UNIVERSITY OF SOUTHAMPTON

Faculty of Social, Human and Mathematical Sciences

Mathematical Sciences

Modelling pathways to diagnosis of breast conditions

by

Christina Emma Saville

submitted for the degree of Doctor of Philosophy

June 2018

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF SOCIAL, HUMAN AND MATHEMATICAL SCIENCES
Mathematical Sciences

Doctor of Philosophy

MODELLING PATHWAYS TO DIAGNOSIS OF BREAST CONDITIONS

by Christina Emma Saville

This thesis describes how logistic regression and discrete-event simulation (DES) can be combined to predict patient risk and evaluate the potential operational impact of implementing risk-based pathways. We test whether using diagnostic information provided by non-specialists to plan diagnostic tests offers benefits in operational performance. We demonstrate our approach on an application area in breast diagnostics with data from the Whittington Hospital breast diagnostic clinic.

Specifically we assess whether GP referral information is complete and accurate enough for use in predicting the risk of an abnormal result (i.e. an abnormality being detected from mammogram, ultrasound or biopsy). The construction of a unique dataset for this purpose is described; it links GP referral information to in-clinic tests and results. This dataset is used to develop two alternative logistic regression scorecards that predict a patient's risk of abnormal breast diagnostic results from their GP referral data (n=179). The simple scorecard uses two referral characteristics while the full scorecard uses seven.

It is usual to base the decision of where to set the cut-off score between low and high risk patients on a scorecard's predictive performance. In contrast, we show how a discrete-event simulation can be used to optimise the cut-off in terms of operational performance. In our example, the performance measure is the daily average proportion of patients' time at the clinic that adds value, called the clinic efficiency. We simulate the potential impacts of introducing the following risk-based pathways. High-risk patients are sent straight for imaging tests and then to a clinician for their results. Low-risk patients are sent to a clinician first (as today) who decides whether imaging is needed. The set of labels that determines a patient's progress through the simulation is modelled empirically for the simple scorecard, since all possible label combinations are present in our sample. However for the full scorecard this is not the case, so using the empirical distribution is not appropriate. Instead we introduce a novel method using Poisson loglinear models to generate representative sets of patient labels.

Contents

List of Figures

List of Tables

Academic Thesis: Declaration Of Authorship

I, Christina Emma Saville, declare that this thesis entitled “Modelling pathways to diagnosis of breast conditions” and the work presented in it are my own and have been generated by me as the result of my own original research.

I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University;
- 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;
- Where I have consulted the published work of others, this is always clearly attributed;
- 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;
- I have acknowledged all main sources of help;
- 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;
- Part of Chapter 3 has been published as Saville, C.E., Smith, H. and Bijak, K. (2018). Operational Research techniques applied throughout cancer care services: a review. *Health Systems* https://doi.org/10.1080/20476965.2017.1414741 Published online: 17 Jan 2018.
- Parts of this work have been presented at the following conferences.

  1. Saville, C, Smith, H and Izady, N. Improving the effciency of a breast cancer diagnostic clinic. Presented at 27th European Conference on Operational Research (EURO), Glasgow, 2015.

2. Saville, C, Smith, H, Izady, N and Bijak, K. Improving the effciency of a breast diagnostic clinic. Presented at 5th Student Conference on Operational Research (SCOR), Nottingham, 2016.
3. Saville, C, Smith, H, Bijak, K and Izady, N. Improving the route to diagnosis of breast conditions. Poster presented at 42nd annual meeting of EURO Working Group on Operational Research Applied to Health Services (ORAHS), Pamplona, 2016.
4. Leonard, P., Saville, C., Smith, H., Bijak, K., Verjee, A. and Izady, N.. How a collaboration between an Academic Maths Department and a local NHS Breast Cancer Service improve patient pathways. Poster presented at National Cancer Research Institute Conference (NCRI), Liverpool, 2016. Abstract published in European Journal of Surgical Oncology, 42(11), S253-S254.
5. Saville, C., Smith, H. and Bijak, K. Generating virtual patients for discrete-event simulation from a small sample with categorical characteristics. Poster presented at 43rd annual meeting of EURO Working Group on Operational Research Applied to Health Services (ORAHS), Bath, 2017.

Acknowledgements

Thank you to my supervisors, Dr Honora Smith and Dr Kasia Bijak (and Dr Navid Izady in the early stages) for their motivation, support and ideas. I acknowledge all the patients who kindly allowed us to use their records in the study, as well as the Whittington clinic and imaging staff for their help and patience. A particular thank you to Dr Pauline Leonard for her continued support and enthusiasm throughout the project.

On a personal note, thanks to my family for inspiring and encouraging me throughout my studies. Last but by no means least, a big thank you to Ed, for help of every imaginable kind.

Chapter 1

Introduction

1.1 Practical motivation

In the UK, the proportion of people affected by cancer is projected to increase from 1 in 3 in 1992, to nearly 1 in 2 by 2020 (Macmillan Cancer Support, 2013). National cancer services must deal with almost 340,000 new cancer cases yearly, in addition to continued care for patients previously diagnosed (Macmillan Cancer Support, 2014). According to 2013 data, annual NHS costs for cancer and tumour services in England are £5.8 billion, making this the third largest NHS spend category after mental disorders and circulation problems (Nuffield Trust, 2016).

The cancer care system is comprised of prevention, diagnosis, staging and treatment services. Prevention includes screening a target population for cancers that are not yet showing symptoms (NHS, 2016) and encouraging lifestyle changes to lower cancer risk. Cancers are diagnosed following an abnormal screening result, development of suspicious symptoms or incidental detection when examining patients for other reasons, for example after being admitted through the emergency department (Elliss-Brookes et al., 2012). Following a cancer diagnosis, *staging* tests help determine the size of the cancer and how much it has spread (National Cancer Institute, 2016b). There are a variety of treatments used in curing cancer, as well as *adjuvant treatments* to lower the risk of the cancer returning and *palliative treatments* to relieve symptoms (National Cancer Institute, 2016b). The type and stage of cancer affect which particular combination of treatments is most appropriate (National Cancer Institute, 2016a).

The most common cancer in the UK is breast cancer, which affects 1 in 8 women and 1 in 870 men (Cancer Research UK, 2016c). Survival is improving, with 87% of women diagnosed in England and Wales surviving the disease for at least five years. However, the stage at which a cancer is detected greatly impacts chances of survival; only 3 in 20 women with final stage disease survive beyond 5 years (Cancer Research UK, 2016c).

Treatment for breast cancer typically consists of surgery to remove some or all of the affected breast, in addition to radiation or drug treatments (Cancer Research UK, 2016a). The physical and mental impacts of both cancer and cancer treatments substantially reduce a patient's quality of life, even after the cancer is cured (Elliott et al., 2011).

The majority (60%) of breast cancer patients are diagnosed following referral to a specialist clinic by their GP (Cancer Research UK, 2016b). In England, annual volumes of patients referred to a specialist with suspected breast cancer or breast symptoms increased by 26% from 2010 to 2015 (calculated from NHS England (2016) statistics). This far exceeds the increase in incidence over the same time period (Cancer Research UK, 2016c) and could instead be partly explained by heightened awareness of breast cancer, for example due to the national "Be Clear on Cancer" campaign that included encouraging women aged over 70 to visit their GP upon noticing any changes to their breasts (National Health Service, 2017). In particular, a survey commissioned by Cancer Research UK (2011) found that cancer is the most feared serious illness, with women fearing breast cancer second most after brain cancer. However, unnecessary diagnostic tests can cause harm, worry and overdiagnosis (Glasziou et al., 2013). Finding the right balance between offering tests and reassurance is a national challenge.

The diagnostic *pathway* (sequence of services) for patients referred by their GPs to breast diagnostic clinics is as follows. There is a two-week wait target between when a patient is referred and their attendance in clinic (Keogh, 2009), recognising the urgency of confirming or eliminating a cancer diagnosis for both physical and mental reasons. It is recommended that breast diagnostic clinics are one-stop, that is, they should offer all necessary diagnostic tests on a single day (Willett et al., 2010). There are two main options for organising the sequence of services within the day. In some clinics, information provided by GPs on referral is used to identify those patients who should be sent straight for imaging tests. The remaining patients are sent to see a clinician who decides whether imaging is required. In other clinics, all patients see a clinician first. In this case, the information provided by GPs may not be used at all.

It is unclear which option for organising clinics is best. There is no evidence that information recorded in primary care is accurate enough to use for triaging patients. Overall proven breast cancer forms a very small proportion of referrals and the disease is difficult to distinguish from other breast conditions without imaging (Harvey et al., 2014). Correspondingly, studies have shown that triaging patients based on GPs' assessments of urgency can lead to cancer cases being diagnosed late (Cant and Yu, 2000; Sauven, 2001; Thrush et al., 2002). This may have increased distrust of referral information among clinicians. However, repeatedly questioning patients in detail about their medical history and symptoms, which are unlikely to have changed substantially over the course of two weeks, is an inefficient use of patients', GPs' and clinicians' time.

From a patient experience perspective, on-the-day waiting times should be as low as possible, especially since this group of patients will be particularly prone to worry and anxiety; arguably this will be exacerbated when surrounded by other patients and in a cancer services environment. Given the increased volumes of patients attending diagnostic services, it is especially important to make the best possible use of resources.

This research investigates whether GP referral information, if used appropriately, could help improve the efficiency of diagnostic clinics. It is not to be expected that it is completely accurate since GPs are by nature not as experienced in breast diagnostics as specialists, so may not be able to assess every case correctly. Hence rather than trying to predict cancer, a more successful use of GP referral information could be to predict the risk of clinic staff detecting any breast abnormality (including other conditions as well as cancer). If GP referral information were used in this way, it could help better tailor patient pathways to the likelihood of risk. It could potentially increase the proportion of patients' time at the clinic that is *value-added* (contributes to their care), as well as alleviating pressure on resources. Timely diagnosis or reassurance are important because all patients are likely to be worrying about a cancer diagnosis during their time at the clinic.

1.2 Research objective and questions

This research project's objective is to use OR modelling to evaluate ways of improving diagnostic pathways for patients with breast symptoms. The methods are demonstrated for a case study at the Whittington Hospital breast diagnostic clinic. We define the following research questions to address our objective:

1. For the Whittington breast diagnostic clinic, is GP referral information accurate and complete enough to be used to predict the risk of an abnormal result (defined later), and so to identify patients who could be sent straight for imaging tests?
2. For the Whittington breast diagnostic clinic, could introducing risk-based pathways increase the proportion of patients' time at the clinic that adds value?
3. What general insights, practical and methodological, can be drawn from the models developed for this case study?

1.3 Theoretical motivation and research contributions

Our methodological contributions can be summarised as follows, and are detailed below. We combine three established techniques (logistic regression, Poisson regression and discrete-event simulation) in a novel way to test whether using diagnostic information provided by non-specialists to plan diagnostic tests offers benefits in operational performance. Firstly, by collecting a unique dataset and fitting logistic regression scorecards, we provide the only existing research evidence (to the best of our knowledge) of the link between GP referral information and detection of breast abnormalities in specialist clinics. Secondly, we provide a new method for deciding where to set the scorecard threshold between low and high risk patients. For this, we use DES to evaluate operational measures for different risk thresholds. Thirdly, we contribute a unique approach to generating patient characteristics for simulation. This involves fitting Poisson regression models to estimate the joint distribution of categorical, inter-dependent patient characteristics, where the data sample may be missing possible combinations of characteristics.

Breast diagnostic clinics are one of many examples of outpatient services which are attended by patients referred by their GP. The GP provides diagnostic information about the patient that due to their non-specialist status may be only partially accurate and incomplete. Specialists in the outpatient service may either use the GP referral information to identify patients at high risk of requiring further diagnostic tests, or reassess all patients to obtain more complete and accurate information. We contribute an approach to test whether using this non-specialist information to plan diagnostic tests can improve the outpatient service's operational performance. Our approach combines logistic regression, discrete-event simulation and Poisson regression.

Statistical and data mining classification techniques are used to predict class membership; unlike linear regression they predict categorical rather than continuous values. There are many existing classification models for predicting breast cancer risks, and these are applicable at different stages of the cancer care system (see Section 3.2). However, our problem, which as far as we are aware has not been previously considered, is using GP referral information to predict the risk of detecting any breast abnormality, including but not limited to cancer, in a specialist clinic. I collected a unique dataset linking patient characteristics recorded on GP referrals with in-clinic tests and results. This was used to fit two alternative logistic regression scorecards using the weights-of-evidence approach popular in credit scoring applications.

The decision of where to set the threshold between low- and high-risk patients is typically informed by the prediction accuracy of classification models at different thresholds. However, this neglects the impact on operational measures such as waiting times and costs. Discrete-event simulation (DES) models are a popular choice for estimating the potential impacts of service changes. Ways in which classification and DES have been combined in previous research are described in Section 3.4. Our novel method is combining logistic regression and DES to evaluate the operational impact of planning pathways based on risk prediction tools with different thresholds.

The way in which patient characteristics are modelled needs to be carefully considered when evaluating risk-based patient management strategies in DES models. Not only do characteristics from GP referral information (*referral characteristics*) affect the risk groups that patients are placed in, but more generally characteristics may also affect the pathways that patients follow through the simulation and how long they spend in each stage. Categorical characteristics that are inter-dependent are common in risk tools. Some possible combinations of characteristics may not appear in a data sample, but may exist in the wider population. The review of simulation input modelling of patient characteristics, in Section 3.3, uncovers a scarcity of methods addressing this situation. We propose a novel approach using Poisson regression models for generating combinations of dependent categorical characteristics, allowing missing combinations to be generated.

1.4 Outline

This thesis proceeds as follows.

Chapter 2 provides background information about breast conditions, breast diagnostic clinics and our case study of the Whittington breast diagnostic clinic.

Chapter 3 is a review of relevant literature. It describes OR applications in breast cancer care, breast cancer risk classification models, ways of modelling patient characteristic inputs in DES pathway models, and how patient classification has been combined with DES.

Chapter 4 describes data collected, firstly concerning the daily operation of the breast clinic, and secondly linking GP referral information to tests and results.

Chapter 5 presents classification models using GP referral characteristics to predict a patient's risk of having a breast abnormality detected at the clinic.

Chapter 6 describes DES development, validation and experimentation. Two of our key contributions are outlined here: Firstly, our method for generating combinations of dependent categorical characteristics when some combinations do not appear in the sample, and secondly our use of DES for testing the impact of different classification thresholds on operational performance.

Chapter 7 concludes the thesis.

Chapter 2

Background to practical motivation

This chapter expands on the practical motivation provided in Section 1.1 with background information firstly about breast conditions in Section 2.1, secondly about breast diagnostic clinics in Section 2.2 and thirdly about our case study in Section 2.3.

2.1 Breast conditions

Here we briefly explain some background to the biology and natural history of breast conditions that patients may be diagnosed with. Many patients attending a breast diagnostic clinic are concerned about breast cancer, which is what may have led them to visit their GP in the first place (Harvey et al., 2014). Fortunately only a small proportion of them do have breast cancer; for example, according to our data the figure is 4% at the Whittington hospital (as explained in Section 4.2.2). The remaining patients attending a diagnostic clinic are diagnosed with either a non-cancerous disease or with normal (healthy) breasts.

Cancer, also called malignancy, develops when abnormal cells keep dividing to form a lump, called a tumour, which grows into neighbouring tissue (Macmillan Cancer Support, 2016). Cancers may also spread to other parts of the body. Non-invasive, or in situ, breast disease are names given to pre-malignant breast disease, when the abnormal cells have not grown into neighbouring tissue (Harvey et al., 2014). This disease is often detected when screening asymptomatic individuals, and there is controversy around whether it should be treated, since it may not develop into invasive breast disease (Rue et al., 2012). There are a range of symptoms that may signify breast cancer, with a lump being the most common (84%) and other possible symptoms including nipple discharge, nipple alteration and pain (Koo et al., 2017). Breast cancer affects many more women than men (Cancer Research UK, 2016c).

Benign breast diseases cover a variety of problems that do not pose high risks to patients (Harvey et al., 2014). These include benign breast lumps such as cysts and fibroadenomas, mastalgia (breast pain), infections (for example mastitis and abscesses), congenital problems, which cause the breast to have an abnormal external appearance, and a condition called gynaecomastia, which affects men (Harvey et al., 2014). Breast cysts may be aspirated (drained) on the day of diagnosis, while some other conditions may require further clinic visits (Harvey et al., 2014).

2.2 Breast diagnostic clinics

In this section we provide some more detail about the challenges facing breast diagnostic clinics, how they are organised and the information provided in GP referrals.

2.2.1 Challenges facing breast diagnostic clinics

As explained in Section 1.1, in England, the numbers of patients referred by their GP to a specialist with suspected breast cancer or breast symptoms increased by 26% from 2010 to 2015 (calculated from NHS England (2016) statistics). Although this increase may be beneficial from the perspective of earlier diagnosis, it means that diagnostic resources are under increased pressure. These resources include consultants (breast surgeons and radiologists), equipment (ultrasound and mammogram machines) and specialist staff (pathologists, radiographers and specialist cancer nurses). Due to costs and sometimes a lack of suitably trained professionals, these resources are capped. For example, there is a reported shortage of cellular pathologists which is projected to worsen in the coming five to ten years (Cancer Research UK, 2016d).

The patients attending breast diagnostic clinics are all likely to be anxious about receiving a cancer diagnosis. The two week wait target, which means that clinics aim to see all symptomatic patients within two weeks of referral (Keogh, 2009), may further increase a patient's sense of their problem's urgency, since this is relatively quick compared to other outpatient waiting lists. When attending the clinic, patients' anxiety may be heightened by the lack of distractions, other nervous patients and visual cues about cancer such as leaflets and posters. Thus it is important to recognise the particular vulnerability of this group of patients, and ensure that on the day of their clinic visit they are not left waiting for unnecessarily long periods in between tests, consultations and results.

2.2.2 Organisation of breast diagnostic clinics

Breast diagnostic clinics may be organised in two main ways, as explained in Section 1.1. Firstly, in some clinics, information provided by GPs on referral is used to identify those patients who should be sent straight for imaging tests. The remaining patients are sent to see a clinician first, who performs a clinical examination and decides whether imaging is required. In other clinics, all patients see a clinician first, before any necessary tests. In both cases it is usual for patients to return to the clinician for their test results.

There are no published statistics about how many clinics are arranged each way. Some examples of clinics arranged the first way are at the Princess Anne (University Hospital Southampton NHS Foundation Trust, 2014), Whipps Cross and Newham Cross hospitals (Verjee, 2015). Sometimes appointments for imaging and with a clinician may be on different days (University Hospital Southampton NHS Foundation Trust, 2014). A few examples of clinics that are organised in the second way are the Medway NHS Foundation Trust breast clinic (Medway NHS Foundation Trust, 2013), Guys and St Thomas' breast clinic (Guy's and St Thomas' NHS Foundation Trust, 2011), and the Park Centre for Breast Care (Brighton and Sussex University Hospitals NHS Trust, 2017). At the latter clinic, specialist nurse practitioners as well as clinicians see patients.

2.2.3 GP referral information

When GPs refer patients to a breast diagnostic clinic, they provide information about the referral. Referral forms differ by region or even hospital, and are updated over time as the guidelines on referral criteria change (National Institute for Health and Care Excellence, 2015). There are standard referral forms available online, for example for London (Healthy London Partnership, 2018), Kent and Medway (NHS Kent and Medway Cancer Collaborative, 2018), and Vale of York clinical commissioning group (Referral support service, 2018). These referral forms typically contain space for details about patients' symptoms (usually in tickbox format), family history, medical history, duration of symptoms, medication and a diagram on which to mark the site of symptoms. There is currently no research evidence (that I am aware of) that referral information provided on these forms is complete and accurate enough to be used to identify patients who require imaging tests.

2.3 Case study: Whittington breast diagnostic clinic

The Whittington Health integrated care organisation is based in North London and provides hospital and community services to a population of 500,000 in Islington, Haringey, Barnet and Camden (Whittington Health NHS, 2015a). One of the services the hospital provides is diagnosis of patients with breast symptoms. A *one-stop clinic* offers patients a series of diagnostic tests on a single day to confirm or exclude a cancer diagnosis (Whittington Health NHS, 2015b).

2.3.1 Case study initialisation and obtaining approvals

The case study was initiated by Dr Pauline Leonard, the then Lead Cancer Clinician at the Whittington Hospital, who was keen to improve patient experiences of cancer services. Over a series of meetings, Pauline explained background information and challenges facing cancer services at the hospital. In return, we suggested potential applications of OR methods. Together we decided to limit the scope of the study to the breast diagnostic clinic and decided on the study objectives. We developed a research protocol, incorporating feedback from Pauline.

I spent a substantial amount of time completing paperwork to obtain permission to carry out observations and data collection at the clinic. The application process for an honorary contract with the Whittington Hospital (needed so that I could spend time in the clinic and access data) is shown in Figure C.1 and took four months from March until June 2015. Since the project involves patient-identifiable data, both ethics and R&D approval were required. The process to request these approvals is shown in Figure C.2 and took almost a year, from December 2014 until November 2015. Ethical approval was granted by the London-Bromley NRES Committee (reference number 15/LO/1335). The patient-identifiable data used in this study were obtained from patients who provided informed consent.

2.3.2 Process flow

A period of observation helped me to become familiar with the Whittington breast diagnostic clinic. In order to appreciate different viewpoints of the service, one day was spent shadowing each of the following: a patient, a consultant breast clinician, a registrar breast clinician and a consultant radiologist. In addition, administrative staff advised on what data are recorded in the computer systems and their experiences of working in the clinic. Based on these observations and conversations, a process map was developed (see Figure 2.1). This diagram maps how patients flow through the one-stop clinic from arrival until their visit is complete.

