential LC symptoms	Pt-2: Other cardiorespiratory presentation	Pt-3: Other alarm symptoms
GP-1: CXR/rapid referral	118 (9.6)	10 (0.8)	4 (0.3)
GP-2: Other tests/investigations (not related to LC)	115 (9.3)	33 (2.7)	85 (6.9)
GP-3: Prescriptions	242 (19.6)	37 (3.0)	29 (2.3)
GP-4: Review/advice	114 (9.3)	241 (19.6)	61 (5.0)
GP-5: Did nothing	40 (3.2)	61 (4.9)	42 (3.4)
Column total	629 (51.1)	382 (31.0)	221 (17.9)

Note: the denominator in this table was the total number of events (N).

6.4.2 Sequence profile 3: LC symptom sequences in subgroup analysis

Individual sequences were sorted by the number of visits related to potential LC symptoms, presented in Figure 6.8 (left). Most patients consulted once (47.8%, 141/295) or twice (24.4%, 72/295) with potential LC symptoms, median 2, IQR [1, 3], but a small number of patients consulted more than six times (3.4%, 10/295), as shown in the state distribution plot in Figure 6.8. For patients having more than two visits of potential LC symptoms, the median interval between two visits of potential LC symptoms was 61.5 days, IQR [16, 217] days. A detailed description of the number and interval of patients' presentations of potential LC symptoms is in Table 6.4.

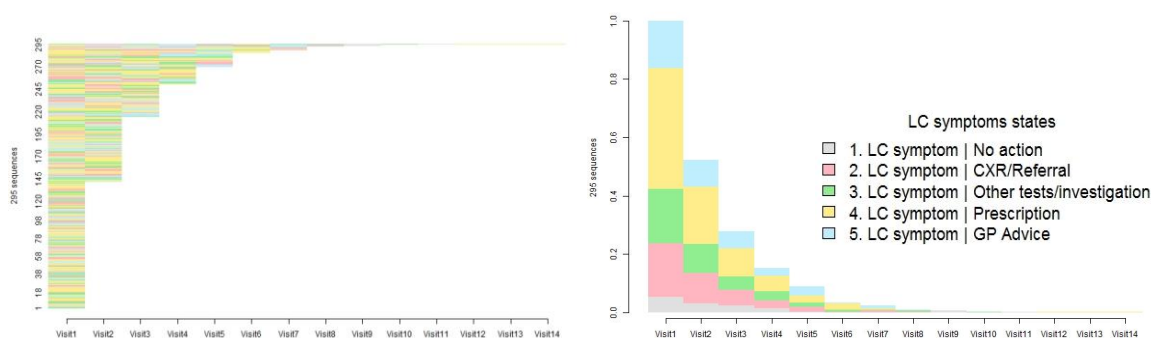


Figure 6.8 – Sequence index plot (left, sorted by the number of visits) and state distribution plot (right) for LC symptom sequences

Table 6.4 – The number of consultations related to potential LC symptoms and the intervals between two visits (days)

Visits	n	Interval between n and (n-1) visit	Visits	n	Interval between n and (n-1) visit
1	141	NA	6	3	28.5 [7, 125] days
2	72	103 [20, 297] Median [IQR]	7	4	30 [7, 132]
3	37	53 [14, 194] days	8	1	73 [4, 78]
4	19	45 [15, 133]	9	1	41 [21, 61]
5	16	50 [20, 139]	≥10	1	8 [6, 8]

Note: n – the number of patients, e.g. 141 patients presented only once with potential LC symptoms; therefore, the interval between two visits was not applicable (NA). For patients with two visits, the descriptive statistics of the interval was between the second (n) and the first (n-1) visit; for patients with three visits, the interval between the first and second visit was already summarised in No. 2; the interval in 3 was between the third (n) and the second (n-1) visits, and so on. There was only one patient with 14 visits. Therefore, interval ≥ 10 was the result of that patient from 10-14 visits.

Prescription was the most common GP action for patients presented with potential LC symptoms, twice as many as ordering CXR/making a referral to a chest physician, other tests or investigation, review or providing health advice. These three states had a similar percentage (Table 6.3), while no GP action was in a small percentage (coloured as grey at the bottom of the figure). For the ten most frequent visitors (presented with potential LC symptoms for ≥ 6 times), prescription (33.8%, 26/77) was still the dominant GP action, followed by CXR/referral and advice (both 19.5%, 15/77). No GP action (14.3%, n=11) was slightly higher than tests for non-LC investigation (13.0%, n=10).

Based on the presentation in Figure 6.8, it is more sensible to construct sequences by visit, rather than to spread the small number of events in a timeline of 29 months, which is not helpful to find patterns with a very dominant 'non-attendance' state in each month. In addition, without a clinical diagnosis of LC, timing in a calendar month is less important. Data-driven and constant substitution costs in OM ($OM_{[1, TR]}$ and $OM_{[1, 2]}$) could be used to analyse sequences of unequal length, and to compare the similarity and differences of typologies between the two ways of cost setting.

6.4.3 Dendrograms and regression trees for LC symptom sequences

6.4.3.1 Substitution cost matrix

The difference between $OM_{[1, TR]}$ and $OM_{[1, 2]}$ was in the substitution cost matrix. The indel cost was 1 in both. In $OM_{[1, 2]}$, the substitution costs were a symmetric matrix of 2, with the value of 0 in the diagonal. IN $OM_{[1, TR]}$, they were proportional to the transition rate among states, between 1.57 and 1.86 in Table 6.5. Transitions between two states that happened more often cost less. It was more economical to use substitution in $OM_{[1, TR]}$ than dual action of deletion and insertion in $OM_{[1, 2]}$. The substitution cost matrix is reported for two reasons: one is to know the substitution costs among the five states related to potential LC symptoms. The differences in substitution costs between $OM_{[1, TR]}$ and $OM_{[1, 2]}$ have implications on the dissimilarity matrix of sequences. The second reason is that these costs could provide empirical references for future studies of this kind to set the substitution cost. Given that the sequences were in unequal lengths, substitution may not be used very often, as direct insertion or deletion may cost less than substitution, to make the sequence with equal length and identical to another.

Table 6.5 – Substitution cost matrix of $OM_{[1, TR]}$ for LC symptom sequences (subgroup analysis-1)

	No action	CXR/Referral	Other tests	Prescriptions	Advice
No action	0				
CXR/Referral	1.86	0			
Other tests	1.76	1.57	0		
Prescriptions	1.58	1.57	1.59	0	
Advice	1.77	1.71	1.58	1.59	0

Note: all patient states were potential LC symptoms in this subgroup analysis. To avoid redundancy, here only put the five GP states in the table heading.

6.4.3.2 Dendrograms – the clustering structure of sequences

There were 95 distinct sequences of all 295 sequences, which meant that the other 200 sequences were identical with one of the distinct sequences. No operations were needed for the identical sequences. Therefore, their distance was 0. The height in the dendrogram was the distance to group sequences in clusters at each step. Due to a large number of identical sequences, the first quartile and the median of distance were 0, and the third quartiles were 1.27 in both $OM_{[1, TR]}$ and $OM_{[1, 2]}$. The maximum distance to group all 295 sequences was 22.58 in $OM_{[1, TR]}$ and 26.51 in $OM_{[1, 2]}$, which was understandable, as substitution costs in $OM_{[1, 2]}$ were bigger than those in $OM_{[1, TR]}$. The two dendrograms in Figure 6.9 demonstrated how the LC symptom sequences were grouped by Ward's method using the dissimilarity matrices calculated by $OM_{[1, TR]}$ and $OM_{[1, 2]}$, respectively. Based on the structures of dendrograms, four clusters may be a good option for both $OM_{[1, TR]}$ and $OM_{[1, 2]}$ algorithms. Otherwise, if the node continued to split, sequence No. 295 would be in a single cluster. The following subsections further explain why four clusters made a good typology.

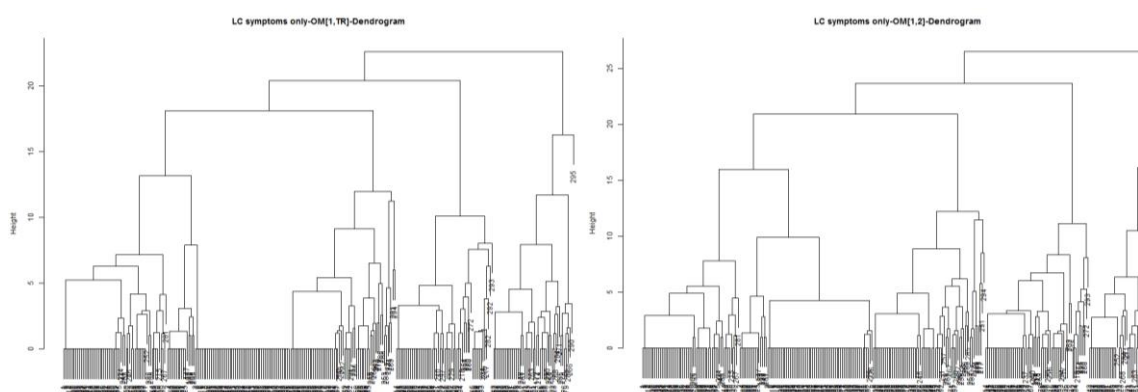


Figure 6.9 – Dendrograms of $OM_{[1, TR]}$ (left) and $OM_{[1, 2]}$ (right) for the LC symptom sequences (subgroup analysis-1)

6.4.3.3 Regression tree – how patterns in clusters changed by splitting the nodes in the dendrogram

Regression tree is a very useful way of graphical presentation to illustrate how the pattern of clusters changes by splitting the nodes in a tree-structured dendrogram. Figure 6.10 illustrates this process for $OM_{[1, TR]}$ and $OM_{[1, 2]}$. We could see new patterns emerge step by step, which is very helpful to decide on the number of clusters empirically. As Figure 6.9 (dendrogram) and

Figure 6.10 (regression tree) show, both trees should stop at Step 4, because there is only one sequence in a cluster at Step 5.

6.4.3.4 Deciding the optimal number of clusters for the typologies

From a statistical perspective, the optimal number of clusters could be chosen from the highest ASW value, by calculating and comparing the ASW values among the number of clusters ranging from 2 to 9. The results returned that 9 clusters had the highest ASW value, 0.321 for $OM_{[1, TR]}$ and 0.371 for $OM_{[1, 2]}$; and the lowest ASW values, 0.263 for 6 clusters in $OM_{[1, TR]}$ and 0.293 for 3 clusters in $OM_{[1, 2]}$, which was not a big difference. The ASW values in $OM_{[1, 2]}$ were larger than those in $OM_{[1, TR]}$. Although ASW values indicated that 9 clusters were the best solution statistically, it would end up with some clusters only having a few sequences. Such a solution may not be optimal in practice and probably limited its further applications. It may be more reasonable to choose the typologies of 4 clusters for both $OM_{[1, TR]}$ and $OM_{[1, 2]}$, based on the subjective criteria – clusters provided distinctive information and also made sense in the research context (interpretation of cluster patterns in later subsections). The clustering structure and typologies from $OM_{[1, TR]}$ and $OM_{[1, 2]}$ were similar, but they had different numbers of sequences in each cluster. For a solution of 4 clusters, the ASW was 0.292 in $OM_{[1, TR]}$ and 0.323 in $OM_{[1, 2]}$. A higher value indicated a better clustering structure. The decision on the best typology can be confirmed after interpreting the meaning of the cluster patterns.

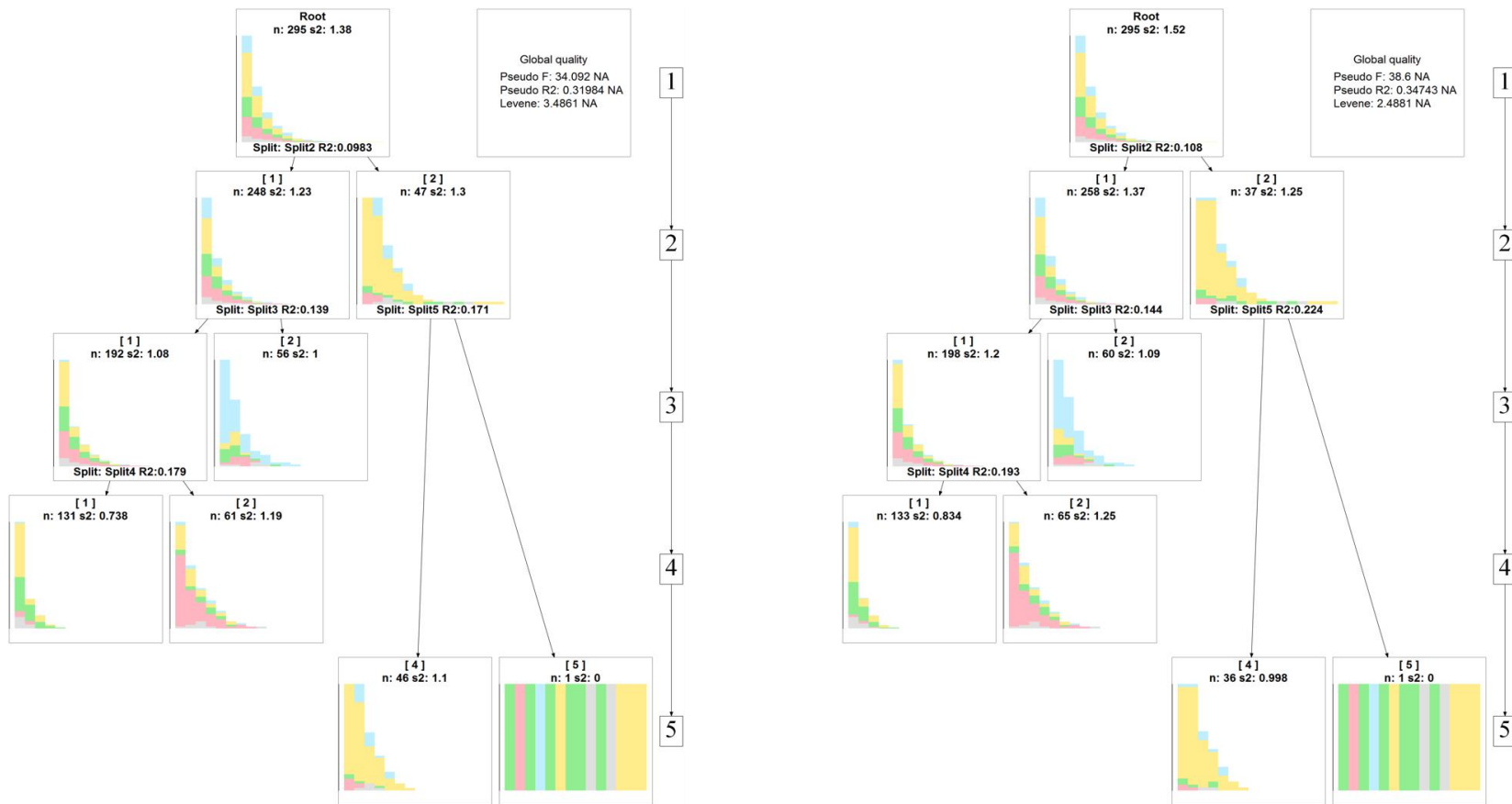


Figure 6.10 – Regression tree demonstrating how the cluster patterns changed step by step in $OM_{[1, TR]}$ (left) and $OM_{[1, 2]}$ (right) for the LC symptom sequences

6.4.4 Comparison and interpretation of the cluster patterns in the typologies of LC symptom sequences

Figure 6.11 presents the typologies of $OM_{[1, TR]}$ (left) and $OM_{[1, 2]}$ (right) for the LC symptom sequences. The patterns for four clusters in $OM_{[1, TR]}$ and $OM_{[1, 2]}$ are generally similar. There are ten more sequences in Cluster 4 by $OM_{[1, TR]}$ than that by $OM_{[1, 2]}$ (47 vs 37), but two sequences less in Cluster 1, and 4 sequences less in Clusters 2 and 3, respectively. The differences between the two typologies were due to the cost setting schemes (data-driven vs constant). Using substitution in $OM_{[1, TR]}$ for sequences with similar lengths cost less, but for sequences with different lengths, direct insertion or deletion may cost less.

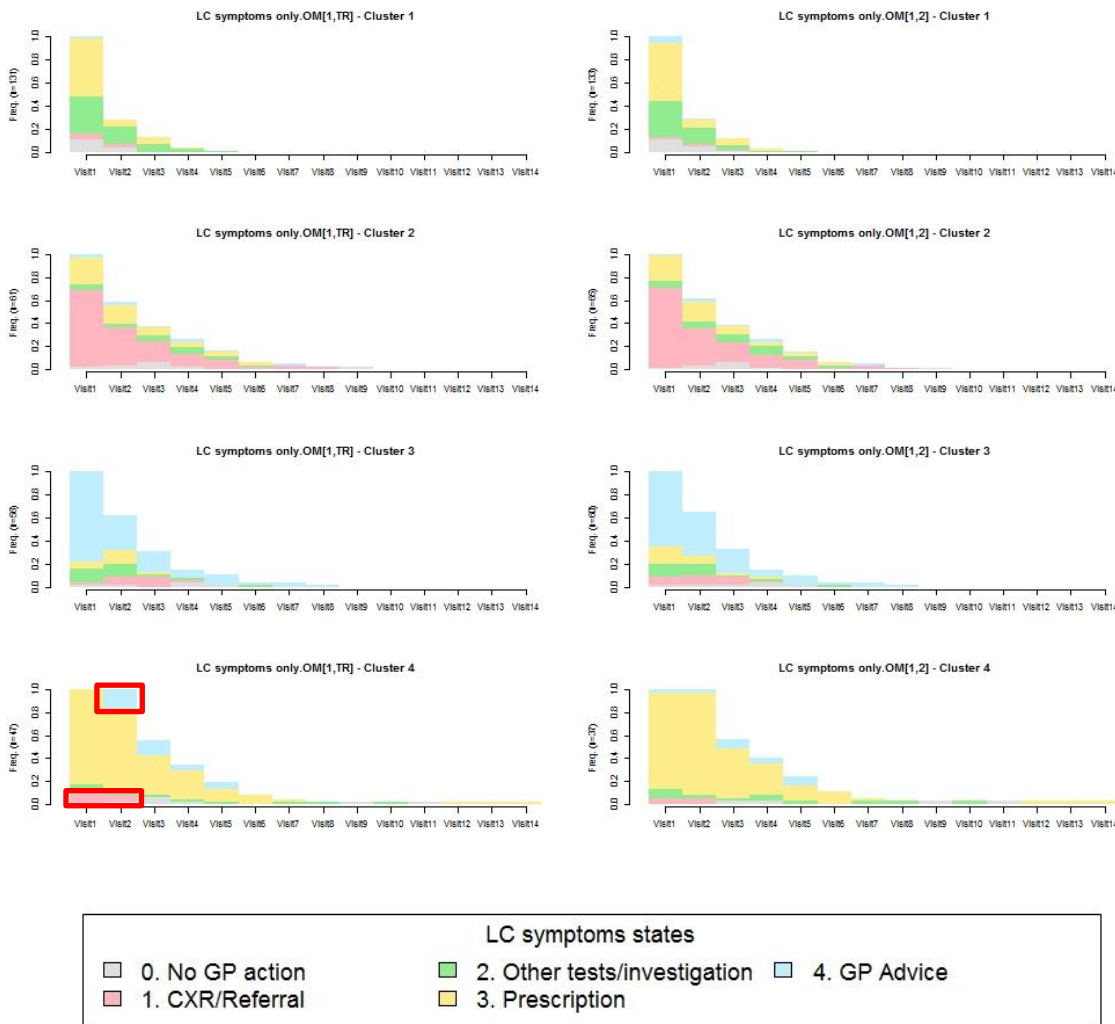


Figure 6.11 – The typologies (4 clusters) of $OM_{[1, TR]}$ (left) and $OM_{[1, 2]}$ (right) for the LC symptom sequences

The ten more sequences in Cluster 4 of $OM_{[1, TR]}$ resulted in a larger proportion of CXR/referral in the first two visits (in pink), and GP advice at Visit 2 (in light blue, both marked with red boxes).

These two parts were likely to be from Clusters 2 and 3 of $OM_{[1, 2]}$ (4 sequences for each), respectively. It was difficult to judge whether the ten sequences were misclassified, because there was no gold standard to classify the cluster membership for sequences. However, it made more sense for those ten sequences to join in respective clusters of $OM_{[1, 2]}$ based on the identified cluster patterns, which may explain why $OM_{[1, 2]}$ had a slightly better clustering structure indicated by ASW. Therefore, the typology of $OM_{[1, 2]}$ was considered a better typology and analysed afterwards.

6.4.4.1 Interpretation of the cluster patterns of $OM_{[1, 2]}$ in the LC symptom sequences

Cluster 1 (n=133): Most patients had only one visit of potential LC symptoms and GP ordered tests (in green), prescribed medications (in yellow), or no specific action (in grey). The percentage dropped rapidly in the second visit;

Cluster 2 (n=65): GP ordered CXR or referred the patients to a chest physician for potential LC symptoms (in pink);

Cluster 3 (n=60): GP reviewed patients' potential LC symptoms and offered advice (in light blue);

Cluster 4 (n=37): Patients had multiple consultations of potential LC symptoms and received repeated medications (in yellow).

6.4.4.2 Further explanations of the mechanism of grouping sequences

In Cluster 1 of $OM_{[1, 2]}$, there was a big proportion of prescriptions (in yellow) at Visit 1. It seemed odd that this part did not join in Cluster 4. But that was why the algorithm is called 'optimal' matching. Sequences in Cluster 1 was the shortest among the four clusters. Most patients had only one visit. It only needs one substitution to make the sequences identical. The distances between these sequences were short, cost=2 in $OM_{[1, 2]}$. Even for patients with two visits, only one insertion or deletion can make the sequences identical, which cost 1. However, sequences were longer in Cluster 4. Sequences in Cluster 1 need more operations to be the same as those in Cluster 4. Therefore, they cost more and the distances were larger. Patients in Cluster 4 had more long-term health problems, while patients in Cluster 1 just had some transient symptoms (e.g. cough) and very few visible parts in pink (CXR/referral for potential LC symptoms) in the figure. The percentages of patients with COPD in Cluster 4 (27.0%, 10/37) and Cluster 2 (26.2%, 17/65) were higher than those in Cluster 1 (18.1%, 24/133) and Cluster 3 (15.0%, 9/60). Therefore, the clustering solution is sensible. The comorbidity status of patients in different clusters is reported and further discussed in the next chapter.

In the four clusters, there were one or two dominant states, and other states were in small percentages. This was because the states at each visit were aggregated from the sequences in respective clusters to allow us to understand the cluster patterns over time. Without such aggregation, it would be impossible to interpret the pattern from individual sequences. But at the individual level, how GP responded to a patient's potential LC symptoms was different. For example, GPs may prescribe antibiotics for the symptom first, before ordering CXR or making a referral. What reflected in the figure of Cluster 2 was the states in yellow at the first visit (GP prescribed medications), and states in pink at the second visit (CXR/referral). The same situation could occur at later visits (yellow in Visit 2, pink in Visit 3, and so on). This example explained why there was a small part of yellow at Visits 1-3 in Cluster 2. This sequential pattern was also identified as one of the most frequent sub-sequences in Cluster 2 by **event sequence analysis** (event SA). Although event SA fragments the sequences, it is still helpful to provide information to facilitate the understanding of **how the algorithm works** and **the sequential event patterns within each cluster**.

6.4.5 The most frequent LC symptom sequences in each cluster of the typology by $OM_{[1,2]}$

Figure 6.12 presents the five most frequent (**whole**) LC symptom sequences in each cluster by $OM_{[1,2]}$, with the coverage of the cluster from around 62% (Clusters 2 and 3) to 80% (Cluster 1). **The height of the bars is proportional to the sequence frequency in each cluster**. Therefore, the height gradually shortens from the bottom to the top. The presentation of **the most frequent sequences supports the analysis and explanation of the cluster patterns in the previous subsection**. Sequences in Cluster 1 were the shortest, where the three most frequent sequences had only one state. Cluster 4 had the longest sequences – patients received multiple prescriptions after presenting with potential LC symptoms (at least twice, up to five times).

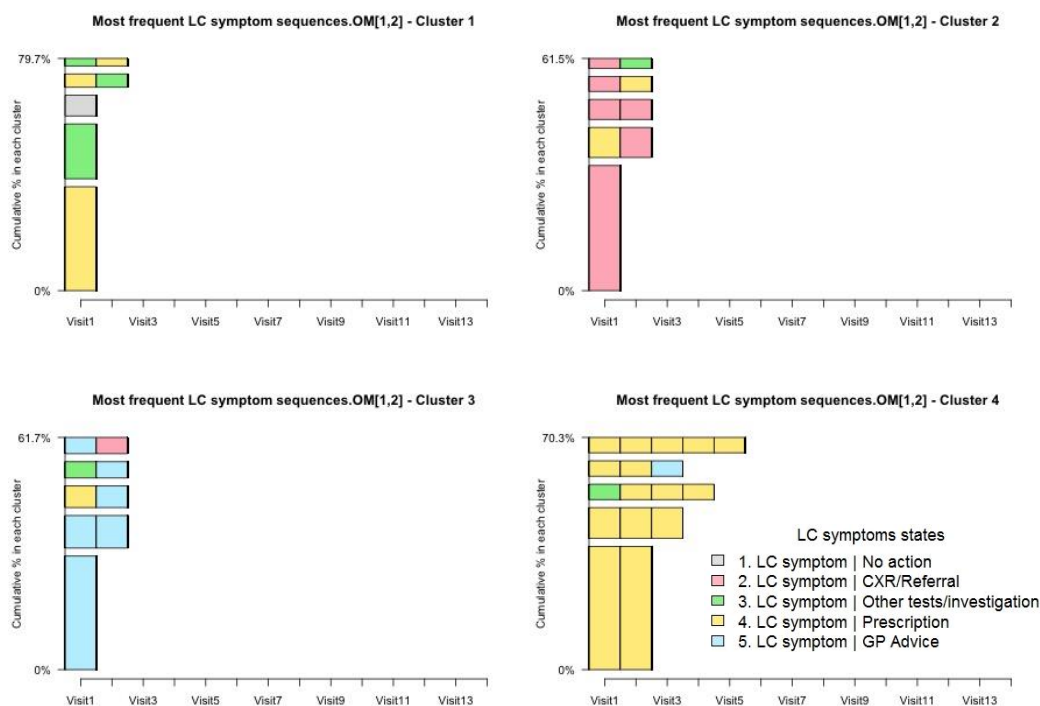


Figure 6.12 – The five most frequent LC symptom sequences in each cluster by OM_[1,2]

6.4.6 Event SA for LC symptom sequences in each cluster of the typology by OM_[1,2]

6.4.6.1 Explanations and examples of the results from event sequence analysis

Event sequence analysis was conducted in each cluster of the typology by OM_[1,2] (see Table 6.6). The table provides useful but not exhaustive or repetitive information to understand the sequential patterns in each cluster. Repetitive patterns were only reported once to avoid redundancy. The table is organised by the frequency of sub-sequences in descending order. Sequential events linked by the symbol “>” are consecutive (e.g. No. 4 sub-sequence in Cluster 2). Those linked by “-” are intermittent events, and may have other events in between (e.g. No. 2 sub-sequence in Cluster 1). Those marked with “>%” means the last event of the sequence. Notes are added in navy blue by the side of the sub-sequences to help readers make sense of the event sequences. Even in small percentages, some interesting event patterns are reported (e.g. Cluster 3, No. 5 and 6).

Here are some short examples of how sub-sequences could assist the interpretation of the cluster patterns in Figure 6.11. In Cluster 1, 75 patients (56.4%) had “Prescription” at the **last position** of the sequences (No. 1 sub-sequence, marked with “>%”), which could be at Visit 1, 2, 3... In Cluster 2, 14 patients (21.5%) had “Prescription” (No. 3 sub-sequence), and 13 of them (20%) had a subsequent “CXR/Referral” (No. 4 sub-sequence). The frequency and percentage were counted by patient, not at the event level.

Table 6.6 – The common sub-sequences in the four clusters of LC symptom sequences by OM_[1,2]

No.	Cluster 1, based on 133 sequences	%	Count
1	(Prescription>%) [Note: ">%" means the last event of the sequence]	56.4%	75
2	(Prescription) - (Prescription>%) [Note: something else may happen between the bar "-"]	44.4%	59
3	(Other tests/investigation>%)	33.1%	44
4	(Other tests/investigation) - (Other tests/investigation>%)	24.8%	33
5	(No GP action) [Note: in first and second visits, in grey colour of the figure]	10.5%	14

Note: the frequency and percentage of sub-sequences were counted by patient, not by event.

No.	Cluster 2, based on 65 sequences	%	Count
1	(CXR/Referral>%)	72.3%	47
2	(CXR/Referral) - (CXR/Referral>%)	52.3%	34
3	(Prescription)	21.5%	14
4	(Prescription > CXR/Referral) [Note: two consecutive events]	20.0%	13
5	(CXR/Referral > Prescription)	18.5%	12
6	(Prescription > CXR/Referral) - (CXR/Referral>%)	18.5%	12

No.	Cluster 3, based on 60 sequences	%	Count
1	(GP Advice>%)	83.3%	50
2	(GP Advice) - (GP Advice>%)	55.0%	33
3	(Prescription > GP Advice)	20.0%	12
4	(Other tests/investigation > GP Advice) - (GP Advice>%)	15.0%	9

5	(GP Advice) - (GP Advice > Other tests/investigation)	10.0%	6
6	(GP Advice > CXR/Referral)	10.0%	6

No.	Cluster 4, based on 37 sequences	%	Count
1	(Prescription)	83.8%	31
2	(Prescription) - (Prescription>%)	73.0%	27
3	(Prescription) - (Prescription > GP Advice)	16.2%	6

6.4.6.2 Similarities and differences between the most frequent sequences and event sequence analysis

Readers may find some similar patterns identified by both the most frequent sequences and event sequence analysis. The main difference is that **the most frequent sequences are the whole sequences**, while **sub-sequences are fragments of the whole sequences**. The whole sequence can be broken down at the event level. A, CE, ABD, and ABDE are sub-sequences of ABCDE. For short sequences like LC symptom sequences, you may find lots of overlapping patterns between the two methods.

Because LC symptom sequences were relatively simple, with only five states and high homogeneity (95 distinctive sequences out of 295). Therefore, it was possible to use a sequence frequency plot to present the most frequent sequences for each cluster. **For highly heterogeneous sequences, it is not always possible to select a few sequences to reach a certain threshold (e.g. 25%) and to represent others.** But event SA can always find sequential fragmented event patterns, from either homogeneous or heterogeneous sequences. Sequential patterns from event SA and the most frequent sequences provide distinctive and complementary information for readers to better understand the patterns in each cluster.

6.4.7 Brief summary and learning points: the analytical process of LC symptom sequences

SA can classify LC symptom sequences and identify the patterns. Although the pattern of the two typologies looks generally similar, the number of sequences is slightly different in respective clusters. Each of the four clusters provides distinctive information and does not overlap with others (good discrimination of cluster pattern). The five most frequent sequences were identified for each cluster by OM_[1, 2]. As a subsidiary method, event SA helped us further understand the

sequential pattern of events in each cluster in text, complementing the graphical presentation.

Interpreting the cluster patterns by triangulating the results from state SA and event SA can get a holistic perspective of the LC symptom sequences.

An important lesson from the exploration in this section is that the highest value of the statistical indicator (ASW) is not necessarily the optimal number of clusters in the empirical context. The differences in ASW values from 2 to 9 clusters were marginal (0.26-0.37). Therefore, other pragmatic considerations were more important, like the number of clusters, the number of sequences in each cluster, **the interpretability of the cluster patterns related to the research context**, and **further application and the implication of clusters in practice**. When it came to choosing the better typology between $OM_{[1, TR]}$ and $OM_{[1, 2]}$, as four clusters were considered the best solution for both algorithms, and the main difference was in the cluster membership, the ten sequences in Cluster 4 of $OM_{[1, TR]}$ made more sense to join in respective clusters of $OM_{[1, 2]}$, as explained in subsection 6.4.4. In addition, $OM_{[1, 2]}$ had a bigger ASW value, which indicated a better clustering structure. All of these made $OM_{[1, 2]}$ the better solution. Similarly, in the study by Hougham et al. (2014)(in the systematic scoping review), the authors did not choose the final solution with the highest ASW (0.26 for 2 clusters), but considering the cluster quality, discrimination between the clusters of sequences, and the sample size for each cluster together, and chose eight clusters (ASW=0.14) as the final solution. **Researchers should go through and compare different solutions, make sense of the cluster patterns, consider thoroughly and exercise judgement to come up with the optimal solution.**

6.5 Subgroup analysis 2 – potential LC symptoms situated in smoking-related comorbidities and other alarm symptoms

6.5.1 Sequence profile 4: high-risk sequences

6.5.1.1 The whole high-risk sequences

The high-risk sequences were more complex than those including potential LC symptoms only. Sequence length ranged from 1 to 27, median 3, IQR [2, 5]. The median interval between two high-risk consultations was 57 days, IQR [20, 161] days. The sequence length and interval of visits were heavily right-skewed. The frequency of the 15 states was already present in Table 6.3. The state distribution plot by visit is in Figure 6.13. The states are in the same order as they are numbered in the legend, from the bottom to the top. Besides the four states about how GP

managed patients with potential LC symptoms, the tenth state “GPs reviewed or offered advice for patients with cardiorespiratory presentations” in ‘sky blue’ and the 13th state “GP ordered tests or investigated other alarm symptoms” in ‘sea-green’ are more visible in the figure, as they have higher frequencies.

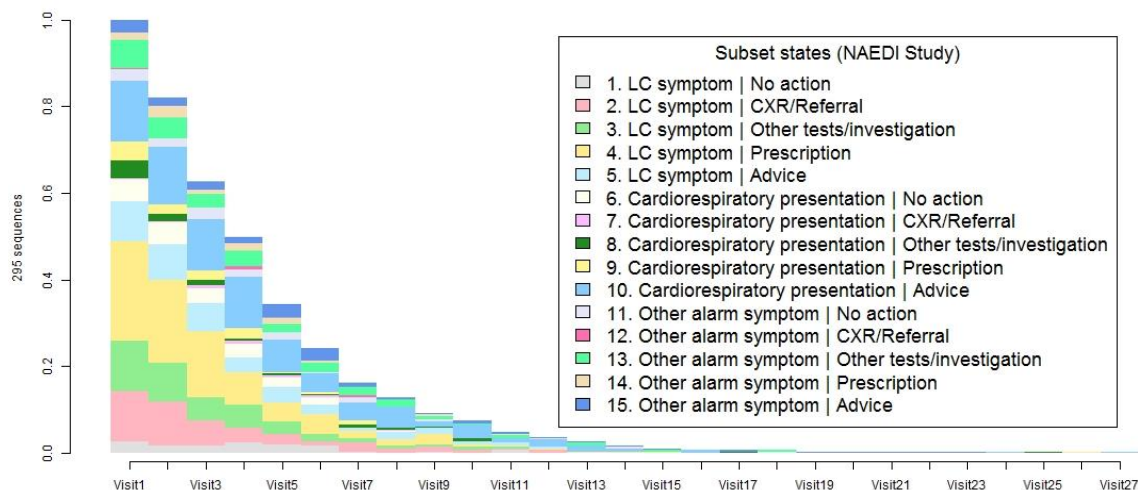


Figure 6.13 – The state distribution plot for the high-risk sequences (subgroup analysis-2)

6.5.1.2 Introducing a common reference point in the sequences – the first observed presentation of potential LC symptoms

The interpretation of the sequences is limited, if the sequences do not start from the same reference point. For example, in Figure 6.13, the state “LC symptoms | CXR/Referral” in light pink is sporadic after Visit 4. Without a reference event, we could not know whether that was the first time patient presented with potential LC symptoms and GP ordered CXR/made a referral (three other non-LC symptom consultations happened before), or that visit was the fourth time patient presented with potential LC symptoms. Introducing a common reference point for all the sequences, i.e. the first observed presentation of potential LC symptoms, and left truncating the sequences, can help us know what happened after the reference point and make the sequences more comparable, as they have the same starting point. Visit 1 is the first observed presentation of potential LC symptoms. The patients may have presentations of potential LC symptoms before the starting point of this study, but this is the best possible way to make sense of the high-risk sequences. After truncation, the new state distribution plot is presented in Figure 6.14. The sequences after the first presentation of potential LC symptoms were used in this set of subgroup analysis. After finalising the clusters, the left-truncated parts are added back in each cluster to obtain a full picture of the whole high-risk sequences.