New patients are those visiting the diagnostic clinic. Unless otherwise specified, when the word “patient” is used in this report, “new patient” is meant. *Follow-up patients* are patients returning for repeat appointments with a clinician or for imaging, for example patients undergoing treatment for breast diseases and cancer. The focus of

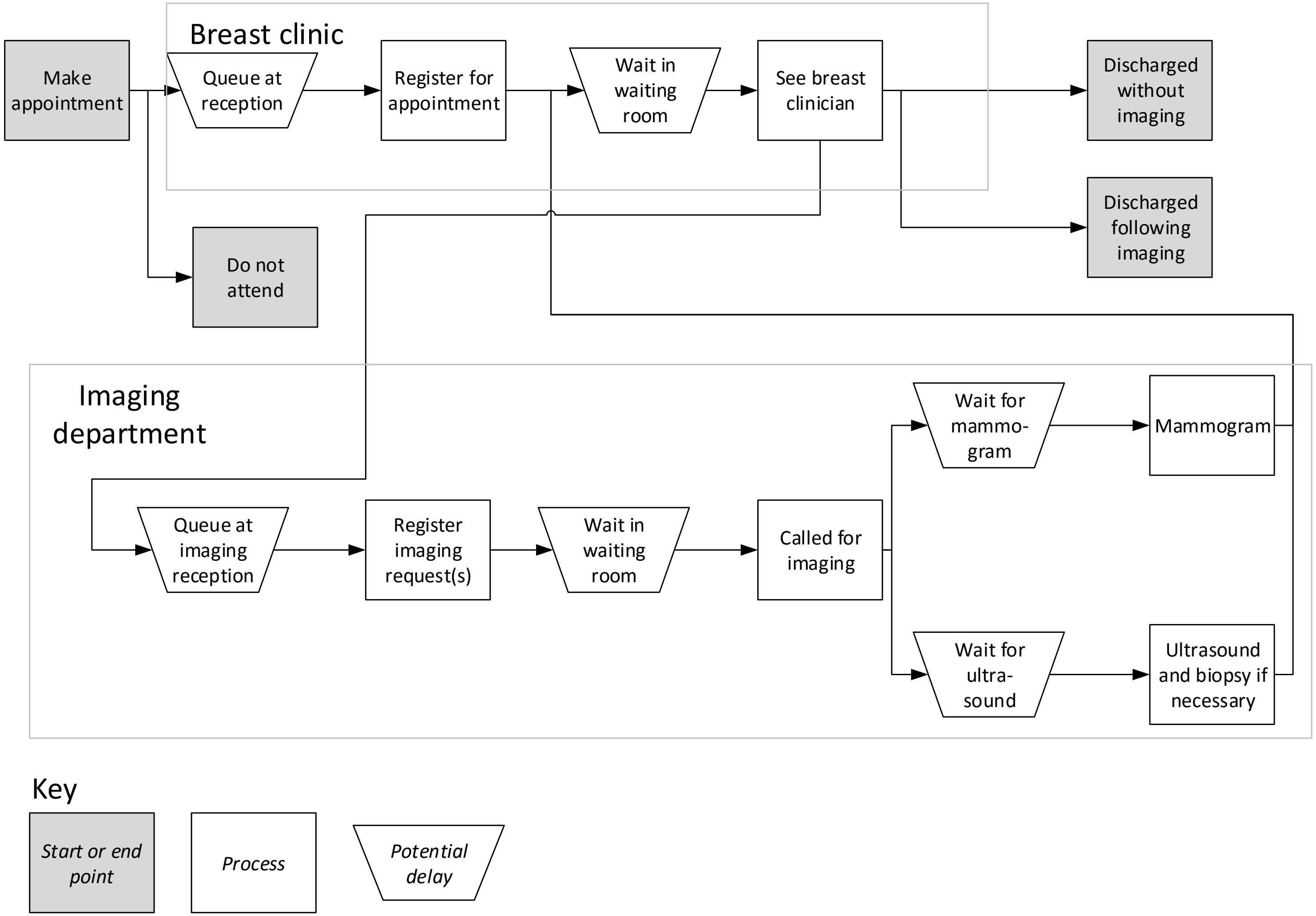


FIGURE 2.1: Process map of the Whittington Hospital breast diagnostic clinic

this research is on diagnostics of new patients. However since follow-up patients share imaging resources with the new patients, these are also considered where necessary.

The process flow through the diagnostic clinic is as follows. When new patients arrive at the clinic, the receptionist gives them a questionnaire (New Patient Assessment Form) to fill in while they wait for their *initial consultation*. During this consultation, a breast clinician discusses the questionnaire and performs a clinical examination. Depending on the outcome of this clinical assessment, the clinician may request a mammogram, an ultrasound, or both to be performed in the imaging department on the floor below.

In the imaging department, there are two rooms used by the breast clinic. These are shared between new and follow-up breast patients. In one of these rooms, a *radiographer* carries out *mammograms* (low dose X-rays). In the adjacent room, a consultant *radiologist* performs *ultrasounds* (imaging scan using ultrasonic waves) as well as reporting both ultrasounds and mammograms. *Reporting* involves interpreting images and dictating findings. If the findings warrant further investigation, the radiologist may decide to carry out an image-led biopsy. This involves removing a sample of cells, using ultrasound to locate the area, for analysis by a pathologist. Unlike imaging results, biopsy results take about a week to process.

Following these tests, all new patients have a *results consultation* where they see the breast clinician again to discuss the imaging results. Some may also undergo a biopsy performed by the clinician, or aspiration to drain a cyst. Finally, patients are either discharged or given a further appointment, for biopsy results, further tests or treatment.

2.3.3 Challenges facing the Whittington breast diagnostic clinic

Mirroring the national situation, the numbers of patients referred by their GP to attend the Whittington breast diagnostic clinic increased substantially from 2010 to 2015 (NHS England, 2016), causing increased pressure on services. Although these patients are symptomatic, the percentage of patients diagnosed with cancer is low. Despite this, all these patients may be anxious about the possibility of a cancer diagnosis. Thus it is desirable to keep waiting times at the clinic short and provide reassurance or testing as appropriate. Clinic staff are lacking data about current in-clinic waiting times and patient experiences, but have observed that some patients wait a long time and spend many hours at the clinic.

Currently the decision of which tests to carry out is made on the day, based on a clinical assessment, hence demand for tests is not known in advance. There is a reported peak in numbers of breast patients arriving for ultrasounds in the imaging department in the late morning, leading to long waiting times. In order to smooth the workload, Dr Pauline Leonard suggested that some new patients could have imaging tests first, before seeing the clinician. This is only possible if it is known in advance which patients are likely to require tests.

2.3.4 The potential of using GP referral information

When GPs refer patients to the clinic, they provide referral information containing individual patient details. There is a standard form for referral to hospitals in the North Central London and West Essex Cancer Commissioning Network, which is shown in Figures C.3, C.4 and C.5. Some GPs use an older version of the form (see Figures C.6 and C.7) or write a letter. The numbers of GPs using each of these methods when referring to the Whittington Hospital is discussed later, in Chapter 4.

Currently the referral information provided by GPs is not used in this clinic. Instead, information about patients' symptoms and medical histories is collected again, through the New Patient Assessment Form and initial consultation. The breast clinicians explained that they doubt the accuracy of the information provided by GPs, who due to their profession have less access to patients with breast symptoms. Therefore our research investigates whether there is any quantitative evidence of a link between referral information and diagnostic results. Harnessing data from GP referrals may provide the key to maximising the proportion of patients' time at the clinic that is value-added, as well as making better use of resources.

2.4 Conclusion to background

In this chapter we have seen that cancer is one of a variety of breast conditions that may be diagnosed in a patient presenting with breast symptoms. Although cancer is rare among patients attending breast diagnostic clinics, imaging tests and sometimes biopsies are required in order to distinguish it from other abnormalities. Therefore knowing in advance which patients are likely to be diagnosed with a breast abnormality of any kind could help with planning imaging tests.

Our case study at the Whittington breast diagnostic clinic provides further practical motivation for the research, as well as offering the opportunity to collect data to populate our models. Observing the process from different viewpoints, talking to staff and developing a process map all helped deepen my understanding of the current diagnostic process and related challenges. In this clinic, GP referral information is not currently used and all patients first see a clinician, since prior to this study there was no evidence that information provided in primary care is accurate and complete enough to make diagnostic decisions. There is a perceived problem with current waiting times for imaging, although there is no existing data to quantify this. Therefore collecting data about the current process at this clinic, as well as assessing the completeness and accuracy of referral information, would be of practical benefit.

Following this expansion on the practical motivation, Chapter 3 deepens the theoretical motivation by providing reviews of relevant literature areas. Chapter 4 describes the data collected in the case study clinic described in this chapter and Chapters 5 and 6 describe the modelling work using these data.

Chapter 3

Literature Review

In this chapter we review related literature to set our work in context. Breast diagnostic clinics form part of cancer care services and there are many OR applications in breast cancer care. As explained in Section 1.2, our first research objective involves predicting patients' risks of breast abnormalities of any kind; to the best of our knowledge there are no such models in the literature, but there is a large body of literature predicting breast cancer risks. As well as affecting the risk groups in which patients are placed, patient characteristics also affect the pathways that patients follow through the diagnostic clinic and how long they spend in each stage. Therefore we need to carefully consider the way in which we sample patient characteristics. Consequently we devote one section of the literature review to different methods for generating patient characteristic inputs for DES. Since our modelling approach involves combining classification and DES, we also look at other research that has combined these techniques to show how our study fits in. Our work also relates to the outpatient scheduling literature, particularly to papers that consider appointments with multiple stages.

This chapter proceeds as follows. In Section 3.1 we review the OR literature applied to breast cancer care. Then we review models predicting breast cancer risk groups in Section 3.2. Next, in Section 3.3, we describe alternative ways that patient characteristic inputs are generated for pathway DES models; both so-called "operational" and "clinical-operational" pathways are considered. Then in Section 3.4 we discuss healthcare papers that combine classification models with DES. Finally, in Section 3.5 we briefly review the literature on outpatient appointment scheduling, focusing on the operational performance measures used.

3.1 OR applications in breast cancer care

There is a large volume of OR literature relating to breast cancer care problems in particular, and to general cancer care problems that are applicable to breast cancer. As described in Chapter 1, the cancer care system consists of prevention, diagnosis, staging and treatment services. OR papers related to each of these services are considered in turn here.

I searched for relevant literature using the University of Southampton's search engine DelphiS (University of Southampton, 2016). This covers a large number of databases including Scopus and JSTOR, which are two of the most widely-used academic databases for OR (Brailsford et al., 2009), as well as the biomedical and health database MEDLINE. I searched for "cancer" in abstracts, and "operations research" or "operational research" in the full text, yielding 1536 papers. I filtered this list by adding search terms (in abstracts) as follows: "screen* OR prevent*" to find papers on prevention, "diag* OR stag*" to find papers on diagnosis and staging papers, and "treat*" to find papers about treatment services. Additional papers found via other routes that met the goals of the review were also included. Papers were checked for relevance based on their abstracts; papers were deemed relevant if they used OR techniques to support decision making in cancer care, they applied to breast cancer specifically or treatments used against breast cancer (for example chemotherapy, surgery and radiotherapy) and the full text was accessible. When finding recent examples of chemotherapy, surgery and radiotherapy papers I filtered the list using the following search terms: "chemo*", "surg*", "radio*" and "radiation".

This review describes 16 examples of OR applied to breast cancer prevention (screening), one to breast cancer staging and 28 to treatments that can be used against breast cancer. No papers on the diagnostic process, aside from screening asymptomatic patients, were found. Since we wish to show the range of problems addressed, we prioritise showing examples of different decision areas rather than many from the same area. That is why, for areas that have existing specialised literature reviews, we direct readers to these and provide only more recent examples.

This section proceeds as follows. We review OR papers addressing breast cancer prevention, staging, and treatment services in turn (see Tables 3.1, 3.2 and 3.3). Within these sections, papers are grouped in terms of what problem they are addressing. When discussing the papers, we describe their goals, the techniques used and performance measures.

3.1.1 Breast cancer prevention

It is estimated that about half the cases of the eleven most common cancers are preventable (Soerjomataram et al., 2007). The risk of developing breast cancer can be reduced by lifestyle changes such as avoiding excess bodyweight and reducing alcohol intake (Cancer Research UK, 2016a). Other risk factors include using certain types of hormone replacement therapy and oral contraceptives. However there are risk factors that are not controllable; for example higher age, a family history of cancer and having dense breasts all increase the chances of developing breast cancer. In recent years genetic testing has started to be available to assess whether a person has inherited so-called cancer susceptibility genes (Macmillan Cancer Support, 2016). This has proved controversial because even if a person has a gene that signals they are at higher risk, this still does not mean they will definitely contract cancer.

At what stage of the disease a person is diagnosed can affect treatment decisions and ultimately survival chances. For this reason, in the UK and other countries, breast screening programs are offered in an attempt to diagnose cases earlier, sometimes at the pre-cancerous stage (NHS, 2016). The harms and benefits of screening tests must be carefully balanced when deciding which screening method to use, who should be invited for screening and what the time interval between screening tests should be (Sense about science, 2017). Screening methods are selected by trading-off cost and accuracy. The target population are chosen because it is predicted that they will benefit most, when taking into account harms. If the screening interval is too long, cases which develop between screens (interval cancers) will be missed. However screening too often increases harms such as increased exposure to radiation from mammograms.

We did not find any papers relating to reducing breast cancer risks but found many examples relating to screening for breast cancer. These relate to screening strategies, locating screening facilities, following up screening tests, scheduling and the measurement of screening effectiveness. Table 3.1 summarises key information about these papers.

Screening strategies

Screening strategies involve defining which screening method to use, the target population and the frequency of testing. Evaluating and comparing screening strategies is a popular topic among OR and statistics researchers, with a series of review papers having been published covering this topic (Pierskalla and Brailer, 1994; Stevenson, 1995; Heidenberger, 1996; Knudsen et al., 2007). Most recently, Alagoz et al. (2011) systematically reviewed those papers not described in detail previously. Screening models can be classified as either simulation or analytical. Both are powerful

Table 3.1: OR applications to breast cancer screening

Problem	Reference	Aim	Techniques
Screening strategies	Arrospide et al. (2015)	Evaluation of screening strategy	Discrete event simulation (DES)
	Ayer et al. (2012)	Optimising risk-based screening policy	Partially observable Markov decision process (POMDP)
	Ayer (2015)	Finding sensitivity and screening values for which a screening policy is optimal	Partially observable Markov chain (POMC), nonlinear program (inverse optimisation), heuristic algorithm
	Ayer et al. (2016)	Optimising risk-based screening policy considering adherence	POMDP
	Brailsford et al. (2012)	Comparing fixed-interval and age-based screening strategies considering adherence	DES, logistic regression
	Madadi et al. (2015)	Comparing wide range of fixed-interval and age-based screening strategies considering adherence	POMC
	O'Mahony et al. (2015)	Optimising risk-based screening policy	Mathematical model
	Tejada et al. (2014)	Comparing fixed interval, risk-based and factor-based screening strategies	DES and system dynamics (SD)
	Tejada et al. (2015)	Development of natural history of cancer model for use in above paper	DES and SD
	Wang et al. (2017)	Optimising risk- and age-based screening policy	Logistic regression, misclassification cost criterion
Locating screening facilities	Haase and Mller (2015)	Optimization of preventive health care facility locations	Multinomial logit model within linear optimisation
	Gu et al. (2010)	Optimization of preventive health care facility locations	Multi-objective optimization, heuristic
Following up screening tests	Alagoz et al. (2013)	Optimising use of biopsies and follow-up mammograms	Bayesian network, Markov decision process (MDP)
	Chhatwal et al. (2010)	Optimising use of biopsies	Bayesian network, MDP
Scheduling screening appointments	Baker and Atherill (2002)	Optimization of appointment schedule given attendance probability	Simulation-optimisation, heuristic
Improving measurement of screening effectiveness	Vieira et al. (2011)	Comparing severity of tumours detected by screening compared to self-detected tumours	Discrete time simulation

techniques but have downsides; simulations can only compare a relatively small number of scenarios, whereas analytical models tend to make some unrealistic simplifying assumptions (Alagoz et al.). Furthermore, Alagoz et al. stress the importance of using reliable data to develop models. Koleva-Kolarova et al. (2015) provide a recent review of papers using simulation to assess breast cancer screening strategies in particular. Here the focus is on models that feature in multiple publications. These models have been used to influence screening decisions in different settings, although Koleva-Kolarova et al. warn that they have not been validated with data outside the setting for which they were originally developed.

Given the vast body of literature in this area, we here only provide examples published since Alagoz et al.'s (2011) review. The majority of these breast cancer screening papers either compare mammogram screening strategies (Brailsford et al., 2012; Tejada et al., 2014; Madadi et al., 2015) or optimise the decision of whether to perform a mammogram or not each year (Ayer et al., 2012; O'Mahony et al., 2015; Ayer et al., 2016; Wang et al., 2018). Of these, Madadi et al. consider an especially large set of screening strategies. There has been a shift in focus away from the fixed-interval policies that are common in practice, to consideration of dynamic screening intervals (Ayer et al., 2012; Brailsford et al., 2012; Tejada et al., 2014; Madadi et al., 2015; O'Mahony et al., 2015; Ayer et al., 2016; Wang et al., 2018). These dynamic intervals may vary based on changing risk, age and adherence to screening guidelines. Adherence is modelled in a range of ways from detailed psychological models of behaviour (Brailsford et al., 2012), to changing physician belief about which patients are regular or irregular screeners (Ayer et al., 2016), and to uncertain adherence probabilities based on age and screening interval (Madadi et al., 2015). Arrospide et al. (2015), on the other hand, evaluate how well a particular screening strategy is likely to perform in the long run based on short term real-world results. When choosing a screening method, there is a trade-off between *sensitivity*, the probability of screening correctly identifying a person with cancer, and *specificity*, the probability of screening correctly identifying a person without cancer. Ayer (2015) assess for what range of sensitivity and specificity values a screening interval is best. If more accurate screening tests are introduced in future, the ongoing appropriateness of the screening interval can be judged by this model.

Both simulation (Brailsford et al., 2012; Tejada et al., 2014, 2015; Arrospide et al., 2015) and analytical (Ayer et al., 2012; Ayer, 2015; Madadi et al., 2015; O'Mahony et al., 2015; Ayer et al., 2016; Wang et al., 2018) techniques are used to model breast cancer progression and the screening process, as in earlier papers. There are some very sophisticated models combining multiple approaches, for example Tejada et al. (2014; 2015) combine system dynamics with detailed discrete-event simulations. Contrastingly, O'Mahony et al. (2015) purposely built a relatively simple mathematical model, validated its results against a more complex simulation and found that their model was detailed enough to demonstrate that different risk levels have different optimal screening intervals. The performance measures assessed by the models are diverse and include mortality measures (Brailsford et al., 2012; Tejada et al., 2014; Arrospide et al., 2015; Madadi et al., 2015; Wang et al., 2018), quality-adjusted life-year measures (Ayer et al., 2012; Tejada et al., 2014; Madadi et al., 2015; Ayer, 2015; Ayer et al., 2016) and cost-effectiveness measures (Tejada et al., 2014; O'Mahony et al., 2015).

Locating screening facilities

Gu et al. (2010) and Haase and Müller (2015) aim to determine the best locations for screening facilities. Both case studies involve locating breast cancer screening services. Haase and Müller extend previous work by Zhang et al. (2012) who incorporated a discrete choice model inside an optimisation. Haase and Müller reformulate this non-linear model so that it is linear and able to solve mid-size instances to optimality or close to optimality within one hour. Gu et al., on the other hand, solve their multi-objective optimisation with a heuristic, which consists of trying to improve upon the position of each facility one at a time.

The authors measure the suitability of location sets in different ways. Haase and Müller fix a minimum demand at each centre in order to ensure quality, and choose the location set that maximises participation in screening. However Gu et al. maximise both the efficiency, which is a measure of fairness and is the sum of weighted accessibility values, and the coverage, which measures how many people are within an acceptable distance of a facility. In their case study, the optimum solution improves both the efficiency and coverage with fewer facilities compared to the current set-up.

Following up screening tests

Two related papers address how patients should be managed following abnormal mammogram results (Chhatwal et al., 2010; Alagoz et al., 2013). The aim is to identify those patients at high enough risk of breast cancer to justify the expense, worry and possible harm caused by carrying out further tests. The risk of cancer is calculated from mammogram results and demographic data using a Bayesian network (Chhatwal et al.), and these risk scores are also used by Alagoz et al.. Both papers formulate the problem as a Markov decision process and aim to maximise total expected quality-adjusted life-years. Chhatwal et al. optimise the decision of whether or not to biopsy, while Alagoz et al. additionally consider follow-up mammograms as an option. Chhatwal et al. find that the risk threshold at which patients should be biopsied depends on age. According to Alagoz et al.'s model, fewer biopsies and follow-up mammograms should be carried out than were recommended by radiologists.

Other screening-related studies

Here we discuss two further examples of OR work related to breast cancer screening. Firstly, Baker and Atherill (2002) schedule screening appointments for breast cancer. They predict individual no-show probabilities based upon previous attendance behaviour, then develop a simulation model and a heuristic procedure to optimise a combination of waiting time, idle time and overtime. It was found that screening sessions should start with the patients who are most likely to attend, and end with some overbooked appointments. This is predicted to increase throughput by 10%.

Secondly, Vieira et al. (2011) aim to improve how screening effectiveness is measured. In particular they compare the severity of breast cancer detected by screening to self-detected breast cancers. This is achieved by fitting distributions to data on tumour size and developing a simulation model of tumour progression, screening and self-detection. The output is tumour doubling times for screen-detected and self-detected cancers. It was found that the increased survival benefits achieved through screening for cancer at fixed intervals have been overestimated, due to so-called "length bias". This means that slower growing tumours are more likely than faster growing tumours to be detected at the time of screening, rather than being self-detected in between screens.

3.1.2 Cancer diagnosis and staging

Cancers are diagnosed following an abnormal screening result, development of suspicious symptoms or incidental detection when examining patients for other reasons, for example after being admitted through the emergency department or during an unrelated medical consultation (Elliss-Brookes et al., 2012). After screening, patients with abnormal results are typically invited for further diagnostic tests to confirm or exclude a cancer diagnosis. In the UK, patients who self-detect suspicious symptoms are encouraged to visit their GP who will refer them to a specialist diagnostic clinic if necessary. Patients who have been diagnosed with cancer may undergo further tests to determine the size of the cancer and how much it has spread (Cancer Research UK, 2016a). This staging process helps to determine the most appropriate treatment.

We did not find any papers relating to diagnosis (other than those concerning the screening of asymptomatic patients, described in Section 3.1.1). We found only one paper on staging, which is summarised in Table 3.2.

Staging accuracy

Ekaette et al. (2006) use a Monte Carlo simulation to model staging and radiation treatment of post-surgery breast cancer patients. The aim is to determine how often a

TABLE 3.2: OR application to breast cancer staging

Problem	Reference	Aim	Technique
Staging accuracy	Ekaette et al. (2006)	What is the chance of mis-staging patients and so providing wrong treatment?	Monte Carlo simulation

mistake is made in determining the stage of cancer, leading to patients not receiving the most effective treatment. Additionally, the expected information value of each combination of tests in identifying metastases (secondary tumours) is calculated. It is found that there is a small chance of patients being mis-staged and hence receiving the wrong type of radiation treatment.

3.1.3 Cancer treatment

The appropriate treatment or combination of treatments depends on the type of breast cancer and stage. Additionally, personal preferences regarding convenience and potential side effects also affect choice of treatment. Breast cancer is often treated by performing surgery, along with other treatments such as chemotherapy and radiotherapy to shrink the tumour or stop it spreading (Macmillan Cancer Support, 2016). Cancer patients may need to attend clinics regularly for treatment, so planning access to treatment is important. Cancer treatment centres are concerned with how best to organise their services to meet performance measures, and indeed what these performance measures should be.

Cancer surgery may remove all or part of the tumour. Patients undergoing surgery will likely need to spend some time in hospital afterwards to recover, which means there is an interaction between operating room workload and workload in inpatient wards. Multidisciplinary teams may be needed to perform these complex surgeries, which means scheduling should take into account the availability of different professionals.

Chemotherapy requires patients to take drugs that are designed to destroy cancer cells. Unfortunately healthy cells may also be damaged, and some of the drugs have harmful side effects. When the treatment is fed into a vein, patients typically visit an outpatient clinic (Cancer Research UK, 2016a). Outpatient chemotherapy visits are complicated to schedule because patients receive this treatment in cycles, where the gap between visits is different for different patients. Since pharmacists prepare drugs specially for each patient, drugs may go to waste if patients are too ill for treatment.

Radiotherapy involves using radiation to destroy cancer cells, and may be internal or external. There is a balance between providing a high enough dose to cancer cells while keeping the dose reaching the surrounding normal tissues and organs low. Internal radiation treatment consists of radioactive sources being placed inside the body (Macmillan Cancer Support, 2016). When the radioactive sources are solid, this is known as brachytherapy, and is delivered through implants. External radiotherapy consists of targeting the cancer site with X-ray beams using a linear accelerator (LINAC) machine (Cancer Research UK, 2016a). Patients must undergo a pre-treatment assessment involving imaging, for the specialists to decide the area to target, the appropriate dosage and how to configure the LINAC. 3D conformal radiotherapy uses multileaf collimators to block parts of the beams to match the shape of the tumour. Intensity modulated radiotherapy (IMRT) is a more advanced method where the beams are divided into beamlets with different intensities.

This section proceeds as follows. First we discuss OR models relating to treatment decisions, then access to treatment followed by performance of cancer treatment centres. Then we address scheduling of cancer surgery, chemotherapy and radiotherapy. Chemotherapy planning and different types of radiotherapy planning are addressed next. We finish the section with other treatment-related studies (deciding which drugs to produce in advance). The papers are summarised in Table 3.3.

Access to treatment

Here we describe OR approaches to improving access to cancer treatment or medicines. Firstly, Cotteels et al. (2012) consider the problem of locating radiotherapy centres in Belgium. They use the p-median method, which minimizes the total demand-weighted distance between each patient and the nearest radiotherapy centre, given that there are p centres. The locations recommended by the model are compared to the actual locations, in order to highlight inequities in provision. They found that the current locations are for the most part near the locations suggested by the model.

Performance of cancer treatment centres

In this section we discuss papers that address how well cancer treatment centres perform. According to Brailsford and Vissers (2011), two stages of developing and managing health services relate to performance: defining performance criteria and managing how well these are met. Some performance measures can be improved by changing the scheduling of patients or resources. OR approaches to optimal scheduling of chemotherapy and radiotherapy services are discussed separately in following sections.

Santos et al. (2007) focus on designing appropriate performance measures for a radiotherapy department. It is reported that the current performance measurement is inadequate because only a few aspects are considered and there is no systematic regular process for measuring performance. The authors use system dynamics and

Table 3.3: OR applications to treatments used against breast cancer

Problem	Reference	Focus	Aim	Technique
Access to treatment	Cotteels et al. (2012)	Radiotherapy	Optimising locations of treatment centres	Optimisation, p-median method
Performance of cancer treatment centres	Santos et al. (2007)	Radiotherapy	Designing appropriate performance criteria	System dynamics (SD) and multi-criteria decision analysis
	Baesler and Seplveda (2001)	Chemotherapy	How many resources (treatment chairs, nurses, laboratory staff and equipment, and pharmacy staff and equipment) are required?	Goal programming simulation-optimisation, genetic algorithm
	Matta and Patterson (2007)	Chemotherapy and radiotherapy	Compare strategies to improve performance of treatment centre	Discrete event simulation (DES)
	Werker et al. (2009)	Radiotherapy planning	Compare strategies to reduce treatment planning time	DES
Scheduling	Mutlu et al. (2015)	Breast surgery	Optimising multidisciplinary team schedules	Integer program, simulation
	Lim et al. (2016)	Surgery	Optimise assignment of nurses to surgery cases and optimise lunch breaks	Multi-objective optimisation (mixed integer program), swap heuristic, column generation approach

Continued on next page

Problem	Reference	Focus	Aim	Technique
	Mobasher et al. (2011)	Surgery	Optimise assignment of nurses to surgery cases	Multi-objective optimisation (mixed integer program), a new version of modified goal programming, solution pool method
	Vanberkel et al. (2011)	Surgery	Comparing impact of different surgical block schedules on workload in other departments	Analytical models involving queuing theory
	Hahn-Goldberg et al. (2014)	Chemotherapy	Optimising patient appointment times within a day	Constraint programming optimisation, “shuffle” algorithm
	Santibez et al. (2012)	Chemotherapy	Comparing changes to booking process, optimising patients appointment times and evaluating real impact of service changes	DES, multi-objective optimisation (integer program)
	Woodall et al. (2013)	Chemotherapy	Optimising nurse shift start times	DES, optimisation
	Bikker et al. (2015)	External radiotherapy	Optimising doctors’ allocation to pre-treatment appointments	Integer program, DES
	Castro and Petrovic (2012)	External radiotherapy	Optimising pre-treatment appointments	Multi-objective optimisation (mixed integer programs), problems solved hierarchically

Continued on next page

Problem	Reference	Focus	Aim	Technique
	Petrovic et al. (2011a)	External radiotherapy	Optimising treatment start days	Multi-objective optimisation, genetic algorithm
	Saur et al. (2012)	External radiotherapy	Optimising treatment start days	Markov decision process, approximate dynamic programming
	Conforti et al. (2010)	External radiotherapy	Optimising treatment start days	Integer program
	Legrain et al. (2015)	External radiotherapy	Optimising treatment start days	Stochastic optimisation, greedy and primal-dual algorithms
Treatment planning	Alam et al. (2013)	Chemotherapy	Optimising treatment plan	Multi-objective optimisation, closed-loop optimal control model, genetic algorithm
	Jalalimanesh et al. (2017)	Intensity-modulated radiation therapy (IMRT)	Optimising number of treatments (fractions) and doses (fraction size)	Agent-based simulation, reinforcement learning
	Dias et al. (2014)	IMRT	Optimising beam angles	Genetic algorithm and neural network
	Mahmoudzadeh et al. (2016)	IMRT, breast cancer	Optimising beamlet intensities under breathing uncertainty	Robust optimisation, conditional value-at-risk, decomposition (constraint generation)
	Chan et al. (2014)	IMRT, breast cancer	Optimising beamlet intensities under breathing uncertainty	Robust optimisation, conditional value-at-risk
	Cabrera et al. (2014)	IMRT	Optimising beamlet intensities	Multi-objective optimisation

Continued on next page

Problem	Reference	Focus	Aim	Technique
	Taskin and Cevik (2013)	IMRT	Optimising leaf sequencing	Mixed-integer program, combinatorial Benders decomposition
Chemotherapy drug production policy	Masselink et al. (2012)	Chemotherapy	Deciding which chemotherapy drugs to prepare in advance	Analytical models involving queuing theory, DES
	Vidal et al. (2010)	Chemotherapy	Deciding which chemotherapy drugs to prepare in advance	AHP
Optimising treatment	Holder and Llagostera (2008)	Photodynamic therapy	Modelling effects of treatment	Linear program, interior-point algorithm

multi-criteria analysis to identify summary indicators (relating to capacity, access, efficiency and outcomes) that better represent overall performance, help managers understand some underlying causes for the observed performance and predict the impact of changes on performance.

Baesler and Sepúlveda (2001), Matta and Patterson (2007) and Werker et al. (2009) all focus on how cancer treatment centre performance can be improved. Matta and Patterson present a framework for combining multiple performance measures across multiple dimensions into a single score. They develop a discrete-event simulation model of a cancer treatment centre offering both chemotherapy and radiotherapy. For this case study, the performance score consists of the average system time and average overtime weighted by throughput and frequency respectively, and stratified by day of week, disease type and patient routing through facilities. A variety of process, resource and scheduling changes are simulated and compared in terms of the performance score. In this way eleven changes that individually improve overall performance are identified and then their joint impact on performance is assessed. As a result of the study, the treatment centre has changed how appointments are scheduled, increased capacity and introduced a separate blood testing area.

Baesler and Sepúlveda (2001) use a goal programming simulation-optimisation method to find the numbers of different resources required in a chemotherapy centre. Multiple objectives are considered (waiting time, chair utilization, closing time and nurse utilization), and weighted based on their importance. A genetic algorithm is used to find possible solutions to the problem and succeeds in finding a configuration of resources that is at least as good as the current configuration in all four objectives. In particular it improves nurse utilization and needs just one extra chair.

Werker et al. (2009) also develop a discrete-event simulation. Specifically they model the radiation therapy pre-treatment process in order to find ways to reduce its length (the total planning time). This consists of multiple stages required to plan treatment including consultations, oncologist input, dose planning and verification stages. Three different types of staff are involved in the planning: oncologists, radiation therapists and medical physicists. It was found that shorter, more consistent delays to oncologists being available would reduce the total planning time.

Surgery scheduling

Mutlu et al. (2015) optimise the individual schedules of members of a multidisciplinary team in order to maximise the time that they are available to work together. They formulate the problem as an integer program that includes restrictions relating to preferences and availability for clinic work. Their case study, optimising the schedules of plastic and oncologic surgeons, succeeded in increasing the number of sessions when teams of two surgeons were available for breast cancer surgery by 94%.

Vanberkel et al. (2011), on the other hand, compare surgical block schedules, that is, which blocks of operating room time are assigned to different specialties. They use an analytic approach involving queuing theory to output the workload in different wards created by patients recovering from surgery. In particular, their model outputs the following ward-level statistics per day: 90th percentile occupancy, expected admissions, expected discharges and expected numbers of patients in each day of recovery. This modelling work helped the Netherlands Cancer Institute-Antoni van Leeuwenhoek Hospital choose a new surgical block schedule after opening an extra operating room. The new schedule succeeded in smoothing the numbers of beds required on the ward compared to alternative schedules considered.

Two related studies focus on the scheduling of nurses in operating rooms (Mobasher et al., 2011; Lim et al., 2016). Mobasher et al. develop a multi-objective mixed integer program, with six soft constraints modelled as penalised objectives. The aim is to assign nurses to surgery cases taking account of their specialties and skills. Lim et al. extend this work by adding a second optimisation to maximise the number of nurses who can take breaks over lunchtime. Four different methods are proposed to solve the assignment problem; the swap heuristic and column generation approach proposed in later work (Lim et al., 2016) find acceptable solutions much more quickly than the earlier ones, a new version of modified goal programming and the solution pool method (Mobasher et al., 2011).

Chemotherapy scheduling

There are papers considering a variety of chemotherapy scheduling problems. Hahn-Goldberg et al. (2014) address the online scheduling problem: how to optimise a daily schedule of patients when new requests must be scheduled as they arrive. First a template of appointment times is created by optimising over a sample of previous appointment requests. As requests arrive they are assigned appointment times within this template, until this is no longer possible, when the schedule is re-optimised based on the requests that have arrived up to that point. The optimisations are solved using constraint programming to minimise the total working time (makespan) on a particular day. Finally an algorithm to shift appointment times is applied to deal with cancellations. Using this approach the authors are able to improve the makespan compared to current scheduling practices by up to 20%.

The goals of the patient scheduling study by Santibáñez et al. (2012), on the other hand, are to notify patients of their appointment at least a week in advance and to reduce the waiting list size. In the first stage of the project at a cancer centre in

British Columbia (Canada), a detailed process review of current booking practices and a patient survey were carried out. A discrete-event simulation is developed to test changes to booking processes. Then a multi-objective optimisation program is presented which creates a daily appointment schedule including nurse allocation. This aims to satisfy patient preferences, balance the numbers and complexity of patients per nurse throughout the day, assign clinical trial patients to specialised nurses and limit pharmacy workload dependent on available resources. Resulting from the modelling work, changes were made to booking practices and software for the optimisation tool was introduced. Following the changes, the median wait list size decreased, the numbers of patients notified of their appointment less than a week in advance decreased and patient satisfaction increased.

Woodall et al. (2013) also describe a discrete-event simulation of a cancer centre, this one being in North Carolina (USA). The simulation enables the authors to identify the service bottleneck: nurses who administer chemotherapy to particular disease groups. Secondly the authors develop a mixed integer program to optimise the weekly and monthly schedules of different nurse types to minimise the hours of unmet demand. Thirdly they use simulation-optimisation to determine the best nurse shift start times so that average patient waiting time is minimised. In particular it was found that shifts should be allowed to start on the half hour and that more nurses were needed. Consequently the cancer centre adjusted shift times and hired more nurses. Additionally the models were used to help plan staffing levels for a new cancer centre.

Radiotherapy scheduling

In this section we focus on examples of papers that have addressed the pre-treatment scheduling and treatment scheduling problems. Vieira et al. (2016) produced a more inclusive review on radiotherapy resource use that includes 18 papers on patient scheduling.

A study based in the Academic Medical Centre in Amsterdam offers potential improvements to radiotherapy treatment access times, which are the time from referral until treatment starts (Bikker et al., 2015). The authors develop an integer linear program to optimise when doctors should be allocated to pre-treatment tasks in order to minimise the access time for all patient types. In order to assess the impact of their schedules in a more realistic stochastic situation, they conduct experiments in a discrete-event simulation model. On the other hand, Castro and Petrovic (2012) model the problem from the patient perspective. Pre-treatment appointments consist of multiple stages requiring different resources. They solve this scheduling problem as a hierarchy of optimisation problems with different waiting time objectives. Since this approach does not yield a feasible solution in a short enough time, they also experiment with six different rules to generate the initial solution to the first problem more quickly.

Several papers schedule patients for radiotherapy treatment, by determining what day each patient should start treatment given available LINAC capacity, and assuming the timing of continuing treatment follows a fixed pattern. Some of these tackle the offline problem (Petrovic et al., 2011), whereas Sauré et al. (2012) and Legrain et al. (2015) solve the online problem, where appointment requests are scheduled as they arrive. Objective functions involve minimising days waiting to start treatment (Conforti et al., 2010; Petrovic et al., 2011; Sauré et al., 2012; Legrain et al., 2015), days overdue to start treatment (Petrovic et al., 2011; Legrain et al., 2015), overtime (Sauré et al., 2012; Legrain et al., 2015) and booking decisions postponed (Sauré et al., 2012). All these papers consider patients with different priorities, for example, curative and palliative groups.

The authors of these treatment scheduling papers use a variety of approaches (Conforti et al., 2010; Petrovic et al., 2011; Sauré et al., 2012; Legrain et al., 2015). Conforti et al. provide an integer program and solve it using exact methods. Petrovic et al. compare the performance of different genetic algorithms to solve their multi-objective optimisation problem and find that the algorithm prioritising emergency patients performs best overall. Sauré et al. formulate the problem as a Markov decision process which is reformulated using approximate dynamic programming to obtain a linear program, for which the dual problem is solved using column generation. They simulate the generated scheduling procedure for a case study, and find that there would be improvements to waiting times compared to current practice. Legrain et al. first solve an offline optimisation problem using two algorithms. Then the uncertainty of treatment duration and new patients arriving is incorporated to produce a stochastic online optimisation. Algorithms to solve this are also presented. Results show that the model improves on current scheduling practice.

Chemotherapy treatment planning

Mathematical approaches to optimising chemotherapy plans have recently been reviewed by Shi et al. (2014), and the only paper our search found that is more recent than the scope of that review is by Alam et al. (2013). For a particular patient, chemotherapy treatment planning involves balancing two objectives: destroying as many cancer cells as possible while minimising the toxicity to normal cells (Cancer Research UK, 2016a). An added complexity is that cancer cells may become resistant if exposed to drugs for a long enough time.

Shi et al. (2014) reports that many papers formulate the problem as an optimal control model and aim to shrink the tumours as much as possible over a fixed time period, given tumour growth rate and limits on the chemotherapy drug dose. These models are single or multi-objective optimisations that involve solving systems of differential equations. Researchers may consider dosages, at what time points to treat patients (for example cyclically or continuously) and single or multiple drugs. Since these differential equations are challenging to solve analytically, a range of approaches including simplifying the model then solving it exactly, approximations and heuristics have been applied (Shi et al.). Some authors add even further complexity by modelling the problem stochastically to capture randomness in the rates of tumour growth and drug-induced shrinkage (Shi et al.). In order to encourage the application of these models in clinical practice, Shi et al. recommend focusing on a specific cancer type, including cost as an objective, modelling how treatment plans are updated and only considering solutions that are feasible in practice.

Alam et al. (2013) provide a recent example of a multi-objective optimisation of chemotherapy plans, and consider multiple drugs. Compartment models are used to describe cancer cell change, where cells in different phases (resting, dividing or dead) are affected by the drugs to a greater or lesser extent. The objectives are: reducing the numbers of both resting and dividing cancer cells, maximising the number of normal cells, reducing the toxicity and keeping the drug concentration within an acceptable limit. This sophisticated model is formulated as a closed-loop optimal control model and solved using a genetic algorithm. An extensive set of experiments with different numbers of drugs is carried out and comparisons to results of other models are made. Unfortunately, attention to Shi et al.'s (2014) recommendations for improving practical relevance are not evident here, but the authors do perform robustness analysis.

Radiotherapy treatment planning

Here we describe OR approaches to planning different types of radiotherapy. First we discuss a review paper on high dose rate brachytherapy, a type of internal radiotherapy. Then we discuss the papers on intensity modulated radiation treatment (IMRT), a type of external radiotherapy.

We refer interested readers to De Boeck et al.'s (2014) review of dose optimisation models for high dose rate brachytherapy between 1990 and 2010. We did not find any more recent papers. These models optimise how long (dwell time) a radioactive source should stay in each position (dwell location). De Boeck et al. found that in the earlier papers, forward planning is the norm, where the dwell times are changed iteratively and the dose is calculated each time. These do not take into account the anatomy of particular patients. In later papers there is a move to inverse planning, where the desired dose is specified in advance, and images of individual patients' anatomies are used in planning. Usually the positions of the catheters containing the radioactive sources is given, but some papers also optimise these positions (De Boeck et al.). The review found that models have been developed for a range of cancers, and both exact and heuristic methods have been used to solve them. De Boeck et al. categorise the multi-objective models depending on whether the importance of each objective is decided in advance, during or after the optimisation. It is recommended that future papers concentrate on making models more clinically relevant as well as incorporating uncertainty.

There is a large body of literature on designing IMRT plans and several review papers (Bortfeld, 2006; Ehrgott et al., 2010; Censor and Unkelbach, 2012). Bortfeld (2006) discusses the mathematical, physical and technological developments relating to IMRT. The author describes the typical problem formulation, which is to calculate the necessary beamlet intensities given prescribed doses that should reach the tumour site (target). This is known as inverse planning. Later Ehrgott et al. (2010) reviewed optimisation approaches to three related problems: (1) fluence map optimisation, which consists of finding the best set of beamlet intensities, (2) beam angle optimisation and (3) the segmentation problem, which is how to configure the multileaf collimators to achieve this. Censor and Unkelbach (2012) describe the two key approaches to solving the inverse problem: continuous analytic techniques and fully-discretised algebraic methods. They explain the change in perspective to considering the problem as an optimisation where the damage to healthy tissue should be minimised.

We compare examples of IMRT treatment planning papers published since Censor and Unkelbach's (2012) paper that were identified from our search strategy. Some of these address the fluence map optimisation problem (Cabrera et al., 2014; Chan et al., 2014; Mahmoudzadeh et al., 2016). Jalalimanesh et al. (2017) optimise the number of fractions (treatment sessions) and dose per fraction, rather than assuming constant doses as is commonly the case. Dias et al. (2014) address the beam angle optimisation problem. The leaf sequencing problem is considered by Takin and Cevik (2013): which sequence of rectangular aperture shapes and intensities to use so that the total planned intensities are achieved.

There are many objectives to consider in radiotherapy planning optimisation. Importantly, the tumour should receive enough radiation, called target coverage, while normal tissues and surrounding organs should receive as little as possible, called organ sparing. For example, some models minimise the conditional value-at-risk, which is the average dose that is received in the parts of an organ that receive the highest dose (Chan et al., 2014; Mahmoudzadeh et al., 2016). Different optimisation, heuristic and simulation methods have been used on these problems. Cabrera et al. (2014) prove theoretical results for solving a particular class of multi-objective optimisations through a series of single-objective optimisation problems. Takin and Cevik (2013) use combinatorial Benders decomposition to break down their mixed integer program into an integer program master problem and a linear program subproblem, then compare heuristic and exact solution procedures. On the other hand, Dias et al. (2014) provide a nonlinear formulation and find solutions with a genetic algorithm incorporating a neural network to estimate the fitness functions quickly. Jalalimanesh et al. (2017) develop an agent-based simulation of tumour growth and use the Q-learning algorithm, a type of reinforcement learning, to find an optimal solution. A series of papers use robust optimisation to capture the uncertainty in patients' breathing patterns (Chan et al., 2014; Mahmoudzadeh et al., 2016). Mahmoudzadeh et al. (2016) use constraint generation (a decomposition method) to solve the robust optimisation first presented by Chan et al. (2014).

Some authors report how their methods improve solutions or computational time compared to other approaches (Takin and Cevik, 2013; Dias et al., 2014). Mahmoudzadeh et al. (2016) find that adding one constraint each time their problem is re-solved is slowest, and it is fastest to add several constraints each time. Compared to the standard treatment planning method at the time, Chan et al.'s (2014) approach better matches the planned dose with the actual dose received, since they incorporate breathing uncertainty. Jalalimanesh et al. (2017) show that the dose should be varied over time as the tumour changes size. Cabrera et al. (2014) demonstrate how to generate infinitely many Pareto-optimal solutions to their multi-objective optimisation problem.