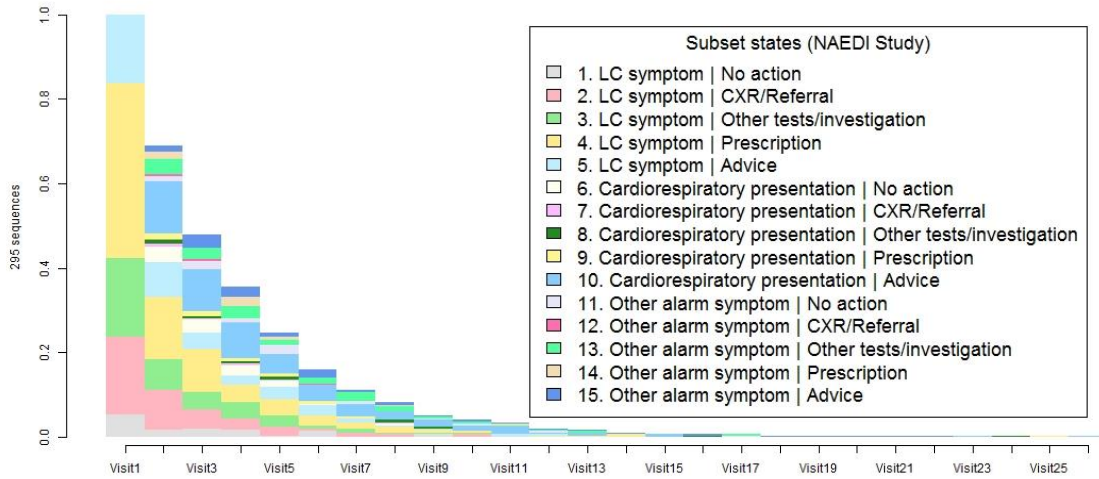


Figure 6.14 – State distribution plot for the high-risk sequences after the first presentation of potential LC symptoms

6.5.2 Dendrograms and regression trees for the high-risk sequences

6.5.2.1 Dendrograms and the outlier sequence

Among the 295 truncated high-risk sequences, there are 181 distinct sequences. Figure 6.16 presents the dendrograms of high-risk sequences in subgroup analysis by $OM_{[1, TR]}$ (top panel) and $OM_{[1,2]}$ (bottom). The two dendrograms on the left include all 295 sequences. Sequence No. 295 at the upper right corner is an outlier (26 visits), shown in Figure 6.15. Most of the states in this outlier sequence were GP offered advice for patients presented with other alarm symptoms. After removing the outlier, the longest sequence consisted of 17 visits. The whole analysis was run again. The new dendrograms by $OM_{[1, TR]}$ and $OM_{[1,2]}$ are presented on the right side of Figure 6.16. To avoid redundancy, the substitution cost matrices of $OM_{[1, TR]}$ are in Appendix D.

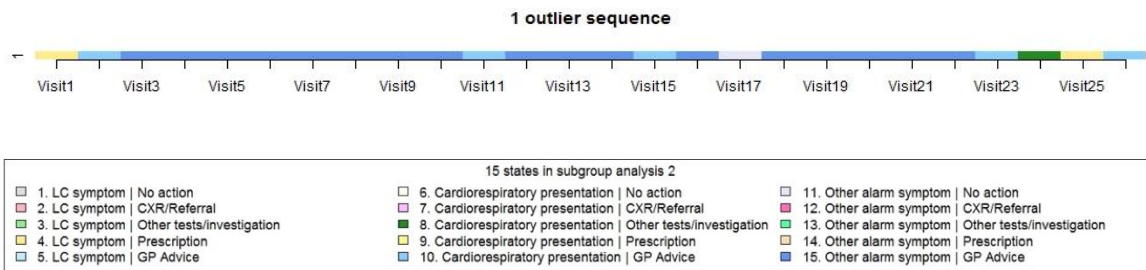


Figure 6.15 – An outlier high-risk sequence in subgroup analysis 2

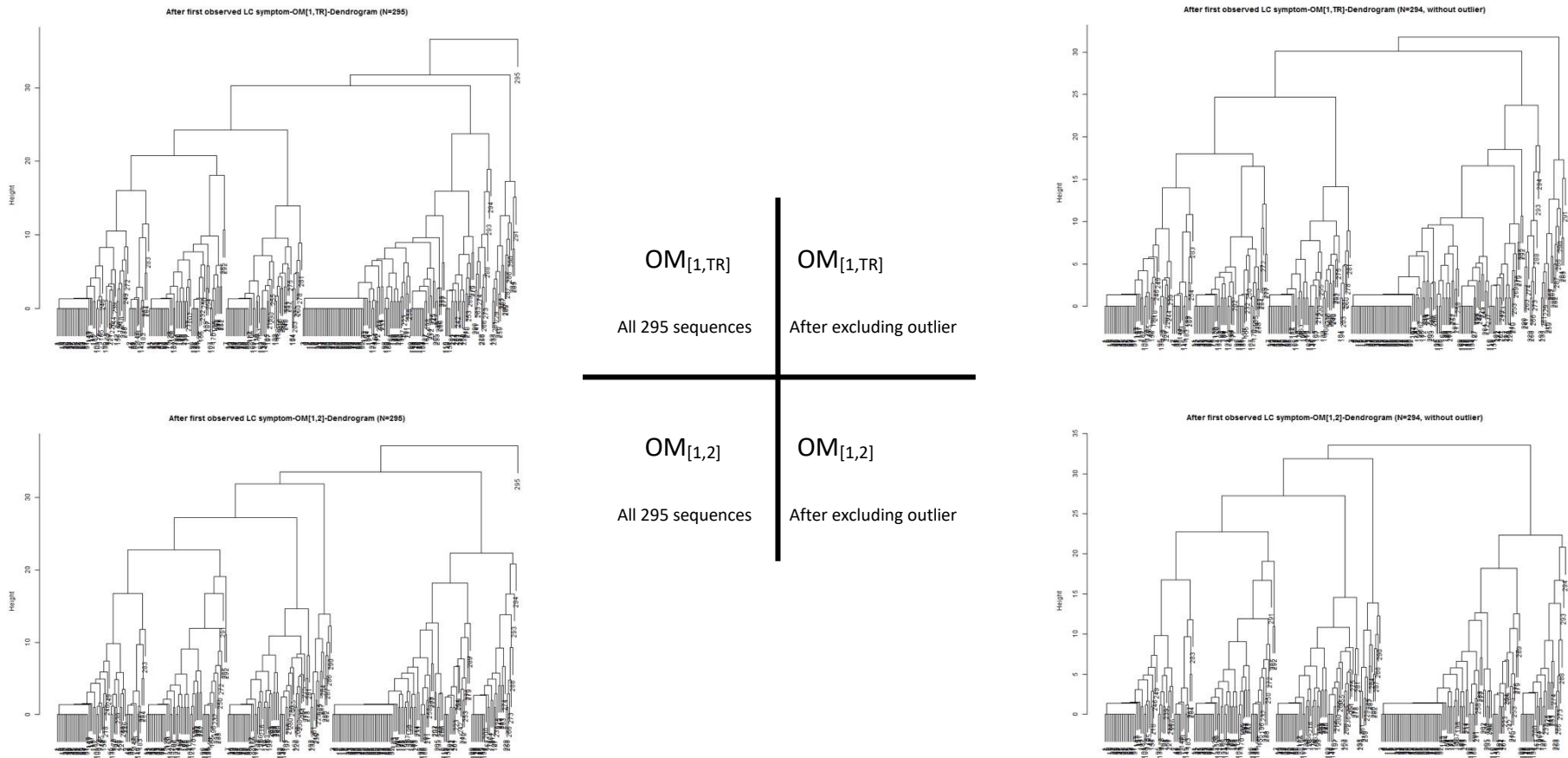


Figure 6.16 – Dendrograms of the high-risk sequences (top: $OM_{[1,TR]}$; bottom: $OM_{[1,2]}$; left: all 295 sequences; right: after excluding the outlier sequence)

6.5.2.2 Regression tree and deciding the number of clusters

Based on the structures of the dendrograms, there are some possible solutions for $OM_{[1, TR]}$ and $OM_{[1, 2]}$ algorithms. Regression trees in Figure 6.17 show how the cluster patterns changed in $OM_{[1, TR]}$ and $OM_{[1, 2]}$ as the nodes in the dendrogram are split. The regression trees stop growing at Step 5 in $OM_{[1, TR]}$ and Step 6 in $OM_{[1, 2]}$ due to repetitive patterns in both. ASW values were calculated and compared among different solutions, and again, the optimal solution based on the highest ASW value (statistical criterion) was not ideal in the empirical context. It is more important that the cluster patterns make sense in the empirical context. Therefore, the ASW values are not reported here.

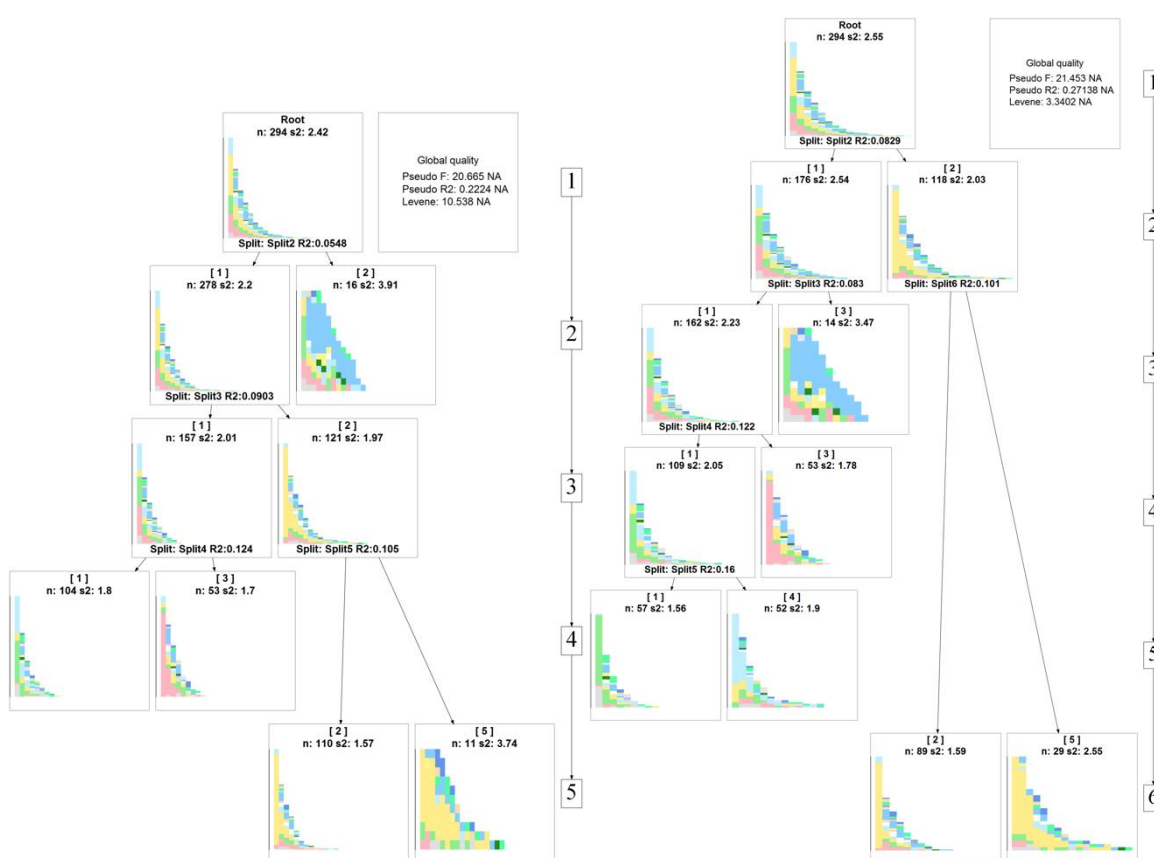


Figure 6.17 – Regression tree of the high-risk sequences by $OM_{[1, TR]}$ (left) and $OM_{[1, 2]}$ (right) in subgroup analysis 2 (after excluding the outlier sequence, $n=294$)

6.5.3 Comparison and interpretation of the cluster pattern in the typologies of high-risk sequences

The typologies of the high-risk sequences by $OM_{[1, TR]}$ and $OM_{[1, 2]}$ are presented in Figure 6.18. It is considered the cluster patterns by $OM_{[1, 2]}$ have better differentiation than those by $OM_{[1, TR]}$. Clusters 1 and 4 in $OM_{[1, 2]}$ are like a split from Cluster 1 in $OM_{[1, TR]}$, while in $OM_{[1, TR]}$, they are

Chapter 6

mixed together. Such differentiation is useful and relevant to the RQ, by separating the GP actions between ordering tests to investigate symptoms and offering advice only. The other three cluster patterns are similar between $OM_{[1, TR]}$ (Clusters 2-4) and $OM_{[1, 2]}$ (Clusters 2, 3, 5). Therefore, the typology by $OM_{[1, 2]}$ is a better result.

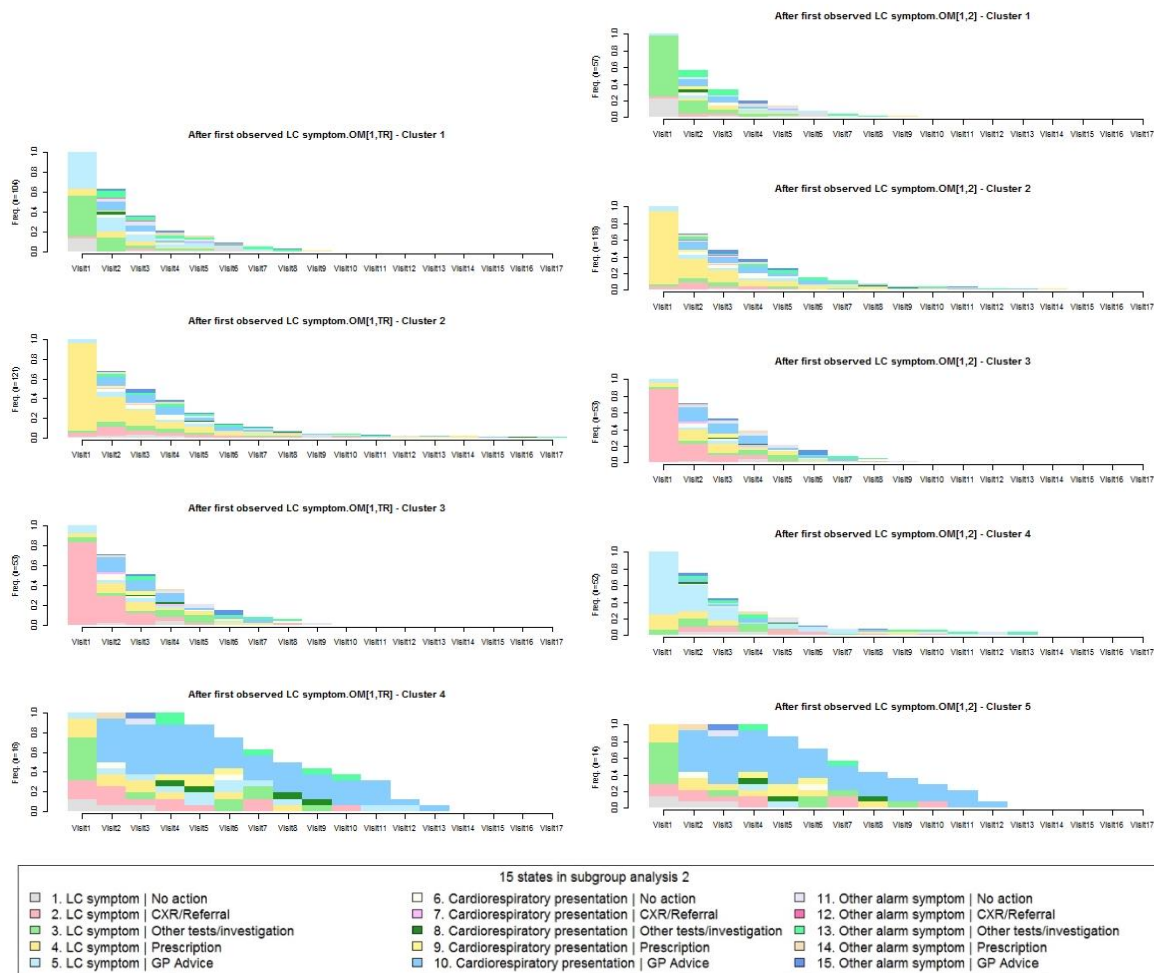


Figure 6.18 – Typologies of high-risk sequences by $OM_{[1, TR]}$ (left) and $OM_{[1, 2]}$ (right) in subgroup analysis 2

The cluster patterns of the high-risk sequences by $OM_{[1, 2]}$ are more complicated than those of potential LC symptoms only in the previous section. Here is the interpretation of **the main patterns** in the five clusters.

- Cluster 1 (n=57): GP requested other tests for potential LC symptoms (in light green, in the first two visits) and no GP action (in grey, first visit);
- Cluster 2 (n=118): GP prescribed medications for patients with potential LC symptoms in the first three visits (in yellow);

- Cluster 3 (n=53): GP ordered CXR and/or referred the patients with potential LC symptoms (in pink, the first two visits), prescribed medications for potential LC symptoms, and GP offered advice for cardiorespiratory presentations (the second and third visits);
- Cluster 4 (n=52): GP offered advice (in light blue, major pattern) and prescribed medications (minor pattern) for potential LC symptoms in the first three visits;
- Cluster 5 (n=14): GP offered advice for cardiorespiratory presentations (after the second visit, in sky blue, the main pattern)

Most states in the outlier sequence (Figure 6.15) were GP offered advice for patients presented with other alarm symptoms, which explained why it strongly deviated from the other sequences and did not fit in any of the five clusters above. The non-LC symptom states only occurred in Cluster 5 as a major pattern. This was the new information that did not appear in the typologies of the LC symptom sequences, as these states were not included in the previous analysis. In addition, the LC symptom states were not just in the first three visits of Cluster 5, but appeared across visits, with non-LC symptom states in between. For example, “LC symptoms | CXR/referral” in pink was visible at Visits 2-4, 7 and 10. This cluster was quite different from the other four clusters. We could not know whether the other health conditions (e.g. COPD) distracted patients’ attention from the potential LC symptoms, when they explained their health problems to their GPs, or that was related to the habit of GP recording health events (e.g. GP only recorded COPD but not the symptoms).

6.5.4 Representative high-risk sequences in each cluster of the typology by $OM_{[1,2]}$

Figure 6.19 presents the five most frequent high-risk sequences in five clusters by $OM_{[1,2]}$. The coverages are from around 36% (Cluster 5) to 54% (Cluster 1), substantially less than those in the LC symptom sequences (62%-80%). The main reason was that the sequences became more complex and heterogeneous because ten more states were added to the sequences. Notably, the five sequences in Cluster 5 were randomly selected by the statistical programme, rather than representative, as $5/14=35.7\%$. Each sequence in Cluster 5 is unique. Despite this, we can still see several consecutive states of “GP offered advice for patients with cardiorespiratory presentations” in ‘sky blue’ at the middle or the end of the five sequences. The 4th and 5th sequences in Cluster 3 are only one sequence for each ($1/53=1.9\%$). These two examples demonstrate the point mentioned in subsection 6.5.2.2 – it could be difficult to select representative sequences for complex and heterogeneous sequences. **The software can give you the number of the most frequent sequences you want, but it cannot guarantee the coverage and representativeness of sequences.** For the rest of the four clusters, most sequences just have 1-2 states, and most of these states are still related to GP actions for potential LC symptoms, with

Chapter 6

some sporadic state of “GP offered advice for patient’s cardiorespiratory presentations” at Visit 2. There are two possible explanations for this. One is because the first state is an LC symptom-related state as a reference point. The second explanation is that LC symptom states still have higher frequencies and percentages than the other ten states. Therefore, it means that situating the potential LC symptoms in ten clinically relevant states does not actually add more new information, but perhaps increases more noise in this study.

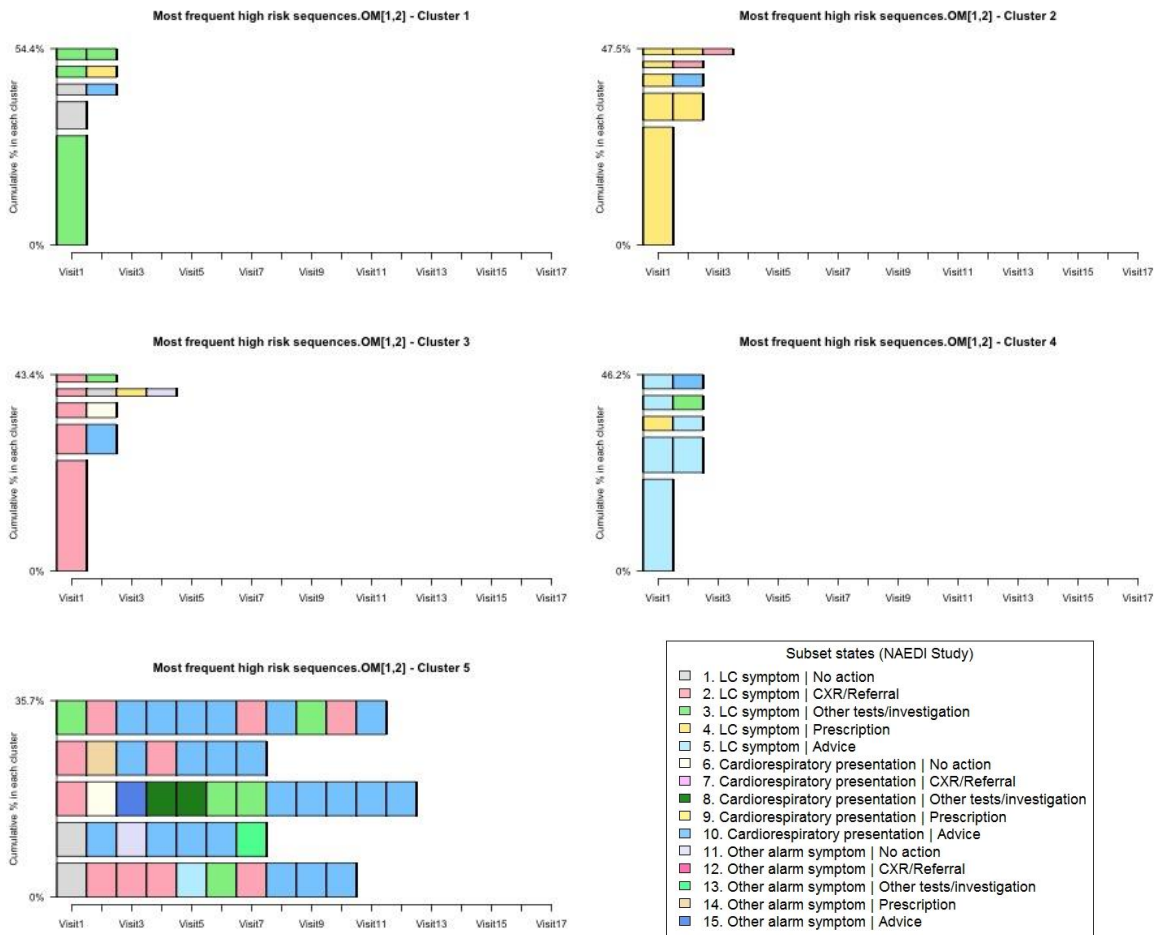


Figure 6.19 – The five most frequent high-risk sequences in each cluster by OM_[1, 2]

6.5.5 Event SA of high-risk sequences in each cluster by OM_[1,2]

Some interesting sequential patterns were revealed by event SA in each cluster. Only informative patterns are reported in Table 6.7, leaving out the repetitive ones. The longer the common subsequences, the smaller the percentages are. Some of the patterns were already reported in Table 6.6, as the high-risk sequences were the expansion of the LC symptom sequences with ten more states. Besides the five GP responses to potential LC symptoms on patients, only one additional high-risk state was picked up by event SA, i.e. GP advice for cardiorespiratory presentation, mostly after the four combined LC symptom states (CXR/referral, other tests, prescription, and GP

advice). Other events were less frequent. Interesting consecutive events are in bold and in navy blue in Table 6.7. Once again, there are some differences in the identified patterns between the most frequent sequences and event SA.

Table 6.7 – The common sub-sequences in the five clusters of the high-risk sequences by OM_[1,2]

No.	Cluster 1, based on 57 sequences	%	Count
1	(LC Other tests)	73.7%	42
2	(LC Other tests) - (LC Other tests>%)	42.1%	24
3	(LC No GP action)	22.8%	13

Note: LC is short for potential LC symptoms; “|” separates the patient and the GP part, i.e. patient | GP parts; “-” means other events may happen between the bar; “>” means two consecutive events. “>%” means the last event/state in the sequence.

No.	Cluster 2, based on 118 sequences	%	Count
1	(LC Prescriptions)	89.0%	105
2	(LC Prescriptions) - (LC Prescriptions>%)	48.3%	57
3	(LC Prescriptions > Cardiorespiratory GP advice)	17.8%	21
4	(LC Prescriptions) - (LC Prescriptions > Cardiorespiratory GP advice)	15.3%	18
5	(Cardiorespiratory GP advice>%)	13.6%	16

No.	Cluster 3, based on 53 sequences	%	Count
1	(LC CXR/Referral)	88.7%	47
2	(LC CXR/Referral) - (LC CXR/Referral>%)	35.8%	19
3	(LC CXR/Referral > LC Prescriptions)	20.8%	11

Chapter 6

4	(LC CXR/Referral > Cardiorespiratory GP advice)	18.9%	10
5	(LC CXR/Referral) - (LC CXR/Referral > LC Prescriptions)	17.0%	9
6	(LC CXR/Referral) - (LC CXR/Referral > Cardiorespiratory GP advice)	17.0%	9
7	(LC CXR/Referral > Cardiorespiratory GP advice) - (Cardiorespiratory GP advice>%)	13.2%	7
8	(LC CXR/Referral) - (LC CXR/Referral > Cardiorespiratory GP advice) - (Cardiorespiratory GP advice>%)	11.3%	6

No.	Cluster 4 , based on 52 sequences	%	Count
1	(LC GP advice)	75.0%	39
2	(LC GP advice) - (LC GP advice>%)	38.5%	20
3	(LC Prescriptions)	19.2%	10
4	(LC Prescriptions > LC GP advice)	17.3%	9
5	(LC Prescriptions) - (LC Prescriptions > LC GP advice)	15.4%	8
6	(LC GP advice > LC Other tests)	15.4%	8
7	(LC GP advice) - (LC GP advice > LC Other tests)	11.5%	6
8	(LC GP advice > LC CXR/Referral)	11.5%	6

No.	Cluster 5 , based on 14 sequences	%	Count
1	(Cardiorespiratory GP advice>%)	64.3%	9
2	(LC Other tests > Cardiorespiratory GP advice)	50.0%	7

3	(LC Other tests) - (LC Other tests > Cardiorespiratory GP advice)	35.7%	5
4	(LC Other tests > Cardiorespiratory GP advice) - (Cardiorespiratory GP advice>%)	28.6%	4
5	(LC Prescriptions > Cardiorespiratory GP advice)	28.6%	4
6	(LC CXR/Referral > Cardiorespiratory GP advice) - (Cardiorespiratory GP advice>%)	21.4%	3
7	(LC Prescriptions) - (LC Prescriptions > Cardiorespiratory GP advice) - (Cardiorespiratory GP advice>%)	21.4%	3
8	(Cardiorespiratory GP advice > LC Prescriptions)	21.4%	3

6.5.6 The left-truncated part – events happened before the first observed presentation of potential LC symptoms

The left-truncated high-risk states (called pre-1st LC SYM sequences) joined the typology of the high-risk sequences by $OM_{[1,2]}$. Based on the cluster membership, the pre-1st LC SYM sequences were matched and presented by cluster to get a full picture of the whole high-risk sequences.

The left-hand side of Figure 6.20 is the pre-1st LC SYM sequences, with the x-axis from Pre10 to Pre1. It means there are up to 10 events before the first observed presentation of potential LC symptoms. The frequency increases from Pre10 to Pre1. However, the frequency of Pre1 adding up from the five clusters is not 100%, as not every patient had a high-risk event before the first observed LC symptom. Among the 245 pre-LC SYM events, 36.7% (n=90, coloured in sky blue) were GP advice for cardiorespiratory presentation, 13.9% (34/245, in sea green) were GP ordered tests for patients with other alarm symptoms, and 11.4% (n=28, coloured in ivory) were no GP action for cardiorespiratory presentation. The other seven states were less frequent (<10%). States only appeared in the last two visits of the pre-1st LC SYM sequences in cluster 5, which was probably due to the small sample size of the cluster membership (n=14). The outlier sequence had only one pre-LC state, which was GP advice for cardiorespiratory presentation (not plotted).

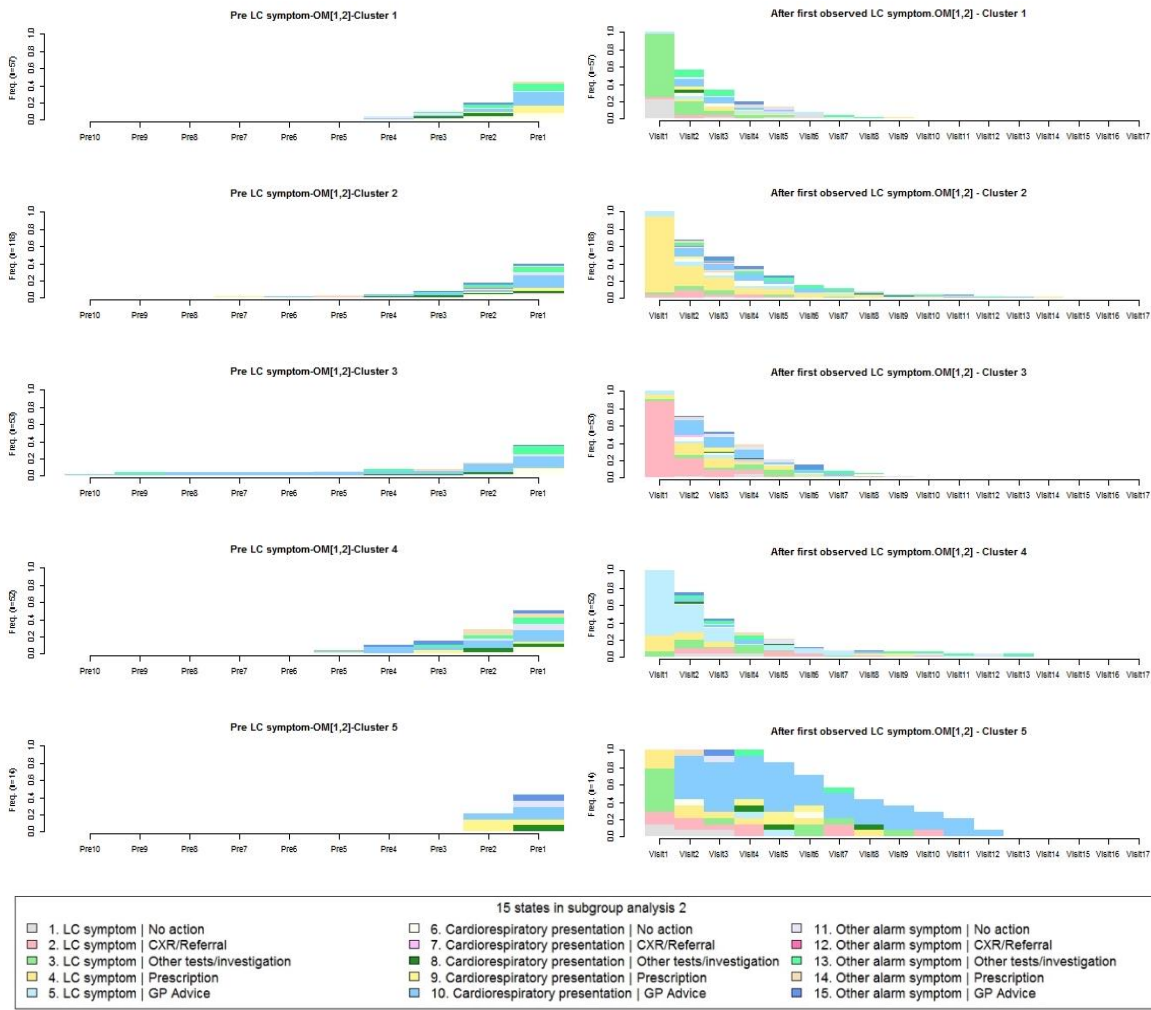


Figure 6.20 – The state distribution plots before (left) and after (right) the first observed presentation of potential LC symptoms in the five clusters by $OM_{[1,2]}$

6.5.7 Comparison of the typologies between the two sets of subgroup analysis

Figure 6.21 presents the typologies of LC symptom sequences (left) and high-risk sequences (right). Between the two cost setting schemes, constant cost $OM_{[1,2]}$ produced better results than $OM_{[1, TR]}$. Comparing the two subgroup analyses, the similarity is that SA could classify different GP actions for potential LC symptoms. The difference is that one extra cluster is identified in high-risk sequences. The major pattern is GP advice for cardiorespiratory presentation. Adding ten more states provide limited extra useful information, as the percentages for most of the high-risk states are too small to influence the typology. Their presence in the typology (thin lines in different colours) increases substantial noises and makes the interpretation of the cluster patterns more difficult. Therefore, the result of LC symptom sequences (five LC states, four clusters) is considered a better solution between the two sets of subgroup analyses. The reasons are:

1. How GP managed patients presented with potential LC symptoms is the research interest. The LC symptom sequences are simpler and the results are more straightforward and with less noise;
2. The typologies from LC symptom sequences and high-risk sequences are not substantially different, other than one more cluster with a small number of sequences, which could be explained by some variables of comorbidities (patients had specific cardiorespiratory diseases, and the total number of comorbidities counted from the GP notes);
3. It is difficult and inappropriate to assign the outlier sequence back to any cluster of the high-risk sequences, as it is essentially different from the current five cluster patterns. It is neither practical to make it as a cluster with only one sequence.