Other treatment-related studies

Holder and LLagostera (2008) model how best to apply photodynamic therapy to deep tissue cancers, and assess whether the only drug approved in the USA for this treatment at the time would yield acceptable results. They develop biological models of the concentration of the drug in different tissues, and the rate at which cellular damage occurs when the drug is activated with light. Their linear model to optimise the application of the light source is adapted from a model for external radiotherapy planning. They find that even under the optimal alignment of the light source, the drug can not target the tumour closely enough, and the surrounding normal tissues are damaged to an unacceptable extent.

Vidal et al. (2010); Masselink et al. (2012) both address the decision of which chemotherapy drugs to produce in advance, rather than on demand directly before use. These drugs are prepared for specific patients, and so become useless if patients are too ill for treatment and the drugs expire before they can be used. Vidal et al. interviewed pharmacists at a French pharmacy, which prepares drugs for medical facilities in the area, to find out criteria that make a drug suitable for preparing in advance. The relative importance of each criterion is determined using the analytical hierarchy process. The most important criteria are found to be drug stability, time between ordering and when needed, as well as total annual volume of the drug. Following the study, a decision support tool, designed to assist in choosing the drugs to produce in advance, was adopted by other pharmacies in France.

Masselink et al. (2012) worked with a pharmacy that is attached to a chemotherapy unit in the Netherlands. Unlike Vidal et al. (2010), they focus on the effect on patient waiting times of preparing some chemotherapy drugs in advance. This is achieved by modelling the drug order queue as well as the linked queue of patients waiting for treatment. For the case study, a discrete-event simulation is developed. Analytical expressions to approximate patient waiting times under different policies are derived, which are also applicable to other settings. The policies considered involve making the cheapest drugs (up to some threshold price) in advance and optionally reallocating drugs when patients are too ill for treatment. Using the results of the modelling work, managers from the pharmacy and chemotherapy unit agreed on which drugs should be made in advance and that some drugs should be reallocated. Model results suggest this will cause waiting times to be halved with only a 1-2% increase in cost.

3.1.4 Conclusion to OR applications in breast cancer care

Our review showcases examples of OR techniques applied to problems throughout breast cancer care services. These problems lend themselves to OR modelling because of conflicting objectives, large numbers of options to be compared and patient-specific parameters. A key strength of many OR methods is that they can make goals, constraints and uncertainties explicit. Using our search strategy, a substantial amount of research on screening strategies was identified, as well as on treatment planning and scheduling. Arguably these areas are particularly suited to OR modelling, since they have the characteristics described at the start of this paragraph. On the other hand, we uncovered comparatively few examples of OR models applied to reducing breast cancer risks, diagnostic services and staging. Our project contributes to research on the diagnosis of symptomatic breast patients.

A limitation of our search strategy is that there may be papers using OR techniques applied to breast cancer care (including breast diagnostic services) that did not mention “operational research” or “operations research”. However we believe that our search provided a roughly representative sample of papers, meaning that there are likely to be relatively few OR applications to breast cancer diagnostic services.

3.2 Breast cancer risk classification models

We next review OR papers that classify patients into breast cancer risk categories. In this section we use the word “risk” to encompass the following measures: the risk of developing breast cancer, the risk of having breast cancer given certain symptoms or test results, the risk of the cancer returning, called *recurrence*, and the risk of death within a certain time of diagnosis. Our focus is on papers predicting risk categories, rather than continuous measures i.e. using classification rather than e.g. linear regression. We also exclude papers using cell, gene or image data.

The aim of this review is not a systematic search but rather to find examples of breast cancer risk classification in the OR literature. The following search strategy was used. To find examples of classification models (apart from logistic regression, which was dealt with separately) I searched for "breast AND cancer AND data mining AND classification AND predict* NOT gene* NOT cell*" in full texts again using the DelphiS search engine. Papers were checked for relevance based on their abstracts. Logistic regression is widely-used to predict cancer risks, so we decided to concentrate solely on recent examples in our review. This was achieved by using the search string "breast AND cancer AND logistic regression AND predict* NOT gene* NOT cell*", then identifying recent examples addressing different phases of cancer care services. These logistic regression papers are described Section 3.2.4.

This section proceeds as follows. First in Section 3.2.1, the types of patient data that can be used as model inputs are outlined. Then in Section 3.2.2 the purpose of each classification model from our initial search is explained. Following this in Section 3.2.3 performance measures are described. Next in Section 3.2.4 methods including those predicting continuous measures, logistic regression and other common classification methods are discussed. Finally in Section 3.2.5 we describe to what extent these models have found practical use, and conclude with Section 3.2.6.

3.2.1 Types of patient data

In general, as a patient spends more time in the cancer care system, more data can be collected about them (see Figure 3.1). Of course, not all the data have predictive value, and part of the process of model building is to ascertain which input variables to include (see Sree et al. (2010) for some feature selection methods). Before using any cancer services, data about a patient's medical history, demographics or family history of cancer may be known from general medical appointments. Further information becomes available when a patient with symptoms visits their GP. Tests performed at the screening, diagnosis, staging, treatment and post-treatment stages provide further data including tumour features, blood test values and information about cell samples obtained via biopsy. The treatments administered can also be used as input variables to predict survival or recurrence.

The datasets and types of input variables used in the breast cancer risk classification models identified by our search are listed in Table 3.4 alongside further details. Mangasarian et al. (1995), for example, use the results of fine-needle aspiration tests as input data. The dataset was donated by Dr. William H. Wolberg and is publicly available by searching on the UCI Machine Learning Repository for "Wisconsin

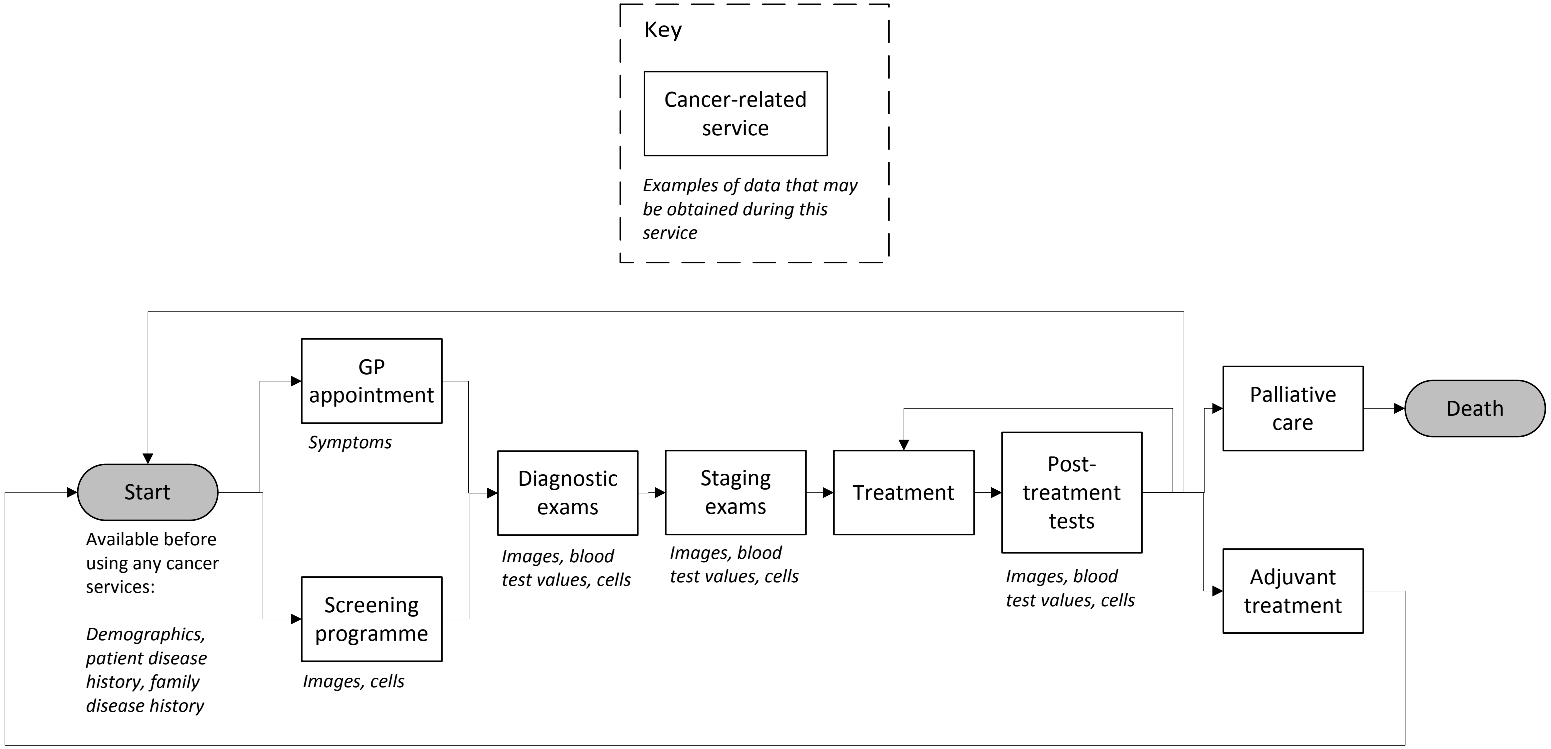

Key
Cancer-related service
Examples of data that may be obtained during this service
GP appointment
Symptoms
Start
Available before using any cancer services:
Demographics, patient disease history, family disease history
Screening programme
Images, cells
Diagnostic exams
Images, blood test values, cells
Staging exams
Images, blood test values, cells
Treatment
Post-treatment tests
Images, blood test values, cells
Palliative care
Death
Adjuvant treatment

Figure 3.1: Pathway through cancer services and data availability

diagnostic data" (Lichman, 2016). Some other datasets available on the UCI repository relate to breast cancer prognosis, also from Wisconsin, and recurrence, from Ljubljana. Many researchers use these datasets to test the performance of different classification algorithms (see for example those in Table 3.4). Risk assessments by clinicians can also be used as input variables. Before performing diagnostic tests, clinicians may provide a prior suspicion score (Sree et al., 2010), while at the treatment stage they may estimate the risk of recurrence or death (Jonsdottir et al., 2008).

3.2.2 Purpose of prediction

Patient data, as described above, when paired with outcomes (e.g. survive or not, cancer or not), are used to predict the cancer risk levels of future patients. The predictions made by each of the papers considered are listed in Table 3.4. Many papers test the performance of algorithms, rather than aiming to change practice (Lundin et al., 1999; Pendharkar et al., 1999; Abbass, 2002; Hung et al., 2002; West et al., 2005; Ryu et al., 2007; Jonsdottir et al., 2008; Jin et al., 2012).

Until a person visits a screening service or a GP due to suspicious symptoms, they are, to the best of their knowledge, unaffected by cancer. Pendharkar et al. (1999) use patient characteristics to predict the risk of developing breast cancer. Such predictions could help to diagnose patients and determine how often particular patients should have screening, although in this paper the primary objective is to test the performance of a predictive method.

Breast cancer screening and diagnostic tests such as mammograms, ultrasounds and biopsies produce results that require expert interpretation. Mangasarian et al. (1995), Tourassi et al. (2001) and Winkler et al. (2013) all predict breast cancer diagnosis from test results. Accurate predictions would remove the need for further, potentially risky and invasive tests. Unlike these models, Jin et al. (2012) predict both whether a lump was benign or malignant, and whether malignancies would recur.

A common staging test for breast cancer is to remove lymph nodes for examination, allowing clinicians to estimate prognosis (National Breast Cancer Foundation, 2016). In order to prevent the need for this procedure, Lundin et al. (1999) predict the five-, ten- and fifteen- year breast-cancer-specific survival of patients from other available information. Çakir and Demirel (2011) predict which treatment breast cancer patients will be prescribed after surgery. The purpose is to decide the most appropriate treatment plan for future patients.

Some authors use the treatment type as an input to predict recurrence or survival. For example, Jonsdottir et al. (2008) predict which breast cancer patients will be disease-free within five years. Other models predict the risk of breast cancer recurrence within four years (Razavi et al., 2007) and within ten years (Štrumbelj et al., 2010).

Table 3.4: Classification models to predict breast cancer risk

Paper	Service	Prediction	Data set (Number of cases)	Variables
Pendharkar et al. (1999)	Before accessing services	Cancer	Patients at surgery centre in Pennsylvania, USA (479)	Age, menopausal status, use of hormones
Abbass (2002), Hung et al. (2002), West et al. (2005)	Diagnostic tests	Cancer	UCI - Breast Cancer Wisconsin (Diagnostic) (569)	Results of fine-needle aspiration tests
Jin et al. (2012)	Diagnostic tests	1.Cancer 2.Recurrence	1.UCI - Breast Cancer Wisconsin (Diagnostic) (569) 2. UCI - Breast Cancer Wisconsin (Prognostic) (198)	Results of fine-needle aspiration tests
Winkler et al. (2013)	Diagnostic tests	Cancer	Patients at General Hospital (AKH) Linz, Austria (20,819)	Blood values, tumour markers
Sree et al. (2010)	Diagnostic tests	Cancer	Dataset from manufacturer of electropotential diagnostic machine (291)	Family history, age, tumour features, electropotentials, prior suspicion score
Ryu et al. (2007)	Diagnostic tests	1.Cancer 2.Recurrence	1. UCI - Breast Cancer Wisconsin (Diagnostic) (569) 2. UCI - Breast Cancer (286)	1. Results of fine-needle aspiration tests 2. Age, menopausal status, tumour features
Tourassi et al. (2001)	Diagnostic tests	Cancer	Mammographic descriptions with biopsy results (500)	Mammographic findings, age, family history, personal history, menopausal status, use of hormones

Continued on next page

Paper	Service	Prediction	Data set (Number of cases)	Variables
Mangasarian et al. (1995)	Diagnostic tests	Cancer	UCI - Breast Cancer Wisconsin (Diagnostic) (569)	Results of fine-needle aspiration tests
Lundin et al. (1999)	Staging tests	5-, 10-, 15- year cancer-specific survival	Breast cancer cases in Turku, Finland (951)	Symptoms, lab results and tumour features
Çakir and Demirel (2011)	Treatment	Treatment type	Breast cancer patients at Ankara Oncology Hospital, Turkey (462)	Treatment, tumour features, age
Delen et al. (2005)	Treatment	5-year survival	Surveillance, Epidemiology, and End Results (SEER) Cancer Incidence Public-Use Database in the USA (433,272)	Sociodemographic data, tumour features, treatment
Štrumbelj et al. (2010)	Treatment	Recurrence within 10 years	Breast cancer cases in Ljubljana (1035)	Tumour features, family history, age
Razavi et al. (2007)	Treatment	Recurrence within 4 years	Sweden breast cancer registry data (3699)	Age, tumour features, lab results
Jonsdottir et al. (2008)	Treatment	Recurrence within 5 years	Breast cancer patients in Iceland (257)	Tumour features, treatment, comorbidities, age, risk of recurrence/death according to doctor

Clinicians could use such models as decision support tools to determine the appropriate level of follow-up and further treatment. Delen et al. (2005) predict the five-year survival of breast cancer sufferers following treatment.

3.2.3 Performance measures

Different measures are available to assess how well methods perform. Some authors calculate only the prediction accuracy, which is the percentage of correct classifications (Pendharkar et al., 1999; Çakir and Demirel, 2011; Winkler et al., 2013). As well as the prediction accuracy, Abbass (2002) also assesses the computational time of different algorithms. West et al. (2005) and Ryu et al. (2007) consider how well classifiers extend to other datasets using the generalisation error. This is the average difference between the prediction errors of the original training dataset and further test datasets. Jonsdottir et al. (2008) calculate kappa, which is the agreement between a model and the truth, adjusted for chance agreement.

Delen et al. (2005), Razavi et al. (2007) and Jonsdottir et al. (2008) calculate the confusion matrix, which consists of the numbers of true positives, true negatives, false negatives and false positives. When predicting cancer, a case is true positive if a patient with cancer is predicted to have cancer, and true negative if a patient without cancer is predicted not to have cancer. Correspondingly false positives are those without cancer but predicted to have cancer, and false negatives are those with cancer but predicted not to have cancer.

In breast cancer prediction, it tends to be most important to correctly classify as many cancer patients as possible, which may be achieved at the expense of classifying non-cancer patients correctly. To this end, the *sensitivity*, (also known as the true positive rate), measures the proportion of cancers that are classified correctly. On the other hand, the *specificity*, or (also known as the true negative rate) measures the proportion of non-cancers that are classified correctly. Many authors report the sensitivity and specificity of various classifiers (Lundin et al., 1999; Delen et al., 2005; Jonsdottir et al., 2008; Sree et al., 2010; Jin et al., 2012). Tourassi et al. (2001) fix the acceptable sensitivity at 95% and report the corresponding specificity value. Relatedly, the *Receiver Operating Characteristic* (ROC) curve is a plot of the true positive rate (sensitivity) against the false positive rate (one minus specificity) at different classification thresholds (cut-offs). The area under this curve (AUROC) gives the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one. Lundin et al. (1999), Tourassi et al. (2001), Razavi et al. (2007) and Sree et al. (2010) all report this statistic.

3.2.4 Methods

In this section we describe methods to predict breast cancer risk. Although models with continuous outcomes, such as regression, are outside our scope, we first briefly describe such models for completeness. It is useful to understand how they are used to predict breast cancer risk so a selection of well-known models that have impacted clinical decisions are discussed. Then we summarise some recent logistic regression papers. After this, we describe the more commonly used methods among the papers in our search, as well as some of their advantages and disadvantages. The papers considered usually compare several classification methods, in order to determine which performs best for the specific problem and dataset (see Table 3.5).

Methods to predict continuous outcomes

In the UK, the charity Macmillan Cancer Support sponsored the development of computer-based decision support tools to prompt GPs when a patient's symptoms indicate an increased risk of cancer (Macmillan Cancer Support, 2016). The risks are calculated from patient medical records contributed by many GP practices. The most recent analysis calculates a patient's risk of developing each of eleven common cancers, including breast cancer, for every month up to 15 years by using Cox proportional hazard models (Hippisley-Cox and Coupland, 2015). Both the five- and ten- year predictions are available to calculate online at `http://qcancer.org/10yr/`. On the other hand Colditz and Rosner (2000) apply a different method, nonlinear Poisson regression, in large-scale studies using nurses' health data to find risk factors for breast cancer.

Notably, the widely-used Nottingham Prognostic Index is the outcome from a multiple regression model (Haybittle et al., 1982). This score gives the stage of breast cancer and is based on just three variables: the tumour size, number of lymph nodes involved and the grade. The grade, in turn, is based on histological factors, that is, the structure of the tumour when looking under a microscope.

Following surgery for breast cancer, patients may have adjuvant treatment to prevent the cancer returning. The online tool called PREDICT (available at `http://www.predict.nhs.uk/predict.html`) calculates breast cancer mortality dependent on different adjuvant treatments (Wishart et al., 2010). This prediction is obtained using Cox proportional hazards. Another tool with the same purpose is called Adjuvant! Online (Ravdin et al., 2001). In this case, survival is calculated using an actuarial life table technique and adjusted conditionally using a Bayesian method.

Table 3.5: Classification methods used to predict breast cancer risk categories. ANN-Artificial neural network, LDA-Linear discriminant analysis, QDA-quadratic discriminant analysis, MSM-multisurface separation, SVM-support vector machine

Paper	Logistic regression	Decision tree	ANN	LDA	QDA	MSM	SVM	Naïve Bayes	Instance-based learning	Other	Ensembles
Pendharkar et al. (1999)			Y	Y						DEA	
Abbass (2002), Hung et al. (2002), West et al. (2005)			Y								
Jin et al. (2012)		Y						Y	Y		Y
Winkler et al. (2013)			Y				Y		Y		Y
Sree et al. (2010)			Y	Y	Y		Y	Y			
Ryu et al. (2007)				Y		Y	Y			Isotonic separation	Y

Continued on next page

Paper	Logistic regression	Decision tree	ANN	LDA	QDA	MSM	SVM	Naïve Bayes	Instance-based learning	Other	Ensembles
Tourassi et al. (2001)			Y								
Mangasarian et al. (1995)						Y					
Lundin et al. (1999)	Y		Y								
Çakir and Demirel (2011)		Y	Y						Y		
Delen et al. (2005)	Y	Y	Y								
Štrumbelj et al. (2010)										Any classifier	
Razavi et al. (2007)		Y									
Jonsdottir et al. (2008)	Y	Y					Y	Y			Y

Logistic regression

Logistic regression is a popular classification technique, commonly used in the medical literature. The prognostic models discussed next predict an outcome based on a collection of factors, and give a patient's absolute risk. On the other hand, aetiological models are those which test whether one particular factor affects a patient's breast cancer risk and measure relative risks (Herbert, 2014); we do not discuss these further. Binomial (or binary) logistic regression models have two possible outcomes, while multinomial logistic regression models have more than two outcomes (Hosmer et al., 2013). To demonstrate the wide-ranging use of logistic regression to predict breast cancer risks, we briefly describe some recent examples relating to different phases of cancer care, some of which are from the medical literature.

Recent examples of using logistic regression to predict a diagnosis of breast cancer include those by Hippisley-Cox and Coupland (2013) and Ahn et al. (2018). In a large-scale study, Hippisley-Cox and Coupland (2013) apply multinomial logistic regression to predict women's risks of being diagnosed with different cancers based on information recorded during GP consultations. Threshold risk values for the top 1%, 5% and 10% of people at risk of each type of cancer, including breast cancer, are given. Unlike Hippisley-Cox and Coupland (2013) who consider the general population, Ahn et al. (2018) develop a scorecard specific to patients diagnosed with intraductal papilloma, a benign breast condition, to predict the chance of malignancy. The scorecard is designed to be used following ultrasound imaging. All variables are binary and the scores are derived from rounded beta values from the logistic regression.

We next describe two recent applications of logistic regression to the breast cancer treatment phase. Mimouni et al. (2015) use binomial logistic regression to predict whether there will be cancer cells remaining after surgery and allocate patients to risk groups. Ryu et al. (2017) also use binomial logistic regression but for patients who have had chemotherapy and biopsies of sentinel lymph nodes (those to which the tumour is most likely to spread). They predict whether cancer has also spread to nonsentinel lymph nodes. Their risk tool is presented in user-friendly nomogram format.

Wang et al. (2013) and García-Laencina et al. (2015) both use binomial logistic regression to predict whether breast cancer patients will survive for five years. Wang et al. investigate how to deal with an imbalanced number of survivors and non-survivors in their dataset, while García-Laencina et al. concentrate on handling missing data.

Decision trees

Decision trees are used in many of the papers from our search, and are usually compared to alternative classifiers (Delen et al., 2005; Razavi et al., 2007; Jonsdottir et al., 2008; Çakir and Demirel, 2011; Jin et al., 2012). Jin et al. (2012) compare different types of decision tree including the one-level decision tree, the random tree (where the next variable to branch on is picked randomly from the best k options) and the functional tree (where more than one path may apply to each patient).

Decision trees provide perhaps the most intuitive output of the different techniques. Another advantage is that, after being trained on initial data, the output can often be given to clinicians in diagrammatic form and used to help make future decisions without needing software. However decision trees are often unstable, meaning a small change in their training data could result in a very different tree. They may also have a high computational cost.

Artificial neural networks (ANNs)

Artificial neural networks, ANNs, are also much-used (Lundin et al., 1999; Pendharkar et al., 1999; Tourassi et al., 2001; Abbass, 2002; Hung et al., 2002; Delen et al., 2005; West et al., 2005; Sree et al., 2010; Çakir and Demirel, 2011; Winkler et al., 2013). Their structures consist of connected input, hidden and output nodes. Some authors compare different types of ANNs, for example West et al. (2005) apply both the multilayer perceptron and the radial basis function network.

Even when ANNs predict accurately, clinicians may mistrust them because they do not automatically give justifications for their predictions (Tourassi et al., 2001). However there are ways of overcoming this. Tourassi et al. extract rules to make the reasoning more transparent and Delen et al. (2005) perform sensitivity analysis to ascertain how different variables contribute to the prediction. Another disadvantage of ANNs is that it can be difficult to find the (global) optimal solution when there are multiple local optima (Delen, 2009).

Methods to separate classes

Several methods involve plotting the input variables of patients, then inserting a separator so that the members of each risk class lie on either side. Depending on the number of dimensions, the separator can be a line, plane or hyperplane. Pendharkar et al. (1999), Ryu et al. (2007) and Sree et al. (2010) use linear discriminant analysis (LDA) to find a linear combination of features that separate two classes. Ryu et al. include a penalty for points on the wrong side of the separator when a perfect separation is not possible. Both Ryu et al. (2007) and Mangasarian et al. (1995) apply multisurface separation to iteratively generate pairs of parallel planes. The aim is for points of one class to be on one side of the first plane and points of the other class to be on the far side of the second plane.

As an alternative to LDA, Sree et al. also try separating the classes with a quadratic rather than linear function (QDA). Another option when points cannot be linearly separated is to use support vector machines (SVMs), which map the inputs onto higher dimensional spaces then construct pairs of hyperplanes to separate the classes. Rather than applying explicit mappings, SVMs use kernel functions to implicitly map the inputs onto the new space.

In addition to often predicting accurately, Delen (2009) recommends SVMs for their geometric interpretability and the fact that they have a unique global solution. However the other separation methods described above are arguably easier to understand.

Naïve Bayes

The naïve Bayes approach calculates the probability of being in each class conditional on the inputs, using Bayes' Theorem. Several papers test this method (Jonsdottir et al., 2008; Sree et al., 2010; Jin et al., 2012), which assumes that input variables are independent and of equal importance in the prediction. Even though these assumptions are unlikely to hold in reality, the approach benefits from being intuitive and is often used as a reference method of these classification papers.

Instance-based learning

Instance-based techniques are used by Çakir and Demirel (2011), Jin et al. (2012) and Winkler et al. (2013). The simplest such algorithm is IB1, which looks at the most similar patient and assumes a new patient will be in the same class (Çakir and Demirel, 2011; Jin et al., 2012). In contrast, the k-nearest neighbour algorithm used by Winkler et al. (2013) looks at the k most similar patients and classifies a new patient as the most commonly occurring class. For this algorithm, similarity is usually measured using the Euclidean distance between input values. Winkler et al. assign more importance to the most similar patients, and vary the number of neighbours, k, between one and ten.

Unlike the other methods described above, instance-based learning is a so-called lazy algorithm, meaning that it is not trained on data, but waits for a new patient before performing computations (Witten et al., 2011). Hence the disadvantage is that the algorithm needs to be run for each new patient. The advantage is that more patients can be included over time, causing the algorithm to become more experienced at predicting.

Ensemble classifiers

To improve predictive performance, the results of individual classifiers are combined to form ensemble classifiers (Ryu et al., 2007; Jonsdottir et al., 2008; Jin et al., 2012; Winkler et al., 2013). Winkler et al. (2013) combine classifiers in two ways: firstly, by taking the mode and secondly, by predicting a malignant outcome if any classifier predicts one. Bagging methods create bootstrap samples on which to train the classifier, then combine results by using majority vote (see for example Jonsdottir et al. (2008)). Boosting methods are used to iteratively develop new classifiers that perform better for those instances for which the previous classifier did not perform well. Examples of boosting algorithms used for breast cancer risk prediction are Adaboost (Ryu et al., 2007; Jonsdottir et al., 2008) and logit boost (Jin et al., 2012).

3.2.5 Implementation

Literature on breast cancer risk classification models ranges from very theoretical, to some consideration of practical issues, to implementation of a model in real life. The majority of the examples considered here are theoretical papers, aiming to test the performance of alternative classification algorithms (Lundin et al., 1999; Pendharkar et al., 1999; Abbass, 2002; Hung et al., 2002; West et al., 2005; Ryu et al., 2007; Jonsdottir et al., 2008; Jin et al., 2012).

Some authors of these example papers consider practical aspects. For example, Çakir and Demirel (2011) describe their user-friendly interface design and Razavi et al. (2007) validate their model predictions against expert opinions. Tourassi et al. (2001) explain why ANNs are often not trusted by clinicians, and demonstrate a method that overcomes some of these blocks to usage. Sree et al. (2010) provide evidence to support the use of electro-potential sensors along with their classification tool as a diagnostic test. Winkler et al. (2013) express the intention to implement their models in practice, and Delen et al. (2005) plan to make their results available on the Internet.

As far as it can be ascertained, the only papers considered that contain evidence of implementation are those by Mangasarian et al. (1995) and Štrumbelj et al. (2010). The diagnostic-prognostic system developed by Mangasarian et al. is in use at the University of Wisconsin. It is reported that oncologists are using the model developed by Štrumbelj et al. to predict recurrence. This model also gives a measure of the reliability of the estimate and explanation of which variables contributed and by how much. Of course, there may be many classifiers in use that are not reported on in the academic literature, and the models from some of the discussed papers might have been implemented since their publication.

3.2.6 Conclusion to breast cancer risk classification papers review

Breast cancer risk classification models have a range of potential uses throughout cancer care services. A variety of datasets are used in these models, including data shared publicly, specifically for testing model performance. Rather than using such secondary data for our project, we created a unique dataset consisting of variables coded from GP referral information (see Section 4.2). Hippisley-Cox and Coupland (2013) have looked at the link between symptoms recorded during GP consultations and the risk of breast cancer, using logistic regression (see Section 3.2.4). However they have not investigated the link between referral information for patients referred to breast clinics and the risk of breast abnormality. Hence compared to our study there are differences in both the patient groups and the variables considered.

In the papers reviewed in this section, performance measures are chosen based on context; computational time or sensitivity may be more important than overall prediction accuracy. There is a wide range of data mining classification algorithms, while traditional logistic regression remains a much-used method, suggesting that it is trusted in the medical community. There is evidence that some cancer risk classification models are in use at different stages of medical decision making, which shows promise for the potential of our models being useful in practice.

3.3 Generating patient characteristic inputs for DES pathway models

In our application, patient characteristics affect patients' predicted risk levels, the pathways that they take through the diagnostic process and how long they spend in some stages. Thus we need to carefully consider how best to generate patient characteristic inputs to our DES. There are patient characteristics that are categorical, for example the presence of symptoms will be binary variables, and there are likely to be dependencies between characteristics. This means that we are particularly interested in methods that can generate dependent, categorical characteristics. Another issue is that we need an approach that will work for a small data sample, where there may be missing (but possible) combinations of characteristics. Therefore in this section we review approaches to generating patient characteristic inputs for DES pathway models. Our aim is to find what methods are in use, and whether any are suited to our problem.

There are different types of pathways occurring in healthcare. Among clinicians, care pathways, also known by a variety of other names, are used to coordinate care for particular groups of patients and are based on clinical guidelines (European Pathway association, 2016). It is important that the pathways proposed from our modelling do satisfy clinical guidelines, as well as meeting our aim to increase the proportion of value-added time for patients (time that contributes to their care). Adeyemi et al. (2013) distinguish between three types of pathway that are modelled in healthcare. The first, an operational pathway, is a sequence of services that a patient receives. Secondly, a clinical pathway represents the progression of disease. The third, a virtual pathway, is the definition of states a patient passes through, which may or may not have a physical interpretation.

The scope of our review is as follows. Operational pathways are the focus of interest here, but combined clinical-operational pathways are also considered. We do not look at virtual pathways, since in our case it is clear what the stages are that patients moves through. A range of techniques are used to model patient pathways: simulation approaches, for example system dynamics, agent-based simulation and discrete-event simulation (DES), and analytic approaches, for example queuing theory, Markov models and compartmental models (Bhattacharjee and Ray, 2014). Here we concentrate on DES models of patient pathways; in particular, how patient characteristics are initially generated in the simulation.

As for the other literature areas we reviewed, we used the DelphiS search engine to find a sample of papers. I searched for the string “pathway AND (hospital OR health*) AND (discrete-event simulation)” in full texts. Relevant papers according to our scope were identified from abstracts. The papers were skimmed to find information relating to how authors model patient characteristic simulation inputs. Papers that do not explain how patient characteristics are generated were excluded. I also searched more widely in the theoretical literature for methods that could be appropriate for our application.

This section proceeds as follows. Firstly, in Section 3.3.1, the selection of papers identified by our search are introduced. Secondly, in Section 3.3.2 the methods used to generate patient characteristics are described. Section 3.3.3 provides information about an additional existing method from the theoretical literature that is not applied in our sample of papers. In Section 3.3.4 we sum up the suitability of existing methods for use in our study.

3.3.1 Results of search

We identified several DES models of operational pathways through emergency departments, considering discharge targets and patient flows to inpatient wards (Eatock et al., 2011; Crawford et al., 2014; Khanna et al., 2016). Further operational pathway models consider stroke services (Bayer et al., 2010; Chemweno et al., 2014; Gillespie et al., 2016; Monks et al., 2016), which vary in scope from considering the diagnostic process to the whole care pathway. There are DES models of clinical-operational pathways for a range of diseases: coronary disease (Cooper et al., 2002), atrial fibrillation (Lord et al., 2013), arthritis (Tran-Duy et al., 2014), major depressive disorder (Vataire et al., 2014), glaucoma (Burr et al., 2012; Crane et al., 2013), skin infections (Revankar et al., 2014) and cancers (Pilgrim et al., 2008; Lord et al., 2013; Wang et al., 2017). Since these combined models track both the disease state and health services used by an individual, they are a useful tool in cost-effectiveness and QALY analyses.

3.3.2 Methods

Patient characteristic inputs to the considered pathway simulations are modelled using the following broad approaches (sometimes in combination): (1) grouping patients with similar characteristics, (2) using empirical data directly, (3a) using statistical distributions that assume different characteristics are independent, and (3b) using statistical distributions that capture correlations between characteristics. We describe methods for each of these approaches in turn.

Firstly we discuss the papers that group patients with similar characteristics (Cooper et al., 2002; Bayer et al., 2010; Crawford et al., 2014; Chemweno et al., 2014; Gillespie et al., 2016; Monks et al., 2016). When grouping patients, authors either specify the probability of belonging to each group (Bayer et al., 2010; Crawford et al., 2014; Chemweno et al., 2014), or use group-specific arrival rates (Cooper et al., 2002; Gillespie et al., 2016; Monks et al., 2016). The relative numbers of patients in each group are sometimes based on expert opinion (Chemweno et al., 2014) or assumed to be the same as in data samples (Cooper et al., 2002; Crawford et al., 2014). When the choice of groups is not obvious, patient data can be analysed to find appropriate groups; Gillespie et al. (2016) group patients with similar lengths of stay using Kaplan-Meier and log-rank tests. In that paper, the patients in each group are characterised by their gender, age, diagnosis and outcome. Elsewhere, different clustering (Isken and Rajagopalan, 2002; Ceglowski et al., 2006) and classification (Harper, 2002) techniques have been used to group similar patients. Thus grouping patients in these ways relies on knowing the prevalence of each group, either from expert opinion or from having sufficient data.

Secondly we consider papers that use empirical data directly; this is achieved in three different ways. One way is putting the information relating to each patient from a data sample directly into the DES (Eatock et al., 2011; Khanna et al., 2016). The second way is bootstrapping, as in Lord et al.'s (2013) atrial fibrillation case study. In this example, missing characteristics were imputed using multivariate regression. The third way, used by Revankar et al. (2014), is to generate three copies of each patient's set of characteristics, to be able to compare different treatment strategies on the same cohort. The drawback of directly using empirical data is that it does not include patients who were not seen in reality, so is most suitable when the data sample is deemed large enough to closely resemble the underlying population.

Thirdly we describe papers that sample characteristics from statistical distributions. In simulations that assume no relationship between characteristics, artificial patients are generated so that each characteristic follows an independent probability distribution (Burr et al., 2012; Crane et al., 2013; Tran-Duy et al., 2014). In contrast, other pathway models do account for relationships between some of a patient's characteristics (Cooper et al., 2002; Pilgrim et al., 2008; Lord et al., 2013; Vataire et al., 2014; Wang et al., 2017); often pairwise relationships between age and other characteristics are considered. These papers either use the empirical conditional distributions present in their data or make assumptions about the relationships when data is not available. Using the empirical conditional distribution relies on the relationships present in the sample being representative of the wider population; combinations of characteristics not present in the sample will not be simulated.

3.3.3 Another method: NORTA (NORmal-To-Anything) distribution

A different method for generating correlated patient characteristics, which was identified from the theoretical literature, is using a NORTA (NORmal-To-Anything) distribution as developed by Cario and Nelson (1997). In this method, the marginal distribution for each characteristic is specified, but the need to explicitly define copulas, the multivariate probability distributions giving the dependence between variables, is avoided (Law and Kelton, 2000). The NORTA method is used to define correlated parameters in several colorectal screening papers (Roberts et al., 2007; Tafazzoli et al., 2009; Li et al., 2015). However it is not applicable to categorical characteristics, which are common in risk tools. Thus, when evaluating the potential impact of risk-based patient management strategies in simulation, a different method is needed to generate a combination of dependent, categorical characteristics for each patient.

3.3.4 Conclusion to generating patient characteristic inputs for DES pathway models

The example DES patient pathway models we considered in this section use three broad approaches to generating patient characteristics: grouping patients with similar characteristics, using empirical data directly, and sampling characteristics from independent or conditional statistical distributions. The NORTA method which was identified from a search of the theoretical literature falls into the last category.

We need a method that generates dependent, categorical patient characteristics from a small data sample where some (possible) combinations are missing. None of the methods we found fulfils these requirements either because they are designed for independent or continuous characteristics, or they require enough data or knowledge to be able to specify underlying distributions correctly. We return to this issue in Section 6.3.3, where we instead propose fitting Poisson loglinear models to our data.

3.4 Combining classification and DES

In this study we are combining classification models and DES so that we can optimise the threshold between classes based on an operational performance measure. Therefore we review existing healthcare literature that combines these techniques. Our focus is on healthcare applications in the OR and management science (MS) literature. We are interested in articles that both fit a classification model to data and combine this with discrete-event simulation.

Our search strategy was as follows. As in the other literature areas, we used DelphiS to look for papers. I searched for simulation and classification technique terms in abstracts, mentions of OR or MS in the full text and references to healthcare in the full text. The full search string is AB (Discrete-event-simulation OR DES OR Simulat*) AND AB (“logistic regression” OR “scorecard” OR “credit scoring” OR “credit score” OR “classification” OR “clustering” OR “artificial intelligence” OR “decision tree” OR “prediction tool” OR “prediction method” OR “prediction model” OR “predictive tool” OR “predictive method” OR “predictive model”) AND TX (“operational research” OR “operations research” OR ‘management science”) AND TX (health* OR hospital OR patient). Papers were checked for relevance based on their abstracts.

This section proceeds as follows. Firstly, in Section 3.4.1, we introduce the papers identified by our search. Secondly in Section 3.4.2 we discuss how and why they have combined classification and DES. We conclude the literature review with Section 3.4.3.

3.4.1 Results of search

We identified only five relevant papers, where both the fitting of classification models and simulation runs are described (Harper, 2002; Harper et al., 2003; Cannon et al., 2013; Bhattacharjee and Ray, 2016; Huang and Hanauer, 2016). These span applications in hospital resource planning (Harper, 2002), imaging and outpatient appointment planning (Bhattacharjee and Ray, 2016; Huang and Hanauer, 2016), prevention and treatment of a long-term complication of diabetes (Harper et al., 2003) and assessing the outcomes and costs of antenatal care provision (Cannon et al., 2013). There are also healthcare OR/MS papers that explain the process of fitting classification models but only suggest that, rather than demonstrate how, the results could be used in simulations, for example Isken and Rajagopalan (2002) and Ceglowski et al. (2006). We do not discuss these types of papers further.

3.4.2 How and why classification and simulation are combined

One way of combining classification and DES is to generate groups of similar patients as simulation inputs (Harper, 2002; Bhattacharjee and Ray, 2016). This allows comparison of resource management strategies including those that deal with patients differently depending on their group. Specifically, Harper (2002) provides a framework for modelling hospital resources, by using classification and regression trees (CART) to group similar patients and then simulating their use of resources. This is exemplified for bed planning and operating room scheduling, where patients are classified according to their length-of-stay and operation durations respectively. In the latter case, the best scheduling policy of those considered (in terms of theatre and bed utilisation) is to schedule the patients with long durations first. Similarly Bhattacharjee and Ray (2016) classify patients using CART and then use DES to evaluate the potential impact of sequencing appointments based on the patient classes. In their context of MRI appointments they find that scheduling the patient classes with shortest service times first is best. Their performance measure consists of waiting times for each patient class and a utilisation measure.

A second way of combining classification and DES is predicting the occurrence of health-related events during a simulation (Harper et al., 2003; Cannon et al., 2013). Harper et al. (2003) use CART to predict patients' risks of developing diabetic retinopathy, a complication of diabetes. This enables them to compare the cost-effectiveness of various detection and treatment options. A different classification technique is used by Cannon et al. (2013), who develop a logistic regression model to predict the occurrence of various pregnancy events which they simulate together with health service visits. Their case study considers the population of Aboriginal women in remote rural areas of western Australia. The likely outcomes and costs of antenatal care for women receiving adequate care versus inadequate care are compared; it is found that receiving inadequate care is more expensive in the long run.

A third way of combining classification and DES is using a simulation to compare the impact of making decisions based on different classification models. This is the approach taken by Huang and Hanauer (2016), who present a series of logistic regression models to predict no-shows at a paediatric clinic. Each model contains information about one more prior attendance than the previous one. The threshold at which to predict a no-show is calculated by minimising misclassifications. DES is used to evaluate the cost (waiting time plus overtime plus idle time) per patient for appointment overbooking policies based on each of these logistic regression models. It is found that information about sixteen previous appointments should be included.

3.4.3 Conclusion to combining classification and DES

We found a small number of healthcare-related papers that combine classification techniques with DES. The applications are spread across a range of healthcare areas. We distinguished between three ways in which these techniques have been combined. The first way is generating groups of similar patients as simulation inputs. The second way is using classification to predict the occurrence of health-related events during a simulation. The third way is using a simulation to compare the impact of making decisions based on different classification models. We are not aware of any examples of using DES to determine the best threshold between classes in terms of operational measures.

3.5 Hospital outpatient scheduling studies

Breast diagnostic clinics are an example of outpatient clinics, which have been widely studied in the OR literature. A characteristic feature of our case study clinic is that there are multiple stages and it is not known in advance which stages a patient will need; all patients see a breast clinician, but the diagnostic tests performed afterwards vary between patients. There are also administrative stages to be completed. Each stage has an associated waiting period in a queue or waiting room.

In this section we look at the different performance measures used in hospital outpatient scheduling studies. There is a large literature addressing outpatient scheduling, which has been reviewed by Cayirli and Veral (2003), and by Gupta and Denton (2008) as part of a more general review of healthcare appointment scheduling. A recent paper by Klassen and Yoogalingam (2018) includes a review of literature on outpatient clinics with multiple stages. We do not review papers concerned with indirect waiting time, which is the delay between an appointment being requested and the scheduled appointment day. There is a national target that patients referred for breast diagnostics should be seen within two weeks (Keogh, 2009), and unlike some other outpatient services there is not a long backlog for appointments.

Our search strategy involved considering key papers (Cayirli and Veral, 2003; Gupta and Denton, 2008; Klassen and Yoogalingam, 2018) and some examples from their reference lists. When reviewing papers our focus is on the choice of performance measures. We review papers considering single-stage and multi-stage appointments separately. In the first case there is only one potential wait, while in the second case patients may wait between each stage of their appointment.

This section proceeds as follows. The literature about on-the-day waits for single-stage outpatient appointments is reviewed first in Section 3.5.1, followed by papers modelling waits during multi-stage appointments in Section 3.5.2.

3.5.1 Single-stage appointments

In an ideal world, outpatients would begin their appointment at the scheduled time, and so direct (physical, on-the-day) waiting times would be zero. However in practice, delays are common and can be caused by previously overrunning appointments, lateness of patients or staff, and prioritising unscheduled, urgent patients over scheduled appointments (Gupta and Denton, 2008). Patients who do not attend their appointments may cause gaps in clinicians' work, leading to the use of overbooking policies in some clinics, as studied by Laganga and Lawrence (2007) and Berg et al. (2013).

Appointment rules govern how appointments are allocated and consist of the following decision variables: number of patients to schedule per slot, length of slot and number of patients to schedule in the first slot of the session (Cayirli and Veral, 2003). These have been investigated using both analytic and simulation approaches. Some authors, for example Vanden Bosch and Dietz (2000), investigate how best to sequence patients, which is relevant if patients can be classified into groups with different appointment characteristics, such as service times or waiting cost (Cayirli and Veral, 2003). In our research project, there are two appointment types: initial and results consultations. We look at how best to prioritise these in Section 6.7.