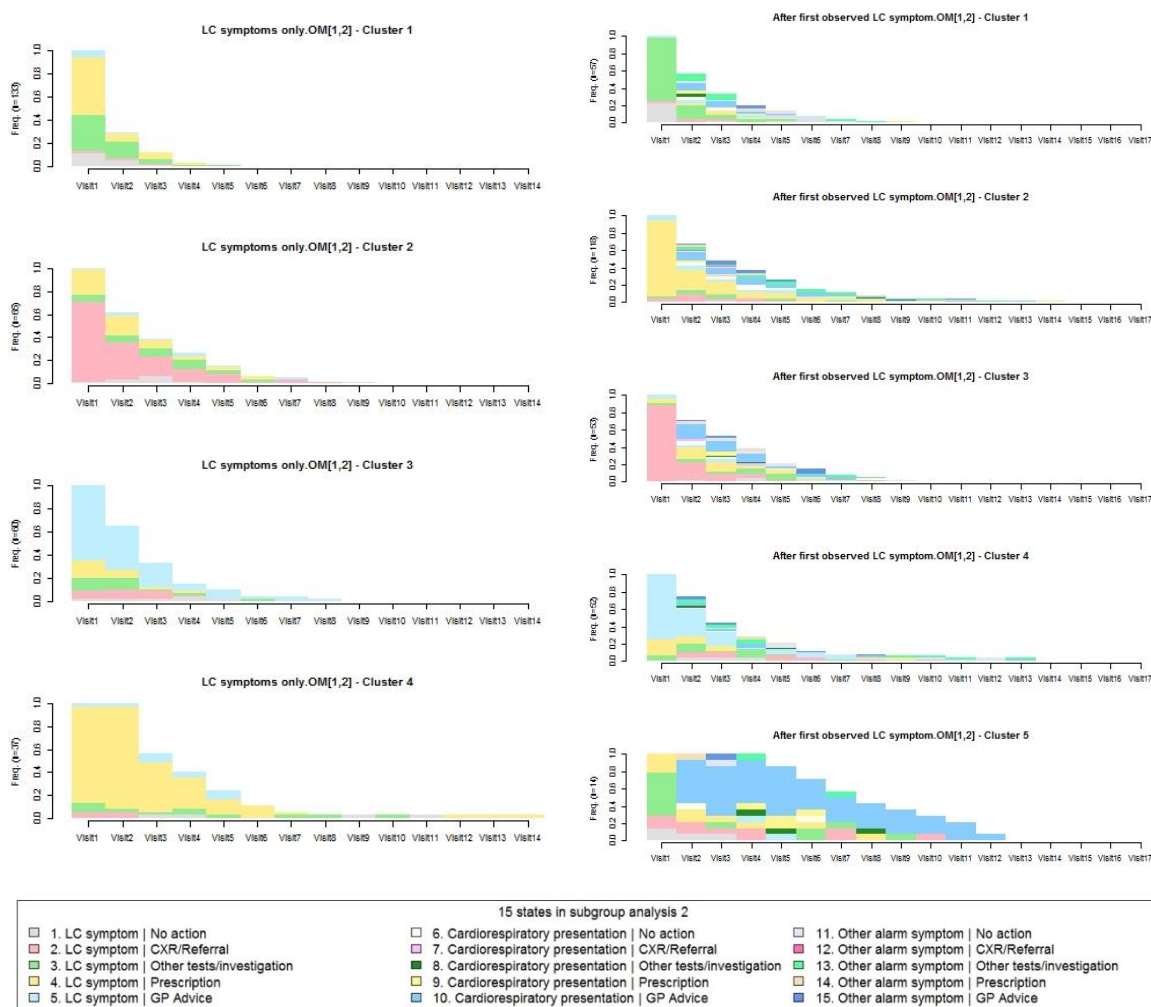


Figure 6.21 – The typologies of LC symptom sequences (left) and high-risk sequences (right) by $OM_{[1,2]}$

An important lesson is that **states with high percentages have more influence on the results**. SA is still a descriptive tool. **Using prior knowledge to specify states**, leaving out less relevant events

or combining less frequent states, **may help get more meaningful patterns**. This point was previously mentioned in the systematic scoping review (see subsection 4.3.3).

6.5.8 A brief summary and learning points: subgroup analysis of high-risk sequences

In this section, the LC symptom states were situated among other high-risk states. There are some learning points gained from the exploration. Firstly, it could be helpful to **use a reference point to construct sequences** in a specific context. It is demonstrated in Figure 6.4 to use the date of cancer diagnosis as a reference point to organise the dates of health events backwards. Here, the first observed potential LC symptoms were used as a reference point to align all the high-risk sequences to make them comparable. The second lesson is that different dissimilarity measures and cost setting schemes should be tried, and the typologies should be compared to choose the one that fits the research context the best.

The third lesson is that **outlier sequences should be identified first, and the whole analysis should be run again, as Ward's method is sensitive to outliers**. One outlier sequence was identified, illustrated, and explained why it deviated from other sequences. Notably, the sequence number was not the same as the Patient ID. The sequences were sorted by the number of visits ascendingly before the graphic presentation. The outlier sequence in this section (patient ID 07/029, presented with multiple alarm symptoms) is not the same patient as the sequence singled out at Step 5 in Figure 6.10 (patient ID 04/063, with 14 visits of potential LC symptoms).

The fourth learning point is that **the optimal number of clusters might not be consistent between the statistical indicator (ASW) and the empirical context. Making sense of the cluster patterns in the empirical context in a parsimonious manner should take precedence over the statistical indicator**. The most frequent sequences and the patterns in text by event SA for each cluster further enhance the understanding of the graphical typology. The fifth lesson is that **states with higher percentages have more influence on the result of typology**. Although some states are important, they could not stand out in the typology due to low frequency. **Grouping relevant states together would be a good strategy**. This is the same as we could group similar categories together in a regression model, when there are small numbers in some categories of a nominal variable. The analytical approach and the lessons learnt from this subsection may be helpful for other researchers to conduct their own research.

6.6 High-risk primary care consultations among patients without potential LC symptoms during the observation period

The previous two sections (6.4 and 6.5) presented the analytical process and the results of 295 patients who presented with potential LC symptoms at least once. For the remaining 563 patients (858-295) without potential LC symptoms, 326 patients (57.9%, 326/563) had 767 primary care consultations about cardiorespiratory diseases and other alarm symptoms (the ten states included in the high-risk sequences of subgroup analysis 2). This section provides a brief overview of the sequence characteristics in this group of patients. The number of visits related to cardiorespiratory diseases and alarm symptoms ranged from 1 to 14, median 2, IQR [1,3]. The state distribution plot is presented in Figure 6.22. The three states with a percentage >10% are still GP advice for cardiorespiratory diseases (38.5%, 295/767), followed by No GP action for cardiorespiratory presentations (16.0%, n=123), and GP ordered tests investigating other alarm symptoms (14.3%, n=110). These three states account for almost 70% of the total consultations. The other seven states had less than 10% of events in each state. It is very rare that GP ordered CXR or referred the patients to a chest physician due to cardiorespiratory presentation 0.9% (n=7) or other alarming symptoms 0.4% (n=3).

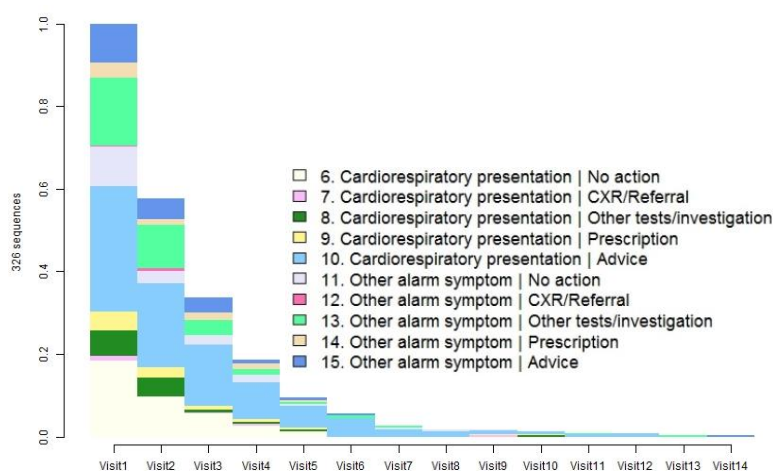


Figure 6.22 – State distribution plot for patients without potential LC symptoms but consulted cardiorespiratory diseases and other alarm symptoms (n=326)

6.7 General discussion of the methodological exploration and the empirical findings of the primary care sequences in the NAEDI study

6.7.1 Summary of the methodological exploration in this chapter

In this chapter, I tried to contextualise the application of SA to study primary care sequences among patients at high risk of developing LC using the NAEDI dataset. I applied the principles of SA to represent the primary care sequences involving interdependent patient-GP events, explained how the methodological decisions were made step by step, reported the whole process in detail, presented the results of SA in different types of figures, interpreted the cluster patterns, compared different approaches of subgroup analyses, and discussed the strengths, limitations, and implications of the methodological decisions in the previous sections. Through this chapter, one can see that SA is an exploratory process. To sum up, the research objectives and the methodological issues mentioned at the beginning of this chapter (section 6.1) have been fully addressed, explained as follows:

1. Two ways to construct primary care sequences, by visits and in a timeline, were introduced. Sequences constructed in a timeline have the same length and interval, and thus have more choices of dissimilarity measures and cost settings (e.g. $OM_{[1, TR]}$, $OM_{[1, 2]}$ HAM, and DHD). In the subgroup analysis, it is more sensible to construct sequences by visits, rather than in a timeline, because spreading the small number of events over 29 months (the observation period) is not helpful to find useful patterns due to the presence of a very dominant state of 'non-attendance' in each month. Although HAM and DHD were planned to compare with OM, it was not possible to conduct such analysis and make a comparison, because HAM and DHD could not be used to analyse sequences of unequal lengths.
2. How different cost setting schemes ($OM_{[1, TR]}$ and $OM_{[1, 2]}$) would influence the clustering structure was explored in two sets of subgroup analyses, which could be considered as sensitivity analysis. After comparing the results, the conclusion is $OM_{[1, 2]}$ in LC symptom sequences yields the best typology in this study.
3. Criteria to help determine the optimal number of clusters were proposed, and further discussed in subsection 6.7.4 below.

6.7.2 Simplification of the primary care events and categorisation into states

States in SA need to be mutually exclusive. Therefore, complicated clinical information needs to be simplified. What happened in a 10-minute consultation becomes a state representing the main reason and outcome of the consultation. Some details are inevitably lost. It is quite challenging to

come up with a reasonable number of states that could represent a wide range of primary care events. Therefore, prior knowledge and judgement are essential when specifying the states, i.e. what events are more important than others in the research context, how to group the events into states that would be informative for the RQs, and how to rank the states in a hierarchy. The states in this study went through multiple steps, from thematic grouping of health events in free text to categories, grouping patients and GP categories into states individually, defining the hierarchy of states, and finally combining the patient and GP states together. The intention to do the final step is because understanding the patterns of interdependent patient-GP events is of interest in this study, but it might not be necessary to combine patient and GP events together in other research contexts. The process of using EHRs for research would be slightly different. Researchers need to decide the variables of interest, and then identify the Read/ICD codes for each variable. But the other steps remain the same. The whole process is time-consuming. However, if using SA in empirical research, time is well spent to consider and choose the right variables, and tailor the states for specific RQs, as the states will directly influence the outcome of the analysis.

Patient's health problem and GP actions are interdependent. It is not easy to distinguish patient events from GP events in health records. One example is the state 'smoking cessation service'. We could not know whether the patients presented with health problems or simply wanted to quit smoking. Without further available information from the transcribed notes, the patient part was left blank and 'smoking cessation service' was coded as GP action. Furthermore, patient's wishes and personal preferences would also influence GP's action, especially in the advocate of "shared decision making" nowadays. Some patients may decline GP's suggestions. In such situations, the transcribed note clearly stated that GP discussed available options with the patient, but the patient preferred specific treatments. For example, some patients presented with flu symptoms (cough, chest complaints) but explicitly expressed that they did not want antibiotics or steroids. Therefore, GP offered advice, and scheduled a follow-up appointment to see whether symptoms were alleviated/resolved or not, while some patients insisted on antibiotics. Very few patients even declined the offer to be referred to a specialist. All of these were anecdotes in the transcribed notes, but we could not know why patients made specific decisions from the notes. Although the outcome of the consultation was reflected on the GP's part, patient's preferences may play a role.

6.7.3 Dissimilarity measures, cost settings, and the implications

Researchers have greater flexibility to choose dissimilarity measures and cost settings when the sequences have equal length. Based on the methodological exploration in this chapter (and the

previous exploration in the HHRAD dataset during my doctoral candidature⁴), results (dendrograms and typologies) from different dissimilarity measures and cost settings can have substantial differences. As Studer and Ritschard (2016) commented, there is no universally optimal distance index. **The choice of dissimilarity measure and cost setting depends on which aspect we want to focus on.** Given that SA is exploratory in nature and there is no gold standard for the final typology, researchers should consider the research context, and try different options based on the given data and make sense of the typology. This process is like using other statistical methods we are more familiar with (e.g. regression). We may need to build and compare different models with different variables, and find one that fits the data and explain the research phenomenon the best.

As mentioned earlier, the timing of primary care consultations relative to diagnosis is of research interest in empirical studies. It is better to use a relative timeline to align the sequences. **A common meaningful reference event and/or time point is important to make the sequences comparable and the cluster patterns meaningful.** For example, the date of cancer diagnosis is ideal as the endpoint of a milestone event, tracking all other relevant care events backwards, and align all the sequences in the same timeline. In the subgroup analysis of this study, the sequences were aligned by the first observed potential LC symptom.

The interval of the timeline should be consistent with the research context. The choice of month as an interval for two years is appropriate, as it can provide an overview of the whole trajectories. The number of contacts in general practice is likely to increase in a period closer to diagnosis (e.g. six months before diagnosis). Week may be a better choice as an interval in such a circumstance, which was explored in the HHRA dataset previously. However, it is not recommended to use week as an interval for a long period before diagnosis, as the percentage of “no visit” could be high, and it is more difficult to identify useful patterns. In addition, states can be changed or more specifically tailored for different periods (e.g. six months or three years before diagnosis) to provide more useful information in the typology, as what I did in presenting the whole primary care sequences for background information, and tailoring the states for the two subgroup analyses. In summary, the choice of states and intervals should consider the research context (e.g. the length of the study period, and the timing related to diagnosis) and be tailored for specific RQ, rather than set in stone.

⁴ The contents and results of the HHRAD dataset was not reported in this chapter. Here only reported the learning points.

6.7.4 Determine the number of clusters and the implications for further statistical analysis

Cluster analysis is an essential connecting step in the whole process of SA. It is based on the dissimilarity matrix of the pairwise sequences generated by SA, and the clustering structure greatly influences the final typology. In hierarchical clustering, the grouping of two sequences is done by searching for the shortest distance, indicating a great similarity between the two sequences, which is a local decision. These local choices, adding up together, may lead to considerable differences at a higher level. The best local choices cannot guarantee the best global clustering structure. Based on the literature review, this study uses a statistical indicator, ASW, to assess the clustering quality. From the dendrograms, heterogeneity could exist even for the sequences grouped in the same cluster, like the outlier sequence singled out at Step 5 of the regression tree in Figure 6.10. The ASW measures how similar a sequence is to its own cluster (homogeneity) compared with other clusters. In some clusters, great heterogeneity lowers the overall ASW values.

As an unsupervised learning method, there is no golden rule to determine the number of clusters. For pragmatic reasons, the number of clusters should not be too large (e.g. over ten). If the number of clusters is over ten, similar clusters could be grouped together to decrease the cluster number. Based on the exploration in two sets of subgroup analysis, **ASW should not be the key factor to determine the number of clusters in the typology**. ASW indicates the quality of clustering, but a good clustering structure from the statistical perspective does not always explain the empirical research context in the best way. This partly explains why in most published articles, the number of clusters is often decided by researchers with informal criteria, i.e. based on the interpretation of cluster patterns in a specific research context. It is not very often to see authors reporting the clustering quality using objective statistical indicators. In empirical research, the interpretation of cluster patterns is more important and meaningful than those found by statistical indicators. The ASW values can provide researchers with objective measurements of the clustering structure among different solutions. If the differences of ASW values are not large, it indicates the quality of clustering is at the same level, it may reassure researchers to choose the one that fits the empirical contexts the best. The exploration in this study reinforces the comment that **SA is a statistical method that sits somewhere between purely narrative and traditional variable analysis** (Pollock, 2007), **with a substantial part of subjective decision and interpretation**.

In some situations, it is possible to have more than one solution that the cluster patterns make sense in the research context. Should the typology be further used in other statistical analyses (e.g. regression model), researchers may pick one with pragmatic considerations such as the

number of clusters, the number of sequences (sample size) in each cluster. If the clusters are used as a dependent variable, more clusters could make the results in multinomial logistic regression more challenging to interpret. Results with a smaller number of clusters may have more advantages. Parsimony is the more favourable here.

6.7.5 The empirical findings and the implications

The intervals between two visits of potential LC symptoms (reported in Table 6.4) provide new information on the characteristics of help-seeking behaviours among high-risk patients. Based on the characteristics of primary care events, the sequences in the study sample were stratified into different subgroups and analysed separately.

Subgroup 1: patients presented with potential LC symptoms (32.8%, 295 out of 899 patients).

There are four clusters in this subgroup. GP actions reflected their perceptions of LC risk in patients.

- 1) Cluster 1 (n=133): GP either ordered test or prescribed medications for transient LC symptoms;
- 2) Cluster 2 (n=65): GP ordered CXR and/or referred the patients to a specialist for potential LC symptoms, which meant that GP suspected patients might have LC;
- 3) Cluster 3 (n=60): GP reviewed patient's symptoms and offered advice;
- 4) Cluster 4 (n=37): GP prescribed medications repeatedly for patients presented with symptoms multiple times over a longer period. It was worth further investigation of this cluster in the empirical analysis (in the next chapter), to find out what patient characteristics could help explain the difference between this cluster and the others.

Subgroup 2: patients visited GP but without potential LC symptoms during the observation period (62.6%, 563/899). Patients in this subgroup may be further divided into two clusters:

- a) Cluster 5 (high-risk consultations, 326/899, 36.3%): patients presented with cardiorespiratory diseases and/or other alarming symptoms indicating severe health problems;
- b) Cluster 6 (minimal care needs, 237/899, 26.4%): patients without high-risk presentations, they used primary care services only for minor health problems.

Subgroup 3: patients did not visit general practices at all (Cluster 7, no visit, 41/899, 4.6%), which could be one of the following three possibilities:

- i. Some patients could afford private health care and did not use public-funded health services. This explains no records for their attendances in the NHS general practices;
- ii. Patients were in good health status and used health services very rarely. They may see a GP in the NHS, but outside the observation period of this study;
- iii. Patients who need care the most did not seek help. They may have a higher risk than patients presenting with potential LC symptoms, due to their unawareness of LC symptoms. They may have a higher risk of ending up with an emergency presentation, late stage diagnosis, and poor survival.

The seven clusters have distinctive cluster patterns (good discrimination), which enables us to know how GP managed high-risk patients in the community over time. Patients in different clusters may be at different levels of risk of developing LC. They may need different strategies to early detect and diagnose LC from primary care. Together with other variables, the typologies from SA could be very useful. For example, traditional statistical methods like parametric or non-parametric tests can be used to compare patients' characteristics (age, sex, SES, comorbidities) in each cluster. From this, we can understand the differences of patient profiles in each cluster. Multinomial logistic regression can be used to investigate the association between patient characteristics and the clusters of care sequences (as categorical dependent variable). All of these results are reported in the next chapter.

6.7.6 Relevance and implications in the broader field of early diagnosis research

The clinical value and benefits of studying how GPs manage patients presenting with potential LC symptoms in primary care is obvious. Identifying the cluster patterns and understanding how GPs manage different patient groups can help us understand which patterns are associated with delays in cancer diagnosis. In addition, such research evidence can help allocate primary care resources and inform GP to better manage patients at high risk, which can increase patients' chance of early diagnosis and intervention of LC.

6.7.7 Strengths and limitations of the methodological exploration

This is the first study of using SA to construct primary care sequences from routine consultations in general practice and analyse sequences from different angles. None of the studies in the systematic scoping review (Chapter 4) used SA to study interdependent GP-patient events. This study addresses some important methodological issues specific to applying SA in primary care sequences. Sensitivity analyses were conducted. Typologies from different cost settings and subgroup analyses were compared. The graphic presentation of the patterns is a unique value of

SA, which is not achievable from other types of statistical analysis. For example, regression tree is helpful to present how the patterns change at each step. State and event SA are used to make sense of the cluster patterns. The similarity and difference of the patterns between the most frequent sequences and event patterns are discussed. **None of the studies included in the systematic scoping review has applied state and event SA in such depth to understand a research problem.** The lessons learnt from the methodological exploration are summarised and discussed, which would be helpful for researchers who are interested in using SA to explore the patterns of primary care sequences in other types of cancer or diseases. All of these are the strengths of the methodological exploration phase.

The technical uncertainties of SA have been discussed in section 3.6. The main limitation is that the algorithm might classify some sequences incorrectly. However, we cannot know the true cluster membership of each patient, as they are unlabeled data (terminology in machine learning literature). In addition, there is no gold standard to verify the cluster membership.

6.8 Conclusion of the methodological exploration phase

SA can be used to represent and analyse primary care sequences, and identify meaningful typologies, which can provide us with a new perspective in early diagnosis research. Stratified analysis could be an analytical option for a **heterogeneous** study sample. Although this study focuses on primary care sequences relevant to LC, the application of SA is not limited to cancer. Its application can be extended to study care pathways or disease trajectories for other health conditions. This will be further discussed in the final chapter.

The next chapter reports the results of the empirical analysis, to understand patient's sociodemographic and clinical characteristics in each cluster, and further explore the association between patient characteristics, different clusters of primary care sequences, and indicators of patient's help-seeking behaviour.

Chapter 7 Results of the empirical analysis

7.1 Introduction

Cluster patterns of primary care sequences have been identified in the previous chapter. This chapter reports the results of the empirical analysis outlined in sections 5.3.2 to 5.3.4 and discusses the implications of clinical practice and health services research.

7.2 Results and interpretations

7.2.1 Sociodemographic and clinical characteristics of the study sample

The sociodemographic characteristics of the study sample ($n=899$) are presented in Table 7.1. The completeness of the variables is good. Only a few patients did not provide information accurate enough to calculate age ($n=16$, 1.8%) and years of smoking ($n=12$, 1.3%). A small proportion (4–5%) of data in marital status and ethnicity was coded as “unknown” in the demographic dataset (by the data collection team), rather than missing. Other variables did not have missing data.

The average age of the patients was 65.6 ± 9.4 years old, ranging from 50 to 95. 56% of them were male ($n=501$). There were 41 patients without primary care visits during the observation period. For those who visited general practices, the total number of visits varied greatly among patients, ranging from 1 to 54, with a median of 9 visits in 29 months. The study sample had a relatively long smoking history, median 34 years, IQR [20, 43] years, and the longest record of smoking history was up to 74 years. 32% of patients ($n=286$) reported themselves as current smokers, and 63% of patients ($n=566$) reported they quit smoking at various times in their lives, but not all successful.

The majority of patients were white (93%, 837/899). Besides 39 (4.4%) patients with unknown ethnicity, only 23 patients were ethnic minority (Black, Asian, mixed-race, or others). The study sample lived around the south coast of England, one of the most affluent areas in the country. The IMD was not equally distributed in quintiles, with a low percentage of the most deprived quintile (7.3%), and a higher percentage in the middle quintile (Q3, 35.2%). 65% of patients ($n=586$) were married, partnered, or cohabited, 30% ($n=266$) were single, divorced, or widowed, and the rest 5% ($n=47$) did not report their marital status. About 60% of patients were retired, 22% were still full-time employed, 9% part-time employed, and the rest 10% were unemployed.

Chapter 7

About 24% patients (n=216) did not obtain any qualification, followed by GCSE/O-level (22.5%, n=202), vocational (18.5%, n=166), and degree or above (17%, n=153).

About 40% patients (n=361) had one comorbidity, 24% patients (n=217) had two comorbidities, and 15% (n=134) had three or more comorbidities. Only about 18% of patients (n=159) had no comorbidity. As to the four specific comorbidities, hypertension had the highest prevalence, about 37.5% (n=337), followed by other heart diseases 10.7% (n=96), which could be broken down as CHD (n=46), CVD (n=24), IHD (n=21), AF (n=13), MI (n=7). Notably, GP may record more than one abbreviation for patient's heart problem, e.g. CVD and AF. Therefore, the sum of the five abbreviations of heart disease (111) was greater than 96. The prevalence of COPD was 10.0% (n=90) and asthma 8.3% (n=75), which were lower than those of cardiovascular diseases.

Table 7.1 – The sociodemographic and clinical characteristics of the study sample (N=899)

Continuous variables	Mean±SD
Age, (n=883)	65.6±9.4 years
Total number of visits (n=858)	10.4±7.9 times
Years of smoking (n=887)	32.2±15.4 years
Binary or categorical variables	N(%)
Sex	
Male	501 (55.7)
Female	398 (44.3)
Ethnicity	
White	837 (93.1)
Mixed	6 (0.7)
Black/ Black British	4 (0.4)
Asian/ British Asian	8 (0.9)
Chinese	1 (0.1)
Other	4 (0.4)

Point 8 - change of presentation style of the table

Point 8 - add female in the table

Unknown	39 (4.4)
IMD quintile	
1 (Most deprived)	66 (7.3)
2	192 (21.4)
3	316 (35.2)
4	165 (18.3)
5 (Less deprived)	160 (17.8)
Marital status	
Married or cohabiting	586 (65.2)
Single or not partnered	266 (29.6)
Unknown	47 (5.2)
Employment status	
Full time employed	198 (22.0)
Part-time employed	80 (8.9)
Unpaid role / unemployed	89 (9.9)
Retired	532 (59.2)
Highest qualification	
Unreported	77 (8.6)
None	216 (24.0)
GCSE/ O-Level	202 (22.5)
Vocational qualification	166 (18.5)
A level	85 (9.4)
Degree or above (MA, PhD)	153 (17.0)

Number of comorbidities	
Unknown	28 (3.1)
0	159 (17.7)
1	361 (40.2)
2	217 (24.1)
≥3	134 (14.9)

Point 7

7.2.2 Results of the association between sociodemographic and health variables

Considering the majority of patients were retired, it was difficult to know their financial situation from the variable of employment status. Therefore, it was not used to explore the pairwise association. Only significant results were reported here, and leaving out the nonsignificant ones. The IMD quintile had a marginal association with qualification (Spearman's rho, $\rho=0.09$, $P=0.01$), but was not associated with other sociodemographic variables. Education level (qualification) was negatively associated with the years of smoking ($\rho=-0.17$, $P<0.001$) and the number of comorbidities ($\rho=-0.10$, $P=0.007$). Although in a weak association, it meant that the higher education the patients received, the fewer smoking years and comorbidities they had. The number of comorbidities had a moderate significant association with the total number of visits ($\rho=0.46$, $P<0.001$). Patients who had more comorbidities had more visits to the general practices. Age was significantly associated with the years of smoking (Pearson's $r=0.16$, $P<0.001$), the total number of visits ($r=0.17$, $P<0.001$), and the number of comorbidities (Spearman's $\rho=0.18$, $P<0.001$), all of which were weak association.

7.2.3 Patient characteristics among the seven clusters

7.2.3.1 Patient profile and statistical tests among the seven clusters

The seven cluster patterns of the primary care sequences were described in the previous chapter (subsection 6.7.5). Patients' characteristics among the seven clusters are reported in Table 7.2. The results of significant variables are presented in front of the nonsignificant ones. Age, the total number of visits (six clusters, except for the "No visit" group), the number of potential LC symptom consultations (4 clusters), the number of comorbidities, COPD, asthma, hypertension, other heart diseases, and current smoking status were the nine statistically significant variables among the seven patient clusters.

Patients in Cluster 2 (LC symptoms | CXR/referral) were the oldest patient group, aged 69.1 ± 9.7 years, 8.1 years older than patients in Cluster 7 (No visit, 61.0 ± 7.4 years old, $P < 0.001$), and 4.4 years older than patients in Cluster 6 (Minimal care needs, 64.6 ± 9.5 years old, $P = 0.018$). Both differences were significant by Šidák's method in post hoc pairwise comparison after ANOVA. Patients in the other five clusters (Clusters 1, 3-6) had similar ages, around 65-66 years.

The variance in the number of visits was not homogeneous (rejected by Bartlett's test) among six groups (not applicable to the group 'No visit'). Non-parametric Kruskal-Wallis test found that the number of visits was significantly different among the six clusters. Patients presented with potential LC symptoms in the four clusters had a much higher number of visits than patients without potential LC symptoms (Clusters 5-6). Patients in Cluster 4 (potential LC symptoms | Multiple prescriptions) had the largest numbers of visits (17.7 ± 9.6 visits in 29 months) and potential LC symptom consultations (3.6 ± 2.3 times), followed by patients in Cluster 2 (potential LC symptoms | CXR/Referral), Cluster 3 (potential LC symptoms | GP advice), and lastly Cluster 1 (transient LC symptoms | Test/prescriptions). For patients without potential LC symptoms, patients with high-risk presentations (Cluster 5, 9.7 ± 6.6 visits) had four more visits to general practice on average than those only with minimal care needs (Cluster 6, 5.6 ± 4.8 visits).

For binary, categorical, and ordinal variables, percentages are reported by column with frequency, for easier comparison of the percentages at each level of the variables among clusters. From the distribution, the comorbidity burden was significantly different among patients in seven clusters. The number of comorbidities (0, 1, 2, ≥ 3) was similar among patients in Clusters 2, 3, and 4. Most patients in these three clusters had at least two comorbidities, while patients in Cluster 2 (potential LC symptom | CXR/Referral) had the highest percentage of ≥ 3 comorbidities (34.9%) than any other 6 clusters. Patients in Clusters 1 and 5 had a similar but smaller comorbidity burden than those in the previous three groups. Most patients in Clusters 1 and 5 had 1 or 2 comorbidities. Almost 70% of patients in Cluster 7 (No visit at all) did not have any comorbidity, better than patients in Cluster 6, the majority of whom had 0 or 1 comorbidity. However, 31.7% (13/41) of patients in Cluster 7 had an unknown comorbidity burden, which could be underestimated due to the unavailable data.

The prevalence of COPD, asthma, and cardiovascular diseases was higher in patients presented with potential LC symptoms, except for hypertension, where patients in Cluster 5 (no LC symptoms, but with cardiorespiratory diseases) had the highest prevalence ($>50\%$). The prevalences of four cardiorespiratory comorbidities (COPD, asthma, hypertension, and other

heart diseases) were significantly different among seven clusters. However, if only comparing the prevalence among the first four clusters (patients with potential LC symptoms), the differences were **not significant**. Patients in Cluster 4 receiving multiple prescriptions were probably due to the highest prevalence of COPD (27%, 10/37) and hypertension (43%, 16/37), while patients in Cluster 2 got a rapid referral to CXR and/or a specialist were probably because they were the oldest group and had more comorbidity burden, reflecting on the number of comorbidities and the prevalence of four specific cardiorespiratory conditions. Patients in Cluster 2 had the highest prevalence of asthma (20%, 13/65) and heart diseases (22%, 14/65), and the second highest prevalence of COPD and hypertension. From the descriptive statistics in Table 7.2, the prevalence of four cardiorespiratory diseases was in three levels: patients with potential LC symptoms (Clusters 1-4), patients with cardiorespiratory diseases and/or other alarm symptoms (Cluster 5), and patients in the rest two clusters (Cluster 6 minimal care needs and Cluster 7 no visit). This probably explained why the differences in comorbidity prevalence were significant among the seven clusters, but not among patients presented with potential LC symptoms.

The percentages of current smoking status were also significantly different. Patients in Clusters 2, 3, and 5 had <30% as current smokers, less than patients in Clusters 1, 4, and 6 (around 35%). More than 50% of patients in Cluster 7 (no visit) were current smokers.

Other patient characteristics, including sex, years of smoking, highest qualification, marital status, IMD quintile, and employment status, were not significant among the seven clusters. The detailed descriptive statistics and the test result are all in Table 7.2.

In summary, the patient characteristics significantly different among patients in seven clusters are in four dimensions, including age (**biological**), the number of visits to general practice and potential LC symptom consultations (**behavioural**), comorbidities (5 variables, **clinical**), and current smoking status (**lifestyle**). These characteristics were distinctively different among three groups of patients without potential LC symptoms (Clusters 5, 6, 7). For the four clusters of patients with potential LC symptoms, it was more difficult to tell the differences between the clusters. Patients in Cluster 2 (potential LC symptoms | CXR/referral) were the oldest, which was the most distinctive characteristic. Other patient characteristics in Cluster 2 were similar to those in Cluster 3 (potential LC symptoms | Advice).

Table 7.2 – Descriptive statistics and tests comparing patient characteristics among the seven clusters

	Cluster 1 (n=133)	Cluster 2 (n=65)	Cluster 3 (n=60)	Cluster 4 (n=37)	Cluster 5 (n=326)	Cluster 6 (n=237)	Cluster 7 (n=41)	P-value
Age, mean±SD	65.5±8.8	69.1±9.7	65.9±9.9	66.1±8.8	66.0±9.4	64.6±9.5	61.0±7.4	<u>0.0013</u>
The total number of visits	13.1±8.2	16.1±8.1	15.6±10.1	17.7±9.6	9.7±6.6	5.6±4.8	0	<u>0.0001</u> [#]
The number of consultations related to LC symptoms	1.4±0.8	2.6±1.8	2.3±1.5	3.6±2.3	0	0	0	<u>0.0001</u> [#]
Number of comorbidities	n (col %)							<u><0.0001</u>
- 0	14 (10.7)	5 (8.0)	8 (13.3)	4 (10.8)	23 (7.1)	86 (37.4)	19 (67.8)	
- 1	54 (41.2)	15 (23.8)	18 (30.0)	11 (29.7)	152 (47.2)	107 (46.5)	4 (14.3)	
- 2	41 (31.3)	21 (33.3)	15 (25.0)	13 (35.2)	95 (29.5)	28 (12.2)	4 (14.3)	
- ≥3	22 (16.8)	22 (34.9)	19 (31.7)	9 (24.3)	52 (16.2)	9 (3.9)	1 (3.6)	
COPD, n (%)	24 (18.1)	17 (26.2)	9 (15.0)	10 (27.0)	29 (8.9)	0	1 (2.44)	<u><0.0001</u>

	Cluster 1 (n=133)	Cluster 2 (n=65)	Cluster 3 (n=60)	Cluster 4 (n=37)	Cluster 5 (n=326)	Cluster 6 (n=237)	Cluster 7 (n=41)	P-value
Asthma, n (%)	17 (12.8)	13 (20.0)	8 (13.3)	5 (13.5)	26 (8.0)	5 (2.1)	1 (2.4)	<0.0001
Hypertension, n (%)	49 (36.8)	24 (36.9)	26 (43.3)	16 (43.2)	183 (56.1)	34 (14.4)	5 (12.2)	<0.0001
Other heart diseases, n (%)	12 (9.0)	14 (21.5)	8 (13.3)	3 (8.1)	41 (12.6)	18 (7.6)	0	0.007
Current smokers, n (%)	47 (35.3)	17 (26.2)	14 (23.7)	13 (35.1)	92 (28.6)	82 (35.3)	22 (53.7)	0.017
Years of smoking	35.3±14.9	33.7±15.1	31.0±15.3	35.0±14.5	31.5±15.9	30.5±15.8	34.3±11.4	0.067
Male sex, n (%)	65 (48.9)	36 (55.4)	31 (51.7)	18 (48.7)	198 (60.7)	131 (55.3)	22 (53.7)	0.312
Marital status								0.060
- Married or cohabiting	82 (61.6)	44 (67.7)	34 (56.7)	15 (40.5)	220 (67.5)	161 (67.9)	30 (73.2)	
- Single or not partnered	42 (31.6)	19 (29.2)	23 (38.3)	21 (56.8)	87 (26.7)	66 (27.9)	8 (19.5)	
- Missing/unknown	9 (6.8)	2 (3.1)	3 (5.0)	1 (2.7)	19 (5.8)	10 (4.2)	3 (7.3)	

	Cluster 1 (n=133)	Cluster 2 (n=65)	Cluster 3 (n=60)	Cluster 4 (n=37)	Cluster 5 (n=326)	Cluster 6 (n=237)	Cluster 7 (n=41)	P-value
Employment status								0.076
- Full time employed	33 (24.8)	8 (12.3)	9 (15.0)	3 (8.1)	70 (21.5)	61 (25.7)	14 (34.2)	
- Part time employed	8 (6.0)	2 (3.1)	5 (8.3)	3 (8.1)	35 (10.7)	23 (9.7)	4 (9.8)	
- Unpaid role/unemployed	14 (10.5)	9 (13.8)	10 (16.7)	5 (13.5)	30 (9.2)	17 (7.2)	4 (9.8)	
- Retired	78 (58.7)	46 (70.8)	36 (60.0)	26 (70.3)	191 (58.6)	136 (57.4)	19 (46.2)	
IMD quintile								0.511
- 1 (Most deprived)	16 (12.1)	5 (7.7)	6 (10.0)	3 (8.1)	20 (6.1)	15 (6.3)	1 (2.4)	
- 2	22 (16.5)	12 (18.5)	8 (13.3)	9 (24.3)	84 (25.8)	48 (20.3)	9 (22.0)	
- 3	45 (33.8)	25 (38.5)	18 (30.0)	14 (37.9)	107 (32.8)	90 (38.0)	17 (41.5)	
- 4	22 (16.5)	11 (16.9)	16 (26.7)	8 (21.6)	60 (18.4)	43 (18.1)	5 (12.2)	
- 5 (Less deprived)	28 (21.1)	12 (18.4)	12 (20.0)	3 (8.1)	55 (16.9)	41 (17.3)	9 (21.9)	

	Cluster 1 (n=133)	Cluster 2 (n=65)	Cluster 3 (n=60)	Cluster 4 (n=37)	Cluster 5 (n=326)	Cluster 6 (n=237)	Cluster 7 (n=41)	P-value
Highest qualification								0.805
- Missing/unknown	13 (9.8)	4 (6.2)	4 (6.7)	7 (18.9)	31 (9.5)	15 (6.3)	3 (7.3)	
- None	34 (25.6)	17 (26.2)	14 (23.3)	10 (27.0)	80 (24.5)	50 (21.1)	11 (26.8)	
- GCSE/ O-Level	30 (22.5)	11 (16.9)	15 (25.0)	8 (21.6)	74 (22.7)	56 (23.6)	8 (19.5)	
- Vocational qualification	19 (14.3)	11 (16.9)	14 (23.3)	4 (10.8)	52 (16.0)	58 (24.5)	8 (19.5)	
- A level	15 (11.3)	7 (10.7)	5 (8.3)	3 (8.1)	31 (9.5)	19 (8.0)	5 (12.2)	
- Degree and above (MA, PhD)	22 (16.5)	15 (23.1)	8 (13.3)	5 (13.5)	58 (17.8)	39 (16.5)	6 (14.6)	

Note: other heart diseases included the abbreviations of CVD (cardiovascular disease), IHD (ischaemic heart disease), CHD (coronary heart disease), AF (atrial fibrillation), and MI (myocardial infarction) in free text from GP notes.