A common approach to tackling direct waiting times is to design appointment scheduling systems to minimise a weighted cost function consisting of patient direct waiting time and clinician idle time or overtime (Cayirli and Veral, 2003). Authors tend to assume that the clinician's time is more important than the patients' and use a linear cost function (Cayirli and Veral, 2003). However, Klassen and Rohleder (1996) question whether a non-linear function would be more realistic, since many short waits are perhaps not equivalent to one long wait. It is conceivable that there is a time threshold that represents an acceptable level of waiting. Under fixed interval appointments, patients seen later in the day are at a disadvantage since congestion tends to build up (Cayirli and Veral, 2003). This has led some authors to consider fairness measures such as the average waiting times according to patients' positions in the appointment list (Bailey, 1952). Other performance measures that have been considered are for example patient total time, also called flow time (Laganga and Lawrence, 2007), and average queue size (Lehaney et al., 1999).

3.5.2 Multi-stage appointments

Outpatient appointments may consist of multiple stages involving resources from different departments within a single visit. In this more complicated situation there may be multiple in-clinic waits. Relatively few scheduling papers have addressed such situations (Klassen and Yoogalingam, 2018). Some examples of multi-stage clinics that have been studied are ear, nose and throat clinics (Harper and Gamlin, 2003) and surgical clinics (Saremi et al., 2013). Swisher et al. (2000) present a versatile model of a general multi-stage outpatient clinic which can be customised to experiment with different set-ups.

It is usual for multi-stage appointments that the first stage is scheduled and patients proceed through the other stages by a priority rule such as First Come First Served (Klassen and Yoogalingam, 2018). As for single stage appointments, researchers have attempted to find good sets of appointment rules for different situations (Klassen and Yoogalingam, 2018). To this end, techniques such as simulation (Swisher et al., 2000; Harper and Gamlin, 2003) and simulation-based optimisation (Saremi et al., 2015) have been applied.

The performance measures assessed in multi-stage appointment planning are the same as, or extended versions of, those used for single-stage appointments. For example, Harper and Gamlin (2003) consider three different performance measures: the average waiting time for the first stage, the percentage of patients waiting more than 30 minutes for the first stage, and the average total time. Swisher et al. (2000) combine several performance measures into a "clinic effectiveness" measure, which is not explicitly defined, but is said to incorporate patient waiting and profit. Both Saremi et al. (2015) and Klassen and Yoogalingam (2018) consider the waiting time at different stages. Specifically, Saremi et al. (2015) find the best surgery schedule by minimising the sum of waiting times and completion time, where the waiting times can be weighted differently depending on the patient type and stage. Klassen and Yoogalingam (2018) minimise the weighted sum of waiting times at each stage, clinician idle time and overtime.

3.5.3 Conclusion to hospital outpatient scheduling review

Single-stage outpatient appointments have benefited from research into the best appointment scheduling rules for different situations. Research on multi-stage outpatient appointments is scarcer. The performance indicators considered include waiting time measures and clinician idle time and overtime measures.

In our case study, the waiting time for ultrasound is a particular concern (see Section 2.3.3), so only considering the waiting time at the first stage (initial consultation) like Harper and Gamlin (2003) is not sufficient. Another possibility is including a sum of (potentially weighted) waiting times and other measures, like Saremi et al. (2015) and Klassen and Yoogalingam (2018). However this would favour scenarios where patients visit fewer stages, i.e. sending all patients to imaging first, without accounting for the benefit of an extra service, i.e. the initial consultation. Since patients visiting a breast diagnostic clinic are likely to be worried and anxious while waiting for a cancer diagnosis to be confirmed or excluded, it is important that as much of their time at the clinic as possible is contributing to their diagnostic care.

3.6 Conclusion to literature review

In this chapter, we reviewed literature on OR applications to breast cancer care, breast cancer risk classification models, generating patient characteristic inputs for DES pathway models, combining classification and DES, and outpatient scheduling studies. The research gaps that we address in this study are summarised below.

Although there is a large body of literature applying OR techniques to breast cancer care problems, the majority of papers address screening and treatment services. On the other hand, our study uses OR techniques to improve diagnostic services for patients with symptoms; we did not find any other examples like this in our review.

Since there are no existing models in the academic literature that we are aware of for predicting the risk of breast abnormalities, we instead focussed on reviewing the related papers on predicting breast cancer risks. Among these we found many examples of logistic regression being applied, which leads us to use this technique in Chapter 5.

We highlighted a research gap in terms of generating categorical, inter-dependent patient characteristics for simulation models where the data sample may be missing possible combinations of characteristics. We address this research gap in Section 6.3.3 by providing a novel way of using Poisson regression to generate combinations of characteristics for simulation.

Classification and DES techniques have rarely been combined in the literature, and our way of combining them is new: selecting the best threshold between low and high risk patients based on operational measures, as described in Chapter 6.

Chapter 4

Data from GP referrals and the breast diagnostic clinic: collection and analysis

This chapter describes the data that I collected for the case study, as well as preliminary analysis of these data. Firstly in Section 4.1, data used to understand the daily operating of the breast diagnostic clinic are described and analysed. These data are required for populating our simulation models, described in Chapter 6. Secondly in Section 4.2, production of a dataset linking GP referral information to in-clinic tests and results is explained, and analysis of this dataset is presented. In particular, we address the first research objective (see Section 1.2) which involves assessing the completeness of GP referral information. The patient-level dataset produced is used to develop classification models in Chapter 5 and for setting patient labels in simulation in Chapter 6.

4.1 Data concerning daily operating of the breast diagnostic clinic

In order to understand how the breast diagnostic clinic currently operates, various data sources were used, including both quantitative and qualitative primary data collected specifically for the project, as well as secondary data. Each source and the results of data analysis are described here. The data sources are summarised in Table 4.1, and details about sample sizes, pre-processing and date ranges are provided in Table D.1. Based on findings from the primary data collection, further process mapping was carried out and is described at the end of Section 4.1.1.

Table 4.1: Sources of data concerning daily operating of the breast diagnostic clinic (sample sizes for individual variables are provided in Table D.1)

	Data source	Data
Primary data	Patient questionnaires	Timestamps (for service times and waiting times)
	Observed times	Timestamps (for service times and turnaround times)
Secondary data	Medway appointment system	Number scheduled appointments, no-show rate, patient punctuality
	Picture Archiving and Communications System (PACS)	Timestamp of when mammogram or ultrasound image taken (to calculate time taken for report to be ready)
	Radiology Information System (RIS)	New patient tests (mammogram, ultrasound, biopsy) and results (normal, abnormal, cancer) matched to GP referral information, number follow-up mammograms, number follow-up ultrasounds, timestamp of when report ready (to calculate time taken for report to be ready)

4.1.1 Primary data collected

Patient questionnaires

I designed a patient questionnaire to audit current waiting and process times at the clinic, as well as asking for patient feedback (see Appendices C.8 and C.9). From 23rd November to 16th December 2015, receptionists handed out copies of the questionnaire to new patients attending the clinic, and collected the completed questionnaires. Patients were asked to record the time at different stages during their visit and also to provide feedback on their experience of the clinic. During the period when questionnaires were being handed out, 249 new patients attended the breast clinic. 111 of these patients, that is 45% of the total, returned the questionnaire. The uptake varied per day, as demonstrated in Chart C.10. I collated the data from the questionnaires in a spreadsheet.

The timestamps that patients recorded were used to calculate the service times of each stage of their appointment (initial consultation, mammogram, ultrasound and results consultation) as well as waiting times between stages. Some of the times were excluded as being invalid, for example when an end timestamp was before its corresponding start timestamp. There were also missing timestamps, either because a patient did not undergo a certain test (NA) or because some of the questionnaire boxes were left blank

(NR). Comparison of the service time samples to those collected in the observation study is described in the next section, along with further analysis. The service times from patient questionnaires informed some distributions in the simulation model as described in Section 6.4.1. The waiting times were used to validate the core simulation model as explained in Section 6.5.

Some strengths and limitations of the patient-collected times are as follows. The benefit of patients recording timestamps is that they are present for the whole diagnostic process. It is cheaper than hiring a team of independent observers, of whom many would be needed to collect the same volume of data. However there may be some bias since patients with shorter waiting times had less time to read and fill in the questionnaire. Likewise, patients with particularly strong or negative opinions about their visit may have been more likely to provide feedback. Patients may not reliably record the time for a variety of reasons, including being understandably more focussed on the reason for their visit, obtaining a diagnosis.

The patient feedback question was asked in an open-ended way so that patients could mention any aspect of their visit to staff, but for this research comments relating to how the service is organised are most relevant. The feedback section was completed in 62 of the 111 questionnaires. Of these 62 patients, 9 patients made negative comments only, 32 positive comments only and 21 both positive and negative comments. The negative comments are summarised in Table 4.2 and the positive in Table 4.3.

Table 4.2: Negative feedback from patient questionnaires

Comments	Number of patients
Long delays/waits/day	11
Long wait for results	5
Long wait for ultrasound	3
Need to return another day for scan	3
Painful mammogram	2
Appointment rebooking process	1
Initial consultation repeat of GP appointment	1
Stressful day	1
Particular staff member	1
Long wait for mammogram	1
Unclear which different stages required	1

The most common negative theme recorded was a long day or long waiting times. In contrast, 8 patients commented that their visit was quick, prompt or efficient. This highlights the variability in the perception or experience of waiting times. More patients commented on the waiting time for results than on other specific waiting times. In particular, one patient fed back that “It seemed quite a long wait for results considering ultrasound Dr had already told me results”. Another issue is that although in theory, patients should complete all diagnostic tests within one day, in practice, this

Table 4.3: Positive feedback from patient questionnaires

Comments	Number of patients
Overall service/experience/clinic	18
Staff in general	9
Quick/efficient/prompt process	8
Particular staff member	5
Completed in one day	3
Buzzers so can get cup of tea	1

is not necessarily the case. Some patients commented how pleased they were to finish the process during a single visit, while others complained that they would have to return another day for an ultrasound scan. One patient mentioned that the initial consultation seemed to be a repeat of their GP appointment. Some patients provided recommendations and ideas on how to improve the clinic, as displayed in Table 4.4. Several of these ideas relate to the importance of managing expectations of waiting times. One patient recommended discharging some patients directly from imaging. This idea is explored further as a simulation scenario in Section 6.7.

Table 4.4: Recommendations from patient questionnaires

Recommendations	Which patient?
Discharge patients after imaging if have no significant abnormality. Send results letter to GP and patient	33
Inform patients of expected length of day, so they can prepare	25, 105
Inform patients of expected wait at various stages	86
Update patients on current waiting time	71, 95
Send New Patient Assessment Form in advance so can complete in privacy	71

Observed times

Given the limitations of the service time samples obtained from questionnaires, I carried out an observation study to obtain a second set of these data. I spent time in different locations to collect the following service times: initial consultation, mammogram, ultrasound and results consultation. Table D.1 shows a summary of the sizes of the samples provided by patients and the samples I collected.

The dotplots in Figures 4.1 to 4.4 compare service times calculated from the patient questionnaires with those I observed. From the dotplots it can be seen that the questionnaire data have peaks at multiples of five minutes. For example, many patients recorded times of 5, 10, 15 and 20 minutes for the initial consultation, while in my sample, these peaks do not exist. For both the initial consultation and mammogram service times, the questionnaire samples and observed times have similar summary statistics, as shown in Table 4.5. If one considers the underlying distributions and ignores the artificial peaks discussed earlier, the dotplots of the questionnaire samples and observed times suggest similar distributions (see Figures 4.1 and 4.2). However for ultrasounds and results consultations, the questionnaire and observed times data are not entirely consistent. A non-negligible number of patients recorded longer service times for ultrasounds (7 patients) and results consultations (5 patients) than I observed. This can be seen in the dotplots in Figures 4.3 and 4.4, as well as by the higher mean, median and upper quartile for the questionnaire data in Table 4.6.

Table 4.5: Service times: Summary statistics 1

	Initial consultation (minutes)		Mammogram (minutes)	
Data sample	Questionnaire	Validation	Questionnaire	Validation
Minimum	3	5	4	5
Lower quartile	10	10	7	7
Median	11	12	10	8
Mean	13	13	9	9
Upper quartile	15	15	10	10
Maximum	45	27	20	14

Table 4.6: Service times: Summary statistics 2

	Ultrasound (minutes)			Results consultation (minutes)	
Data sample	Questionnaire	Original validation	Extended validation	Questionnaire	Validation
Minimum	4	4	3	1	1
Lower quartile	6	7	7	2	2
Median	10	9	9	5	3
Mean	12	11	12	6	3
Upper quartile	15	12	15	6	4
Maximum	60	25	36	25	12

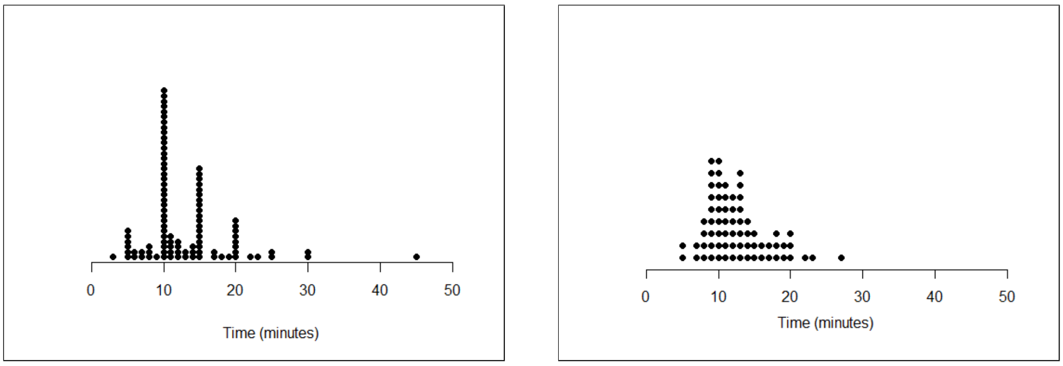


FIGURE 4.1: Initial consultation service time frequencies: Comparison of questionnaire (left) and observed times data (right)

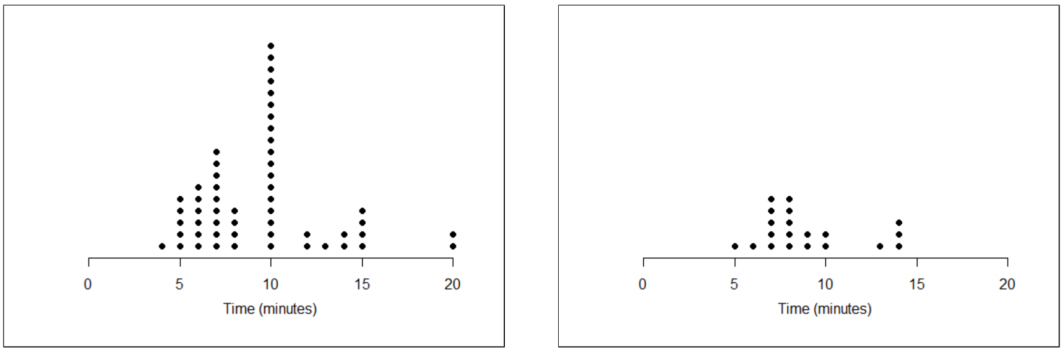


FIGURE 4.2: Mammogram service time frequencies: Comparison of questionnaire (left) and observed times data (right)

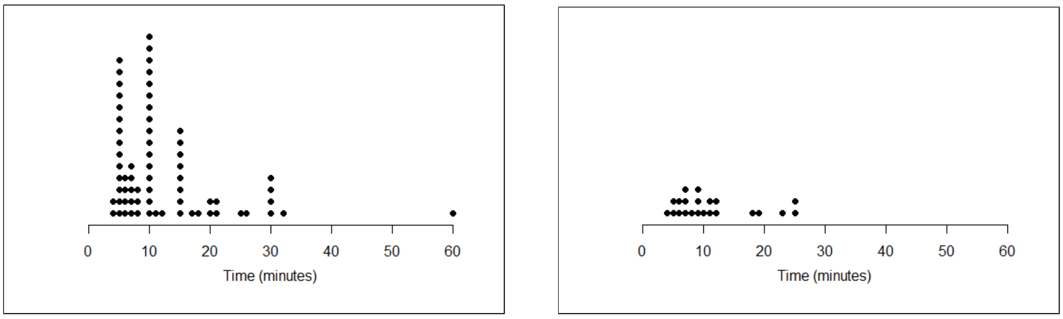


FIGURE 4.3: Ultrasound service time frequencies: Comparison of questionnaire (left) and observed times data (right)

Some reasons for the differences between the questionnaire and observed times distributions are as follows. Unlike the patients who filled in the questionnaire in addition to attending consultations and tests, I was dedicated to the single task of timing. So patients may have filled in some times in hindsight or been less accurate in recording. Given the short data collection period for each service time, it is possible that not all staff were observed, potentially introducing some bias. On the other hand, the questionnaires do capture the variation between staff, since they were filled in over a month. One reason I may have missed longer times is due to censoring (Banks et al., 2010a). That is, some ultrasounds and results consultations may have begun before or ended after the observation period, and this is more likely to happen with particularly

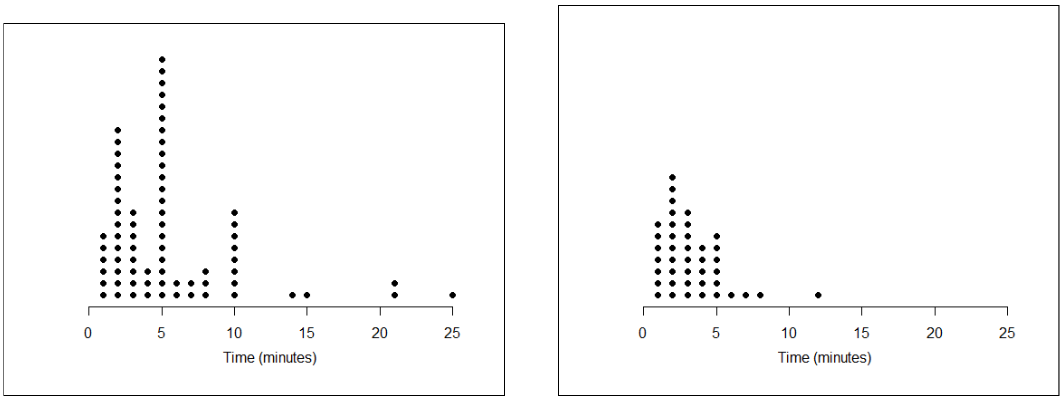


FIGURE 4.4: Results consultation service time frequencies: Comparison of questionnaire (left) and observed times data (right)

long times. Another possibility is that patients perceived the time as longer than it was. I discussed the anomalous results with clinic staff and dealt with the differences as follows.

The radiographer explained that longer times in the ultrasound room correspond to patients having biopsies following their ultrasound. Further times were collected in an attempt to capture more of these long times, and the full sample of observed times is shown in the dotplot in Figure 4.5.

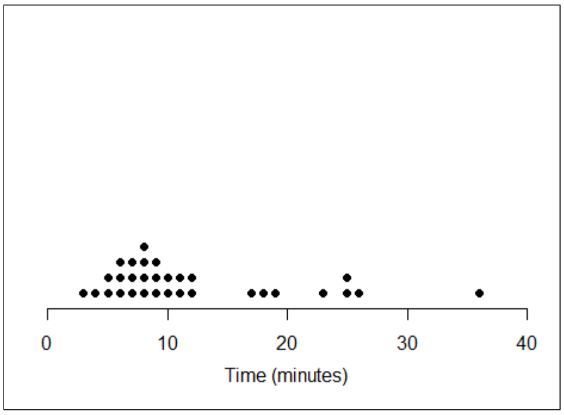


FIGURE 4.5: Ultrasound service time frequencies: Full observed times data

Nurses accounted for the long results consultation times by revealing that patients who have cancer spend about 20 minutes in the results consultation. Since this is a very small percentage of patients (about 4% of patients as explained in Section 4.2.2), it was decided that it would be infeasible to collect enough times to validate this further. Instead, assumptions were made about the shape of the results consultation distribution, which will be discussed in Section 6.4.

Mapping the results transmission process

At the beginning of the project, I spent time observing processes in the clinic and developing a process map, as described in Section 2.3.2. Since in the questionnaires patients commented particularly on the long wait for results consultations, I decided to carry out further investigation of the process for transmitting imaging results to patients, from the time that imaging tests are done. During the data collection period, I observed patients asking at the reception desk whether there was an option to receive their results by telephone rather than from the breast clinician in person. This provides further support to the idea that some patients may prefer to forego a results consultation, for example if they need to leave the clinic for practical reasons.

When investigating why the results transmission process is sometimes long, it was discovered that there are four different staff types in different locations dealing with imaging results, as shown in the flowchart in Figure 4.6. First, a radiologist dictates a report about the images (either while the patient is in the room or in between patients). So there is a potential delay until the radiologist has time to do this. Next, a receptionist in the clinic prints the report. However, since the receptionist is not alerted when the report is ready, there is another potential delay here. A nurse then collects the report from the printer in the back office. Again, nurses do not necessarily know when a receptionist has printed a report, so this may cause another delay. Finally, the breast clinician sees the patient to discuss the results, in between seeing other patients. So in total there are four potential delays in the results transmission process, explaining why some patients may experience long waits for their results consultations.