χ^2 (chi-square) test was used for binary and categorical variables. ANOVA was used to compare continuous variables among clusters. For those marked with #, Kruskal-Wallis test was used instead, as the hypothesis of homogeneity of variance was rejected by Bartlett's test.

7.2.4 Results of multinomial logistic regression model

7.2.4.1 The reference outcome cluster and the independent variables

Multinomial logistic regression was performed to investigate the associations between patients' characteristics and the clusters. Cluster 2 "LC symptom | CXR/Referral" was served as the reference cluster, as it was the research interest in this study. The other six clusters were compared with the reference cluster. Sex, marital status, qualification, IMD, and years of smoking were not significant predictors among the seven clusters at any level in multinomial logistic regression, and therefore not reported here.

7.2.4.2 Significant results in the univariable multinomial logistic regression model and the interpretations

Significant patient characteristics associated with the seven clusters in the univariable multinomial logistic regression model are reported in Table 7.3. To avoid redundancy, constants for each variable are not reported. Unadjusted relative risk ratio (RRR) and the corresponding 95% CI are presented for each cluster compared with the reference cluster, rather than the original coefficients (β), for more intuitive interpretation.

The results in Table 7.3 are consistent with the patient profile presented in Table 7.2 and described in subsection 7.2.3.1. What the results from the multinomial logistic regression added is the relative risk of each variable among different patient clusters. For example, patients in Cluster 1 had a smaller number of potential LC symptom consultations than patients in Cluster 2; therefore, the $RRR < 1$; and patients in Cluster 4 had a larger number of potential LC symptom consultations, the $RRR > 1$. For significant results, the 95% CI of RRR does not include 1, while for nonsignificant results, the 95% CI of RRR include 1 (coloured in light grey in the table). In addition, for categorical variables, what multinomial logistic regression is superior to χ^2 (chi-square) test is that multinomial logistic regression can not only identify significant results between pairwise comparison, but also quantify the strength of RRR, while χ^2 test only tells you the percentages are different in a categorical variable across clusters. For example, patients in Cluster 6 and 7 had significantly less comorbidity burden than patients in Cluster 2 ($RRR < 0.1$, no comorbidity as the reference category). Due to the small sample size in Cluster 7, the 95% CI of RRR had very wide intervals in some variables (e.g. employment status, current smokers, asthma).

Point 2

Table 7.3 – Significant results in univariable multi-nominal logistic regression (**unadjusted** relative risk ratio and 95% CI)

	Cluster 1 (n=133)	Cluster 3 (n=60)	Cluster 4 (n=37)	Cluster 5 (n=326)	Cluster 6 (n=237)	Cluster 7 (n=41)
Age	0.96 (0.93, 0.99) *	0.97 (0.93, 1.00)	0.97 (0.93, 1.01)	0.97 (0.94, 0.99) *	0.95 (0.92, 0.98)***	0.90 (0.86, 0.95) ***
Total number of visits	0.96 (0.93, 0.99) *	0.99 (0.96, 1.03)	1.01 (0.98, 1.05)	0.90 (0.87, 0.93) ***	0.76 (0.73, 0.80) ***	N.A. (0)
Consultations of LC symptoms	0.43 (0.32, 0.57) ***	0.91 (0.72, 1.14)	1.28 (1.03, 1.59)*	N.A. (0)	N.A. (0)	N.A. (0)
Number of comorbidities	No comorbidity as reference					
- 1	1.29 (0.40, 4.14)	0.75 (0.20, 2.78)	0.92 (0.20, 4.22)	2.20 (0.73, 6.64)	0.41 (0.14, 1.19)	0.07 (0.02, 0.31) ***
- 2	0.70 (0.22, 2.20)	0.45 (0.12, 1.64)	0.77 (0.18, 3.42)	0.98 (0.34, 2.89)	0.08 (0.03, 0.22) ***	0.05 (0.01, 0.21) ***
- ≥3	0.36 (0.11, 1.16)	0.54 (0.15, 1.93)	0.51 (0.11, 2.35)	0.51 (0.17, 1.53)	0.02 (0.01, 0.08) ***	0.01 (0.001, 0.11) ***
Full time VS Retired (reference)	2.43 (1.04, 5.71) *	1.44 (0.50, 4.10)	0.66 (0.16, 2.72)	2.11 (0.95, 4.69)	2.58 (1.15, 5.79) *	4.24 (1.53, 11.75) **
Current smokers (No smoking as reference)	1.54 (0.80, 2.98)	0.88 (0.39, 1.99)	1.53 (0.64, 3.66)	1.13 (0.62, 2.07)	1.54 (0.83, 2.86)	3.27 (1.43, 7.47) **

	Cluster 1 (n=133)	Cluster 3 (n=60)	Cluster 4 (n=37)	Cluster 5 (n=326)	Cluster 6 (n=237)	Cluster 7 (n=41)
COPD	0.62 (0.31, 1.26)	0.50 (0.20, 1.22)	1.05 (0.42, 2.60)	0.28 (0.14, 0.54) ***	N/A	0.07 (0.01, 0.55) *
Asthma	0.59 (0.27, 1.30)	0.62 (0.24, 1.61)	0.63 (0.20, 1.92)	0.35 (0.17, 0.72) **	0.09 (0.03, 0.25) ***	0.10 (0.01, 0.80) *
Hypertension	1.00 (0.54, 1.84)	1.30 (0.64, 2.68)	1.30 (0.57, 2.96)	2.19 (1.26, 3.79) **	0.29 (0.15, 0.53) ***	0.24 (0.08, 0.69) **
Other heart diseases	0.36 (0.16, 0.84) *	0.56 (0.22, 1.45)	0.32 (0.09, 1.20)	0.52 (0.27, 1.03)	0.30 (0.14, 0.64) **	N/A

Note: Cluster 2 “CXR/Referral (n=65)” as the reference category, other clusters were compared with Cluster 2. Nonsignificant results were coloured in light grey.

Other heart diseases included abbreviations of CVD (cardiovascular disease), IHD (ischaemic heart disease), CHD (coronary heart disease), AF (atrial fibrillation), and MI (myocardial infarction). N/A meant there was no event in that cluster.

*** indicated $P < 0.001$, ** $P < 0.01$, * $P < 0.05$. [Point 2, please find the stars in the table above.](#)

7.2.4.3 Differentiate clusters by significant patient characteristics

Based on the results in Table 7.3, patients in Cluster 2 and 3 had a very similar profile. The current patient characteristics could not differentiate the two clusters. There may be some other significant but **unmeasured characteristics** (unavailable information) between the two patient clusters. At least one significant patient characteristic was able to differentiate patients in the other five clusters from patients in Cluster 2.

- Cluster 1: age, the numbers of potential LC symptom consultations and total visits, full-time employment, other heart diseases (5 variables);
- Cluster 4: the number of total visits (1 variable);
- Cluster 5: age, the number of total visits, no LC symptom consultations, COPD, asthma, hypertension (6 variables);
- Cluster 6: age, the number of total visits and comorbidities, no LC symptom consultations, full-time employment, asthma, hypertension, other heart diseases (8 variables);
- Cluster 7: age, the number of comorbidities, current smoking status, no primary care visit (thus no LC symptom consultations), full-time employment, COPD, asthma, hypertension (8 variables).

7.2.4.4 Multinomial logistic regression model with multiple predictors

Based on the results of single predictors, models with different combinations of significant predictors were tried, but yielded **inconsistent results**. The possible explanations are discussed in subsection 7.3.1.2. **Overfitting** was the main problem, which means fitting a model that has too many parameters. For a categorical outcome with seven clusters, it needs to establish six models.

One possible way to address the problem of overfitting is reducing the number of parameters that need estimating in the multinomial logistic regression model. As the research interest was to understand the differences in GP management for patients presented with potential LC symptoms, it is reasonable to establish the model in these patients as a start. The result is reported in Table 7.4. Only two variables were significant – age and the number of potential LC symptom consultations. In the Model of Cluster 1 vs Cluster 2, the RRRs of both variables were <1. The other two significant variables [Point 2](#) **in the previous uni-variable analysis** (Table 7.3), i.e. the total number of visits and full-time employment, **were not significant**. The RRR of age and the number of potential LC symptom consultations remained **unchanged from uni-variable to multivariable model**. While in the Model of Cluster 4 vs Cluster 2, although age was not significant, the RRR of the number of potential LC symptom consultations slightly increased from 1.28 in the univariable

model to 1.31 in the multivariable model, and a slightly wider 95% CI. Future studies with a larger sample size in each cluster may increase the statistical power and improve the model fitting.

Table 7.4 – Multinomial logistic regression model with multiple variables comparing the four clusters of GP management of patients presented with potential LC symptoms

	Cluster 1. Test/MED (n=133)			Cluster 3 (60)		Cluster 4. MED-L (n=37)		
	β	SE	RRR [95% CI]	β	SE	β	SE	RRR [95% CI]
Constant	4.73***	1.23		2.57	1.38	1.19	1.55	
Age	-0.04*	0.02	0.96 (0.93, 0.99)	-0.03	0.02	-0.04	0.02	0.96 (0.92, 1.01)
No. LC consultation	-0.83***	0.16	0.43 (0.32, 0.59)	-0.13	0.13	0.27*	0.12	1.31 (1.04, 1.64)

Note: Cluster 2 (LC symptoms | CXR/referral) was the reference category; No. LC consultation means the number of potential LC symptom consultations. SE – Standard Error, *** indicated $P < 0.001$, ** $P < 0.01$, * $P < 0.05$.

Another possible approach to improve overfitting is combining similar outcome categories. A sensible way is to combine Clusters 1-4 together, as “patients with potential LC symptoms” and continue to serve as the reference category. Clusters 5-6 as “patients without potential LC symptoms”, and Cluster 7 “patients did not visit the general practice at all”. Table 7.5 presents the results of the multivariable multinomial logistic regression model in this approach. Combining the outcome categories also improved model fitting. Comorbidity was the only significant predictor when comparing Clusters 5-6 with Clusters 1-4. [Point 4](#) The results suggested it was still quite difficult to distinguish patients with potential LC symptoms from other primary care attenders, which could be due to the limited available information in this study. Age, comorbidity, and current smoking status were the three significant predictors between Cluster 7 and Clusters 1-4. [Point 3](#) All interaction terms stated in subsection 5.3.2 were explored but non-significant, and therefore not reported in the table.

Before finalising the model with three groups, two other modelling strategies with four groups were tried, by separating Cluster 1 from Clusters 2-4 (transient VS more LC symptoms), and separating Cluster 5 and 6 (with and without cardiorespiratory comorbidities and/or other alarming symptoms). All the variables were nonsignificant when Cluster 1 compared with Clusters 2-4 (reference category). The conclusion also did not change when splitting Clusters 5 and 6. The results did not improve at the cost of estimating more parameters by adding one more model. In

the spirit of parsimony, the current multinomial logistic regression model with three clusters was considered the best. The two reference points to differentiate the three clusters were whether the patients presented with potential LC symptoms or not and visited general practice or not (help-seeking behaviours).

Table 7.5 – Multivariable multinomial logistic regression comparing the three patient groups

	Clusters 5-6 (n=563)			Cluster 7 (n=41)		
	β	SE	RRR [95% CI]	β	SE	RRR [95% CI]
Constant	1.53**	0.58		3.26	1.84	
Age	-0.005	0.008	1.00 (0.98, 1.01)	-0.07*	0.03	0.93 (0.88, 0.99)
Comorbidities	No comorbidity as reference					
- 1	-0.21	0.24	0.81 (0.51, 1.30)	-2.56***	0.60	0.08 (0.02, 0.25)
- 2	-0.89***	0.25	0.41 (0.25, 0.67)	-2.45***	0.60	0.09 (0.03, 0.28)
- ≥ 3	-1.35***	0.27	0.26 (0.15, 0.44)	-3.31**	1.06	0.04 (0.00, 0.29)
Current smoker	-0.05	0.17	0.95 (0.68, 1.32)	1.12*	0.44	3.06 (1.29, 7.24)

Note: Clusters 1-4 “patients presented with potential LC symptoms” as the reference category, Clusters 5-6 “patients presented in general practice without potential LC symptoms”, and Cluster 7 “patients did not visit the general practice at all”. SE – Standard Error, *** indicated $P < 0.001$, ** $P < 0.01$, * $P < 0.05$.

7.2.5 Results of modelling count data

7.2.5.1 Model of primary care attendance and the interpretation of results

The null hypothesis that the dispersion parameter alpha was equal to zero was rejected ($P < 0.001$). This suggested that the number of primary care consultations was overdispersed and not suitable to use Poisson regression. Employment status, marital status, qualification, and ethnicity were the first batch of nonsignificant variables removed from the model. For the six known influencing factors from the published studies (age, sex, the number of comorbidities, IMD quintile, current smoking status, and the years of smoking), the two smoking variables were nonsignificant. The coefficients (β) and the standard errors (SE) are reported in Model 1 of Table 7.6.

The model was run again with the four significant predictors (age, sex, the number of comorbidities, and IMD quintile). The new coefficients and SE are updated in Model 2 of Table 7.6. Age and the number of comorbidities had a positive association with the number of primary care visits. For males (compared with females) and patients in the second quintile (compared with the bottom quintile – the most deprived), the number of primary care visits decreased. Patients in the upper three quintiles of IMD did not have a significant difference. More comorbidities were associated with higher incidence rate ratios (IRR). Compared with patients without comorbidity, the IRR for patients with one comorbidity to present in general practice was 1.54 times higher (95% CI [1.35, 1.78]), 2.1 times and 2.8 times higher for patients with 2 and ≥ 3 comorbidities, holding all the other variables constant in the model. Interactions between sex, the IMD quintile, and the number of comorbidities were tested, but the results were nonsignificant. Therefore, the interaction terms were not included in the final model and presented in Table 7.6.

Table 7.6 – Results of the negative binomial regression model for the number of primary care attendances (n=829)

	β [Model 1]	SE	β [Model 2]	SE	IRR [95% CI]
Constant	1.08 ***	0.20	1.09 ***	0.19	
Age	0.01 ***	0.003	0.01 ***	0.002	1.01 [1.01, 1.02]
Male sex	-0.12 *	0.05	-0.12 **	0.05	0.88 [0.81, 0.97]
Number of comorbidities	No comorbidity as reference				
- 1	0.43 ***	0.07	0.44 ***	0.07	1.54 [1.35, 1.78]
- 2	0.72 ***	0.08	0.72 ***	0.08	2.06 [1.78, 2.39]
- ≥ 3	1.03 ***	0.08	1.03 ***	0.08	2.81 [2.40, 3.31]
IMD	Most deprived quintile as reference				
- 2	-0.19 *	0.10	-0.19 *	0.10	0.83 [0.68, 0.99]
- 3	-0.13	0.09	-0.13	0.09	0.87 [0.73, 1.04]
- 4	-0.001	0.10	-0.001	0.10	1.00 [0.83, 1.21]
- 5 (Less deprived)	-0.03	0.10	-0.04	0.10	0.96 [0.79, 1.16]
Current smoker [Yes]	-0.007	0.06			

Years of smoking	0.001	0.002		
Alpha	0.33	0.02	0.32	0.02

Note: SE – Standard Error; IRR – Incidence Rate Ratio; the results of all six variables are reported in Model 1; Model 2 presents the results for the four significant variables in Model 1. *** indicated $P < 0.001$, ** $P < 0.01$, * $P < 0.05$.

7.2.5.2 Model of potential LC symptoms consultations and the interpretation of results

The number of potential LC symptom consultations was right-skewed and over-dispersed. Only one-third of patients (295/858, 34.4%) consulted with potential LC symptoms. Due to excessive zero and the over-dispersion of the count data, zero-inflated negative binomial regression model was used to explore the predictors for the number of potential LC symptom consultations.

Patients' age, sex, IMD, current smoking status, and years of smoking were all **non-significant** factors. Two clinical characteristics, comorbidity and having asthma, were significant variables in the full model, but not in the inflate model (model predicting whether it contains zero or not). For comorbidity, patients having **≥ 3 comorbidities** had an IRR of 2.38, 95% CI [1.28, 4.41], $\beta = 0.87$, $P = 0.006$, compared with patients without comorbidity in the full model. Patients who had **asthma** had an increased IRR of consulting potential LC symptoms, IRR=2.11, 95% CI [1.42, 3.13], $\beta = 0.75$, $P < 0.001$, compared with those without asthma. COPD was NOT a significant factor for the number of potential LC symptom consultations, which may be explained by the following reasons. The first explanation may be due to the underdiagnosis of COPD in community patients. A previous study reported that many patients did not get the diagnosis of COPD several months before the diagnosis of LC. The diagnosis of COPD was the result of suspicion of LC, and the diagnosis was made after more intensive clinical investigation and assessment (Powell et al., 2013). The second possibility could be due to GP's recording habit. If the patient had a diagnosis of COPD, GPs may record the consultation as COPD related, rather than symptoms, which could lead to underreporting of symptoms and probably result in a lack of association between potential LC symptoms and COPD. The third explanation could be due to different distributions among patients with COPD and asthma. For patients with ≥ 5 times of potential LC symptom consultations, 5.6% (5/90) patients had COPD, and 10.7% (8/75) had asthma. This may explain why asthma was a significant predictor but not COPD, although the prevalence of COPD was slightly higher than that of asthma.

7.2.6 Practice effect

Table 7.7 presents the difference in patient numbers between practices at each stage, including the numbers of eligible patients, responders, non-responders, response rate (Wagland et al., 2016), and the final number of patients included in this study from each practice. Practice 7 had a particularly low number of eligible patients at the beginning, the lowest response rate and the final number of patients included in this study. There could be several reasons. The first one is that it is a small practice (e.g. a single-handed GP and thus had limited patients). The second reason could be that most of the patients registered in that practice were young (e.g. a high proportion of university students) and thus fewer patients met with the age criterion. The third possibility was something happened in that practice during the implementation of the study (e.g. shortage of staff) and thus affected patient recruitment. But I do not have the information to know the real situation. However, one thing is clear – the response rate was significantly different among practices at the beginning ($\chi^2(7)=19.3$, $P=0.007$).

Table 7.7 – The number of eligible patients, responders, non-responders, response rate, and patients included in this study from each participating practice

GP Practice	Eligible	Non-responders	Responders	Response rate (%)	This study
Practice 1	441	346	95	21.5	85
Practice 2	459	329	130	28.3	88
Practice 4	679	501	178	26.2	145
Practice 5	745	555	190	25.5	144
Practice 6	693	500	193	27.8	149
Practice 7	166	135	31	<u>18.7</u>	24
Practice 8	884	687	197	22.3	152
Practice 9	554	396	158	28.5	112
Total	4,621	3,449	1,172	Average 25.4	899

7.2.6.1 Patient characteristics among practices

Patients' age, sex, smoking years, current smoking status, SES (IMD quintiles), and the number of comorbidities may be the possible explanatory variables for GP's vigilance of potential LC

symptoms. Table 7.8 presents descriptive statistics for some key patient characteristics by practice, and Figure 7.1 presents the distribution of patient's comorbidities and IMD quintile by practice. Patients in practice 4 were the oldest, aged 68.4 ± 9.6 years, significantly older than patients in practice 9 (62.0 ± 8.2 years, the youngest), practices 1 and 2. The mean age of patients in the other four practices was between 65 and 67 years. Even so, patients in practice 7 had the longest smoking history (41.7 ± 14.5 years), 15 more years than patients in practices 5 and 6 (26 years, the shortest smoking history). The percentages of current smokers were significantly different among practices ($\chi^2(7)=134.7$, $P<0.001$), ranging from 14% (practice 6) to 61% (practice 9). The percentage of male patients varied from 46% (practice 9) to 62% (practice 6), but the difference was non-significant. Two indicators of patients' help-seeking behaviours were investigated and reported separately in the next subsection.

Table 7.8 – Descriptive statistics of the key patient characteristics by practices

Practice *	Age	Male (%)	Total visits	LC symptom consultations	Smoking years	Current smoker (%)
1 (n=85)	63.1 ± 9.1	43 (50.6%)	11.6 ± 9.0	2.2 ± 1.8	39.4 ± 12.0	42 (53.2%)
2 (n=88)	63.8 ± 10.4	50 (56.8%)	10.9 ± 8.9	2.4 ± 1.5	35.4 ± 13.4	51 (58.0%)
4 (n=145)	<u>68.4 ± 9.6</u>	85 (58.6%)	9.6 ± 6.7	2.2 ± 2.1	31.9 ± 15.8	26 (18.2%)
5 (n=144)	66.8 ± 9.3	75 (52.1%)	<u>7.9 ± 4.8</u>	2.1 ± 1.3	26.8 ± 16.0	27 (19.0%)
6 (n=149)	65.3 ± 7.2	93 (<u>62.4%</u>)	10.8 ± 8.3	2.4 ± 1.8	<u>26.2 ± 15.0</u>	21 (<u>14.1%</u>)
7 (n=24)	64.9 ± 8.3	14 (58.3%)	<u>12.4 ± 13.7</u>	2.1 ± 1.2	<u>41.7 ± 14.5</u>	11 (45.8%)
8 (n=152)	66.9 ± 10.4	89 (58.6%)	11.9 ± 8.8	2.0 ± 1.3	31.3 ± 15.5	41 (27.0%)
9 (n=112)	<u>62.0 ± 8.2</u>	52 (<u>46.4%</u>)	10.2 ± 7.0	1.8 ± 1.1	38.9 ± 12.5	68 (<u>60.7%</u>)

Note: * there was no practice numbered as 3 in the NAEDI dataset.

Although the distribution of the number of comorbidities was significant among the eight practices ($P=0.008$), the differences were not as easily detected as those in IMD. The distribution of IMD was quite different among the eight practices. The majority of patients in practice 4 were in Q3 and patients in practice 5 in Q2. Patients in the other six practices were from more diverse socioeconomic backgrounds. The IMD was spread across all five quintiles. Patients in practices 1, 2, 6, and 9 had an increasing proportion from the most to the least deprived quintile. Patients in

practice 7 were mostly in Q1 and Q4 but had the highest proportion in Q4 (54%, more affluent). About 30% (46/152) of patients in practice 8 were in the most deprived quintile, which was the highest percentage among all the practices. About 80% of patients (n=121) were in Q1-Q3 (more deprived quintiles), compared with 82% patients in practice 6 (122/149) were in the two least deprived quintiles (Q4-Q5). Patients in practice 6 (higher SES) had the lowest proportion of current smokers.

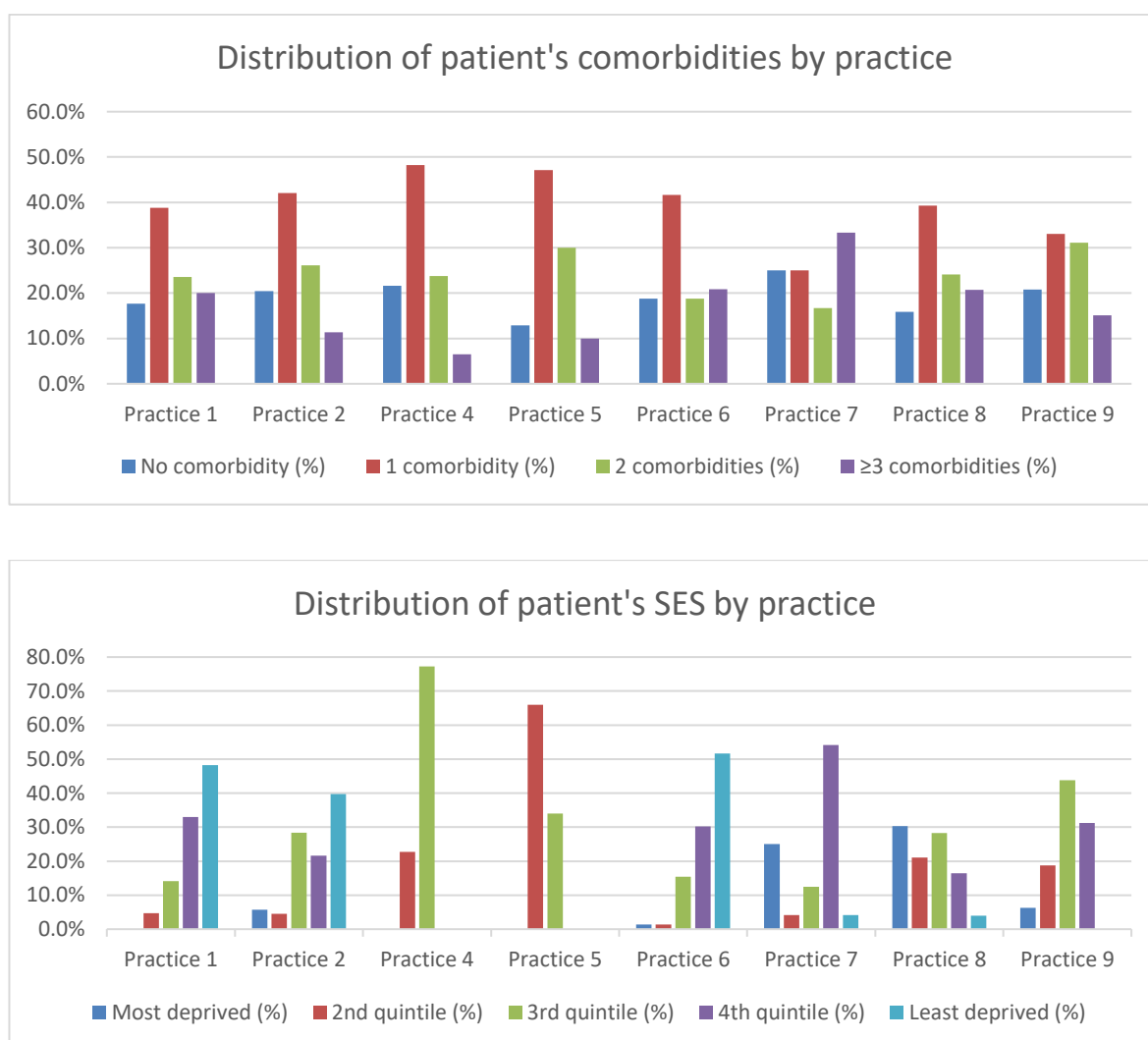


Figure 7.1 – Distribution of patient's comorbidities and IMD quintile by practices

7.2.6.2 Practice effect in patient's help-seeking behaviours

The numbers of total visits and potential LC symptom consultations were reported in Chapter 6, as part of the sequence characteristics. Table 7.8 presents both indicators by practice. Although the data were right-skewed, mean and standard deviation were used to describe the central tendency and dispersion, rather than median and IQR, as the practices almost had the same median and IQR and thus not able to distinguish from each other.

Following the findings in subsection 7.2.5.1, negative binomial regression was used again to explore whether there was a practice effect on the number of primary care visits, accounting for patient's other characteristics. When introducing GP practice in the model, IMD became a nonsignificant factor, probably because IMD is postcode based, and the GP practices are likely to be located in the same geographical area. The results of age, sex, and comorbidities remained stable. After adjusting for these confounding factors and taking Practice 1 as a reference, patients in **practice 5** had a significantly lower IRR in the number of primary care visits (IRR=0.69, 95% CI 0.58-0.83), $P<0.001$ (observed **practice effect** of lower patient attendance rate). Patients in the other six practices had a nonsignificant difference, compared with patients in practice 1. The number of primary care visits among patients in practice 5 was 7.9 ± 4.8 on average, while the mean number of GP visits among patients in the other seven practices was ranged from 9.6 (practice 4) to 12.4 (practice 7).

If only based on the descriptive and inferential statistics, we cannot find any significant difference in the number of potential LC symptom consultations among practices (Kruskal-Wallis rank test, $P=0.684$). However, a practice effect was observed in the inflated part of the zero-inflated negative binomial regression (predicting whether it contains zero or not, following subsection 7.2.5.2). Only 13.2% of patients (19/144) in **Practice 5** had potential LC symptoms recorded, significantly lower than the reference practice (Practice 1, 31.8%, second lowest, $\beta=1.55$, $P=0.028$). 42.8% of patients (65/152) in Practice 8 had LC symptoms recorded, which was the highest (further discussed in the next subsection). The proportions in the other six practices were around 32%-39%.

7.2.6.3 Cluster membership among eight practices

Table 7.9 presents the distribution of cluster membership among the eight general practices, which is significantly different ($P<0.001$). The practice with the highest percentage in each cluster is underscored. Practice 8 had high percentages in Clusters 1-3. If considering Clusters 1-4 together, GPs in practice 8 seemed to have higher vigilance of recognising potential LC symptoms and were most likely to investigate potential LC symptoms. Among 295 patients who presented with potential LC symptoms, 22% were from practice 8 ($n=65$). For those who got CXR or were referred to a specialist (Cluster 2), about 30% of patients (19/65) were from practice 8. Such differences could not be explained by GP's recording habits, because it was established in Section 6.6 that GP ordered CXR or referred the patients to a chest physician due to non-LC symptom presentations was very rare ($n=10$, 1.3%). Interestingly, patients in Practice 8 had the second lowest frequency in the number of potential LC symptom consultations (2.0 ± 1.3 times on average). The comorbidity burden among patients in Practice 8 was close to the average level.

Therefore, GPs in practice 8 were more vigilant, even with a relatively low frequency of patients consulting with potential LC symptoms, probably could be explained by a **larger proportion of patients from lower SES background**, but not older age, or a higher percentage of current smokers, nor a higher comorbidity burden.

GPs in practice 4 were also vigilant with potential LC symptoms, recognising 56 patients with potential LC symptoms (19.0%, the second highest after practice 8), probably because patients in this practice were the oldest. Patients in practice 5 had the highest percentages of Clusters 5-6 (visited GP, with and without cardiorespiratory and/or other alarming symptoms). The implications of practice effect were discussed in subsection 7.3.4.3.

Table 7.9 – Distribution of cluster membership among patients from the eight general practices, n (column %)

Practice*	Cluster 1 (n=133)	Cluster 2 (n=65)	Cluster 3 (n=60)	Cluster 4 (n=37)	Cluster 5 (n=326)	Cluster 6 (n=237)	Cluster 7 (n=41)
1 (n=85)	16 (12.0)	2 (3.1)	6 (10.0)	3 (8.1)	32 (9.8)	24 (10.1)	2 (4.9)
2 (n=88)	9 (6.8)	9 (13.8)	7 (11.7)	4 (10.8)	29 (8.9)	26 (11.0)	4 (9.8)
4 (n=145)	25 (18.8)	12 (18.5)	11 (18.3)	8 (<u>21.6</u>)	38 (11.7)	40 (16.9)	11 (<u>26.8</u>)
5 (n=144)	7 (5.3)	4 (6.2)	4 (6.7)	4 (10.8)	75 (<u>23.0</u>)	46 (<u>19.4</u>)	4 (9.8)
6 (n=149)	25 (18.8)	10 (15.4)	13 (21.7)	3 (8.1)	57 (17.5)	33 (13.9)	8 (19.5)
7 (n=24)	2 (1.5)	2 (3.1)	2 (3.3)	3 (8.1)	8 (2.5)	5 (2.1)	2 (4.9)
8 (n=152)	26 (<u>19.6</u>)	19 (<u>29.2</u>)	14 (<u>23.3</u>)	6 (16.2)	51 (15.6)	35 (14.8)	1 (2.4)
9 (n=112)	23 (17.3)	7 (10.8)	3 (5.0)	6 (16.2)	36 (11.0)	28 (11.8)	9 (21.9)

Note: The highest % in each cluster is underscored.

7.3 Discussion

7.3.1 Statistical methods in the empirical analysis

7.3.1.1 Summary of the whole analysis process

After methodological exploration, the empirical analysis aims to understand how patient characteristics could help explain the differences in the cluster patterns. Descriptive statistics provided the characteristics of patient profile in each cluster, with traditional parametric and non-parametric tests to investigate whether each variable was significantly different among the seven clusters or not. Multinomial logistic regression further quantified the relative risk ratio (RRR). The association between patient characteristics and help-seeking behaviours (the numbers of primary care visits and potential LC symptom consultations) were investigated. Finally, practice effect on patient's response rate, help-seeking behaviours, and cluster patterns of primary care sequences was explored.

7.3.1.2 Different statistical techniques provide different angles to find useful information from the data

Two main models, multinomial logistic regression and (zero-inflated) negative binomial regression, were used for two types of outcomes—categorical clusters and count data. They are in the family of generalised linear models, but with different link functions for different types of outcomes. It was difficult to build a model with multiple predictors in the zero-inflated negative binomial regression model, due to overdispersion of the count data in a small sample size and overfitting.

Before building the multinomial logistic regression model, parametric and non-parametric statistical tests were used to find out whether each patient characteristic was significantly different among the seven clusters or not. In the univariable model, significant differences were quantified by RRR and 95% CI. Inconsistent results occurred when attempting to build a multivariable model with seven clusters as a categorical outcome. There were four possible explanations for the failed attempt. The first one was the patient characteristics to differentiate the seven clusters were different. It was not even possible to differentiate patients in Clusters 2 and 3. Therefore, it was unrealistic to expect the same variables could effectively predict the cluster membership. The second explanation was in the variables. For patients in Clusters 5-6, the number of potential LC symptom consultations was 0; for patients in Cluster 7, they did not visit the general practice, the values for these two variables were 0. Introducing these variables made the model not able to converge. Collinearity between variables was the third explanation. As

reported in subsections 7.2.2, some variables were correlated. The number of total visits and the number of potential LC symptom consultations are the intermediate outcomes, also associated with some patient's characteristics. The fourth explanation was that the number of clusters might be too large for a multinomial logistic regression model with multiple predictors. It was possible to use univariable to explore the association and calculate the unadjusted RRR between clusters, but more difficult to establish a multivariable model. There were also some categorical variables as independent variables. Too many parameters that need to be estimated could cause convergence failure in the multinomial logistic regression model (Steyerberg, 2019)(Page 96). A relatively small sample size for a large categorical outcome, and a big difference of sample size in each cluster, these two factors could also be the reasons for overfitting. When the number of clusters reduced from seven to three or four, the model fit improved. But if further reduced to two clusters, i.e. whether the patients presented with potential LC symptoms or not and using logistic regression, the simple dichotomisation would make no difference from the results of the inflated part of the zero-inflated negative binomial regression model.

The QCancer studies using EHRs with a **very large sample size** (1.26 million males and 1.24 million females) **can use multivariable multinomial logistic regression** to identify potential symptoms and risk factors for **nine types of cancer** for men (Hippisley-Cox and Coupland, 2013a) and **eleven types of cancer** for women (Hippisley-Cox and Coupland, 2013b), comparing with patients without cancer. The findings from these two papers have been validated and transformed into clinical decision support tools, approved by the NHS and integrated into the EMIS primary care EHR system for GP to use. This example demonstrates **the importance of sample size for multinomial logistic regression models with a large number of categories**. If the sample size of sequences is large and within the limit of storage for a distance matrix in R (about 46,340 sequences, section 3.5), it may be possible to perform a multivariable multinomial logistic regression model with several clusters.

7.3.2 Summary and discussion of the findings from empirical analysis

Patients' sociodemographic and clinical characteristics could explain part of the variation of the primary care sequences and GP management of potential LC symptoms. Ten patient characteristics were useful to characterise and differentiate patients in the seven clusters, including age, the numbers of primary care visits, potential LC symptom consultations, and comorbidities, COPD, asthma, hypertension, and other heart diseases (4 specific comorbidities), current smoking status, full-time employment. Sex, years of smoking, marital status, SES (IMD quintiles), education (highest qualification) were nonsignificant factors. Therefore, the

unknown/unreported status in ethnicity, marital status, and qualification had minimal impact, as these factors were not in the regression model.

Three clusters of patients with potential LC symptoms had **similar** patient characteristics but received different GP actions – GP ordered CXR or made a referral (Clusters 2), offered advice (Clusters 3), and gave repeated prescriptions (Clusters 4). No significant difference in patient characteristics was found between Clusters 2 and 3, although patients in Cluster 2 were older. The only significant difference between patients in Cluster 2 and 4 was the number of comorbidities, but not any of the four specific cardiorespiratory comorbidities. From another perspective, this result means that **primary care sequences provide new information to categorise patients by SA, even in a relatively small sample size, while traditional statistical methods could not differentiate** the three clusters based on patient characteristics. This is **the added value and the benefit of using SA to analyse holistic information of primary care sequences**, which could be useful in early diagnosis research.

Patients in Cluster 1 just presented with **transient** LC symptoms, and GP ordered tests to investigate the symptom or prescribed medications, which was likely to be benign presentations. They had a similar age as patients in Clusters 3 and 4 but were younger than patients in Cluster 2. They had a higher percentage of full-time employment, and a smaller number of total visits.

Patients in Cluster 5 did not present with potential LC symptoms, had a lower prevalence of COPD and asthma, but had the highest prevalence of hypertension among the seven clusters, which explained why the cluster pattern was “cardiorespiratory diseases and/or other alarm symptoms”. Patients in Cluster 6 were younger, had a higher percentage of full-time employment, and had a less comorbidity burden. None of the patients in this cluster was diagnosed with COPD. The prevalence of cardiorespiratory conditions was also low in this group, which may explain why they had fewer attendances to the general practice, and only had minor care needs. Patients not visiting their GP at all (Cluster 7) were the youngest, had the highest percentages of full-time employment and current smokers, but had fewer comorbidities burden, which would be due to incomplete records, as they did not use primary care services.

Age and comorbidities were the two significant patient characteristics, consistently found in different models. Age had a significant but weak association with the number of comorbidities. Older age and a larger number of comorbidities increased the number of visits to general practice, while males and patients in the second quintile of SES (compared with the most deprived quintile) had a smaller number of visits to general practice. Patients with ≥ 3 comorbidities (compared with

no comorbidity) and asthma had a higher number of LC symptom consultations, but not COPD. The possible reasons were explained and discussed in subsection 7.2.5.2.

Practice effect was explored and observed with patients' participation in the questionnaire survey, help-seeking behaviours, and the cluster patterns of primary care sequences. GPs in practice 8 seemed to be more vigilant of potential LC symptoms, followed by GPs in practice 4. A higher proportion of patients presented with potential LC symptoms in these practices were investigated, referred, and managed (prescriptions and advice). Patients in practice 5 had a lower attendance rate in that practice, and fewer events of potential LC symptoms, but higher rates of presentations related to cardiorespiratory diseases and minor care needs.

7.3.3 Boarder connection with other literature

Research evidence about patient's help-seeking behaviours and possible influencing factors were summarised in subsections 2.5.5 (literature review). The findings in this study are generally consistent with those reported in other studies (Moffat et al., 2015, Whitaker et al., 2015) in this field. Patients not working (because of illness), married/cohabiting (compared with single), and older people were more likely to seek help in primary care for alarm symptoms in the British populations (Whitaker et al., 2016, Hannaford et al., 2020). As to education, smoking status and duration, patient's SES and household income, the results were less consistent, depending on the study sample and how the variables were operationalised in studies (binary, categorical, and the cutoff values).

7.3.3.1 Symptom recognition and the help-seeking behaviours

Whether patients can recognise the symptoms they experience are alarm LC symptoms or not, is a prerequisite of making the decision to visit their GP. An Australian study found no significant difference in symptom recognition between current and former smokers (Crane et al., 2016). In the community settings, current smokers were more likely to experience cough, breathlessness, and tiredness (Walabyeki et al., 2017). A Danish web-based survey reported that 39.6% of patients (3,080/7,870) with at least one respiratory alarm symptom contacted their GPs, but the percentage of consultation for specific respiratory alarm symptoms varied, from 27.4 % of prolonged hoarseness to 49.7 % of shortness of breath (Sele et al., 2016). This study observed a **34.4%** (295/858) consultation rate of potential LC symptoms from GP notes. The published article of the NAEDI study (Wagland et al., 2016) reported that 53.7% (629/1,172)⁵ of total participants

⁵ Not all the participants completing the questionnaire gave consent to researchers to review their health records.

reported ≥ 1 symptom in the IPCARD questionnaire, and 35.1% (411/1,172) reported ≥ 2 symptoms. There was a **difference in the proportions between patient-reported symptom experience (53.4%) and help-seeking for potential LC symptoms (34.4%, actual behaviour)**.

There are several possible explanations for the observed difference. Alarm symptoms indicative of LC in clinical guidelines are not necessarily considered as alarming by the general population, who may instead normalise such symptoms as part of the fluctuations in daily life (Smith et al., 2009, Brindle et al., 2012), attributed to the long-term smoking habit or the other diagnosed cardiorespiratory comorbidities (McCutchan et al., 2019). Non-consulters may also have a higher tolerance for symptoms, and are more likely to manage symptoms by themselves (Wagland et al., 2016). Some patients may not realise the symptoms they experienced were 'alarming' until they filled in the IPCARD questionnaire, which triggered their help-seeking behaviours. This may explain the observed increase in primary care consultations after the questionnaire survey (Wagland et al., 2016). The 'symptom iceberg' theory (Last, 1963, Hannay, 1980) discussed in subsection 2.5.5 may also explain the observed difference. Besides doing nothing and consulting with GP, patients may apply some lay care strategies to manage the symptoms by themselves, which included taking OTC medications, seeking advice from pharmacists, complementary therapists, NHS 111, but this assumption could not be ascertained from the study data.

7.3.3.2 Patient's characteristics and the patterns of primary care consultations

Women and older age were significantly associated with a higher frequency of GP contact. Generally, women are more aware of their health and the uncommon bodily changes. When noticing new symptoms, the emotional response (the level of worry and anxiety) is different between men and women (Briscoe, 1987). Furthermore, the degree that the symptoms interfere with daily activities may also differ in sex, all of which may explain the difference in help-seeking behaviours between sexes. Despite having a higher risk of developing LC, smokers were less likely to seek help for common respiratory symptoms of LC than non-smokers in community settings (Chatwin and Sanders, 2013, Friedemann Smith et al., 2016, Walabyeki et al., 2017). Long-term smokers were more likely to delay presentation in primary care with symptoms (Smith et al., 2009). This study sample had a relatively long smoking history, 32.2 ± 15.4 years on average. The proportion of current smokers was much higher in patients in Cluster 7 (no visit at all) than the other six patient clusters ($P=0.017$), suggesting that the current smoking behaviour is a barrier for patients to see their GPs. Sele et al. (2016) reported that current smoking status and alcohol consumption were significantly associated with lower odds of contacting GP in both sexes among the Danish patients. People choosing healthy lifestyles may be more likely to take actions when they experience symptoms, while individuals with excessive tobacco and/or alcohol intake may be

more willing to take a risk, or not realise they have an increased risk of disease (Weinstein et al., 2005) than those choose a healthy lifestyle.

7.3.4 Clinical relevance and implications

7.3.4.1 Potential opportunity to encourage patients at high risk to seek help more promptly in primary care

Although the study sample had a higher risk of developing LC, a promising note was that about 95% (858/899) of patients had at least one attendance in general practice for various reasons. This could be an opportunity for primary care professionals to provide health education or accessible information (like leaflets, brochures) for these patients. LC awareness campaigns could target heavy and long-term smokers, encouraging them to initiate contact with GP if they experience any alarm symptoms. Based on the current findings from this study, more efforts should be put to raise the awareness among patients in Clusters 6-7, as patients in the other five clusters have already sought help for potential LC symptoms and/or cardiorespiratory diseases. The Australian CHEST trial (Emery et al., 2019) observed a significant 40% relative increase in consultations about respiratory symptoms among patients at increased risk of developing LC in the intervention arm, who received a self-help manual to improve knowledge of respiratory symptoms, followed by patient preferred reminders (SMS/email reminders, postcards, phone calls, or fridge magnets) to encourage help-seeking when experiencing symptoms, compared with those in the control group having a brief discussion about lung health with a trained researcher. It would be helpful to conduct studies to explore effective and friendly ways to get specific and tailored health messages to the populations at high risk in social campaigns. Theory-based interventions incorporating behavioural change techniques have the potential to prompt earlier consultation in symptomatic patients at high risk.

7.3.4.2 New campaign to improve smoker's lung health

The Targeted Lung Health Check is a new service provided by NHS England. As mentioned in subsection 5.4.2, people aged 55-75 years old that have ever smoked and registered with a GP (patients at high risk) are eligible for a free lung health check in some parts of England. Southampton is one of the pilot sites. This campaign aims to early detect cardiorespiratory problems and help diagnose LC at an earlier stage when treatments are likely to be more successful. Neutral language was used for the branding of the service, as "Lung Health Check", rather than "Lung Cancer Check" to avoid negativity and to increase the uptake of this service.

A community-based lung health check programme in Manchester identified a high prevalence (37.4%, 944/2525) of airflow obstruction (Balata et al., 2020), where half (49.7%, 469/944) of the patients with airflow obstruction had no previous diagnosis of COPD. This means that airway obstruction is common in smokers in the 55-75 age group and confirms that COPD is underdiagnosed in English community-based patients (also discussed in subsection 7.2.5.2). 53.3% (250/469) of those without a prior diagnosis of COPD were symptomatic, which meant patients did not seek help even with symptoms. The NAEDI study also had the same finding (subsection 7.3.3.1). The authors found that male sex, younger age, lower smoking duration, fewer cigarettes per day (three continuous variables) were associated with the detection of airflow obstruction without a prior COPD diagnosis in multivariable analysis. These findings indicated the importance of health education in patients, to raise their awareness of symptoms and to encourage them to see their GP timely when experiencing symptoms. Early detection of COPD is important for LC surveillance (Sekine et al., 2012). If the lung health check is proved cost-effective and rolled out in the whole country in the near future, it is promising to early detect COPD in patients at high risk and monitor the disease progression in primary care.

7.3.4.3 The practice effect and clinical implications

Based on the preliminary exploration, practice effect was observed in this study. When reviewing the transcribed GP notes, cervical smear test was recorded in practice 2 only, but not in other practices. Cervical smear test is part of the NHS services and offered to all women and people with a cervix aged 25 to 64. This example reflects the difference in GP's recording habits among different practices. Similarly, the number of potential LC symptom consultations could be influenced by GP's recording habits. However, it was very rare that GPs referred patients to CXR and/or specialists due to non-LC symptom presentations. GPs in practice 8 seemed to be more vigilant in patients presenting with potential LC symptoms. The diverse patient makeup and a large proportion of patients in that practice from lower socioeconomic backgrounds may explain GP's vigilance. Patients in practice 7 had a lower participation rate in this study. Patients in practice 5 had a lower frequency of visiting their GP, being recorded and investigated of potential LC symptoms. This indicates that there may be a structural problem in some practices, but we cannot ascertain this assumption with the current study data. Fear of being judged or blamed for their smoking and/or drinking behaviours by HCPs was reported as a barrier to prompt help-seeking (Scott et al., 2015, McCutchan et al., 2019). A positive attitude from the HCPs during the consultation and creating a supportive environment is vital to building trust with the patients.

We can observe the difference in management strategies by GP for patients with potential LC symptoms. However, it is not possible to know the diagnostic reasoning based on the GP actions

from the health records. It may be possible to conduct interviews to understand GP's diagnostic reasoning and ask GP to comment on the differences in the cluster patterns of primary care sequences. However, this is out of the scope of the current study. Even if asking GP to comment back why they managed their patients in a particular way, it may be difficult for them to remember why they made that decision at that time.

Distinguishing symptoms due to benign respiratory conditions from LC is difficult. GP should consider the patient's background health conditions, medical and family history, and use decision support tools, like Qcancer (lung)(Hippisley-Cox and Coupland, 2011) embedded in the NHS primary care computer system, to estimate individual patient's risk. For those identified as "high-risk" and qualified for the two-week wait referral, GP should refer them for further investigation without undue delay.

7.3.5 Strengths and limitations of the empirical analysis

Data on some known factors that could influence patients' help-seeking behaviour in primary care were collected in the NAEDI study, and the original information was used as much as possible in the analysis. Different statistical techniques were employed to analyse the data to understand patients' help-seeking behaviour and the primary care sequences from different angles. The results were fully reported, compared, and discussed. All of these are the strengths of the empirical analyses.

There are several limitations of the empirical analyses. Overfitting was a problem in the multivariable multinomial logistic regression model, probably due to a small sample size for a large number of outcome categories. There is currently no algorithm available to conduct post hoc power calculation for multinomial logistic regression in mainstream statistical software. The small sample size also limited a full exploration of the practice effect. Practice 7 only had 22 patients. Future studies with a larger sample size per practice could increase the statistical power and be in a better position to further explore the heterogeneity of patients' help-seeking behaviours and the cluster patterns of primary care sequences, and the practice effect.

Another limitation is the scope of data collection. Only data at the patient level were available. It is possible to get more insight of the practice effect, if practice level data and GP characteristics were available, like the size of practice (the number of registered patients), the number of GP in each practice, the ratio of patients per GP, the numbers of full-time GP/salaried GP/locum GP/female GP/home-trained GP (UK qualified), GP age, the location of the practices (urban/rural area), whether the practice is a training practice or not, and the QOF points. These indicators

were used in a study of primary care health services research for early cancer diagnosis (Maclean et al., 2015). Getting some of these indicators would be very helpful to analyse how GP and practice factors could explain the practice effect.

The empirical findings may also relate to the limitation of SA. The algorithm might classify some sequences incorrectly. This is another explanation why it was difficult to use the patient characteristics to explain the variation of the cluster patterns. However, we cannot know the true cluster membership for each patient, and there is no gold standard to verify this. Cluster analysis belongs to unsupervised learning, which allows the method to discover the patterns of unlabelled data. This topic is further discussed in subsection 8.7.2.

7.4 Conclusion of the empirical analysis phase

Age and comorbidities were the two significant patient characteristics in different models. It was possible to distinguish patients with transient LC symptoms (Cluster 1, probably benign presentation) from the other three clusters of patients with potential LC symptoms, with four patient characteristics (age, the numbers of potential LC symptom consultations and total visits, and full-time employment). However, the available patient characteristics were still unable to explain the variation of GP management among Clusters 2-4. Practice effect on the attendance rate and patient management was explored and observed. Patients at high risk of developing LC but not yet diagnosed with LC may be at different levels of risk, particularly among patients who did not visit general practices at all. This study provides some preliminary empirical findings, which needs to be corroborated by future studies with a well-designed study with a bigger sample size.

Chapter 8 General discussion and conclusion

8.1 The central argument of this thesis

The motivation to conduct this PhD study is to address a fundamental health problem in England – late diagnosis and poor survival of LC. In this thesis, I argue that **studying primary care sequences (pathways) is an important research direction and should be integrated in the field of early diagnosis research, especially for the ‘harder to suspect’ cancer like LC.** Through extensive literature reviews, it is established that primary care sequence is still uncharted territory, probably due to the lack of proper statistical methods to cope with the complexity of interdependent patient-GP events (help-seeking and clinical management) over time. SA is the statistical method proposed to study primary care sequences. Through the whole study, I try to demonstrate the value of using SA to classify different cluster patterns of care sequences. Additionally, I discuss how SA could advance knowledge, promote earlier diagnosis, and improve LC survival in this thesis.

8.2 Key messages and findings from this PhD study

8.2.1 The research gap this study addresses

The systematic scoping review concluded that SA has not been used to study complex primary care sequences, nor in the field of early diagnosis research, which is the research gap this PhD thesis addresses. In addition, this study focuses on a less investigated but the target population for early cancer diagnosis – patients at high risk.

8.2.2 Characteristics of patients at high-risk and their help-seeking behaviours

This study sample had a relatively long smoking history (32.2 ± 15.4 years on average, patient-reported outcome), compared with their age (65.6 ± 9.4 years). About 32% of patients (286/899) were still active smokers. Their comorbidity burden was high. About 80% of patients had at least one comorbidity. Over 95% of patients (858/899) visited general practices at least once during the observation period, median 9 visits in 29 months, ranging from 1 to 54. About one-third of patients (295/899) presented with potential LC symptoms in primary care. Most patients just consulted once (47.8%, 141/295) or twice (24.4%, 72/295), but ten patients (3.4%) consulted ≥ 6 times. For patients having ≥ 2 visits of potential LC symptoms, the median interval between

presentations was 61.5 days, IQR [16, 217] days. These descriptive statistics provide new knowledge on high-risk patients' help-seeking behaviours in primary care.

8.2.3 Cluster patterns of primary care sequences in the NAEDI study

The added value of SA is the cluster patterns containing clinical features summarised the longitudinal information from individual primary care sequences. This study sample was categorised into seven clusters, and these cluster patterns are also new knowledge from this study.

- 1) GP ordered tests or prescribed medications for patients with transient LC symptoms (probably due to benign presentations, n=133/899, 14.8%);
- 2) Patients presented with potential LC symptoms, GP ordered CXR or referred them to the specialists (n=65, 7.2%);
- 3) GP offered health advice for patients presented with potential LC symptoms (n=60, 6.7%);
- 4) Patients presented with potential LC symptoms multiple times, and received repeated prescriptions from GPs (n=37, 4.1%);
- 5) Patients without potential LC symptoms, but had consultations related to cardiorespiratory comorbidities and/or other alarming symptoms indicating severe health problems (n=326, 36.3%);
- 6) Patients only had minor care needs (without potential LC symptoms, nor cardiorespiratory comorbidities or severe health problems, n=237, 26.4%);
- 7) Patients did not visit GP at all (n=41, 4.6%).

8.2.4 The association between patient characteristics and cluster patterns

Ten patient characteristics were useful to differentiate the seven clusters, including age, the numbers of visits to general practices and potential LC symptom consultations, comorbidities, current smoking status, full-time employment, COPD, asthma, hypertension, and other heart diseases. Sex, the years of smoking, marital status, SES (IMD quintiles), education (highest qualification) were nonsignificant factors. In the four clusters of patients with potential LC symptoms, patients in Cluster 2 were the oldest. There was no significant difference in patient characteristics between Cluster 2 (CXR/referral) and Cluster 3 (GP advice). Patients with transient LC symptoms (Cluster 1) were younger and had a higher percentage in full-time employment. Patients receiving multiple prescriptions (Cluster 4) had a higher comorbidity burden, higher prevalence of COPD and hypertension.

Although all patients had an increased risk of LC, **heterogeneity still existed between patients in different subgroups. Patients may be at different levels of risk of developing LC.** Unfortunately, it was not possible to quantify the risk, nor verify this hypothesis based on the available study data. This could be a research direction for future work. Overfitting in the regression model was a challenge, mainly due to the small sample size for a categorical outcome, which may be underpowered to detect a potentially significant difference.

8.2.5 The gap between patients experiencing symptoms and help-seeking behaviours

There was a gap in the proportions between patient-reported symptom experience in the postal questionnaire survey (53.4%) and the actual help-seeking behaviour for potential LC symptoms from GP notes (34.4%) in the same study sample. Some patients may not realise that the symptoms they experienced were 'alarm' symptoms until they filled in the IPCARD questionnaire. Some may apply to lay care strategies to self-manage the symptoms. Some may not do anything at all for their symptoms. However, there was no information available to confirm these hypotheses. But a promising note was that about 95% of patients had some contact with general practice for various reasons. This could be an opportunity for primary care professionals to provide health education or accessible information for patients at high risk. Social campaigns for LC symptom awareness could target at heavy and long-term smokers and the most deprived groups, to encourage them to contact GP if they experience any usual symptoms.

8.2.6 Practice effect

The practice effect was explored and observed. Practices had their own ways to manage their patients. GPs in practice 8 were probably more vigilant of patients presenting with potential LC symptoms and more proactively managed their patients than GPs in other practices, probably because patients in that practice were from lower and more complex socioeconomic backgrounds. Patients in practice 5 had a lower attendance rate and a lower proportion of potential LC symptoms being recorded by their GPs than patients in other practices. Patients in practice 7 had a substantial lower participation rate of the study. All these indicate there may be structural problems in some practices. It would be interesting to explore the practice effect of primary care sequences in future studies, which has the implications of the health services audit and quality improvement in primary care.

8.3 Reflection and further discussion

8.3.1 The face validity of the study findings

This study is the first of using SA to study primary care sequences in patients with a high risk of developing LC. There are no findings from other studies for this study to compare with. Instead, this study provides some initial findings. The thesis title points out this study is “exploratory”. The purpose of this PhD study/thesis is more to tackle the essential issues, establish the analytical procedures, gather the initial findings, generate hypotheses based on observations and results, and pave the way for future studies. Therefore, I would like to more focus on the face validity of the findings, rather than overstating result generalisability.

Face validity originally means the extent to which a test is subjectively viewed as covering the concept it purports to measure in psychometrics. A test can be considered having face validity if it “looks like” it is going to measure what it is supposed to measure (Holden, 2010). In this study, face validity means the extent that SA and cluster analysis can classify sequences and create meaningful typology. In this sense, the face validity is good, as the cluster patterns of LC symptom sequences make sense in the empirical context. In addition, patient characteristics can explain part of the cluster patterns (discussed in subsection 7.3.2). The difficulty in explaining the three clusters of patients with potential LC symptoms but managed differently by GPs is probably due to the small sample size and other unavailable information.

Generalisability means the results from a sample can be extended to the population from which the sample is drawn, also known as external validity (Murad et al., 2018). This concept relates to sampling theory and can be evaluated by examining the size, characteristics, and representativeness of the study sample. Sample representativeness was previously discussed in subsection 5.4.2. The limitations of the sample selection method and sample size were acknowledged and discussed in the data quality sections. Sample representativeness is a necessary but not sufficient condition to generalise study findings. Another concern of generalisability is data timeliness. *Representativeness is time- and place-specific and will therefore always be a historical concept* (Nohr and Olsen, 2013). The study period is from 2010 to 2012, which is already ten years from now. It is fine to use historical data for methodological exploration. But there may be some changes in coding, clinical practice, guidelines, and documenting requirements in EHRs by the health authorities in the last decade. If we want to generalise the findings, it is better to use contemporary EHRs data (further discussed in subsection 8.7.1) and representative study samples for SA in future studies.

8.3.2 Reflection on the strengths and limitations of the whole study

The discussion of the strengths and limitations for individual aspects of this study are placed in each chapter as the thesis develops, including the systematic scoping review (subsection 4.4.7), the study sample (subsection 5.4.3), the primary care data (subsection 5.4.4), the methodological exploration (subsection 6.7.7), and the empirical analysis (subsection 7.3.5). By reflection, I summarise the strengths and limitations of the whole thesis in this subsection, rather than repeating the previous contents here. This study was carried out in a systematic approach by conducting thorough reviews to identify the research gaps, tailoring the methodological exploration for a unique study sample and the specific research context, using a wide range of statistical techniques to answer all the RQs, and integrating the methodological and empirical works. These are the strengths of the whole thesis. Such strengths enable me to make original contributions to the early diagnosis research field (further discussed in section 8.4).

There are three main limitations in this study. The first one is data. The sample size is small, especially for patients presented with potential LC symptoms. Some important patient and practice level variables are not available. The small sample size and the data availability restricts the scope of exploration. The second limitation is the study sample. The original study design limits the sample size and the sample selection. The study sample was representative of being high risk, but they were from a handful of participating practices. Therefore, the study sample was less likely to be representative of the whole high-risk population in England. The third limitation is methodology. The algorithm might misclassify some sequences, and there is no gold standard to verify the cluster membership for an unsupervised learning method. A large number of clusters make running a multivariable multinomial logistic regression difficult. I also make some recommendations for future research (in section 8.6), based on my reflection on the limitations of this study and asking myself “how can I design better studies if I have better resources?”

8.3.3 Rethinking the conceptual models of the cancer care pathway

An important question has emerged during the exploration of the LC symptom sequences – is it possible to know when the patient presents with ‘the first symptom’ of LC and identify ‘the first clinical presentation’ from health records? The two popular conceptual models of cancer care pathways were presented in Figure 1.1 and Figure 2.5, respectively. Many studies in this field report different intervals based on these two conceptual models. The second model (Walter et al., 2012) recognises that ‘patient appraisal’, ‘help seeking’, ‘diagnostic’, and ‘pre-treatment’ is an iterative process (marked as circles in the model), rather than a linear and straightforward process in the first model (Olesen et al., 2009). This study sample was not diagnosed with LC, but one-

third of them presented with potential LC symptoms. Some patients even had several presentations. It is more likely for the patients to remember when ‘the first symptom’ appears for some unusual or alarm symptoms, like breast lump or testicular lump through self-examination, or coughing up with blood (haemoptysis), blood in stool (rectal bleeding), blood in urine (haematuria) by self-observation, and seek help from an HCP. Therefore, it is more likely to get accurate information about ‘the first presentation’ of more specific symptoms for cancer from EHRs. But for cancer without specific symptoms, like LC, it may be difficult to identify the ‘first symptom’. The ‘first presentation’ could be just the ‘first observed’ presentation in the study period defined by researchers. Patients may present similar symptoms before or after the study period. These two conceptual models are helpful to simplify the complex care processes and pathways. However, when applying them to study specific cancer type, researchers need careful consideration and make adaptations.

8.4 Original contributions of this PhD study

This is the first study investigating primary care sequences among a group of high-risk patients for early diagnosis research. Most published studies use patients with a confirmed diagnosis of LC, or focus on selecting patients at high risk for LC screening. High-risk patients are the target population for early diagnosis and have significant clinical implications, but they are less investigated. Therefore, **the characteristics of the study sample** and **studying primary care sequences** are the two aspects of **novelty** in this PhD study.

The first original contribution of this thesis is in the **conceptual and methodological aspects**. I argue the importance of studying longitudinal primary care sequences for early cancer diagnosis and provide a new research perspective for this field. **This is an original idea**, which is **very different from the existing theoretical framework and empirical studies** focused on identifying risk factors or calculating the diagnostic interval using the conceptual models in this field. Secondly, I **propose a novel statistical method** (sequence analysis) **to study** interdependent patient-GP events in **primary care sequences**. No study has ever used SA to study interdependent GP-patient events. Therefore, this is original. Thirdly, I contextualise SA in primary care sequences for health services research and propose several methodological solutions. For example, I propose two approaches to construct primary care sequences – by visit and in a timeline. A “no visit” state could be used to fill the gaps between events when constructing sequences in a timeline. I also propose two ways to reflect the interdependent patient-GP events in the sequences, i.e. combining patient and GP states together and using traditional SA, and creating an extended alphabet and using MCSA. I argue the former approach is more fit for the research purpose of this study. I also propose and argue the importance of a reference event/point in the sequences.

These original methodological innovations are applicable and generalisable to the other studies. Fourthly, I establish the whole analytical process and explain how the findings can be presented in different ways (figures and tables). None of the studies has applied state and event SA in such depth or has integrated both methods to make the results as informative as what I have done in this thesis. Finally, I also make some recommendations for using SA in health research for other researchers. This study demonstrates the potential and the value of using SA to study primary care sequences leading to the diagnosis of LC and paves the way for future research. The contribution has been made through the **original application of existing knowledge to implement a novel research idea and to enhance the understandings in the field of early diagnosis research.**

The empirical findings (summarised in section 8.2) reveal the heterogeneity in both patients' help-seeking behaviours and the patterns of primary care sequences. These patterns may be helpful for GPs to manage patients with a high risk of developing LC. In addition, the heterogeneity at practice level was explored. Most studies include patient's characteristics as explanatory variables or confounding factors, but very few explore the practice effect, which is a novelty of this study. As an exploratory study, this study produces some initial findings and clues that contribute to informing how to design future studies to gain a deeper understanding of the heterogeneous primary care pathways to LC diagnosis among patients at high risk (section 8.6 below).

8.5 Recommendations for other researchers on how to use SA in health research

Based on the methodological exploration in this study, this section summarises my experiences, learning points, and some recommendations that I think would be helpful for other researchers if they consider using SA to study health trajectories in other diseases in their studies.

8.5.1 Data preparation

If using EHRs to conduct a large scale population-based study, it is sensible to include as many relevant variables as possible and have comprehensive code lists in the study design phase for data extraction. It is more convenient, economical, and time-saving to use part of a comprehensive dataset, rather than to prepare for the codes and extract the EHRs multiple times from the database when finding the data is insufficient in the analysis phase. However, **selecting the right information (variables) to answer the RQ is the key.** Data source is further discussed in subsection 8.7.1.

8.5.2 State specification

SA has a substantial part of subjective decision and interpretation. It is essential to apply prior knowledge, judgement, and have expert input when specifying the states. It needs careful consideration in the study design phase and deciding what variables should be included in the sequences. Time is well spent to choose the right variables, and tailor the states for specific RQs, as the **states will directly influence the outcome of the analysis. Do not include more states in the sequences than necessary**, as including more irrelevant elements could make the sequences more complex and may bring more noise into the sequences and dilute the states of interest in smaller percentages. This may lead to failure to identify useful patterns. For multiple correlated dimensions, whether to combine the dimensions together or create an extended alphabet and use MCSA is context-dependent.

8.5.3 Construct and analyse sequences

There are many possible ways to construct and analyse sequences, depending on the RQ, study population, specific context, and the available information and dataset. I propose and demonstrate two approaches of constructing primary care sequences – **by visits** and **in a timeline**, and discuss the implications for each approach. It is more sensible to construct the sequences by visits in this study, with reasons explained in subsection 6.4.2 (P136). In addition, **a common meaningful reference event and/or time point is important to make the sequences comparable and assist the interpretation of the findings** (cluster patterns). The start or end reference point for the sequences is determined by the RQ. For example, when studying sequences leading to cancer diagnosis, constructing sequences in a timeline and making the date of diagnosis as the endpoint is a good strategy. But in this study, the first observed potential LC symptoms was made as the start reference point for the sequences. In addition, **the interval of the timeline could be adjusted to fit with the research context** (e.g. shorter interval when it is closer to the diagnosis).

8.5.4 The choices of dissimilarity measures and cost setting

Sequences with the same length have more choices of dissimilarity measures (algorithms). **The choice of dissimilarity measure and cost setting depends on which aspect we want to focus on.** Each dissimilarity measure has its characteristics and advantages to identify patterns on a specific aspect. If researchers have theoretical or empirical support from existing references and would like to focus on some specific traits of the sequences, then choose the dissimilarity measure directly. If not, it is recommended to explore different dissimilarity measures and cost setting schemes, as sensitivity analyses, and compare the outputs and results to find the optimal solution.

8.5.5 Clustering structure and deciding the optimal number of clusters

SA is exploratory in nature and there is no gold standard to determine the final typology. I proposed some criteria to determine the optimal number of clusters and explored the usefulness of these criteria in this study. It is recommended to check whether there are any outlier sequences first, and then decide whether to include or exclude the outlier sequences in the analysis. For pragmatic reasons, the number of clusters should not be too large (e.g. over ten). A statistical indicator like the ASW value could be used as a reference to assess the clustering quality for SA. However, it is not recommended to use ASW as the key determinant to choose the optimal number of clusters, as the solution with the highest ASW value may not be the best one in the empirical contexts. Making sense of the clusters in context (**face validity**) is more important than the results based on a statistical indicator. However, different clustering solutions with similar ASW values may reassure researchers to choose the one that fits the empirical contexts the best, as similar ASW values indicate the quality of different clustering is at the same level (discussed in subsection 6.7.4). In summary, the interpretability of cluster patterns within a specific research context, clustering quality assessed by statistical indicators, the number of clusters in the typology, and the number of sequences (sample size) in each cluster, should be considered together when making the decision.

Should the clusters be further used in other statistical analyses (e.g. regression model), researchers should consider the pragmatic criteria above, especially when the clusters will be the dependent variable in multinomial logistic regression. More clusters could make the results in the model more challenging to interpret. Results with a smaller number of clusters may have more advantages.

8.5.6 Presentation of the cluster patterns

The intuitive graphic presentation can make the communication of results easier. Sequence index plots can be used to present individual sequences and the outlier sequences, state distribution plots to present the cluster patterns, and state frequency plots for the most frequent sequences. The tree-structured dendrogram can tell us how the sequences are grouped and the clustering structure. Regression tree can show how the cluster patterns change step by step in a figure when splitting the nodes, making it easier for researchers to communicate the process with the readers. Event SA can provide extra information in text for readers to understand the event patterns, which is complementary to the graphic presentation. These tools are all useful to present simplified information from complex sequences on different aspects. The authors should provide some interpretations in text to help readers understand what each cluster means in the typology.

8.5.7 A final reminder

As Pollock (2007) pointed out, SA sits somewhere between purely narrative and traditional variable analysis, and it does involve a substantial part of subjective interpretation at each step. However, this is not necessarily the weakness of the method. It is more related to the nature of applied health research. Even using established statistical methods like regression models in health studies, the final models often involve subject knowledge and researchers' judgment on the meaning of variables in the research context, not just based on some statistical indicators like adjusted R^2 , Akaike information criterion (AIC), or Bayesian information criterion (BIC). SA could be a helpful method for exploratory studies, and we may require a new mindset, different from other statistical methods (e.g. regression), to embrace it and use it.

8.6 Recommendations for future research and potential implications for policy and practice

8.6.1 Recommendations for future empirical studies for early diagnosis research

This study investigated primary care sequences using GP notes in free text. The same research process can be performed in a population to study the primary care pathways with a confirmed diagnosis of primary LC (different study sample) using EHRs from validated databases (different data source), which could increase the sample size, geographical coverage, and the representativeness of study population to better characterise the primary care pathways to LC diagnosis. The comparison of research evidence in different clinical periods and populations can help us better understand patients' care needs and gain a holistic perspective of the disease trajectory, which may have implications in health services planning, and proactively monitoring patient's disease progress. Due to a small sample size and only several practices in this study, traditional statistical methods were used to explore the practice effect. However, it is unrealistic to use the same analytical approach for a population-based study with a large sample size and patients from hundreds of practices. Multilevel multinomial logit model may be more suitable. In addition, practice level data and GP characteristics (summarised in subsection 7.3.5) are also helpful for exploring the practice effect and explaining the heterogeneity.

Another possible research direction to investigate the primary care sequences in patients at high risk is to select a retrospective cohort from a big EHR database and perform a baseline risk assessment using validated risk prediction models developed from the British population, e.g. the LLP risk model (Cassidy et al., 2008) or QCancer (lung)(Hippisley-Cox and Coupland, 2011). Researchers could select patients at the same level of baseline risk at the start point (e.g. 5-year

absolute risk threshold of $\geq 2.5\%$ in the LLP_{v2} model as a “high-risk” study sample in 2015), and then use EHR data to understand their primary care sequences onwards (e.g. 2016-2020). Some patients may get LC diagnosis during the follow-up, some may not. This study design allows us to understand the heterogeneity of disease development for patients at the same level of risk at baseline and identify the phenotypes of disease trajectories. Such research evidence could inform the design of different intervention strategies for different patient subgroups. This is an improvement of the current study, as patients in the NAEDI study, even though they were long-term smokers, were probably at different levels of risk. We do not know the absolute risk of individual patients and it is difficult to explain the heterogeneity of the sequences in this study sample. The proposed study is conceptually simple, but this project needs a lot of resources, including a large database with extensive linkage to the clinical outcomes (cancer case ascertainment and staging information from cancer registration, and preferably, mortality data from ONS), skilled statisticians to process the data and perform the analysis. It is also computationally demanding, as it needs to calculate the baseline risk for the whole cohort in the EHR database to assess subject eligibility to meet the inclusion criteria. It needs teamwork and several years of fundings to conduct such a big scale of study.