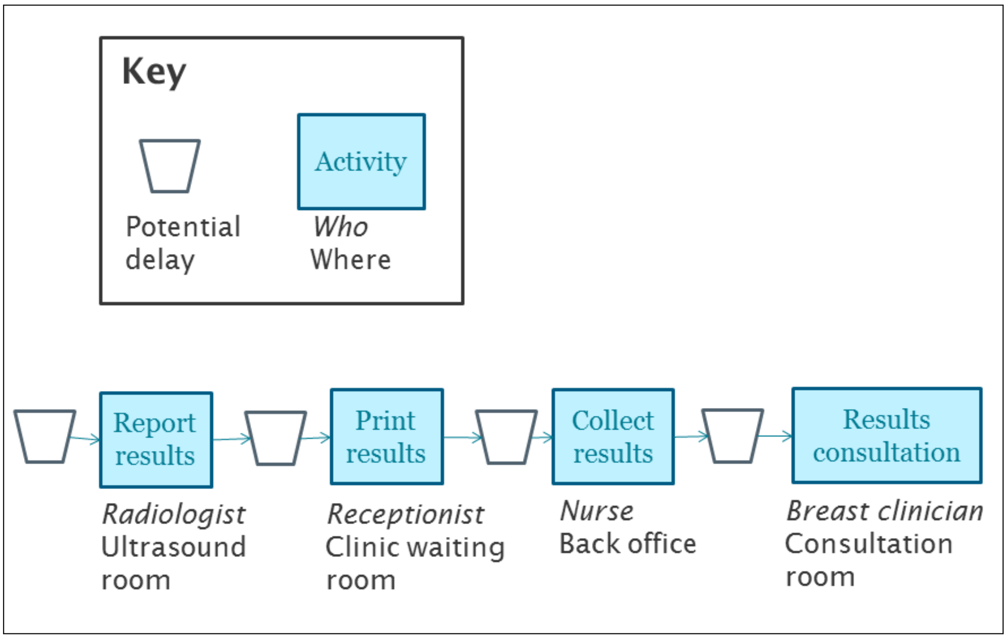


FIGURE 4.6: Results transmission process

4.1.2 Secondary data used

Medway appointment system

The Medway appointment system is used in the Whittington hospital to track patient appointments. I used this system to find the number of appointments scheduled per day and the no-show rate. As well as containing patients' scheduled appointment times, the times patients registered at the clinic reception desk are also recorded. I calculated the registration punctuality for a sample of patients, that is, how long before or after their scheduled appointment time each patient registered. Only the initial consultations are scheduled, while the imaging tests and results consultations are not.

The number of scheduled appointments varies by day of the week. Typical Monday, Tuesday and Wednesday schedules are provided in Appendix D.2. On Mondays, clinic starts later after the multi-disciplinary team meeting.

The average daily no-show rate is 10% and the punctuality of patients attending the clinic is shown in the histogram in Figure 4.7. Since patients may have been queuing, it is possible that they arrived on time but registered later than their appointment time. For this reason, we only consider patients registering more than 5 minutes after their appointment time as late. Using this definition of lateness, 22% of patients were late.

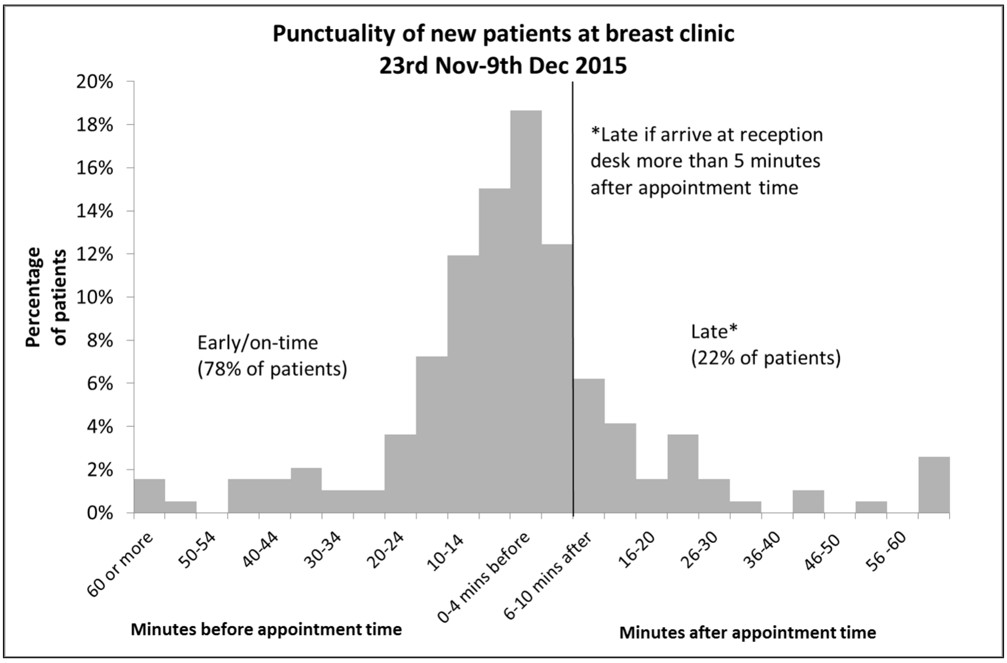


FIGURE 4.7: Punctuality of patients

PACS and RIS

The Picture Archiving and Communications System (PACS) contains the images from mammograms and ultrasounds, as well as timestamps of when the images were taken. These, together with the report timestamps from the Radiology Information System (RIS), were used to work out the length of time taken for a report to be ready. RIS was also used to obtain the numbers of mammograms and ultrasounds performed on follow-up patients, since this affects the availability of imaging resources for new patients, the focus of our study.

The distributions of the time taken for patients' reports to be ready are shown in Figures 4.8 and 4.9. The mean time taken for reports to be ready for patients having only a mammogram is fifteen minutes, while the time is shorter on average for those having an ultrasound (with or without a mammogram), at nine minutes. This is likely because the radiologist can dictate reports for patients having ultrasounds while they are still in the room. Both distributions are skewed to the left, suggesting that reporting is usually done soon after imaging. Regarding follow-up patients, there is a wide range in the numbers of tests per day, with between one and ten mammograms and between zero and seven ultrasounds. On average five mammograms and four ultrasounds were performed on follow-up patients each day.

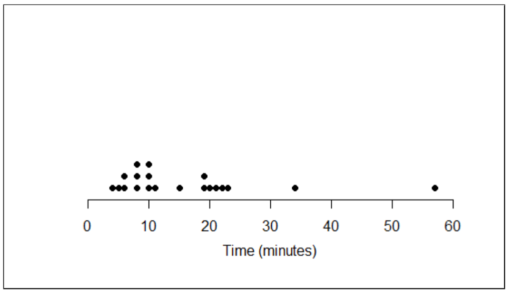


FIGURE 4.8: Time for report to be ready following mammogram

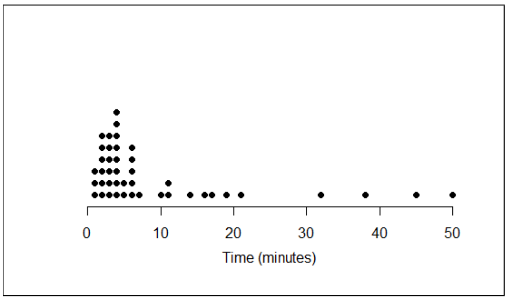


FIGURE 4.9: Time for report to be ready following ultrasound

4.2 Data linking GP referral information to clinic tests and results

In order to understand how GP referral information links to clinic tests and results, further data collection was necessary. Since this is patient identifiable information, patient consent was required to collect these data for research. From January until March 2016, I handed out study information sheets and consent forms to patients on arrival at the clinic and was available to answer questions about the study. Due to the lengthy ethics process, there was limited time available to collect data.

A total of 234 patients, that is 29% of patients who attended the clinic, consented to their information being used for the study. There were very few males so these were excluded. Follow-up patients and patients not referred by a GP were also excluded, leaving 224 patients. The referral information of 179 patients (80% of the 224 eligible patients) had been scanned in, so was available to access on the hospital computer network via the EDMS patient clinical viewer. The data fields obtained from GP referral information and clinic records of these 179 patients are explained in Section 4.2.1 and 4.2.2 respectively.

4.2.1 GP referral information

When a GP refers a patient to the breast diagnostic clinic, they fax or post a referral form or letter to the hospital. The standard form is shown in Figures C.3, C.4 and C.5, and the older version is in Figures C.6 and C.7. The majority of referrals were on the standard form, while some GPs used the older version or wrote letters. Table 4.7 shows how many referrals were made using each method.

Table 4.7: Referral methods

Referral method	*n* (%)
Standard form	137 (77)
Outdated standard form	22 (12)
Letter	17 (9)
Other form	1 (<1)
Standard form and letter	1 (<1)
Outdated standard form and letter	1 (<1)

Completeness of information

For research question 1, we first need to ascertain whether the GP referral information is complete enough to be of use in classification models to predict abnormality. Both versions of the referral form contain tickboxes to indicate the reason for referral (a patient's symptoms and whether there is a family history of cancer). Table 4.8 shows the completeness of tickboxes in our sample. The referrals with no boxes ticked are either referrals to the Family History clinic (so the referral reason is clearly family history of cancer) or information about symptoms is provided elsewhere on the form. The high rate of completeness of the forms suggests that it is feasible to use this information in classification models, particularly if the information provided within text is used in addition to the tickboxes.

Table 4.8: Completeness of GP referral information

Completeness of tickboxes	n (%)
At least one box ticked	152 (85)
No tickboxes available (letters)	17 (9)
No boxes ticked	8 (4)
Page with tickboxes missing	2 (1)

Data pre-processing

Information from the referral forms and letters was coded into data fields in an Excel spreadsheet. As well as looking at the tickboxes, information provided in text was also considered. For example if a lump was recorded anywhere on a form or letter, the data field *lump* was given the value *Yes.* If it was recorded that there was no lump or there was no mention of a lump, the data field *lump* was given the value *No or NR.* These two cases were grouped together because no record of a lump is likely to mean no lump was present.

The concept of referral urgency is recorded differently on the two types of form. "Nature of referral" on the standard form has fewer categories than the older version's "GP assessment of priority" (see Tables 4.9 and 4.10). There is no direct mapping between the categories on the two forms. Since most referrals fall into the "two week wait - suspected cancer" and "symptomatic - not suspected cancer" categories, a new data field called *urgency* was defined with possible values of *suspected cancer*, *symptomatic* or *NR or other.* The last category was used for all referrals on the old form and in letters.

Table 4.9: Referral urgency - standard form

Nature of Referral	Number of patients
Two week wait - suspected cancer	68
Symptomatic - not suspected cancer	56
Referral to Family History Clinic	4
Other	1
NR	5
Two week wait and symptomatic	2
Two week wait and other	1
Symptomatic and other	1
Total standard forms	138

Table 4.10: Referral urgency - old form

GP Assessment of Priority	Number of patients
5) Diagnosis of cancer	0
4) Suspected cancer	5
3) Probably benign	16
2) Benign but needs to be seen	2
1) Routine e.g. family history referral	0
0) Other - please attach explanatory letter	0
Total	23

TABLE 4.11: GP referral information (continuous variables) - Descriptive statistics

Referral information	Median (range)
Age, years	38 (18-82)
Duration of symptoms, days	61 (1-2920)

TABLE 4.12: GP referral information (categorical variables) - Descriptive statistics

Referral information	n (%)
NA or NR	46 (26)
Urgency	
Suspected cancer	68 (38)
Symptomatic	56 (31)
Other or NR	55 (31)
Family history of cancer	
Yes	57 (32)
No or NR	122 (68)
Asymptomatic	
Yes	9 (5)
No or NR	170 (95)

Descriptive statistics

By pre-processing the data, I obtained a dataset with one set of variables regardless of which format referral information was provided in. Descriptive statistics for continuous variables are shown in Table 4.11, categorical variables (except symptoms) in Table 4.12 and recording of symptoms in Table 4.13.

There are two continuous variables: *age* and *duration of symptoms*. The age of patients ranges from 18 to 82 years with a median of 38 years. In 26% of cases, the *duration of symptoms* is not recorded, although in 5% of cases the duration is irrelevant since there are no symptoms (the patient is asymptomatic). The recorded *duration of symptoms* ranges considerably from one day up to eight years, with a median duration of two months.

The *urgency* is a categorical variable. The split is fairly even between the three possible values, with 38% of patients referred as *suspected cancer* cases, 31% as *symptomatic* and 31% as *other or NR*, which includes all referrals on letters or the outdated form, as explained previously. The remaining variables are all binary. A substantial proportion of patients (32%) have a family history of cancer recorded.

In terms of symptoms, the most commonly recorded is a lump, in 53% of cases, followed by unilateral (one-sided) pain, in 40% of cases. The following symptoms were more rarely recorded, in 8% or fewer cases: spontaneous bloody or clear nipple discharge, new nipple alteration, skin dimpling and persistent unilateral nodularity.

Table 4.13: Symptoms recorded by GPs - Descriptive statistics

Symptoms	n (%)
Lump	
Yes	95 (53)
No or NR	84 (47)
Unilateral pain	
Yes	71 (40)
No or NR	108 (60)
Spontaneous bloody or clear nipple discharge	
Yes	5 (3)
No or NR	174 (97)
New nipple alteration	
Yes	15 (8)
No or NR	164 (92)
Skin dimpling	
Yes	4 (2)
No or NR	175 (98)
Persistent unilateral nodularity	
Yes	14 (8)
No or NR	165 (92)
Further symptoms	
Yes	32 (18)
No or NR	147 (82)

Symptoms that do not fall into these categories (both unspecified symptoms and different symptoms) are classed as *further symptoms*, occurring in 18% of referrals.

4.2.2 Patients' clinic records

I collected further information about the 179 eligible consenting patients from patients' clinic records stored on hospital computer systems. This consisted of what tests a patient had during their visit to the diagnostic clinic, the result scores for these tests and the outcome of their appointment. Some descriptive statistics are shown in Table 4.14.

Of the 179 patients, most (73%) had an ultrasound, while fewer than half (42%) had a mammogram. Of the patients who had an ultrasound, 27% also had a biopsy. Some (12%) patients were discharged straight after their initial consultation without any tests.

We define *normal results* and *abnormal results* in the following way. In diagnostic breast clinics, patients are routinely given a score between 1 and 5 for each test and each breast where applicable. A score of 1 means “normal”, 2 “benign”, 3 “unusual/uncertain but probably benign”, 4 “suspicious” and 5 “malignant” (Willett et al., 2010). I calculated an overall test result by taking the highest (worst) score of

Table 4.14: Patients' clinic records - Descriptive statistics

Clinic visit information	n (%)
Imaging tests	
Ultrasound only	83 (46)
Mammogram only	26 (15)
Both ultrasound and mammogram	49 (27)
Neither ultrasound nor mammogram	21 (12)
Worst test result (mammogram, ultrasound, biopsy)	
N/A - No tests	21 (12)
1 - Normal	75 (42)
2 - Benign	68 (38)
3 - Unusual/uncertain but probably benign	4 (2)
4 - Suspicious	2 (1)
5 - Malignant	7 (4)
Missing some results	2 (1)
Outcome	
Discharged	134 (75)
Follow-up or open appointment	38 (21)
Cancer	7 (4)

mammogram, ultrasound and biopsy. If the overall test result is 2 or higher, we say the patient has an *abnormal result*. If the overall test result is 1, we say the patient has a *normal result*. Patients who do not have tests are also said to have *normal results*, since it is assumed that their initial consultation confirmed that their breasts were normal.

More than half of the patients in the sample (97, 54%) have normal results, while the rest (80, 45%) have abnormal results. Figure 4.10 demonstrates the composition of these two groups. The remaining 2 (1%) patients have some missing results so were not assigned an overall test result. From other clinic information it was deduced that the most likely results for these two patients were normal and abnormal respectively. Including these two patients, 55% of patients have a normal result and 45% have an abnormal result.

The majority (75%) of patients were discharged, 21% were offered follow-up or open appointments, while 4% were diagnosed with cancer.

4.3 Conclusion to data from GP referrals and the breast diagnostic clinic

By collecting data on different aspects of how the clinic currently operates, many sources of variability were identified including patient punctuality, service times, test requirements and numbers of follow-up patients requiring tests. These suggest simulation as an appropriate approach to test changes to the organisation of the clinic;

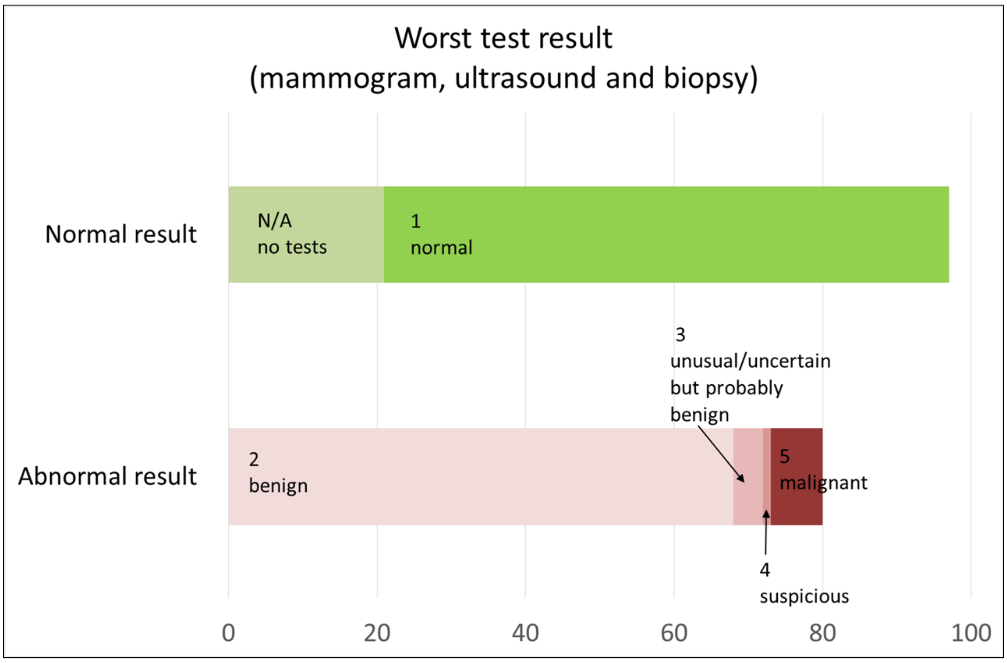


Figure 4.10: Highest test result (mammogram, ultrasound, biopsy)

in Chapter 6 we describe the development of simulation models. Analysis of the patient feedback highlights variation in patients' perceptions or experiences of waiting times, with some recording a day of long waits and others a smooth efficient process. In particular, the process to transmit results from the imaging department to the clinic seems unnecessarily complex. Several patients suggested they would prefer to receive their results in the post or by telephone.

There is potential for GP referral information to be used for operational planning, since we find that in our sample it is relatively complete, particularly when one considers information included in text fields and letters in addition to tickboxes. Since the numbers of patients receiving a cancer diagnosis at the clinic is very small, we are not attempting to predict a cancer diagnosis and this is beyond the scope of our research. However given that our sample contains large numbers both of patients with normal results and of patients with abnormal results, it may be feasible to predict from GP referral information whether patients will receive a normal or abnormal result. This is investigated in Chapter 5 and could help with identifying higher risk patients who could be sent straight for imaging tests. As explained in Section 1.1, there is currently no evidence that GP referral information is suitable to be used to decide which patients to send straight for tests.

GP referral forms are necessarily changed over time as referral guidelines and policies develop. Likewise forms for referral to different geographical areas vary (for example, the standard form in our case study is used for referrals to hospitals in the North Central London and West Essex Cancer Commissioning Network only). Nonetheless, this work is still relevant to other areas of the UK since similar information (if not exactly the same or asked for in exactly the same way) is required by each diagnostic clinic. We have demonstrated that information from referral letters and two versions of referral form including text fields can be coded into a dataset with common variables for all referral methods.

Chapter 5

Predicting abnormal diagnostic results from GP referral information

As concluded in Chapter 4, it could be useful to clinic managers to know in advance which patients are likely to have abnormal results (i.e. for which patients a breast abnormality is likely to be detected from mammogram, ultrasound or biopsy). This could help with identifying higher risk patients who could be sent straight for imaging tests. Initial data analysis in Section 4.2.1 showed that GP referral information is relatively complete, confirming that it may be feasible to use it in operational planning. Next, continuing to address research objective 1, we investigate whether the information provided is useful in predicting abnormality. Crucially, we are not attempting to predict cancer (only 4% of our sample), but the much wider group of patients who have abnormal results (45% of our sample).

As explained in Chapter 2, there are two ways in which breast diagnostic clinics are typically organised. At the Whittington clinic, for example, all patients first see a clinician who decides whether imaging tests are required. In some other clinics, GP referral information is used to identify patients to be sent straight for imaging tests. There is a lack of evidence for the link between GP referral information and diagnostic results to lend support to either approach; this research aims to fill this gap.

This chapter proceeds as follows. First in Section 5.1 we describe our methodology: using logistic regression to develop predictive scorecards. Then in Section 5.2 we present the results including descriptive analysis of referral characteristics (patient characteristics obtained from GP referral information), the scorecards and a comparison of their performance measures. The chapter concludes with Section 5.3.

5.1 Methodology

We chose to use logistic regression scorecards for predicting abnormal results from GP referral information for the following reasons. Classification models are appropriate because the outcome variable is binary rather than continuous, taking values *normal* or *abnormal*. Logistic regression was chosen because it is a standard technique, popular in the medical literature (see Section 3.2.4) and so more likely to be familiar to clinic staff and GPs than some other classification methods. We transform the logistic regression models into scorecards for ease of interpretation using the *weights of evidence (WoE)* approach common in financial credit scoring applications (Thomas, 2009). Such scorecards can be represented on an arbitrary linear scale, are particularly suitable when the concept of risk is involved and make the same predictions as the original logistic regression models. Scorecards are popular in healthcare; other examples of scorecards that were discussed in Section 3.2.4 are the widely-used Nottingham Prognostic Index, which is based on a linear regression, and is used to calculate the stage of cancer, and a recently published scorecard for predicting malignancy after ultrasound that provides an approximation to the original logistic regression model.

The dataset used for fitting models consists of data from the GP referral information and clinic records of 179 patients. This dataset is described in Section 4.2. We considered the following characteristics from referral information for inclusion as predictors in our models: *family history of cancer*, *lump*, *unilateral pain* (one-sided pain), *other symptom*, *urgency*, *duration of symptoms* and *age*. *Other symptom* refers to any of the rarer symptoms (with 15 or fewer cases) and those in the *further symptoms* category. We first examined the relationship between each referral characteristic and the outcome variable (normal or abnormal result). For logistic regression, we need predictors to be not too highly correlated, so we tested for interactions between each pair of referral characteristics using Cramér's V, chosen because most variables are categorical (Cramér, 1946).

Values of the two continuous referral characteristics, *duration of symptoms* and *age* were grouped into attributes using the Interactive Grouping feature in SAS Enterprise Miner. This feature automatically generates attributes starting with quantiles then using a decision tree to split or combine attributes to achieve the largest reduction in entropy. Zero entropy for an attribute would mean that all patients with that attribute had the same diagnostic result. Such grouping helps build more parsimonious and robust models that can capture non-monotonic relationships.