Together with the existing conceptual model of the cancer care pathway and the ‘route to diagnosis’, the cluster patterns have great potential to produce more new knowledge. The associations between the cluster patterns, different intervals in the pathway, routes to diagnosis, and important clinical outcomes such as stage at diagnosis and cancer survival, could be explored, adjusting for relevant variables at patient or practice level in the model. Furthermore, the predictive value of different patterns of sequences in diagnostic and prognostic models could be investigated and compared. Again, this type of study needs a larger sample size and extensive data linkage in the EHR database to make such investigation possible.

8.6.2 Research is needed to assess the impact of COVID-19 on early cancer diagnosis

Before Coronavirus disease (COVID-19), primary care in England was already under great pressure. The way GP manages patients has changed dramatically during COVID-19. For example, when the UK government announced the national ‘lockdown’ on 23rd March 2020 to contain the spread of coronavirus and mitigate the negative impact, the majority of consultations in primary care were delivered by telephone or video consultation without face-to-face contact with the patients. There may be indirect consequences relating to the changes in the access and delivery of health services. This may cause extra barriers to patient-GP communication as non-verbal cues are unavailable, which may influence GP eliciting symptoms from patients. Meanwhile, patient’s help-seeking behaviours also changed, as patients were told to “protect the NHS” and may not

want to “bother the doctors” at this critical time. Therefore, patients may delay help-seeking even with symptoms, or misinterpret the symptoms were related to COVID-19. In addition, patients with existing respiratory conditions may be more vulnerable to COVID-19. All of these may affect how and when the diagnoses are made. The COVID-19 could result in a delayed diagnosis of cancer and exacerbate health inequalities if certain subgroups of the population (e.g. the elderly, deprived, or ethnic minority) are less able to access health services. How COVID-19 influenced patient’s symptom interpretations, help-seeking behaviours, pathways to LC diagnosis, waiting time to start the first treatment, and the survival outcomes would be an interesting research topic for future studies in the next several years.

8.6.3 Recommendations for future methodological work

The systematic scoping review identified two research gaps, which could be opportunities and the research directions for future studies. One is developing a critical appraisal tool to assess the quality of studies using SA (subsection 4.4.7). The reporting of using SA still needs strengthening. A guideline for clear reporting of the technical details related to the analytical process. These works will be helpful for both authors and readers (subsection 4.5.1).

The third possible research direction is to explore other statistical methods that are capable of analysing interdependent patient-GP events together over time, and to compare the strengths and limitations between different statistical methods.

8.7 Connect this thesis to the wider research fields and disciplines

8.7.1 Data source: big data, real-world data, and EHRs

There is no clear definition for ‘big data’, but the term is usually used to refer to datasets with many participants and/or variables using large scale record linkage. Epidemiologic principles and statistical techniques are often applied in big data to answer RQs in biomedical research, to provide evidence of health profiles and disease trajectories, build prediction models for individual patients, infer associations and risk factors, and stratify patient groups (Lawlor, 2019).

EHR database contains rich and ‘live’ data from thousands and millions of patients, which becomes a trending data source for population-based research to provide real-world evidence and inform health policy. Compared with clinical trials with a relatively small sample size and strict inclusion and exclusion criteria, EHR can have a broader population coverage to observe patients’ trajectories (Frohlich et al., 2018). Although the strength of evidence from clinical trials is higher than that from observational studies in the evidence hierarchy (Guyatt, 1995), some RQs in health

research cannot be addressed by interventions or randomised controlled trials, because of the feasibility, practical and ethical concerns. Observational studies using EHRs could be an alternative approach.

The most commonly used databases in the UK primary care setting include the Clinical Practice Research Datalink (CPRD)(Herrett et al., 2015, Wolf et al., 2019), The Health Improvement Network (THIN) database, the QResearch database, and the Secure Anonymised Information Linkage (SAIL) Databank in Wales (Lyons et al., 2009). Primary care records could be linked with secondary care records (hospital admissions, A&E attendances, and outpatient appointments in NHS hospitals), national cancer registration and analysis service, mental health services data set (MHSDS), and ONS death registration (Herbert et al., 2017). The linked data have great potential to provide a full picture of the whole patient care pathway, which is suitable for study designs like retrospective cohort studies and (nested) case-control studies. Despite its importance as a clinical outcome, there are substantial missing data in cancer staging in the current British healthcare databases (e.g. CPRD and HES), even in the cancer registry. The completeness of cancer staging data has been improving in recent years. However, up to 40% of staging data is still missing in the most recent available cancer data (in 2017, personal work experience of using the cancer registry data linked to the QResearch database). The lack of complete and reliable staging data has a significant impact on the scope of empirical works.

Clinical and administrative information is usually recorded in the database in a timely manner. The events and dates of referrals and investigations are generally accurate. Patients continue to contribute information as long as they remain consuming care services in the health system and do not decline a trusted third party to use their data for research purposes (the choice of 'opt-out'). Compared with collecting data directly from subjects, using data from the EHR database avoids recall bias, greatly reduces costs, time, human resources, and attrition rate. All these are the advantages of using EHRs for secondary analysis. However, the records are input by thousands of HCPs across different care facilities and settings. The heterogeneity, data quality, and the mechanism of missingness should be carefully considered and assessed, as I have done in this study (subsection 5.4.5). When using EHR for research, researchers need to consider different possible codes (e.g. Read codes, ICD codes) for the same health condition, and prepare code lists as inclusive and complete as possible to extract the health events. When reporting the study findings, the limitations of the coding systems used in the database should be discussed (Weller et al., 2012). Researchers should also conduct a thorough assessment of the quality of the database and its capacity to capture valid information (e.g., the completeness and accuracy of the encounter, dates), and the coverage and representativeness of the study sample relative to the wider population, in terms of age, sex, ethnicity, geographical area, and the SES of the patients.

The process of data quality assurance is to minimise the two main sources of bias in observational studies, i.e. selection bias and information bias, as bias could have a significant impact on the reliability of results and the generalisability of findings to a wider population.

Two different data sources, EHRs and GP notes in free text, were used to study primary care sequences in my PhD journey, although the research process and the results from EHRs (HHRAD) were not reported in this thesis. Structured data are health events indexed and retrievable by codes (Read, ICD, British National Formulary), generally organised in a hierarchical structure (e.g. organ or body system). Unstructured data, like free text entries or letters received by GPs from specialists, are usually more difficult for researchers to access, as they may contain identifiable and specific patient information that may breach confidentiality and anonymity. But if available, natural language processing (Shah et al., 2018) has the potential to extract useful information and establish prediction models using EHRs (Liao et al., 2015).

8.7.2 Unsupervised and supervised learning

Supervised and unsupervised learning are two techniques in machine learning. Unsupervised learning is often used to identify patterns or clusters through feature elicitation and visualisation of complex data without labelled responses (meaningful tags, labels, or class of the observations). Clustering and dimension reduction techniques are unsupervised learning methods. The outcome, such as the number of clusters in this study, is unknown, while regression with a known outcome belongs to supervised learning. Semi-supervised learning is an approach between unsupervised and supervised learning, which has a small amount of labelled data and a large amount of unlabelled data for training (Sidey-Gibbons and Sidey-Gibbons, 2019).

In this study, the exploratory process of SA to identify clusters of primary care sequences is essentially an unsupervised learning technique. It is not possible to know how many clusters (the outcome) would be before the analysis, and it is not appropriate to arbitrarily set a fixed number of clusters beforehand. The optimal number of clusters was decided by comparing several possible clustering solutions and making sense of the cluster patterns in the empirical context. One example in this field is the use of unsupervised cluster analysis to understand COPD heterogeneity and the attempt to create COPD subtypes from disease-related clinical characteristics (FEV1, FVC, FEV1/FVC, BMI, modified Medical Research Council score, asthma, and cardiovascular comorbid disease)(Castaldi et al., 2017).

After identifying the clusters, multinomial logistic regression was used to investigate the association between patient's characteristics and cluster membership, which is a supervised learning technique, as the outcome was known at the point of analysis. In machine learning

literature, **classification** is often referred to prediction of categorical outcomes, while **regression** refers to prediction of continuous outcomes. Many machine learning algorithms developed for classification could be adapted to address problems using regression and vice versa (Bi et al., 2019).

Machine learning algorithms (e.g. artificial neural network, support vector machine, random forest, gradient boosting machine, k-nearest neighbour) have the advantages of handling an enormous amount of data, selecting useful predictors from a great number of variables, and looking for the combination of variables that can predict the outcomes reliably. Machine learning has great potential to provide new insight and solve complex problems in medical sciences. There is an increasing number of studies using machine learning algorithms in big data (linked health records) to develop and validate prediction models for different diseases in recent years, as machine learning algorithms require millions of observations to train the prediction model to an acceptable performance level (Obermeyer and Emanuel, 2016). The application of machine learning algorithms in large health data could be a trend in early diagnosis research in the next few years.

8.7.3 Phenotypes and stratified medicine

Electronic phenotyping is known as utilising EHR data to identify patients with specific characteristics of interest (either exposures or outcomes), which could be further used to identify clinical risk factors and protective factors, establish prediction models, and support clinical decisions (Banda et al., 2018). From the findings of this study, patients were probably at different risk levels of developing LC. Future population studies using big data and machine learning techniques may be in a better position to stratify patients at different levels of risk more precisely. Different health education programmes and social campaigns could be developed and more targeted at patients at different levels of risk, to encourage behavioural changes and guide the patients to seek help more promptly. Patients with the highest risk are likely to benefit the most from screening and preventive programmes. Stratifying patients based on risk could be a cost-effective way to manage patients and make good use of resources in clinical practice. Furthermore, the idea of stratified medicine is not limited to diagnosis, but also could be extended to treatment. One similar application was treatment allocation for patients diagnosed with prostate cancer in different tiers of risk (risk stratification), using the Cambridge Prognostic Group classification (Parry et al., 2020).

8.8 Conclusion and final message

The application of SA in research has been evolving and expanding to different disciplines and research areas in the last several decades. But it is still a relatively new method in health services research. This study explores the use of SA to understand primary care sequences, with the hope to find new research angles and shed new light on providing new research evidence to promote early diagnosis of LC and improve cancer survival. Through the initial exploration, this study demonstrates that SA can identify meaningful cluster patterns from complex primary care sequences. It is possible to design a study with a diverse population in a larger sample size and geographical coverage, and use SA to fully uncover the heterogeneous primary care sequences leading to LC diagnosis in the near future.

Appendix A Search strategy for the systematic scoping review (Chapter 4)

A.1 Search strategy in the EBSCO platform

The three databases, MEDLINE, PsycINFO, and CINAHL (Cumulative Index to Nursing and Allied Health Literature) Plus with Full Text, were integrated into the EBSCOhost Research Database, available through the online library of the University of Southampton.

In order to be more inclusive, the keywords below were searched in the field of full text (TX), not limited to the title (TI) and abstract (AB), as some keywords (like sequence analysis) may not be necessary in the title and/or abstract. The following results were from the three databases, with restriction of publication date until 31 December 2018 and articles in English. Exact duplicates (overlap among the three databases) have already been removed from the results by EBSCO.

No.	Search strategy	Results
S1	TX "sequence analysis"	295,933
	MeSH: subject heading of different care settings	
S2	MH "primary health care"	79,733
S3	MH "secondary care"	427
S4	MH "tertiary healthcare"	831
	Free text words of care settings	
S5	TX "primary care"	
S6	TX "primary health care"	
S7	TX "primary healthcare"	
S8	TX "general practice"	
S9	TX "family practice"	

Appendix A

S10	OR/S5-S9	354,330
S11	TX “secondary care”	
S12	TX “secondary healthcare”	
S13	TX “secondary health care”	
S14	TX “hospital care”	
S15	TX “inpatient care”	
S16	TX “acute care”	
S17	TX “emergency care”	
S18	OR/S10-S17	150,730
S19	TX “tertiary care”	
S20	TX “tertiary healthcare”	
S21	TX “tertiary health care”	
S22	OR/S19-S22	69,307
S23	OR/S2, S3, S4, S10, S18, S22 MeSH and free text words of care settings	545,502
	Pathways and trajectories	
S24	TX pathway*	1,142,399
S25	TX trajector*	93,399
	All possible applications of SA to study care pathways, disease trajectories and health services research	
S26	S23 AND S24	20,844
S27	S23 AND S25	6,930
S28	TX “care pathway*”	8,923
S29	TX “clinical pathway*”	5,271

S30	TX “care trajector*”	655
S31	TX “disease trajector*”	2,122
S32	MH “health services”	17,582
S33	TX “health service*”	616,075
S34	OR/S26-33	640,679
S35	S1 AND S34	706





Notes and further explanations for the search strategy:

- 1) TX means words were searched in the “full text” field;
- 2) MH means words were searched in Medical Subject Heading (MeSH terms);
- 3) S1: “Sequence analysis” includes other relevant terms like "state sequence analysis" and "multichannel sequence analysis";
- 4) S24 and S25: pathway* and trajector* include both singular and plural forms of pathway and trajectory. In addition, pathway is also commonly used in genetics and microbiology.
- 5) The rationale for S26 and S27: pathway (S26) and trajectory (S27) in different care settings (primary/secondary/tertiary), in full text;
- 6) The rationale for S28-S31: retrieved other expressions like “critical care pathways” (S28) or “palliative care trajectories” (S30). Difference between S26/27 and S28-S31: S26/S27 used “AND” to retrieve records of pathways/trajectories in different care settings; while S28-S31 expanded the search for care pathways/trajectories in more specific disciplines/areas, e.g. dental care pathway, critical care trajectories, terminal care pathways... Any adjective in front of care pathways can be searched and retrieved. It aimed to include as many relevant records as possible.

A.2 Search screenshots from the EBSCOhost platform

Search History

#	Query	Limiters/Expanders	Last Run Via	Results	Action
S20	S1 AND S19	Expanders - Apply equivalent subjects Search modes - Find all my search terms	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - MEDLINE;APA PsycInfo;CINAHL Plus with Full Text	706	Edit
S19	S11 OR S12 OR S13 OR S14 OR S15 OR S16 OR S17 OR S18	Expanders - Apply equivalent subjects Search modes - Find all my search terms	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - MEDLINE;APA PsycInfo;CINAHL Plus with Full Text	640,679	Edit
S18	TX "clinical pathway"	Limiters - Date of Publication: -20181231; Abstract Available; English Language Expanders - Apply equivalent subjects Search modes - Find all my search terms	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - MEDLINE;APA PsycInfo;CINAHL Plus with Full Text	5,271	Edit
S17	MH "health services"	Limiters - Date of Publication: -20181231; Abstract Available; English Language Expanders - Apply equivalent subjects Search modes - Find all my search terms	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - MEDLINE;APA PsycInfo;CINAHL Plus with Full Text	17,582	Edit
S16	TX "health service"	Limiters - Date of Publication: -20181231; Abstract Available; English Language Expanders - Apply equivalent subjects Search modes - Find all my search terms	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - MEDLINE;APA PsycInfo;CINAHL Plus with Full Text	616,074	Edit
S15	S3 AND S10	Expanders - Apply equivalent subjects Search modes - Find all my search terms	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - MEDLINE;APA PsycInfo;CINAHL Plus	20,842	Edit
S14	S2 AND S10	Expanders - Apply equivalent subjects Search modes - Find all my search terms	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - MEDLINE;APA PsycInfo;CINAHL Plus with Full Text	6,930	Edit
S13	TX "disease trajectory"	Limiters - Date of Publication: -20181231; Abstract Available; English Language Expanders - Apply equivalent subjects Search modes - Find all my search terms	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - MEDLINE;APA PsycInfo;CINAHL Plus with Full Text	2,121	Edit
S12	TX "care trajectory"	Limiters - Date of Publication: -20181231; Abstract Available; English Language Expanders - Apply equivalent subjects Search modes - Find all my search terms	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - MEDLINE;APA PsycInfo;CINAHL Plus with Full Text	655	Edit
S11	TX "care pathway"	Limiters - Date of Publication: -20181231; Abstract Available; English Language Expanders - Apply equivalent subjects Search modes - Find all my search terms	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - MEDLINE;APA PsycInfo;CINAHL Plus with Full Text	8,924	Edit
S10	S4 OR S5 OR S6 OR S7 OR S8 OR S9	Expanders - Apply equivalent subjects Search modes - Find all my search terms	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - MEDLINE;APA PsycInfo;CINAHL Plus with Full Text	545,321	Edit
S9	TX "tertiary care" OR TX "tertiary healthcare" OR TX "tertiary health care"	Limiters - Date of Publication: -20181231; Abstract Available; English Language Expanders - Apply equivalent subjects Search modes - Find all my search terms	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - MEDLINE;APA PsycInfo;CINAHL Plus with Full Text	69,305	Edit
S8	TX "secondary care" OR TX "secondary healthcare" OR TX "secondary health care" OR TX "hospital care" OR TX "inpatient care" OR TX "acute care" OR TX "emergency care"	Limiters - Date of Publication: -20181231; Abstract Available; English Language Expanders - Apply equivalent subjects Search modes - Find all my search terms	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - MEDLINE;APA PsycInfo;CINAHL Plus with Full Text	150,548	Edit
S7	TX "primary care" OR TX "primary health care" OR TX "primary healthcare" OR TX "general practice" OR TX "family practice"	Limiters - Date of Publication: -20181231; Abstract Available; English Language Expanders - Apply equivalent subjects Search modes - Find all my search terms	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - MEDLINE;APA PsycInfo;CINAHL Plus with Full Text	354,324	Edit
S6	MH "tertiary healthcare"	Limiters - Date of Publication: -20181231; Abstract Available; English Language Search modes - Find all my search terms	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - MEDLINE;APA PsycInfo;CINAHL Plus with Full Text	831	Edit
S5	MH "secondary care"	Limiters - Date of Publication: -20181231; Abstract Available; English Language Expanders - Apply equivalent subjects Search modes - Find all my search terms	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - MEDLINE;APA PsycInfo;CINAHL Plus with Full Text	427	Edit

S4	MH "primary health care"	Limiters - Date of Publication: -20181231; Abstract Available; English Language Search modes - Find all my search terms	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - MEDLINE;APA PsycInfo;CINAHL Plus with Full Text	79,733	 Edit
S3	TX pathway*	Limiters - Date of Publication: -20181231; Abstract Available; English Language Search modes - Find all my search terms	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - MEDLINE;APA PsycInfo;CINAHL Plus with Full Text	1,142,490	 Edit
S2	TX trajectory*	Limiters - Date of Publication: -20181231; Abstract Available; English Language Search modes - Find all my search terms	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - MEDLINE;APA PsycInfo;CINAHL Plus with Full Text	93,399	 Edit
S1	TX "sequence analysis"	Limiters - Date of Publication: -20181231; Abstract Available; English Language Expanders - Also search within the full text of the articles Search modes - Find all my search terms	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - MEDLINE;APA PsycInfo;CINAHL Plus with Full Text	295,936	 Edit

A.3 Search strategy in Ovid

Ovid hosts the Embase database. It has very specific and detailed search fields, up to a total number of 118 fields. Keywords and subject headings were searched in the following fields: ab: abstract; hw: heading word; kw: keyword; ot: original title; sh: subject headings; sl: summary language; tw: text work; ti: title; mp: search as keywords, often together with subject heading; exp: explode, include more relevant terms.

No.	Search strategy	Result
1	"sequence analysis".ab,hw,kw,ot,sh,ti,sl,tw.	197,560
2	"primary care".ab,hw,kw,ot,sh,ti,sl,tw.	157,472
3	"primary health care".ab,hw,kw,ot,sh,ti,sl,tw.	77,939
4	"primary healthcare".ab,hw,kw,ot,sh,ti,sl,tw.	7,376
5	general practice.mp. or exp general practice/	101,286
6	"family practice".ab,hw,kw,ot,sh,ti,sl,tw.	9,951
7	primary health care.mp. or exp primary health care/	177,899
8	2 or 3 or 4 or 5 or 6 or 7	315,956
9	"secondary care".ab,hw,kw,ot,sh,ti,sl,tw.	12,288
10	"secondary healthcare".ab,hw,kw,ot,sh,ti,sl,tw.	441
11	"secondary health care".ab,hw,kw,ot,sh,ti,sl,tw.	6,717

Appendix A

12	"hospital care".ab,hw,kw,ot,sh,ti,sl,tw.	31,959
13	"inpatient care".ab,hw,kw,ot,sh,ti,sl,tw.	8,681
14	"acute care".ab,hw,kw,ot,sh,ti,sl,tw.	30,987
15	"emergency care".ab,hw,kw,ot,sh,ti,sl,tw.	52,755
16	secondary care.mp. or exp secondary health care/	13,952
17	9 or 10 or 11 or 12 or 13 or 14 or 15 or 16	121,978
18	"tertiary care".ab,hw,kw,ot,sh,ti,sl,tw.	112,364
19	"tertiary healthcare".ab,hw,kw,ot,sh,ti,sl,tw.	701
20	"tertiary health care".ab,hw,kw,ot,sh,ti,sl,tw.	39,798
21	tertiary healthcare.mp. or exp tertiary health care/	97,653
22	18 or 19 or 20 or 21	125,360
23	8 or 17 or 22	543,920
24	"pathway*".ab,hw,kw,ot,sh,ti,sl,tw.	1,441,653
25	"trajector*".ab,hw,kw,ot,sh,ti,sl,tw.	86,871
26	23 and 24	8,258
27	23 and 25	1,637
28	"care pathway* ".ab,hw,kw,ot,sh,ti,sl,tw.	8,204
29	clinical pathway/ or clinical pathway*.mp.	11,445
30	"clinical pathway* ".ab,hw,kw,ot,sh,ti,sl,tw.	11,445
31	"care trajector* ".ab,hw,kw,ot,sh,ti,sl,tw.	423
32	"disease trajector* ".ab,hw,kw,ot,sh,ti,sl,tw.	1,780
33	health service/ or health service*.mp.	594,578
34	26 or 27 or 28 or 29 or 30 or 31 or 32 or 33	618,494
35	1 and 34	144

36 limit 35 to (abstracts and English language and yr="1902 - 2018") 86

Abstract must be available for screening, with restriction of language in English and publication period.

A.4 Search screenshots from Ovid

▼ Search History (36) View S						
# ▲	Searches	Results	Type	Actions	Annotations	
<input type="checkbox"/>	1	"sequence analysis".ab,hw,kw,ot,sh,ti,sl,tw.	197560	Advanced	Display Results More ▼	
<input type="checkbox"/>	2	"primary care".ab,hw,kw,ot,sh,ti,sl,tw.	157472	Advanced	Display Results More ▼	
<input type="checkbox"/>	3	"primary health care".ab,hw,kw,ot,sh,ti,sl,tw.	77939	Advanced	Display Results More ▼	
<input type="checkbox"/>	4	"primary healthcare".ab,hw,kw,ot,sh,ti,sl,tw.	7376	Advanced	Display Results More ▼	
<input type="checkbox"/>	5	general practice.mp. or exp general practice/	101286	Advanced	Display Results More ▼	
<input type="checkbox"/>	6	"family practice".ab,hw,kw,ot,sh,ti,sl,tw.	9951	Advanced	Display Results More ▼	
<input type="checkbox"/>	7	primary health care.mp. or exp primary health care/	177899	Advanced	Display Results More ▼	
<input type="checkbox"/>	8	2 or 3 or 4 or 5 or 6 or 7	315956	Advanced	Display Results More ▼	
<input type="checkbox"/>	9	"secondary care".ab,hw,kw,ot,sh,ti,sl,tw.	12288	Advanced	Display Results More ▼	
<input type="checkbox"/>	10	"secondary healthcare".ab,hw,kw,ot,sh,ti,sl,tw.	441	Advanced	Display Results More ▼	
<input type="checkbox"/>	11	"secondary health care".ab,hw,kw,ot,sh,ti,sl,tw.	6717	Advanced	Display Results More ▼	
<input type="checkbox"/>	12	"hospital care".ab,hw,kw,ot,sh,ti,sl,tw.	31959	Advanced	Display Results More ▼	
<input type="checkbox"/>	13	"inpatient care".ab,hw,kw,ot,sh,ti,sl,tw.	8681	Advanced	Display Results More ▼	
<input type="checkbox"/>	14	"acute care".ab,hw,kw,ot,sh,ti,sl,tw.	30987	Advanced	Display Results More ▼	
<input type="checkbox"/>	15	"emergency care".ab,hw,kw,ot,sh,ti,sl,tw.	52755	Advanced	Display Results More ▼	
<input type="checkbox"/>	16	secondary care.mp. or exp secondary health care/	13952	Advanced	Display Results More ▼	
<input type="checkbox"/>	17	9 or 10 or 11 or 12 or 13 or 14 or 15 or 16	121978	Advanced	Display Results More ▼	
<input type="checkbox"/>	18	"tertiary care".ab,hw,kw,ot,sh,ti,sl,tw.	112364	Advanced	Display Results More ▼	
<input type="checkbox"/>	19	"tertiary healthcare".ab,hw,kw,ot,sh,ti,sl,tw.	701	Advanced	Display Results More ▼	
<input type="checkbox"/>	20	"tertiary health care".ab,hw,kw,ot,sh,ti,sl,tw.	39798	Advanced	Display Results More ▼	
<input type="checkbox"/>	21	tertiary healthcare.mp. or exp tertiary health care/	97653	Advanced	Display Results More ▼	
<input type="checkbox"/>	22	18 or 19 or 20 or 21	125360	Advanced	Display Results More ▼	
<input type="checkbox"/>	23	8 or 17 or 22	543920	Advanced	Display Results More ▼	
<input type="checkbox"/>	24	"pathway".ab,hw,kw,ot,sh,ti,sl,tw.	1441653	Advanced	Display Results More ▼	
<input type="checkbox"/>	25	"trajector".ab,hw,kw,ot,sh,ti,sl,tw.	86871	Advanced	Display Results More ▼	
<input type="checkbox"/>	26	23 and 24	8258	Advanced	Display Results More ▼	
<input type="checkbox"/>	27	23 and 25	1637	Advanced	Display Results More ▼	
<input type="checkbox"/>	28	"care pathway" .ab,hw,kw,ot,sh,ti,sl,tw.	8204	Advanced	Display Results More ▼	
<input type="checkbox"/>	29	clinical pathway/ or clinical pathway".mp.	11445	Advanced	Display Results More ▼	
<input type="checkbox"/>	30	"clinical pathway" .ab,hw,kw,ot,sh,ti,sl,tw.	11445	Advanced	Display Results More ▼	
<input type="checkbox"/>	31	"care trajector" .ab,hw,kw,ot,sh,ti,sl,tw.	423	Advanced	Display Results More ▼	
<input type="checkbox"/>	32	"disease trajector" .ab,hw,kw,ot,sh,ti,sl,tw.	1780	Advanced	Display Results More ▼	
<input type="checkbox"/>	33	health service/ or health service".mp.	594578	Advanced	Display Results More ▼	
<input type="checkbox"/>	34	26 or 27 or 28 or 29 or 30 or 31 or 32 or 33	618494	Advanced	Display Results More ▼	
<input type="checkbox"/>	35	1 and 34	144	Advanced	Display Results More ▼	
<input type="checkbox"/>	36	limit 35 to (abstracts and english language and yr="1902 - 2018")	86	Advanced	Display Results More ▼	

Appendix B Joanna Briggs Institute (JBI) Critical Appraisal Checklist for Case Series

Reviewer: _____ Date: _____

Author: _____ Year: _____ Record Number: _____

	Yes	No	Unclear	Not applicable
1. Were there clear criteria for inclusion in the case series?	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
2. Was the condition measured in a standard, reliable way for all participants included in the case series?	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
3. Were valid methods used for identification of the condition for all participants included in the case series?	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
4. Did the case series have consecutive inclusion of participants?	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
5. Did the case series have complete inclusion of participants?	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
6. Was there clear reporting of the demographics of the participants in the study?	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
7. Was there clear reporting of clinical information of the participants?	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
8. Were the outcomes or follow up results of cases clearly reported?	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
9. Was there clear reporting of the presenting site(s)/clinic(s) demographic information?	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
10. Was the statistical analysis appropriate?	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Appendix B

Overall appraisal: Include Exclude Seek further info

Comments (Including reason for exclusion)

Reference source: Moola S, Munn Z, Tufanaru C, Aromataris E, Sears K, Sfetcu R, Currie M, Qureshi R, Mattis P, Lisy K, Mu P-F. Chapter 7: Systematic reviews of aetiology and risk. In: Aromataris E, Munn Z (Editors). Joanna Briggs Institute Reviewer's Manual. The Joanna Briggs Institute, 2017. Available from <https://reviewersmanual.joannabriggs.org/>

Tool Guidance

Answers: Yes, No, Unclear or Not applicable (N/A)

1. Were there clear criteria for inclusion in the case series?

The authors should provide clear inclusion (and exclusion criteria where appropriate) for the study participants. The inclusion/exclusion criteria should be specified (e.g., risk, stage of disease progression) with sufficient detail and all the necessary information critical to the study.

2. Was the condition measured in a standard, reliable way for all participants included in the case series?

The study should clearly describe the method of measurement of the condition. This should be done in a standard (i.e. same way for all patients) and reliable (i.e. repeatable and reproducible results) way.

3. Were valid methods used for identification of the condition for all participants included in the case series?

Many health problems are not easily diagnosed or defined, and some measures may not be capable of including or excluding appropriate levels or stages of the health problem. If the outcomes were assessed based on existing definitions or diagnostic criteria, then the answer to this question is likely to be yes. If the outcomes were assessed using observer reported, or self-reported scales, the risk of over- or under-reporting is increased, and objectivity is compromised. Importantly, determine if the measurement tools used were validated instruments as this has a significant impact on outcome assessment validity.

4. Did the case series have consecutive inclusion of participants?

Studies that indicate a consecutive inclusion are more reliable than those that do not. For example, a case series that states 'we included all patients (24) with osteosarcoma who presented to our clinic between March 2005 and June 2006' is more reliable than a study that simply states 'we report a case series of 24 people with osteosarcoma.'

5. Did the case series have complete inclusion of participants?

The completeness of a case series contributes to its reliability (1). Studies that indicate a complete inclusion are more reliable than those that do not. As stated above, a case series that states 'we included all patients (24) with osteosarcoma who presented to our clinic between March 2005 and June 2006' is more reliable than a study that simply states 'we report a case series of 24 people with osteosarcoma.'

6. Was there clear reporting of the demographics of the participants in the study?

The case series should clearly describe relevant participant's demographics, such as the following information where relevant: participant's age, sex, education, geographic region, ethnicity, time period, education.

7. Was there clear reporting of clinical information of the participants?

There should be clear reporting of clinical information of the participants such as the following information where relevant: disease status, comorbidities, stage of the disease, previous interventions/treatment, results of diagnostic tests.

8. Were the outcomes or follow-up results of cases clearly reported?

The results of any intervention or treatment should be clearly reported in the case series. A good case study should clearly describe the clinical condition post-intervention in terms of the presence or lack of symptoms. The outcomes of management/treatment, when presented as images or figures, can help in conveying the information to the reader/clinician. It is important that adverse events are clearly documented and described, particularly a new or unique condition is being treated or when a new drug or treatment is used. In addition, unanticipated events, if any that may yield new or useful information should be identified and clearly described.

9. Was there clear reporting of the presenting site(s)/clinic(s) demographic information?

Certain diseases or conditions vary in prevalence across different geographic regions and populations (e.g. women vs men, sociodemographic variables between countries). The study sample should be described in sufficient detail so that other researchers can determine if it is comparable to the population of interest to them.

10. Was the statistical analysis appropriate?

As with any consideration of statistical analysis, consideration should be given to whether there was a more appropriate alternate statistical method that could have been used. The methods section of studies should be detailed enough for reviewers to identify which analytical techniques were used and whether these were suitable.

Appendix C National Heart, Lung, and Blood Institute Quality Assessment Tool for Case Series Studies

Criteria	Yes	No	Other (CD, NR, NA)*
1. Was the study question or objective clearly stated?			
2. Was the study population clearly and fully described, including a case definition?			
3. Were the cases consecutive?			
4. Were the subjects comparable?			
5. Was the intervention clearly described?			
6. Were the outcome measures clearly defined, valid, reliable, and implemented consistently across all study participants?			
7. Was the length of follow-up adequate?			
8. Were the statistical methods well-described?			
9. Were the results well-described?			

* CD, cannot determine; NA, not applicable; NR, not reported

Appendix D Substitution cost matrix of OM_[1, TR] for high-risk sequences in subgroup analysis-2

Table D.1 – Substitution cost matrix of OM_[1, TR] for high-risk sequences in subgroup analysis-2 (N=295)

	1->	2->	3->	4->	5->	6->	7->	8->	9->	10->	11->	12->	13->	14->	15->
1->	0														
2->	1.92	0													
3->	1.87	1.74	0												
4->	1.75	1.75	1.79	0											
5->	1.90	1.81	1.71	1.77	0										
6->	1.92	1.87	1.92	1.80	1.92	0									
7->	1.67	1.64	2	1.98	2	2	0								
8->	1.97	2	1.72	1.87	1.98	2	2	0							
9->	1.97	1.98	1.99	1.59	1.99	1.82	2	2	0						
10->	1.82	1.76	1.76	1.65	1.80	1.92	2	1.58	1.47	0					
11->	1.88	1.98	1.91	1.92	1.94	2	2	2	2	1.81	0				
12->	2	2	2	2	1.99	1.67	2	2	2	2	2	0			
13->	1.85	1.92	1.92	1.90	1.93	2	2	1.88	1.97	1.86	1.81	1.61	0		
14->	2	1.98	1.79	1.87	1.99	1.85	2	2	2	1.79	2	2	1.84	0	
15->	2	1.96	1.97	1.88	1.92	1.95	1.64	1.97	2	1.76	1.76	1.67	1.63	1.87	0

Table D.2 – Substitution cost matrix of $OM_{[1, TR]}$ for high-risk sequences in subgroup analysis-2
(N=294, after excluding the outlier sequence)

	1->	2->	3->	4->	5->	6->	7->	8->	9->	10->	11->	12->	13->	14->	15->
1->	0														
2->	1.92	0													
3->	1.87	1.74	0												
4->	1.74	1.75	1.79	0											
5->	1.90	1.81	1.71	1.77	0										
6->	1.92	1.87	1.92	1.80	1.92	0									
7->	1.67	1.64	2	1.98	2	2	0								
8->	1.97	2	1.69	1.99	1.98	2	2	0							
9->	1.97	1.98	1.99	1.59	1.99	1.82	2	2	0						
10->	1.82	1.76	1.76	1.65	1.80	1.92	2	1.53	1.47	0					
11->	1.88	1.98	1.91	1.92	1.93	2	2	2	2	1.80	0				
12->	2	2	2	2	1.99	1.67	2	2	2	2	2	0			
13->	1.85	1.92	1.92	1.90	1.93	2	2	1.86	1.97	1.86	1.80	1.61	0		
14->	2	1.98	1.79	1.87	1.99	1.85	2	2	2	1.79	2	2	1.84	0	
15->	2	1.94	1.97	1.79	1.90	1.95	1.62	1.95	2	1.78	1.83	1.67	1.51	1.85	0

List of References

- AASSVE, A., BILLARI, F. C. & PICCARRETA, R. 2007. Strings of adulthood: A sequence analysis of young British women's work-family trajectories. *European Journal of Population-Revue Europeenne De Demographie*, 23, 369-388.
- ABBOTT, A. 1983. Sequences of Social Events: Concepts and Methods for the Analysis of Order in Social Processes. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 16, 129-147.
- ABBOTT, A. 1990. A Primer on Sequence Methods. *Organization Science*, 1, 375-392.
- ABBOTT, A. 1992. From Causes to Events: Notes on Narrative Positivism. *Sociological Methods & Research*, 20, 428-455.
- ABBOTT, A. & FORREST, J. 1986. Optimal Matching Methods for Historical Sequences. *Journal of Interdisciplinary History*, 16, 471-494.
- ABBOTT, A. & HRYCAK, A. 1990. Measuring Resemblance in Sequence Data - an Optimal Matching Analysis of Musicians Careers. *American Journal of Sociology*, 96, 144-185.
- ABBOTT, A. & TSAY, A. 2000. Sequence Analysis and Optimal Matching Methods in Sociology: Review and Prospect. *Sociological Methods & Research*, 29, 3-76.
- ABDEL-RAHMAN, M., STOCKTON, D., RACHET, B., HAKULINEN, T. & COLEMAN, M. P. 2009. What if cancer survival in Britain were the same as in Europe: how many deaths are avoidable? *Br J Cancer*, 101 Suppl 2, S115-24.
- ABEL, G. A., SHELTON, J., JOHNSON, S., ELLISS-BROOKES, L. & LYRATZOPOULOS, G. 2015. Cancer-specific variation in emergency presentation by sex, age and deprivation across 27 common and rarer cancers. *Br J Cancer*, 112 Suppl 1, S129-36.
- AGRESTI, A. 2002. *Categorical Data Analysis*.
- AISENBREY, S. & FASANG, A. E. 2010. New Life for Old Ideas: The "Second Wave" of Sequence Analysis Bringing the "Course" Back Into the Life Course. *Sociological Methods & Research*, 38, 420-462.
- AN DER HEIDEN, W. & HAFNER, H. 2015. Investigating the long-term course of schizophrenia by sequence analysis. *Psychiatry Res*, 228, 551-9.
- ARKSEY, H. & O'MALLEY, L. 2005. Scoping studies: towards a methodological framework. *International Journal of Social Research Methodology*, 8, 19-32.
- ARNOLD, M., RUTHERFORD, M. J., BARDOT, A., FERLAY, J., ANDERSSON, T. M. L., MYKLEBUST, T. Å., TERVONEN, H., THURSFIELD, V., RANSOM, D., SHACK, L., WOODS, R. R., TURNER, D., LEONFELLNER, S., RYAN, S., SAINT-JACQUES, N., DE, P., MCCLURE, C., RAMANAKUMAR, A. V., STUART-PANKO, H., ENGHOLM, G., WALSH, P. M., JACKSON, C., VERNON, S., MORGAN, E., GAVIN, A., MORRISON, D. S., HUWS, D. W., PORTER, G., BUTLER, J., BRYANT, H., CURROW, D. C., HIOM, S., PARKIN, D. M., SASIENI, P., LAMBERT, P. C., MØLLER, B., SOERJOMATARAM, I. & BRAY, F. 2019. Progress in cancer survival, mortality, and incidence in seven high-income countries 1995–2014 (ICBP SURVMARK-2): a population-based study. *The Lancet Oncology*, 20, 1493-1505.
- BACH, P. B., MIRKIN, J. N., OLIVER, T. K., AZZOLI, C. G., BERRY, D. A., BRAWLEY, O. W., BYERS, T., COLDITZ, G. A., GOULD, M. K., JETT, J. R., SABICHI, A. L., SMITH-BINDMAN, R., WOOD, D. E., QASEEM, A. & DETTERBECK, F. C. 2012. Benefits and harms of CT screening for lung cancer: a systematic review. *JAMA*, 307, 2418-29.
- BALATA, H., HARVEY, J., BARBER, P. V., COLLIGAN, D., DUERDEN, R., ELTON, P., EVISON, M., GREAVES, M., HOWELLS, J., IRION, K., KARUNARATNE, D., MELLOR, S., NEWTON, T., SAWYER, R., SHARMAN, A., SMITH, E., TAYLOR, B., TAYLOR, S., TONGE, J., WALSHAM, A., WHITTAKER, J., VESTBO, J., BOOTON, R. & CROSBIE, P. A. 2020. Spirometry performed as part of the Manchester community-based lung cancer screening programme detects a high prevalence of airflow obstruction in individuals without a prior diagnosis of COPD. *Thorax*, 75, 655-660.
- BANDA, J. M., SENEVIRATNE, M., HERNANDEZ-BOUSSARD, T. & SHAH, N. H. 2018. Advances in Electronic Phenotyping: From Rule-Based Definitions to Machine Learning Models. *Annu*

List of References

- Rev Biomed Data Sci*, 1, 53-68.
- BARTLETT, M. S. 1937. Properties of sufficiency and statistical tests. *Proceedings of the Royal Society of London. Series A - Mathematical and Physical Sciences*, 160, 268-282.
- BENSON, R., GLASER, K., CORNA, L. M., PLATTS, L. G., DI GESSA, G., WORTS, D., PRICE, D., MCDONOUGH, P. & SACKER, A. 2017. Do work and family care histories predict health in older women? *Eur J Public Health*, 27, 1010-1015.
- BERGLUND, A., LAMBE, M., LUCHTENBORG, M., LINKLATER, K., PEAKE, M. D., HOLMBERG, L. & MOLLER, H. 2012. Social differences in lung cancer management and survival in South East England: a cohort study. *BMJ Open*, 2.
- BI, Q., GOODMAN, K. E., KAMINSKY, J. & LESSLER, J. 2019. What is Machine Learning? A Primer for the Epidemiologist. *Am J Epidemiol*, 188, 2222-2239.
- BJERAGER, M., PALSHOF, T., DAHL, R., VEDSTED, P. & OLESEN, F. 2006. Delay in diagnosis of lung cancer in general practice. *Br J Gen Pract*, 56, 863-8.
- BLAKELY, T., BARENDREGT, J. J., FOSTER, R. H., HILL, S., ATKINSON, J., SARFATI, D. & EDWARDS, R. 2013. The association of active smoking with multiple cancers: national census-cancer registry cohorts with quantitative bias analysis. *Cancer Causes Control*, 24, 1243-55.
- BLANCHARD, P. 2011. Sequence Analysis for Political Science. *Committee on Concepts and Methods Working Paper Series*. 32. [Online].
- BLANDIN KNIGHT, S., CROSBIE, P. A., BALATA, H., CHUDZIAK, J., HUSSELL, T. & DIVE, C. 2017. Progress and prospects of early detection in lung cancer. *Open Biol*, 7.
- BOWER, J. K., BOLLINGER, C. E., FORAKER, R. E., HOOD, D. B., SHOBEH, A. B. & LAI, A. M. 2017. Active Use of Electronic Health Records (EHRs) and Personal Health Records (PHRs) for Epidemiologic Research: Sample Representativeness and Nonresponse Bias in a Study of Women During Pregnancy. *EGEMS (Wash DC)*, 5, 1263.
- BRADLEY, S. H., ABRAHAM, S., CALLISTER, M. E., GRICE, A., HAMILTON, W. T., LOPEZ, R. R., SHINKINS, B. & NEAL, R. D. 2019a. Sensitivity of chest X-ray for detecting lung cancer in people presenting with symptoms: a systematic review. *Br J Gen Pract*, 69, e827-e835.
- BRADLEY, S. H., KENNEDY, M. P. T. & NEAL, R. D. 2019b. Recognising Lung Cancer in Primary Care. *Adv Ther*, 36, 19-30.
- BRAIN, K., CARTER, B., LIFFORD, K. J., BURKE, O., DEVARAJ, A., BALDWIN, D. R., DUFFY, S. & FIELD, J. K. 2017. Impact of low-dose CT screening on smoking cessation among high-risk participants in the UK Lung Cancer Screening Trial. *Thorax*, 72, 912-918.
- BRINDLE, L., POPE, C., CORNER, J., LEYDON, G. & BANERJEE, A. 2012. Eliciting symptoms interpreted as normal by patients with early-stage lung cancer: could GP elicitation of normalised symptoms reduce delay in diagnosis? Cross-sectional interview study. *BMJ Open*, 2.
- BRISCOE, M. E. 1987. Why do people go to the doctor? Sex differences in the correlates of GP consultation. *Social Science & Medicine*, 25, 507-513.
- BROEKHUIZEN, B. D., SACHS, A. P. & VERHEIJ, T. J. 2012. COPD in primary care: from episodic to continual management. *Br J Gen Pract*, 62, 60-1.
- BRZINSKY-FAY, C., KOHLER, U. & LUNIAK, M. 2006. Sequence analysis with Stata. *Stata Journal*, 6, 435-460.
- BUCCHERI, R. K. & SHARIFI, C. 2017. Critical Appraisal Tools and Reporting Guidelines for Evidence-Based Practice. *Worldviews Evid Based Nurs*, 14, 463-472.
- CAMPBELL, S. M. & ROLAND, M. O. 1996. Why do people consult the doctor? *Fam Pract*, 13, 75-83.
- CANCER RESEARCH UK. 2021. *Lung cancer statistics* [Online]. Available: <https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/lung-cancer> [Accessed].
- CASSIDY, A., MYLES, J. P., VAN TONGEREN, M., PAGE, R. D., LILOGLOU, T., DUFFY, S. W. & FIELD, J. K. 2008. The LLP risk model: an individual risk prediction model for lung cancer. *Br J Cancer*, 98, 270-6.
- CASTALDI, P. J., BENET, M., PETERSEN, H., RAFAELS, N., FINIGAN, J., PAOLETTI, M., MARIKE BOEZEN, H., VONK, J. M., BOWLER, R., PISTOLESI, M., PUHAN, M. A., ANTO, J., WAUTERS, E., LAMBRECHTS, D., JANSSENS, W., BIGAZZI, F., CAMICIOTTOLI, G., CHO, M. H., HERSH, C. P., BARNES, K., RENNARD, S., BOORGULA, M. P., DY, J., HANSEL, N. N., CRAPO, J. D., TESFAIGZI,

- Y., AGUSTI, A., SILVERMAN, E. K. & GARCIA-AYMERICH, J. 2017. Do COPD subtypes really exist? COPD heterogeneity and clustering in 10 independent cohorts. *Thorax*, 72, 998-1006.
- CHAPPLE, A., ZIEBLAND, S. & MCPHERSON, A. 2004. Stigma, shame, and blame experienced by patients with lung cancer: qualitative study. *BMJ*, 328, 1470.
- CHATWIN, J. & SANDERS, C. 2013. The influence of social factors on help-seeking for people with lung cancer. *Eur J Cancer Care (Engl)*, 22, 709-13.
- CLARK, M. E., YOUNG, B., BEDFORD, L. E., DAS NAIR, R., ROBERTSON, J. F. R., VEDHARA, K., SULLIVAN, F., MAIR, F. S., SCHEMBRI, S., LITTLEFORD, R. C. & KENDRICK, D. 2018. Lung cancer screening: does pulmonary nodule detection affect a range of smoking behaviours? *J Public Health (Oxf)*.
- CLARKE, M., OXMAN, A., PAULSEN, E., HIGGINS, J. & GREEN, S. 2011. Appendix A: Guide to the contents of a Cochrane Methodology protocol and review. In: Higgins JP, Green S, eds. *Cochrane Handbook for Systematic Reviews of Interventions (Version 5.1.0 edition)*, The Cochrane Collaboration.
- COLEMAN, M. P., FORMAN, D., BRYANT, H., BUTLER, J., RACHET, B., MARINGE, C., NUR, U., TRACEY, E., COORY, M., HATCHER, J., MCGAHAN, C. E., TURNER, D., MARRETT, L., GJERSTORFF, M. L., JOHANNESSEN, T. B., ADOLFSSON, J., LAMBE, M., LAWRENCE, G., MEECHAN, D., MORRIS, E. J., MIDDLETON, R., STEWARD, J., RICHARDS, M. A. & GROUP, I. M. W. 2011. Cancer survival in Australia, Canada, Denmark, Norway, Sweden, and the UK, 1995-2007 (the International Cancer Benchmarking Partnership): an analysis of population-based cancer registry data. *Lancet*, 377, 127-38.
- CORNER, J., HOPKINSON, J. & ROFFE, L. 2006. Experience of health changes and reasons for delay in seeking care: a UK study of the months prior to the diagnosis of lung cancer. *Soc Sci Med*, 62, 1381-91.
- CRANE, M., SCOTT, N., O'HARA, B. J., ARANDA, S., LAFONTAINE, M., STACEY, I., VARLOW, M. & CURROW, D. 2016. Knowledge of the signs and symptoms and risk factors of lung cancer in Australia: mixed methods study. *BMC Public Health*, 16, 508.
- CROSBIE, P. A., BALATA, H., EVISON, M., ATACK, M., BAYLISS-BRIDEAUX, V., COLLIGAN, D., DUERDEN, R., EAGLESFIELD, J., EDWARDS, T., ELTON, P., FOSTER, J., GREAVES, M., HAYLER, G., HIGGINS, C., HOWELLS, J., IRION, K., KARUNARATNE, D., KELLY, J., KING, Z., LYONS, J., MANSON, S., MELLOR, S., MILLER, D., MYERSCOUGH, A., NEWTON, T., O'LEARY, M., PEARSON, R., PICKFORD, J., SAWYER, R., SCREATON, N. J., SHARMAN, A., SIMMONS, M., SMITH, E., TAYLOR, B., TAYLOR, S., WALSHAM, A., WATTS, A., WHITTAKER, J., YARNELL, L., THRELFALL, A., BARBER, P. V., TONGE, J. & BOOTON, R. 2019. Second round results from the Manchester 'Lung Health Check' community-based targeted lung cancer screening pilot. *Thorax*, 74, 700-704.
- DARAK, S., MILLS, M., KULKARNI, V., KULKARNI, S., HUTTER, I. & JANSSEN, F. 2015. Trajectories of Childbearing among HIV Infected Indian Women: A Sequence Analysis Approach. *Plos One*, 10, e0124537-e0124537.
- DE ANGELIS, R., SANT, M., COLEMAN, M. P., FRANCISCI, S., BAILI, P., PIERANNUNZIO, D., TRAMA, A., VISSER, O., BRENNER, H., ARDANAZ, E., BIELSKA-LASOTA, M., ENGHOLM, G., NENNECKE, A., SIESLING, S., BERRINO, F., CAPOCACCIA, R. & GROUP, E.-W. 2014. Cancer survival in Europe 1999-2007 by country and age: results of EUROCORE-5-a population-based study. *Lancet Oncol*, 15, 23-34.
- DEL GIUDICE, M. E., YOUNG, S. M., VELLA, E. T., ASH, M., BANSAL, P., ROBINSON, A., SKRASTINS, R., UNG, Y., ZELDIN, R. & LEVITT, C. 2014. Guideline for referral of patients with suspected lung cancer by family physicians and other primary care providers. *Can Fam Physician*, 60, 711-6, e376-82.
- DEPARTMENT FOR COMMUNITIES AND LOCAL GOVERNMENT 2015. The English indices of deprivation 2015. In: DEPARTMENT FOR COMMUNITIES AND LOCAL GOVERNMENT (ed.). London.
- DEPARTMENT OF HEALTH 2000. The NHS Cancer plan: a plan for investment, a plan for reform. London.
- DEPARTMENT OF HEALTH 2011a. The Likely Impact of Earlier Diagnosis of Cancer on Costs and

List of References

Benefits to the NHS.

- DEPARTMENT OF HEALTH 2011b. Review of waiting times standards. London.
- DI GIROLAMO, C., WALTERS, S., BENITEZ MAJANO, S., RACHET, B., COLEMAN, M. P., NJAGI, E. N. & MORRIS, M. 2018. Characteristics of patients with missing information on stage: a population-based study of patients diagnosed with colon, lung or breast cancer in England in 2013. *BMC Cancer*, 18, 492.
- DIXON-WOODS, M., CAVERS, D., AGARWAL, S., ANNANDALE, E., ARTHUR, A., HARVEY, J., HSU, R., KATBAMNA, S., OLSEN, R., SMITH, L., RILEY, R. & SUTTON, A. J. 2006. Conducting a critical interpretive synthesis of the literature on access to healthcare by vulnerable groups. *BMC Med Res Methodol*, 6, 35.
- DLOUHY, K. & BIEMANN, T. 2015. Optimal matching analysis in career research: A review and some best-practice recommendations. *Journal of Vocational Behavior*, 90, 163-173.
- DOBSON, C. M., RUSSELL, A. J. & RUBIN, G. P. 2014. Patient delay in cancer diagnosis: what do we really mean and can we be more specific? *BMC Health Serv Res*, 14, 387.
- DURHAM, A. L. & ADCOCK, I. M. 2015. The relationship between COPD and lung cancer. *Lung Cancer*, 90, 121-7.
- DURRANT, G. B., MASLOVSKAYA, O. & SMITH, P. W. F. 2018. Investigating call record data using sequence analysis to inform adaptive survey designs. *International Journal of Social Research Methodology*, 22, 37-54.
- EDEN, M., HARRISON, S., GRIFFIN, M., LAMBE, M., PETERSSON, D., GAVIN, A., BREWSTER, D. H., LIN, Y., JOHANNESSEN, T. B., MILNE, R. L., FARRUGIA, H., NISHRI, D., KING, M. J., HUWS, D. W., WARLOW, J., TURNER, D., EARLE, C. C., PEAKE, M. & RASHBASS, J. 2019. Impact of variation in cancer registration practice on observed international cancer survival differences between International Cancer Benchmarking Partnership (ICBP) jurisdictions. *Cancer Epidemiol*, 58, 184-192.
- EDWARDS, B. K., NOONE, A. M., MARIOTTO, A. B., SIMARD, E. P., BOSCOE, F. P., HENLEY, S. J., JEMAL, A., CHO, H., ANDERSON, R. N., KOHLER, B. A., EHEMAN, C. R. & WARD, E. M. 2014. Annual Report to the Nation on the status of cancer, 1975-2010, featuring prevalence of comorbidity and impact on survival among persons with lung, colorectal, breast, or prostate cancer. *Cancer*, 120, 1290-314.
- ELDER, G. H. J. 1985. *Perspectives on the life course.*, Ithaca, New York, Cornell University Press.
- ELLIOTT, A. M., MCATEER, A. & HANNAFORD, P. C. 2011. Revisiting the symptom iceberg in today's primary care: results from a UK population survey. *BMC Fam Pract*, 12, 16.
- ELLISS-BROOKES, L., MCPHAIL, S., IVES, A., GREENSLADE, M., SHELTON, J., HIOM, S. & RICHARDS, M. 2012. Routes to diagnosis for cancer - determining the patient journey using multiple routine data sets. *British Journal of Cancer*, 107, 1220-1226.
- ELNEGAARD, S., ANDERSEN, R. S., PEDERSEN, A. F., LARSEN, P. V., SONDERGAARD, J., RASMUSSEN, S., BALASUBRAMANIAM, K., SVENDSEN, R. P., VEDSTED, P. & JARBOL, D. E. 2015. Self-reported symptoms and healthcare seeking in the general population--exploring "The Symptom Iceberg". *BMC Public Health*, 15, 685.
- EMERY, J. D., MURRAY, S. R., WALTER, F. M., MARTIN, A., GOODALL, S., MAZZA, D., HABGOOD, E., KUTZER, Y., BARNES, D. J. & MURCHIE, P. 2019. The Chest Australia Trial: a randomised controlled trial of an intervention to increase consultation rates in smokers at risk of lung cancer. *Thorax*, 74, 362-370.
- ESTIVILL-CASTRO, V. 2002. Why so many clustering algorithms. *ACM SIGKDD Explorations Newsletter*, 4, 65-75.
- EVANS, J., ZIEBLAND, S., MACARTNEY, J. I., BANKHEAD, C. R., ROSE, P. W. & NICHOLSON, B. D. 2018. GPs' understanding and practice of safety netting for potential cancer presentations: a qualitative study in primary care. *Br J Gen Pract*, 68, e505-e511.
- EXARCHAKOU, A., RACHET, B., BELOT, A., MARINGE, C. & COLEMAN, M. P. 2018. Impact of national cancer policies on cancer survival trends and socioeconomic inequalities in England, 1996-2013: population based study. *BMJ*, 360, k764.
- FEDER, S. L. 2018. Data Quality in Electronic Health Records Research: Quality Domains and Assessment Methods. *West J Nurs Res*, 40, 753-766.

- FEINSTEIN, A. R. 1970. The Pre-Therapeutic Classification of Co-Morbidity in Chronic Disease. *J Chronic Dis*, 23, 455-68.
- FIELD, J. K., DUFFY, S. W., BALDWIN, D. R., WHYNES, D. K., DEVARAJ, A., BRAIN, K. E., EISEN, T., GOSNEY, J., GREEN, B. A., HOLEMANS, J. A., KAVANAGH, T., KERR, K. M., LEDSON, M., LIFFORD, K. J., MCRONALD, F. E., NAIR, A., PAGE, R. D., PARMAR, M. K., RASSL, D. M., RINTOUL, R. C., SCREATON, N. J., WALD, N. J., WELLER, D., WILLIAMSON, P. R., YADEGARFAR, G. & HANSELL, D. M. 2016. UK Lung Cancer RCT Pilot Screening Trial: baseline findings from the screening arm provide evidence for the potential implementation of lung cancer screening. *Thorax*, 71, 161-70.
- FORBES, L. J., WARBURTON, F., RICHARDS, M. A. & RAMIREZ, A. J. 2014. Risk factors for delay in symptomatic presentation: a survey of cancer patients. *Br J Cancer*, 111, 581-8.
- FORREST, L. F., ADAMS, J., RUBIN, G. & WHITE, M. 2015. The role of receipt and timeliness of treatment in socioeconomic inequalities in lung cancer survival: population-based, data-linkage study. *Thorax*, 70, 138-45.
- FORREST, L. F., ADAMS, J., WAREHAM, H., RUBIN, G. & WHITE, M. 2013. Socioeconomic inequalities in lung cancer treatment: systematic review and meta-analysis. *PLoS Med*, 10, e1001376.
- FORREST, L. F., ADAMS, J., WHITE, M. & RUBIN, G. 2014a. Factors associated with timeliness of post-primary care referral, diagnosis and treatment for lung cancer: population-based, data-linkage study. *Br J Cancer*, 111, 1843-51.
- FORREST, L. F., SOWDEN, S., RUBIN, G., WHITE, M. & ADAMS, J. 2017. Socio-economic inequalities in stage at diagnosis, and in time intervals on the lung cancer pathway from first symptom to treatment: systematic review and meta-analysis. *Thorax*, 72, 430-436.
- FORREST, L. F., WHITE, M., RUBIN, G. & ADAMS, J. 2014b. The role of patient, tumour and system factors in socioeconomic inequalities in lung cancer treatment: population-based study. *Br J Cancer*, 111, 608-18.
- FOWLKES, E. B. & MALLOWS, C. L. 1983. A Method for Comparing Two Hierarchical Clusterings. *Journal of the American Statistical Association*, 78, 553-569.
- FRIEDEMANN SMITH, C., WHITAKER, K. L., WINSTANLEY, K. & WARDLE, J. 2016. Smokers are less likely than non-smokers to seek help for a lung cancer 'alarm' symptom. *Thorax*, 71, 659-61.
- FROHLICH, H., BALLING, R., BEERENWINKEL, N., KOHLBACHER, O., KUMAR, S., LENGAUER, T., MAATHUIS, M. H., MOREAU, Y., MURPHY, S. A., PRZYTYCKA, T. M., REBHAN, M., ROST, H., SCHUPPERT, A., SCHWAB, M., SPANG, R., STEKHOVEN, D., SUN, J., WEBER, A., ZIEMEK, D. & ZUPAN, B. 2018. From hype to reality: data science enabling personalized medicine. *BMC Med*, 16, 150.
- GABADINHO, A., RITSCHARD, G., MULLER, N. S. & STUDER, M. 2011a. Analyzing and Visualizing State Sequences in R with TraMineR. *Journal of Statistical Software*, 40, 1-37.
- GABADINHO, A., RITSCHARD, G., STUDER, M. & MÜLLER, N. S. 2011b. *Extracting and Rendering Representative Sequences*, Springer-Verlag.
- GALOBARDES, B., SHAW, M., LAWLOR, D. A., LYNCH, J. W. & DAVEY SMITH, G. 2006a. Indicators of socioeconomic position (part 1). *J Epidemiol Community Health*, 60, 7-12.
- GALOBARDES, B., SHAW, M., LAWLOR, D. A., LYNCH, J. W. & DAVEY SMITH, G. 2006b. Indicators of socioeconomic position (part 2). *J Epidemiol Community Health*, 60, 95-101.
- GAUTHIER, J.-A., WIDMER, E. D., BUCHER, P. & NOTREDAME, C. 2009. How Much Does It Cost? *Sociological Methods & Research*, 38, 197-231.
- GAUTHIER, J. A., WIDMER, E. D., BUCHER, P. & NOTREDAME, C. 2010. Multichannel Sequence Analysis Applied to Social Science Data. *Sociological Methodology*, Vol 40, 40, 1-38.
- GREEN, T., ATKIN, K. & MACLEOD, U. 2015. Cancer detection in primary care: insights from general practitioners. *Br J Cancer*, 112 Suppl 1, S41-9.
- GROSE, D., DEVEREUX, G. & MILROY, R. 2011. Comorbidity in lung cancer: important but neglected. a review of the current literature. *Clin Lung Cancer*, 12, 207-11.
- GROSE, D., MORRISON, D. S., DEVEREUX, G., JONES, R., SHARMA, D., SELBY, C., DOCHERTY, K., MCINTOSH, D., LOUDEN, G., NICOLSON, M., MCMILLAN, D. C., MILROY, R. & SCOTTISH LUNG CANCER, F. 2014. Comorbidities in lung cancer: prevalence, severity and links with socioeconomic status and treatment. *Postgrad Med J*, 90, 305-10.

List of References

- GULDBRANDT, L. M., MOLLER, H., JAKOBSEN, E. & VEDSTED, P. 2017. General practice consultations, diagnostic investigations, and prescriptions in the year preceding a lung cancer diagnosis. *Cancer Med*, 6, 79-88.
- GUYATT, G. H. 1995. Users' Guides to the Medical Literature. *Jama*, 274.
- HALPIN, B. 2003. Tracks Through Time and Continuous Processes: Transitions, Sequences, and Social Structure. 'Frontiers in Social and Economic Mobility' Conference. Cornell University, New York.
- HALPIN, B. 2010. Optimal matching analysis and life-course data: The importance of duration. *Sociological Methods & Research*, 38, 365-388.
- HALPIN, B. 2012. Sequence analysis of life-course data: a comparison of distance measures. *Department of Sociology, University of Limerick* [Online]. Available: <http://www.ul.ie/sociology/pubs/wp2012-02.pdf>.
- HALPIN, B. 2017. SADI: Sequence Analysis Tools for Stata. *The Stata Journal*, (in press).
- HAMILTON, W., PETERS, T. J., ROUND, A. & SHARP, D. 2005. What are the clinical features of lung cancer before the diagnosis is made? A population based case-control study. *Thorax*, 60, 1059-65.
- HAMILTON, W. & SHARP, D. 2004. Diagnosis of lung cancer in primary care: a structured review. *Fam Pract*, 21, 605-11.
- HAMPSHIRE HEALTH RECORD. 2017. Available: <http://www.hantshealthrecord.nhs.uk/> [Accessed June 3 2017].
- HANNAFORD, P. C., THORNTON, A. J., MURCHIE, P., WHITAKER, K. L., ADAM, R. & ELLIOTT, A. M. 2020. Patterns of symptoms possibly indicative of cancer and associated help-seeking behaviour in a large sample of United Kingdom residents-The USEFUL study. *PLoS One*, 15, e0228033.
- HANNAY, D. 1980. The 'iceberg' of illness and 'trivial' consultations. *J R Coll Gen Pract.*, 30, 551-554.
- HART, C. L., HOLE, D. J., GILLIS, C. R., SMITH, G. D., WATT, G. C. & HAWTHORNE, V. M. 2001. Social class differences in lung cancer mortality: risk factor explanations using two Scottish cohort studies. *Int J Epidemiol*, 30, 268-74.
- HERBERT, A., WIJLAARS, L., ZYLBERSZTEJN, A., CROMWELL, D. & HARDELID, P. 2017. Data Resource Profile: Hospital Episode Statistics Admitted Patient Care (HES APC). *Int J Epidemiol*, 46, 1093-1093i.
- HERRETT, E., GALLAGHER, A. M., BHASKARAN, K., FORBES, H., MATHUR, R., VAN STAA, T. & SMEETH, L. 2015. Data Resource Profile: Clinical Practice Research Datalink (CPRD). *Int J Epidemiol*, 44, 827-36.
- HIGGINS, J. P. & THOMPSON, S. G. 2002. Quantifying heterogeneity in a meta-analysis. *Stat Med*, 21, 1539-58.
- HILBE, J. M. 2011. *Negative Binomial Regression*.
- HIOM, S. C. 2015. Diagnosing cancer earlier: reviewing the evidence for improving cancer survival. *Br J Cancer*, 112 Suppl 1, S1-5.
- HIPPISLEY-COX, J. & COUPLAND, C. 2011. Identifying patients with suspected lung cancer in primary care: derivation and validation of an algorithm. *Br J Gen Pract*, 61, e715-23.
- HIPPISLEY-COX, J. & COUPLAND, C. 2013a. Symptoms and risk factors to identify men with suspected cancer in primary care: derivation and validation of an algorithm. *Br J Gen Pract*, 63, e1-10.
- HIPPISLEY-COX, J. & COUPLAND, C. 2013b. Symptoms and risk factors to identify women with suspected cancer in primary care: derivation and validation of an algorithm. *Br J Gen Pract*, 63, e11-21.
- HOLDEN, R. B. 2010. Face validity. In: WEINER, I. B. & CRAIGHEAD, W. E. (eds.) *The Corsini Encyclopedia of Psychology (4th ed.)*. Hoboken, New Jersey: Wiley.
- HOLLISTER, M. 2009. Is Optimal Matching Suboptimal? *Sociological Methods & Research*, 38, 235-264.
- HOUGHAM, G. W., HAM, S. A., RUHNKE, G. W., SCHULWOLF, E., AUERBACH, A. D., SCHNIPPER, J. L., KABOLI, P. J., WETTERNECK, T. B., GONZALEZ, D., ARORA, V. M. & MELTZER, D. O. 2014. Sequence patterns in the resolution of clinical instabilities in community-acquired pneumonia and association with outcomes. *J Gen Intern Med*, 29, 563-71.
- HUANG, R., WEI, Y., HUNG, R. J., LIU, G., SU, L., ZHANG, R., ZONG, X., ZHANG, Z.-F., MORGENSTERN,

- H., BRÜSKE, I., HEINRICH, J., HONG, Y.-C., KIM, J. H., COTE, M., WENZLAFF, A., SCHWARTZ, A. G., STUCKER, I., MCLAUGHLIN, J., MARCUS, M. W., DAVIES, M. P. A., LILOGLOU, T., FIELD, J. K., MATSUO, K., BARNETT, M., THORNQUIST, M., GOODMAN, G., WANG, Y., CHEN, S., YANG, P., DUELL, E. J., ANDREW, A. S., LAZARUS, P., MUSCAT, J., WOLL, P., HORSMAN, J., DAWN TEARE, M., FLUGELMAN, A., RENNERT, G., ZHANG, Y., BRENNER, H., STEGMAIER, C., VAN DER HEIJDEN, E. H. F. M., ABEN, K., KIEMENEY, L., BARROS-DIOS, J., PÉREZ-RÍOS, M., RUANO-RAVINA, A., CAPORASO, N. E., BERTAZZI, P. A., LANDI, M. T., DAI, J., SHEN, H., FERNANDEZ-TARDON, G., RODRIGUEZ-SUAREZ, M., TARDON, A. & CHRISTIANI, D. C. 2015. Associated Links Among Smoking, Chronic Obstructive Pulmonary Disease, and Small Cell Lung Cancer: A Pooled Analysis in the International Lung Cancer Consortium. *EBioMedicine*, 2, 1677-1685.
- INSTITUTE OF HEALTH ECONOMICS. 2016. *IHE Quality Appraisal Checklist for Case Series Studies* [Online]. Alberta, Canada. Available: <https://www.ihe.ca/publications/ihe-quality-appraisal-checklist-for-case-series-studies> [Accessed October 1 2019].
- ISLAM, K. M., JIANG, X., ANGGONDOWATI, T., LIN, G. & GANTI, A. K. 2015. Comorbidity and Survival in Lung Cancer Patients. *Cancer Epidemiol Biomarkers Prev*, 24, 1079-85.
- JACOBSEN, M. M., SILVERSTEIN, S. C., QUINN, M., WATERSTON, L. B., THOMAS, C. A., BENNEYAN, J. C. & HAN, P. K. J. 2017. Timeliness of access to lung cancer diagnosis and treatment: A scoping literature review. *Lung Cancer*, 112, 156-164.
- JAKOBSEN, J. C., GLUUD, C., WETTERSLEV, J. & WINKEL, P. 2017. When and how should multiple imputation be used for handling missing data in randomised clinical trials - a practical guide with flowcharts. *BMC Med Res Methodol*, 17, 162.
- JEFFERSON, T., RUDIN, M., BRODNEY FOLSE, S. & DAVIDOFF, F. 2007. Editorial peer review for improving the quality of reports of biomedical studies. *Cochrane Database Syst Rev*, MR000016.
- JEMAL, A., THUN, M. J., RIES, L. A. G., HOWE, H. L., WEIR, H. K., CENTER, M. M., WARD, E., WU, X. C., EHEMAN, C., ANDERSON, R., AJANI, U. A., KOHLER, B. & EDWARDS, B. K. 2008. Annual Report to the Nation on the Status of Cancer, 1975-2005, Featuring Trends in Lung Cancer, Tobacco Use, and Tobacco Control. *Jnci-Journal of the National Cancer Institute*, 100, 1672-1694.
- JEPPESEN, S. S., HANSEN, N. G., SCHYTTE, T., NIELSEN, M. & HANSEN, O. 2016. Comparison of survival of chronic obstructive pulmonary disease patients with or without a localized non-small cell lung cancer. *Lung Cancer*, 100, 90-95.
- JONAS, D. E., REULAND, D. S., REDDY, S. M., NAGLE, M., CLARK, S. D., WEBER, R. P., ENYIOHA, C., MALO, T. L., BRENNER, A. T., ARMSTRONG, C., COKER-SCHWIMMER, M., MIDDLETON, J. C., VOISIN, C. & HARRIS, R. P. 2021. Screening for Lung Cancer With Low-Dose Computed Tomography: Updated Evidence Report and Systematic Review for the US Preventive Services Task Force. *JAMA*, 325, 971-987.
- KAHN, M. G., CALLAHAN, T. J., BARNARD, J., BAUCK, A. E., BROWN, J., DAVIDSON, B. N., ESTIRI, H., GOERG, C., HOLVE, E., JOHNSON, S. G., LIAW, S. T., HAMILTON-LOPEZ, M., MEEKER, D., ONG, T. C., RYAN, P., SHANG, N., WEISKOPF, N. G., WENG, C., ZOZUS, M. N. & SCHILLING, L. 2016. A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data. *EGEMS (Wash DC)*, 4, 1244.
- KAUFMAN, L. & ROUSSEUW, P. 1990. *Finding groups in data. an introduction to cluster analysis*, New York, Wiley.
- KRUSKAL, J. 1983. *An Overview of Sequence Comparison*, Toronto, Canada, Addison-Wesley.
- KRUSKAL, W. H. & WALLIS, W. A. 1952. Use of Ranks in One-Criterion Variance Analysis. *Journal of the American Statistical Association*, 47, 583-621.
- LAST, J. M. 1963. The Iceberg "Completing the Clinical Picture" in General Practice. *The Lancet*, 282, 28-31.
- LAWLOR, D. A. 2019. Fifteen years of epidemiology in BMC Medicine. *BMC Med*, 17, 177.
- LE MEUR, N., GAO, F. & BAYAT, S. 2015. Mining care trajectories using health administrative information systems: the use of state sequence analysis to assess disparities in prenatal care consumption. *BMC Health Services Research*, 15, 200-200.
- LE MEUR, N., VIGNEAU, C., LEFORT, M., LEBBAH, S., JAIS, J. P., DAUGAS, E. & BAYAT, S. 2019.

List of References

- Categorical state sequence analysis and regression tree to identify determinants of care trajectory in chronic disease: Example of end-stage renal disease. *Stat Methods Med Res*, 28, 1731-1740.
- LESNARD, L. 2010. Setting Cost in Optimal Matching to Uncover Contemporaneous Socio-Temporal Patterns. *Sociological Methods & Research*, 38, 389-419.
- LESNARD, L. & KAN, M. Y. 2011. Investigating scheduling of work: a two-stage optimal matching analysis of workdays and workweeks. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 174, 349-368.
- LIAO, K. P., CAI, T., SAVOVA, G. K., MURPHY, S. N., KARLSON, E. W., ANANTHAKRISHNAN, A. N., GAINER, V. S., SHAW, S. Y., XIA, Z., SZOLOVITS, P., CHURCHILL, S. & KOHANE, I. 2015. Development of phenotype algorithms using electronic medical records and incorporating natural language processing. *BMJ*, 350, h1885.
- LIM, S., GAO, Q., STAZESKY, E., SINGH, T. P., HARRIS, T. G. & LEVANON SELIGSON, A. 2018. Impact of a New York City supportive housing program on Medicaid expenditure patterns among people with serious mental illness and chronic homelessness. *BMC Health Serv Res*, 18, 15.
- LONG, J. S. & FREESE, J. 2014. *Regression Models for Categorical Dependent Variables Using Stata (Third Edition)*, Stata Press.
- LYONS, R. A., JONES, K. H., JOHN, G., BROOKS, C. J., VERPLANCKE, J. P., FORD, D. V., BROWN, G. & LEAKE, K. 2009. The SAIL databank: linking multiple health and social care datasets. *BMC Med Inform Decis Mak*, 9, 3.
- LYRATZOPOULOS, G., ABEL, G. A., MCPHAIL, S., NEAL, R. D. & RUBIN, G. P. 2013. Measures of promptness of cancer diagnosis in primary care: secondary analysis of national audit data on patients with 18 common and rarer cancers. *Br J Cancer*, 108, 686-90.
- LYRATZOPOULOS, G., VEDSTED, P. & SINGH, H. 2015. Understanding missed opportunities for more timely diagnosis of cancer in symptomatic patients after presentation. *Br J Cancer*, 112 Suppl 1, S84-91.
- LYRATZOPOULOS, G., WARDLE, J. & RUBIN, G. 2014. Rethinking diagnostic delay in cancer: how difficult is the diagnosis? *BMJ*, 349, g7400.
- MACLEAN, R., JEFFREYS, M., IVES, A., JONES, T., VERNE, J. & BEN-SHLOMO, Y. 2015. Primary care characteristics and stage of cancer at diagnosis using data from the national cancer registration service, quality outcomes framework and general practice information. *BMC Cancer*, 15, 500.
- MACLEOD, U., MITCHELL, E. D., BURGESS, C., MACDONALD, S. & RAMIREZ, A. J. 2009. Risk factors for delayed presentation and referral of symptomatic cancer: evidence for common cancers. *Br J Cancer*, 101 Suppl 2, S92-S101.
- MARCUS, M. W., CHEN, Y., DUFFY, S. W. & FIELD, J. K. 2015. Impact of comorbidity on lung cancer mortality - a report from the Liverpool Lung Project. *Oncol Lett*, 9, 1902-1906.
- MARINGE, C., PASHAYAN, N., RUBIO, F. J., PLOUBIDIS, G., DUFFY, S. W., RACHET, B. & RAINE, R. 2018. Trends in lung cancer emergency presentation in England, 2006-2013: is there a pattern by general practice? *BMC Cancer*, 18, 615.
- MATTIJSEN, L. & PAVLOPOULOS, D. 2017. A multichannel typology of nonstandard employment careers. Available: http://www.arbeidsconferentie.nl/uploads/submission/document_1/197/Mattijssen_Pavlopoulos_2017_-_A_multichannel_typology_of_non-standard_employment_careers.pdf.
- MCCUTCHAN, G., HISCOCK, J., HOOD, K., MURCHIE, P., NEAL, R. D., NEWTON, G., THOMAS, S., THOMAS, A. M. & BRAIN, K. 2019. Engaging high-risk groups in early lung cancer diagnosis: a qualitative study of symptom presentation and intervention preferences among the UK's most deprived communities. *BMJ Open*, 9, e025902.
- MCPHAIL, S., BARBIERE, J. M., GREENBERG, D. C., WRIGHT, K. A. & LYRATZOPOULOS, G. 2011. Population-based trends in use of surgery for non-small cell lung cancer in a UK region, 1995-2006. *Thorax*, 66, 453-5.
- MCPHAIL, S., ELLISS-BROOKES, L., SHELTON, J., IVES, A., GREENSLADE, M., VERNON, S., MORRIS, E. J. & RICHARDS, M. 2013. Emergency presentation of cancer and short-term mortality. *Br J Cancer*, 109, 2027-34.

- MCPHAIL, S., JOHNSON, S., GREENBERG, D., PEAKE, M. & ROUS, B. 2015. Stage at diagnosis and early mortality from cancer in England. *Br J Cancer*, 112 Suppl 1, S108-15.
- MCRONALD, F. E., YADEGARFAR, G., BALDWIN, D. R., DEVARAJ, A., BRAIN, K. E., EISEN, T., HOLEMANS, J. A., LEDSON, M., SCREATON, N., RINTOUL, R. C., HANDS, C. J., LIFFORD, K., WHYNES, D., KERR, K. M., PAGE, R., PARMAR, M., WALD, N., WELLER, D., WILLIAMSON, P. R., MYLES, J., HANSELL, D. M., DUFFY, S. W. & FIELD, J. K. 2014. The UK Lung Screen (UKLS): demographic profile of first 88,897 approaches provides recommendations for population screening. *Cancer Prev Res (Phila)*, 7, 362-71.
- METS, O. M., BUCKENS, C. F., ZANEN, P., ISGUM, I., VAN GINNEKEN, B., PROKOP, M., GIETEMA, H. A., LAMMERS, J. W., Vliegenthart, R., OUDKERK, M., VAN KLAVEREN, R. J., DE KONING, H. J., MALI, W. P. & DE JONG, P. A. 2011. Identification of chronic obstructive pulmonary disease in lung cancer screening computed tomographic scans. *JAMA*, 306, 1775-81.
- MINICOZZI, P., INNOS, K., SANCHEZ, M. J., TRAMA, A., WALSH, P. M., MARCOS-GRAGERA, R., DIMITROVA, N., BOTTA, L., VISSER, O., ROSSI, S., TAVILLA, A., SANT, M. & GRP, E.-W. 2017. Quality analysis of population-based information on cancer stage at diagnosis across Europe, with presentation of stage-specific cancer survival estimates: A EURO CARE-5 study. *European Journal of Cancer*, 84, 335-353.
- MITCHELL, E. D., PICKWELL-SMITH, B. & MACLEOD, U. 2015. Risk factors for emergency presentation with lung and colorectal cancers: a systematic review. *BMJ Open*, 5, e006965.
- MITCHELL, E. D., RUBIN, G. & MACLEOD, U. 2013. Understanding diagnosis of lung cancer in primary care: qualitative synthesis of significant event audit reports. *Br J Gen Pract*, 63, e37-46.
- MOFFAT, J., BENTLEY, A., IRONMONGER, L., BOUGHEY, A., RADFORD, G. & DUFFY, S. 2015. The impact of national cancer awareness campaigns for bowel and lung cancer symptoms on sociodemographic inequalities in immediate key symptom awareness and GP attendances. *Br J Cancer*, 112 Suppl 1, S14-21.
- MOHER, D., LIBERATI, A., TETZLAFF, J., ALTMAN, D. G. & GROUP, P. 2009. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *BMJ*, 339, b2535.
- MOLLER, H., GILDEA, C., MEECHAN, D., RUBIN, G., ROUND, T. & VEDSTED, P. 2015. Use of the English urgent referral pathway for suspected cancer and mortality in patients with cancer: cohort study. *BMJ*, 351, h5102.
- MOOLA, S., MUNN, Z., TUFANARU, C., AROMATARIS, E., SEARS, K., SFETCU, R., CURRIE, M., QURESHI, R., MATTIS, P., LISY, K. & MU, P.-F. 2017. *Chapter 7: Systematic reviews of etiology and risk. In: Aromataris E, Munn Z (Editors). Joanna Briggs Institute Reviewer's Manual, Adelaide, Australia.*
- MORENO-BLACK, G., BOLES, S., JOHNSON-SHELTON, D. & EVERS, C. 2016. Exploring Categorical Body Mass Index Trajectories in Elementary School Children. *J Sch Health*, 86, 495-506.
- MOURONTE-ROIBAS, C., LEIRO-FERNANDEZ, V., FERNANDEZ-VILLAR, A., BOTANA-RIAL, M., RAMOS-HERNANDEZ, C. & RUANO-RAVINA, A. 2016. COPD, emphysema and the onset of lung cancer. A systematic review. *Cancer Lett*, 382, 240-244.
- MUNN, Z., PETERS, M. D. J., STERN, C., TUFANARU, C., MCARTHUR, A. & AROMATARIS, E. 2018a. Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. *BMC Med Res Methodol*, 18, 143.
- MUNN, Z., STERN, C., AROMATARIS, E., LOCKWOOD, C. & JORDAN, Z. 2018b. What kind of systematic review should I conduct? A proposed typology and guidance for systematic reviewers in the medical and health sciences. *BMC Med Res Methodol*, 18, 5.
- MURAD, M. H., KATABI, A., BENKHADRA, R. & MONTORI, V. M. 2018. External validity, generalisability, applicability and directness: a brief primer. *BMJ Evid Based Med*, 23, 17-19.
- MURCHIE, P., SMITH, S. M., YULE, M. S., ADAM, R., TURNER, M. E., LEE, A. J. & FIELDING, S. 2017. Does emergency presentation of cancer represent poor performance in primary care? Insights from a novel analysis of linked primary and secondary care data. *Br J Cancer*, 116, 1148-1158.
- NATIONAL LUNG SCREENING TRIAL RESEARCH TEAM, ABERLE, D. R., ADAMS, A. M., BERG, C. D., BLACK, W. C., CLAPP, J. D., FAGERSTROM, R. M., GAREEN, I. F., GATSONIS, C., MARCUS, P. M. & SICKS, J. D. 2011. Reduced lung-cancer mortality with low-dose computed tomographic

List of References

- screening. *N Engl J Med*, 365, 395-409.
- NEAL, R. D. 2009. Do diagnostic delays in cancer matter? *Br J Cancer*, 101 Suppl 2, S9-S12.
- NEAL, R. D. & ALLGAR, V. L. 2005. Sociodemographic factors and delays in the diagnosis of six cancers: analysis of data from the "National Survey of NHS Patients: Cancer". *Br J Cancer*, 92, 1971-5.
- NEAL, R. D., ROBBE, I. J., LEWIS, M., WILLIAMSON, I. & HANSON, J. 2015a. The complexity and difficulty of diagnosing lung cancer: findings from a national primary-care study in Wales. *Prim Health Care Res Dev*, 16, 436-49.
- NEAL, R. D., THARMANATHAN, P., FRANCE, B., DIN, N. U., COTTON, S., FALLON-FERGUSON, J., HAMILTON, W., HENDRY, A., HENDRY, M., LEWIS, R., MACLEOD, U., MITCHELL, E. D., PICKETT, M., RAI, T., SHAW, K., STUART, N., TORRING, M. L., WILKINSON, C., WILLIAMS, B., WILLIAMS, N. & EMERY, J. 2015b. Is increased time to diagnosis and treatment in symptomatic cancer associated with poorer outcomes? Systematic review. *Br J Cancer*, 112 Suppl 1, S92-107.
- NEEDLEMAN, S. B. & WUNSCH, C. D. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48, 443-453.
- NHS ENGLAND. 2019. *Evaluation of the Targeted Lung Health Check programme* [Online]. Available: <https://www.england.nhs.uk/contact-us/privacy-notice/how-we-use-your-information/our-services/evaluation-of-the-targeted-lung-health-check-programme/> [Accessed].
- NIKSIC, M., RACHET, B., WARBURTON, F. G., WARDLE, J., RAMIREZ, A. J. & FORBES, L. J. 2015. Cancer symptom awareness and barriers to symptomatic presentation in England--are we clear on cancer? *Br J Cancer*, 113, 533-42.
- NOHR, E. A. & OLSEN, J. 2013. Commentary: Epidemiologists have debated representativeness for more than 40 years--has the time come to move on? *Int J Epidemiol*, 42, 1016-7.
- NUR, U., LYRATZOPOULOS, G., RACHET, B. & COLEMAN, M. P. 2015a. The impact of age at diagnosis on socioeconomic inequalities in adult cancer survival in England. *Cancer Epidemiol*, 39, 641-9.
- NUR, U., QUARESMA, M., DE STAVOLA, B., PEAKE, M. & RACHET, B. 2015b. Inequalities in non-small cell lung cancer treatment and mortality. *J Epidemiol Community Health*, 69, 985-92.
- O'DOWD, E. L., MCKEEVER, T. M., BALDWIN, D. R., ANWAR, S., POWELL, H. A., GIBSON, J. E., IYEN-OMOFOMAN, B. & HUBBARD, R. B. 2015. What characteristics of primary care and patients are associated with early death in patients with lung cancer in the UK? *Thorax*, 70, 161-8.
- OBERMEYER, Z. & EMANUEL, E. J. 2016. Predicting the Future - Big Data, Machine Learning, and Clinical Medicine. *N Engl J Med*, 375, 1216-9.
- OKEN, M. M., HOCKING, W. G., KVALE, P. A., ANDRIOLE, G. L., BUYS, S. S., CHURCH, T. R., CRAWFORD, E. D., FOUAD, M. N., ISAACS, C., REDING, D. J., WEISSFELD, J. L., YOKOCHI, L. A., O'BRIEN, B., RAGARD, L. R., RATHMELL, J. M., RILEY, T. L., WRIGHT, P., CAPARASO, N., HU, P., IZMIRLIAN, G., PINSKY, P. F., PROROK, P. C., KRAMER, B. S., MILLER, A. B., GOHAGAN, J. K., BERG, C. D. & TEAM, P. P. 2011. Screening by chest radiograph and lung cancer mortality: the Prostate, Lung, Colorectal, and Ovarian (PLCO) randomized trial. *JAMA*, 306, 1865-73.
- OLESEN, F., HANSEN, R. P. & VEDSTED, P. 2009. Delay in diagnosis: the experience in Denmark. *Br J Cancer*, 101 Suppl 2, S5-8.
- PARRY, M. G., COWLING, T. E., SUJENTHIRAN, A., NOSSITER, J., BERRY, B., CATHCART, P., AGGARWAL, A., PAYNE, H., VAN DER MEULEN, J., CLARKE, N. W. & GNANAPRAGASAM, V. J. 2020. Risk stratification for prostate cancer management: value of the Cambridge Prognostic Group classification for assessing treatment allocation. *BMC Med*, 18, 114.
- PEACOCK, J. & PEACOCK, P. 2010. *Oxford Handbook of Medical Statistics*, Oxford University Press.
- PEDERSEN, P., LUND, T., LINDHOLDT, L., NOHR, E. A., JENSEN, C., SOGAARD, H. J. & LABRIOLA, M. 2016. Labour market trajectories following sickness absence due to self-reported all cause morbidity--a longitudinal study. *BMC Public Health*, 16, 337.
- PETERS, M. D. 2016. In no uncertain terms: the importance of a defined objective in scoping reviews. *JBI Database System Rev Implement Rep*, 14, 1-4.
- PETERS, M. D., GODFREY, C. M., KHALIL, H., MCINERNEY, P., PARKER, D. & SOARES, C. B. 2015. Guidance for conducting systematic scoping reviews. *Int J Evid Based Healthc*, 13, 141-6.

- PHAM, M. T., RAJIC, A., GREIG, J. D., SARGEANT, J. M., PAPADOPOULOS, A. & MCEWEN, S. A. 2014. A scoping review of scoping reviews: advancing the approach and enhancing the consistency. *Res Synth Methods*, 5, 371-85.
- POLLOCK, G. 2007. Holistic trajectories: a study of combined employment, housing and family careers by using multiple-sequence analysis. *Journal of the Royal Statistical Society Series a-Statistics in Society*, 170, 167-183.
- POWELL, H. A., IYEN-OMOFOMAN, B., BALDWIN, D. R., HUBBARD, R. B. & TATA, L. J. 2013. Chronic obstructive pulmonary disease and risk of lung cancer: the importance of smoking and timing of diagnosis. *J Thorac Oncol*, 8, 6-11.
- QUAIFE, S. L., FORBES, L. J., RAMIREZ, A. J., BRAIN, K. E., DONNELLY, C., SIMON, A. E. & WARDLE, J. 2014. Recognition of cancer warning signs and anticipated delay in help-seeking in a population sample of adults in the UK. *Br J Cancer*, 110, 12-8.
- RACHET, B., QUINN, M. J., COOPER, N. & COLEMAN, M. P. 2008. Survival from cancer of the lung in England and Wales up to 2001. *British Journal of Cancer*, 99, S40-S42.
- RAINE, R., WONG, W., SCHOLE, S., ASHTON, C., OBICHERE, A. & AMBLER, G. 2010. Social variations in access to hospital care for patients with colorectal, breast, and lung cancer between 1999 and 2006: retrospective analysis of hospital episode statistics. *BMJ*, 340, b5479.
- RAO, A., BOTTLE, A., BICKNELL, C., DARZI, A. & AYLIN, P. 2018a. Common Sequences of Emergency Readmissions among High-Impact Users following AAA Repair. *Surgery Research And Practice*, 2018, 5468010-5468010.
- RAO, A., KIM, D., DARZI, A., MAJEED, A., AYLIN, P. & BOTTLE, A. 2018b. Long-term trends of use of health service among heart failure patients. *European Heart Journal. Quality Of Care & Clinical Outcomes*, 4, 220-231.
- RAVIV, S., HAWKINS, K. A., DECAMP, M. M., JR. & KALHAN, R. 2011. Lung cancer in chronic obstructive pulmonary disease: enhancing surgical options and outcomes. *Am J Respir Crit Care Med*, 183, 1138-46.
- REDANIEL, M. T., MARTIN, R. M., RIDD, M. J., WADE, J. & JEFFREYS, M. 2015. Diagnostic intervals and its association with breast, prostate, lung and colorectal cancer survival in England: historical cohort study using the Clinical Practice Research Datalink. *PLoS One*, 10, e0126608.
- RICHARDS, M. A. 2009a. The National Awareness and Early Diagnosis Initiative in England: assembling the evidence. *Br J Cancer*, 101 Suppl 2, S1-4.
- RICHARDS, M. A. 2009b. The size of the prize for earlier diagnosis of cancer in England. *Br J Cancer*, 101 Suppl 2, S125-9.
- RIDD, M. J., FERREIRA, D. L., MONTGOMERY, A. A., SALISBURY, C. & HAMILTON, W. 2015. Patient-doctor continuity and diagnosis of cancer: electronic medical records study in general practice. *Br J Gen Pract*, 65, e305-11.
- ROUND, T., STEED, L., SHANKLEMAN, J., BOURKE, L. & RISI, L. 2013. Primary care delays in diagnosing cancer: what is causing them and what can we do about them? *J R Soc Med*, 106, 437-40.
- ROUSSEEUW, P. J. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20, 53-65.
- ROUX, J., GRIMAUD, O. & LERAY, E. 2018. Use of state sequence analysis for care pathway analysis: The example of multiple sclerosis. *Statistical Methods In Medical Research*, 962280218772068-962280218772068.
- ROYAL COLLEGE OF GENERAL PRACTITIONERS (RCGP). 2019. *Fit for the future: a vision for general practice* [Online]. Available: <https://www.rcgp.org.uk/-/media/Files/News/2019/RCGP-fit-for-the-future-report-may-2019.ashx?la=en> [Accessed 4 July 2019].
- RUBIN, G., BERENDSEN, A., CRAWFORD, S. M., DOMMETT, R., EARLE, C., EMERY, J., FAHEY, T., GRASSI, L., GRUNFELD, E., GUPTA, S., HAMILTON, W., HIOM, S., HUNTER, D., LYRATZOPOULOS, G., MACLEOD, U., MASON, R., MITCHELL, G., NEAL, R. D., PEAKE, M., ROLAND, M., SEIFERT, B., SISLER, J., SUSSMAN, J., TAPLIN, S., VEDSTED, P., VORUGANTI, T., WALTER, F., WARDLE, J., WATSON, E., WELLER, D., WENDER, R., WHELAN, J., WHITLOCK, J., WILKINSON, C., DE WIT, N. & ZIMMERMANN, C. 2015. The expanding role of primary care in cancer control. *The Lancet Oncology*, 16, 1231-1272.
- RUBIN, G., WALTER, F., EMERY, J., NEAL, R., HAMILTON, W. & WARDLE, J. 2014. Research into practice:

List of References

- prompt diagnosis of cancer in primary care. *Br J Gen Pract*, 64, 428-30.
- RUBIN, G. P., MCPHAIL, S. & ELLIOT, K. 2011. National Audit of Cancer Diagnosis in Primary Care. London: Royal College of General Practitioners.
- SARFATI, D. 2012. Review of methods used to measure comorbidity in cancer populations: no gold standard exists. *J Clin Epidemiol*, 65, 924-33.
- SARFATI, D., KOCZWARA, B. & JACKSON, C. 2016. The impact of comorbidity on cancer and its treatment. *CA Cancer J Clin*, 66, 337-50.
- SCOTT, N., CRANE, M., LAFONTAINE, M., SEALE, H. & CURROW, D. 2015. Stigma as a barrier to diagnosis of lung cancer: patient and general practitioner perspectives. *Prim Health Care Res Dev*, 16, 618-22.
- SCOTT, S. & WALTER, F. 2010. Studying Help-Seeking for Symptoms: The Challenges of Methods and Models. *Social and Personality Psychology Compass*, 4, 531-547.
- SEKINE, Y., KATSURA, H., KOH, E., HIROSHIMA, K. & FUJISAWA, T. 2012. Early detection of COPD is important for lung cancer surveillance. *Eur Respir J*, 39, 1230-40.
- SELE, L. M., ELNEGAARD, S., BALASUBRAMANIAM, K., SONDERGAARD, J. & JARBOL, D. E. 2016. Lifestyle factors and contact to general practice with respiratory alarm symptoms-a population-based study. *BMC Fam Pract*, 17, 47.
- SHAH, N. D., STEYERBERG, E. W. & KENT, D. M. 2018. Big Data and Predictive Analytics: Recalibrating Expectations. *JAMA*, 320, 27-28.
- SHAPIRO, S. S. & WILK, M. B. 1965. An analysis of variance test for normality (complete samples). *Biometrika*, 52, 591-611.
- SHAPLEY, M., MANSELL, G., JORDAN, J. L. & JORDAN, K. P. 2010. Positive predictive values of $\geq 5\%$ in primary care for cancer: systematic review. *Br J Gen Pract*, 60, e366-77.
- SHARABIANI, M. T. A., AYLIN, P. & BOTTLE, A. 2012. Systematic Review of Comorbidity Indices for Administrative Data. *Medical Care*, 50, 1109-1118.
- SHERINGHAM, J., SEQUEIRA, R., MYLES, J., HAMILTON, W., MCDONNELL, J., OFFMAN, J., DUFFY, S. & RAINE, R. 2017. Variations in GPs' decisions to investigate suspected lung cancer: a factorial experiment using multimedia vignettes. *BMJ Qual Saf*, 26, 449-459.
- SHIM, J., BRINDLE, L., SIMON, M. & GEORGE, S. 2014. A systematic review of symptomatic diagnosis of lung cancer. *Fam Pract*, 31, 137-48.
- ŠIDÁK, Z. 1967. Rectangular Confidence Regions for the Means of Multivariate Normal Distributions. *Journal of the American Statistical Association*, 62, 626-633.
- SIDEY-GIBBONS, J. A. M. & SIDEY-GIBBONS, C. J. 2019. Machine learning in medicine: a practical introduction. *BMC Med Res Methodol*, 19, 64.
- SIMPSON, C. R., HIPPISELEY-COX, J. & SHEIKH, A. 2010. Trends in the epidemiology of chronic obstructive pulmonary disease in England: a national study of 51 804 patients. *Br J Gen Pract*, 60, 277-84.
- SINGH-MANOUX, A. & MARMOT, M. 2005. Role of socialization in explaining social inequalities in health. *Soc Sci Med*, 60, 2129-33.
- SKRONDAL, A. & RABE-HESKETH, S. 2003. Multilevel logistic regression for polytomous data and rankings. *Psychometrika*, 68, 267-287.
- SMITH, L. K., POPE, C. & BOTHA, J. L. 2005. Patients' help-seeking experiences and delay in cancer presentation: a qualitative synthesis. *The Lancet*, 366, 825-831.
- SMITH, S. M., CAMPBELL, N. C., MACLEOD, U., LEE, A. J., RAJA, A., WYKE, S., ZIEBLAND, S. B., DUFF, E. M., RITCHIE, L. D. & NICOLSON, M. C. 2009. Factors contributing to the time taken to consult with symptoms of lung cancer: a cross-sectional study. *Thorax*, 64, 523-31.
- SMITS, S., MCCUTCHAN, G., WOOD, F., EDWARDS, A., LEWIS, I., ROBLING, M., PARANJOTHY, S., CARTER, B., TOWNSON, J. & BRAIN, K. 2018. Development of a Behavior Change Intervention to Encourage Timely Cancer Symptom Presentation Among People Living in Deprived Communities Using the Behavior Change Wheel. *Ann Behav Med*, 52, 474-488.
- SPEARMAN, C. 1904. The Proof and Measurement of Association between Two Things. *The American Journal of Psychology*, 15.
- STAPLEY, S., SHARP, D. & HAMILTON, W. 2006. Negative chest X-rays in primary care patients with lung cancer. *Br J Gen Pract*, 56, 570-3.

- STEYERBERG, E. W. 2019. *Clinical Prediction Models*.
- STOVEL, K., SAVAGE, M. & BEARMAN, P. 1996. Ascription into achievement: Models of career systems at Lloyds Bank, 1890-1970. *American Journal of Sociology*, 102, 358-399.
- STUDER, M. 2013. WeightedCluster Library Manual: A practical guide to creating typologies of trajectories in the social sciences with R. LIVES Working Papers, 24. Available: <http://dx.doi.org/10.12682/lives.2296-1658.2013.24>.
- STUDER, M. & RITSCHARD, G. 2016. What matters in differences between life trajectories: a comparative review of sequence dissimilarity measures. *Journal of the Royal Statistical Society Series a-Statistics in Society*, 179, 481-511.
- STUDER, M., RITSCHARD, G., GABADINHO, A. & MÜLLER, N. S. 2011. Discrepancy Analysis of State Sequences. *Sociological Methods & Research*, 40, 471-510.
- SYRIOPOULOU, E., BOWER, H., ANDERSSON, T. M., LAMBERT, P. C. & RUTHERFORD, M. J. 2017. Estimating the impact of a cancer diagnosis on life expectancy by socio-economic group for a range of cancer types in England. *Br J Cancer*, 117, 1419-1426.
- TAMMEMÄGI, M. C., KATKI, H. A., HOCKING, W. G., CHURCH, T. R., CAPORASO, N., KVALE, P. A., CHATURVEDI, A. K., SILVESTRI, G. A., RILEY, T. L., COMMINS, J. & BERG, C. D. 2013. Selection criteria for lung-cancer screening. *N Engl J Med*, 368, 728-36.
- TATARU, D., JACK, R. H., LIND, M. J., MOLLER, H. & LUCHTENBORG, M. 2015. The effect of emergency presentation on surgery and survival in lung cancer patients in England, 2006-2008. *Cancer Epidemiol*, 39, 612-6.
- THE NATIONAL HEART LUNG AND BLOOD INSTITUTE (NHLBI). *Quality Assessment Tool for Case Series Studies* [Online]. Available: <https://www.nhlbi.nih.gov/health-topics/study-quality-assessment-tools> [Accessed October 1 2019].
- THE NATIONAL INSTITUTE FOR HEALTH AND CARE EXCELLENCE. 2018. *Developing NICE guidelines: the manual. Process and methods [PMG20] Appendix H: Appraisal checklists, evidence tables, GRADE and economic profiles* [Online]. Available: <https://www.nice.org.uk/process/pmg20/resources/developing-nice-guidelines-the-manual-appendices-2549710189/chapter/appendix-h-appraisal-checklists-evidence-tables-grade-and-economic-profiles> [Accessed October 1 2019].
- THE NATIONAL INSTITUTE FOR HEALTH AND CARE EXCELLENCE (NICE) 2019. Lung cancer: diagnosis and management (Clinical guideline 122).
- THOMSON, C. S. & FORMAN, D. 2009. Cancer survival in England and the influence of early diagnosis: what can we learn from recent EUROCARE results? *Br J Cancer*, 101 Suppl 2, S102-9.
- TORRING, M. L., FRYDENBERG, M., HANSEN, R. P., OLESEN, F. & VEDSTED, P. 2013. Evidence of increasing mortality with longer diagnostic intervals for five common cancers: a cohort study in primary care. *Eur J Cancer*, 49, 2187-98.
- VEDSTED, P. & OLESEN, F. 2015. A differentiated approach to referrals from general practice to support early cancer diagnosis - the Danish three-legged strategy. *Br J Cancer*, 112 Suppl 1, S65-9.
- VOGT, V., SCHOLZ, S. M. & SUNDMACHER, L. 2018. Applying sequence clustering techniques to explore practice-based ambulatory care pathways in insurance claims data. *Eur J Public Health*, 28, 214-219.
- WAGLAND, R., BRINDLE, L., EWINGS, S., JAMES, E., MOORE, M., RIVAS, C., ESQUEDA, A. I. & CORNER, J. 2016. Promoting Help-Seeking in Response to Symptoms amongst Primary Care Patients at High Risk of Lung Cancer: A Mixed Method Study. *PLoS One*, 11, e0165677.
- WAGLAND, R., BRINDLE, L., JAMES, E., MOORE, M., ESQUEDA, A. I. & CORNER, J. 2017. Facilitating early diagnosis of lung cancer amongst primary care patients: The views of GPs. *Eur J Cancer Care (Engl)*, 26.
- WALABYEKI, J., ADAMSON, J., BUCKLEY, H. L., SINCLAIR, H., ATKIN, K., GRAHAM, H., WHITAKER, K., WARDLE, J. & MACLEOD, U. 2017. Experience of, awareness of and help-seeking for potential cancer symptoms in smokers and non-smokers: A cross-sectional study. *PLoS One*, 12, e0183647.
- WALTER, F., WEBSTER, A., SCOTT, S. & EMERY, J. 2012. The Andersen Model of Total Patient Delay: a systematic review of its application in cancer diagnosis. *J Health Serv Res Policy*, 17, 110-8.

List of References

- WALTER, F. M., RUBIN, G., BANKHEAD, C., MORRIS, H. C., HALL, N., MILLS, K., DOBSON, C., RINTOUL, R. C., HAMILTON, W. & EMERY, J. 2015. Symptoms and other factors associated with time to diagnosis and stage of lung cancer: a prospective cohort study. *Br J Cancer*, 112 Suppl 1, S6-13.
- WALTERS, S., MARINGE, C., COLEMAN, M. P., PEAKE, M. D., BUTLER, J., YOUNG, N., BERGSTROM, S., HANNA, L., JAKOBSEN, E., KOLBECK, K., SUNDSTROM, S., ENGHOLM, G., GAVIN, A., GJERSTORFF, M. L., HATCHER, J., JOHANNESSEN, T. B., LINKLATER, K. M., MCGAHAN, C. E., STEWARD, J., TRACEY, E., TURNER, D., RICHARDS, M. A., RACHET, B. & GROUP, I. M. W. 2013. Lung cancer survival and stage at diagnosis in Australia, Canada, Denmark, Norway, Sweden and the UK: a population-based study, 2004-2007. *Thorax*, 68, 551-64.
- WALTON, L., MCNEILL, R., STEVENS, W., MURRAY, M., LEWIS, C., AITKEN, D. & GARRETT, J. 2013. Patient perceptions of barriers to the early diagnosis of lung cancer and advice for health service improvement. *Fam Pract*, 30, 436-44.
- WARD, J. H., JR 1963. Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, 58, 236-244.
- WEINSTEIN, N. D., MARCUS, S. E. & MOSER, R. P. 2005. Smokers' unrealistic optimism about their risk. *Tob Control*, 14, 55-9.
- WEISKOPF, N. G., BAKKEN, S., HRIPCSAK, G. & WENG, C. 2017. A Data Quality Assessment Guideline for Electronic Health Record Data Reuse. *EGEMS (Wash DC)*, 5, 14.
- WEISKOPF, N. G. & WENG, C. 2013. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *J Am Med Inform Assoc*, 20, 144-51.
- WELLER, D., VEDSTED, P., RUBIN, G., WALTER, F. M., EMERY, J., SCOTT, S., CAMPBELL, C., ANDERSEN, R. S., HAMILTON, W., OLESEN, F., ROSE, P., NAFEEES, S., VAN RIJSWIJK, E., HIOM, S., MUTH, C., BEYER, M. & NEAL, R. D. 2012. The Aarhus statement: improving design and reporting of studies on early cancer diagnosis. *Br J Cancer*, 106, 1262-7.
- WHILLANS, J., NAZROO, J. & MATTHEWS, K. 2016. Trajectories of vision in older people: The role of age and social position. *European Journal of Ageing*, 13, 171-184.
- WHITAKER, K. L., SCOTT, S. E. & WARDLE, J. 2015. Applying symptom appraisal models to understand sociodemographic differences in responses to possible cancer symptoms: a research agenda. *Br J Cancer*, 112 Suppl 1, S27-34.
- WHITAKER, K. L., SMITH, C. F., WINSTANLEY, K. & WARDLE, J. 2016. What prompts help-seeking for cancer 'alarm' symptoms? A primary care based survey. *Br J Cancer*, 114, 334-9.
- WHITE, I. R., ROYSTON, P. & WOOD, A. M. 2011. Multiple imputation using chained equations: Issues and guidance for practice. *Stat Med*, 30, 377-99.
- WIGGINS, R. D., ERZBERGER, C., HYDE, M., HIGGS, P. & BLANE, D. 2007. Optimal Matching Analysis Using Ideal Types to Describe the Lifecourse: An Illustration of How Histories of Work, Partnerships and Housing Relate to Quality of Life in Early Old Age. *International Journal of Social Research Methodology*, 10, 259-278.
- WOLF, A., DEDMAN, D., CAMPBELL, J., BOOTH, H., LUNN, D., CHAPMAN, J. & MYLES, P. 2019. Data resource profile: Clinical Practice Research Datalink (CPRD) Aurum. *Int J Epidemiol*, 48, 1740-1740g.
- WORLD HEALTH ORGANIZATION 1998. *Framework for professional and administrative development of general practice/family medicine in Europe.*, Copenhagen, World Health Organization Regional Office for Europe.
- WU, L. L. 2000. Some Comments on "Sequence Analysis and Optimal Matching Methods in Sociology: Review and Prospect". *Sociological Methods & Research*, 29, 41-64.
- WUERKER, A. K. 1996. The changing careers of patients with chronic mental illness: a study of sequential patterns in mental health service utilization. *J Ment Health Adm*, 23, 458-70.
- YEATMAN, S. & TRINITAPOLI, J. 2017. Awareness and perceived fairness of Option B+ in Malawi: A population-level perspective. *J Int AIDS Soc*, 20, 21467.
- YTTERSTAD, E., MOE, P. C. & HJALMARSEN, A. 2016. COPD in primary lung cancer patients: prevalence and mortality. *Int J Chron Obstruct Pulmon Dis*, 11, 625-36.
- ZHAI, R., YU, X., SHAFER, A., WAIN, J. C. & CHRISTIANI, D. C. 2014. The impact of coexisting COPD on survival of patients with early-stage non-small cell lung cancer undergoing surgical resection.

- Chest*, 145, 346-353.
- ZHOU, Y., ABEL, G. A., HAMILTON, W., PRITCHARD-JONES, K., GROSS, C. P., WALTER, F. M., RENZI, C., JOHNSON, S., MCPHAIL, S., ELLISS-BROOKES, L. & LYRATZOPOULOS, G. 2017. Diagnosis of cancer as an emergency: a critical review of current evidence. *Nat Rev Clin Oncol*, 14, 45-56.