Models were developed in SAS Enterprise Miner 14.1 software. The relative risks associated with different attributes of the same referral characteristic were assessed using WoE. The *WoE* of an attribute, for example *young*, of a characteristic, for example *age*, is the strength of evidence that patients who are young will have an abnormal result. Logistic regression parameters for each WoE-coded referral characteristic were estimated using maximum likelihood estimation. These parameters correspond to the WoE-coded referral characteristic's importance in predicting abnormality in the presence of other referral characteristics. The conversion from the logistic regression to scorecard formulation is described in Appendix A. For the conversion, two values are specified: the odds of an abnormal result at a particular score, and the score points to double the odds.

To aid with selecting predictors, the information values, which measure how much each referral characteristic contributes to the abnormality prediction, were calculated. We present two logistic regression scorecards: the *full scorecard*, which contains all predictive referral characteristics, and the *simple scorecard*, which contains only the most predictive referral characteristics.

The two scorecards were compared in terms of goodness-of-fit, discrimination between normal and abnormal results, and performance measures that depend on the choice of *cut-off score*. The cut-off score is the threshold between low- and high-risk patients. Patients with a score below the cut-off are predicted a normal result and those with a score at least as high as the cut-off score are predicted an abnormal result. Since the sample size is relatively small ($n = 179$ patients), the scorecards were estimated using the entire dataset to make use of all the available data, rather than removing some to use in validation. Instead, a bootstrapping technique (sampling with replacement), implemented in Microsoft Excel, was used to internally validate the models. For the original data set and each of the bootstrap datasets, the models' performance were assessed in terms of a Hosmer-Lemeshow goodness-of-fit test, area under the receiver operating characteristic curve (AUROC) and the Kolmogorov-Smirnov (KS) statistic. Cut-off specific performance measures were also calculated for the original data set.

5.2 Results

5.2.1 Preliminary analysis: Comparing referral characteristics of normal and abnormal result patients

Here we discuss the relationships between each referral characteristic and patients' results (normal or abnormal). A chart showing the urgency of referrals according to results is shown in Figure 5.1. A higher proportion of the abnormal results patients were categorised as *suspected cancer* than of the normal results patients, although the proportion of normal results patients referred as *suspected cancer* is still substantial (33%). The chart in Figure 5.2 shows that a similar proportion of patients with normal and abnormal results have a family history of cancer (33% and 31% respectively). This variable may be less relevant for our purpose since we are predicting abnormality rather than cancer. It can be seen from the boxplots in Figure 5.3 that the median, upper quartile and maximum age of patients with abnormal results are higher than for patients with normal results. Ages are less varied for those with normal results.

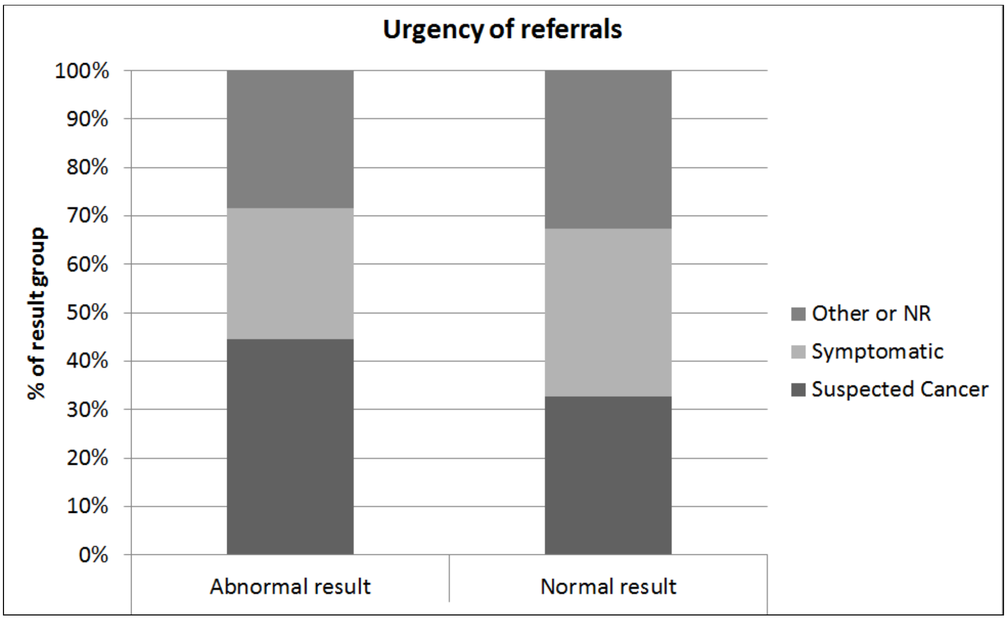


FIGURE 5.1: Urgency assigned by GPs to patients with abnormal/normal results

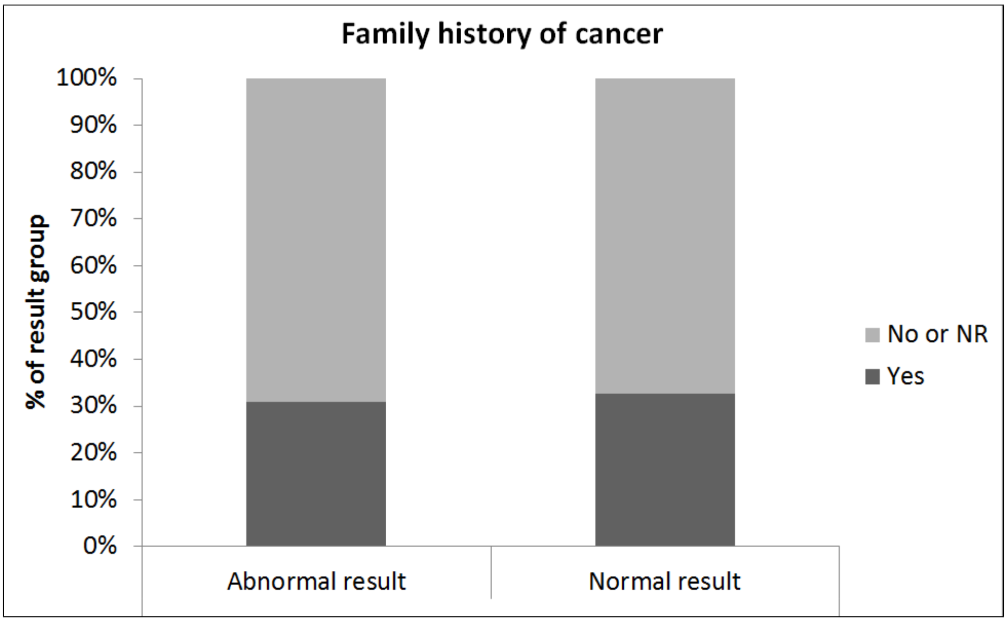


FIGURE 5.2: Proportion of patients with abnormal/normal results who have a family history of cancer recorded

Next we compare information about symptoms for the two patient groups. The chart in Figure 5.4 compares the proportions of patients with abnormal and normal results who have a lump, unilateral pain and other symptoms recorded. In particular it can be seen that a higher proportion of patients with abnormal results than patients with normal results had a lump recorded by their GP. Interestingly, a higher proportion of patients with normal results had unilateral pain recorded than patients with abnormal

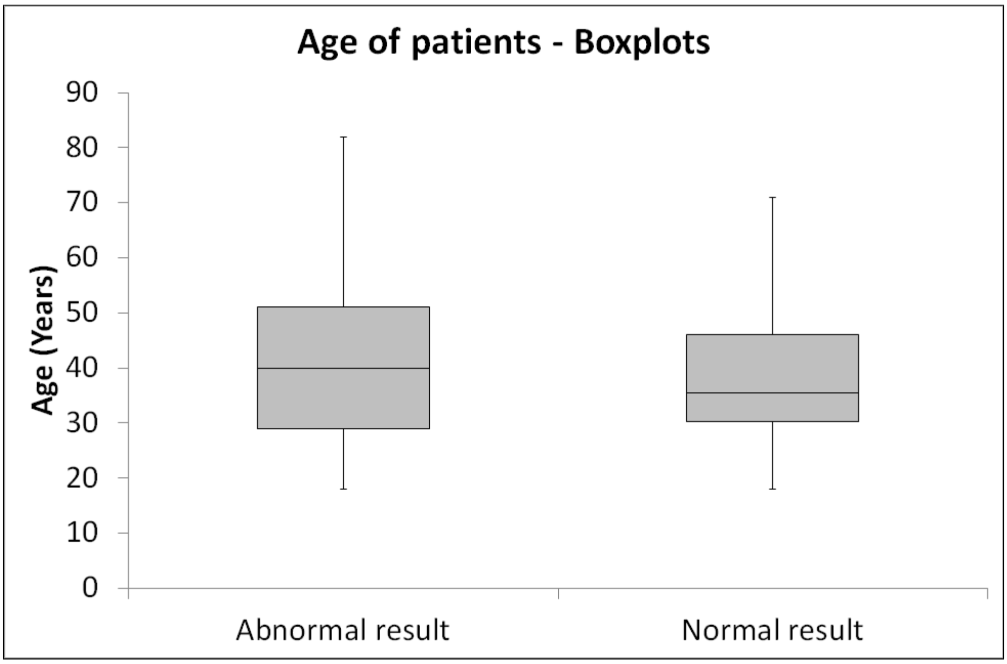


FIGURE 5.3: Ages of patients with abnormal/normal results

results. Pain can be experienced by women with healthy breasts, for example pain related to hormonal changes (Harvey et al., 2014). Symptoms other than lump or unilateral pain are recorded in a higher proportion of normal cases than of abnormal cases. This is surprising and it is possible that this variable may be confounded by other referral characteristics. The duration of symptoms was not recorded or not applicable for a higher proportion of patients with normal results than of patients with abnormal results (28% compared to 23%). The boxplots in Figure 5.5 are for those referrals where the duration was recorded. The range in *duration of symptoms* before being referred to the clinic is very large, from 1 day to 8 years. The upper quartile and maximum *duration of symptoms* for patients with abnormal results is longer than for those with normal results.

5.2.2 Predictor selection and grouping results

The Cramér's V between each pair of referral characteristics is less than 0.6, which means that the associations are not strong, so no interactions will be included in the logistic regression.

The attributes produced by Interactive Grouping for *duration of symptoms* are in days and can be interpreted as roughly equivalent (due to different month lengths) to *under 2 weeks*, *2 weeks to 2 months*, *2 - 5 months*, *over 5 months* and *NA or NR*. The attributes generated for *age* are $Age<29$, $29\leq Age<42$, $42\leq Age<47$, $47\leq Age<52$ and $52\leq Age$. This Interactive Grouping result suggests that further segmenting the last *age* attribute would not improve the grouping of similar patients (in terms of normal

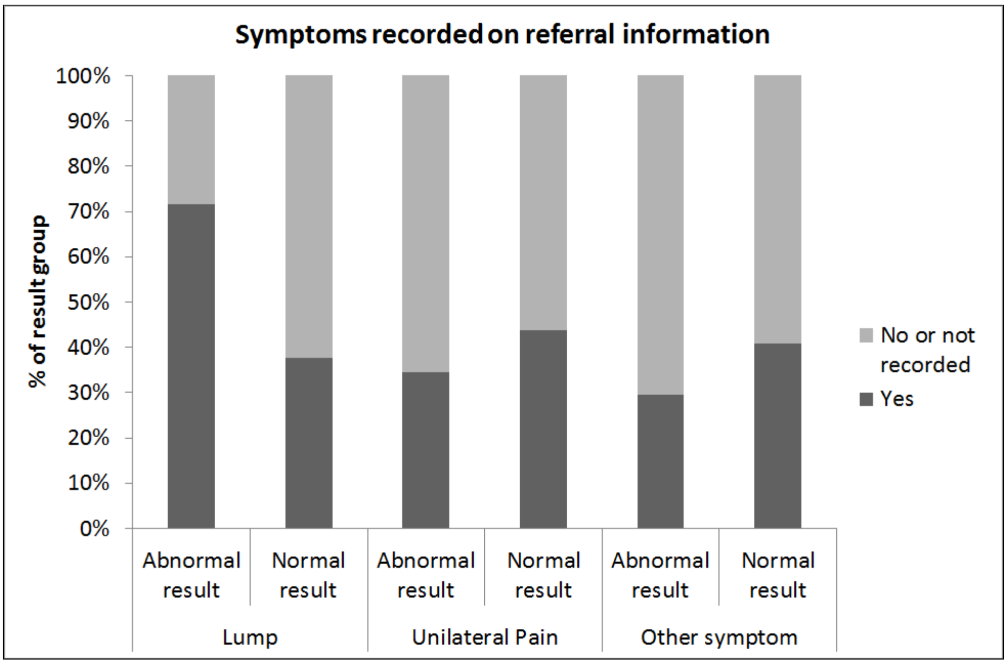


Figure 5.4: Proportion of patients with abnormal/normal results with each symptom recorded

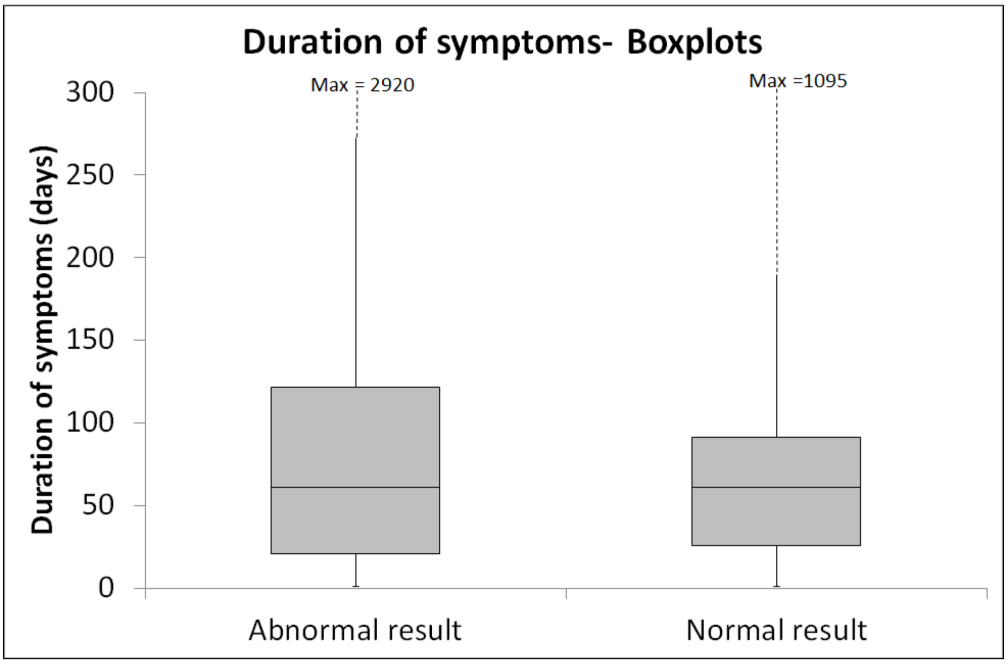


Figure 5.5: Duration of symptoms for patients with abnormal/normal results (for patients where this was recorded)

and abnormal results). These *age* attributes are used in the full scorecard, but for the simple scorecard we slightly adapt them for ease of use: *Age<30*, *30≤Age<40*, *40≤Age<50* and *50≤Age*.

The WoE-coded referral characteristics with the highest information values are *lump* (0.48) and *age* (0.15 or 0.22 depending on grouping). The simple scorecard contains only these two most predictive referral characteristics, which are strong and medium predictors of abnormal results respectively (SAS, 2009). The information value for *duration of symptoms* is 0.08, for *urgency* is 0.06, for *other symptom* is 0.06, for *unilateral pain* is 0.04 and for *family history of cancer* is 0.001. These are all weak predictors of abnormal results and are included in the full scorecard, along with *lump* and *age*, to utilise all the information available.

5.2.3 Specification of models and interpretation

The scorecard ranges were set in the following way. To set a score range for the simple scorecard, odds of 3:1 of having an abnormal result for a score of 25 points, and 5 points to double the odds were assumed. For the full scorecard, a broader score range is used, as odds of 12:1 for a score of 300 points and 20 points to double the odds were specified. The scorecard points for the simple and full scorecards are presented in Tables 5.1 and 5.2 respectively. A new patient's score is calculated by adding up the points corresponding to the attributes recorded in their referral information. The higher the score is, the higher the risk that a patient has an abnormal result. The choice of cut-off score below which a patient is predicted to have a normal result is elaborated on in Section 5.3.

TABLE 5.1: Simple scorecard (points and sample sizes)

		Scorecard points	Sample size, n (%)
Age	Age $<$ 30	10	47 (26)
	30$\leq$ Age$<$ 40	3	51 (28)
	40$\leq$ Age$<$ 50	8	44 (25)
	50$\leq$ Age	12	37 (21)
Lump	Yes	13	95 (53)
	No or not recorded	2	84 (47)

For the simple scorecard, the relationships between scores, odds and probabilities are given in Figure 5.3. A patient aged over 50 with a lump has the maximum score of 25 and odds of 3:1 of having an abnormal result. A patient aged between 30 and 40 without a lump has the minimum score of 5 and odds of 3:16. The scoring for *age* is non-linear and we discuss possible explanations for this later in this section.

For the full scorecard, the relationships between scores, odds and probabilities for a subset of scores are given in Table 5.4. A patient with all the highest risk attributes

Table 5.2: Full scorecard (points and sample sizes)

		Scorecard points	Sample size, n (%)
Age	Age $<$ 29	43	38 (21)
	29$\leq$ Age$<$ 42	16	69 (39)
	42$\leq$ Age$<$ 47	33	24 (13)
	47$\leq$ Age$<$ 52	21	21 (12)
	52$\leq$ Age	64	27 (15)
Lump	No or NR	5	84 (47)
	Y	54	95 (53)
Duration of symptoms	NA or NR	29	46 (26)
	Less than 2 weeks	42	17 (9)
	2 weeks - 2 months	35	63 (35)
	2 - 5 months	26	25 (14)
	Over 5 months	35	28 (16)
Family history of cancer	No or NR	29	122 (68)
	Y	38	57 (32)
Other symptom	No or NR	29	115 (64)
	Y	36	64 (36)
Unilateral pain	No or NR	31	108 (60)
	Y	33	71 (40)
Urgency	Suspected cancer	32	68 (38)
	Symptomatic	31	56 (31)
	Other or NR	32	55 (31)

Table 5.3: The relationship between scores and odds for the simple scorecard

Score	Odds of abnormal result	Probability of abnormal result
25	3:1	0.75
23	$\approx$ 16:7	0.69
21	$\approx$ 12:7	0.63
16	$\approx$ 6:7	0.46
14	$\approx$ 2:3	0.40
12	$\approx$ 1:2	0.33
10	3:8	0.27
5	3:16	0.16

(aged over 52, symptoms for less than 2 weeks, family history of cancer, a lump, unilateral pain, another symptom and classed as suspected cancer) has the maximum score of 299, and the odds of having an abnormal result are about 12:1. The minimum score is 167 and corresponding odds are 1:8. As for the simple scorecard, there is non-linear scoring for the *age* characteristic, but also for the *duration of symptoms* characteristic.

Table 5.4: The relationship between scores and odds for the full scorecard

Score	Odds of abnormal result	Probability of abnormal result
299	≈ 12:1	0.92
280	6:1	0.86
260	3:1	0.75
240	3:2	0.60
228	≈ 1:1	0.50
220	3:4	0.43
200	3:8	0.27
180	3:16	0.16
167	≈ 1:8	0.11

In both scorecards, the oldest age category has the highest score, followed by the youngest age category. This means that according to the scorecards, all other referral characteristics being equal, patients of these ages attending the clinic are most likely to be diagnosed with breast conditions. Although the risk of breast cancer increases with age, the same is not necessarily true of abnormality, which covers many conditions. Age is taken into account when referring patients to the clinic. For example, National Institute for Health and Care Excellence (2015) urges urgent (two-week-wait) referral for patients aged 30 and over with an unexplained breast lump, but suggests GPs consider a non-urgent referral or discussion with a specialist to decide on whether a referral is needed for those under 30 with an unexplained breast lump. Thus it may be that many under 30s with normal breast changes are reassured by their GP rather than attending clinic, making the risk of abnormal results for young clinic attenders high.

We suggest a possible explanation for the non-linear scoring of *duration of symptoms* in the full scorecard. Patients with symptoms for less than two weeks (before being referred to a specialist) have the highest score for *duration of symptoms*, with the score decreasing as *duration of symptoms* increases up to 2-5 months. A possible interpretation is the link to severity of symptoms; patients with less severe symptoms may have waited longer before visiting their GP, or their GP may have managed their care for longer before deciding a referral was necessary. For the group with longest *duration of symptoms*, the risk increases again, perhaps because any symptoms lasting a very long time are more likely to relate to an abnormality.

5.2.4 Goodness-of-fit

In order to visually inspect whether the scorecards are correctly ranking patients with abnormal results above patients with normal results, I plotted cumulative distributions of scores separately for these two types of patients. Figure 5.6 and Figure 5.7 show the results when applying the simple scorecard and full scorecard respectively to the original data set. The cumulative distributions for the full scorecard are plotted at intervals of 10 points.

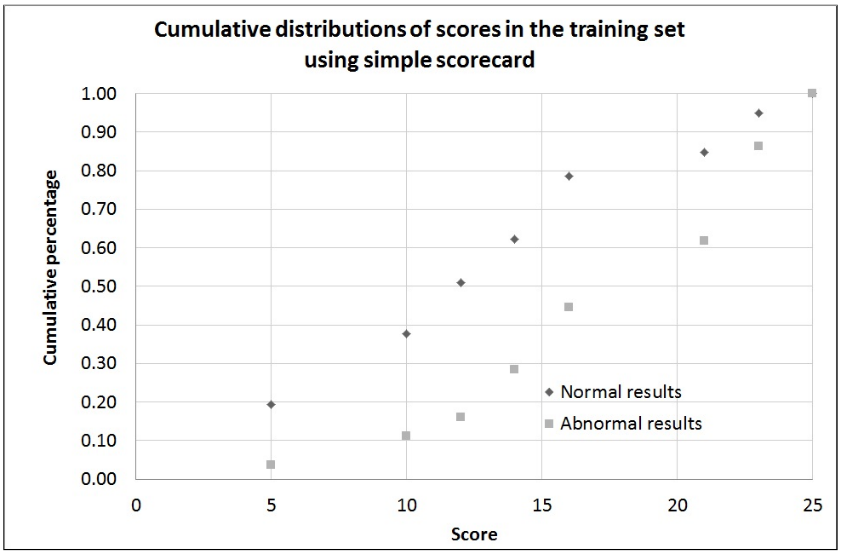


FIGURE 5.6: Cumulative distributions of scores in the original data set using the simple scorecard

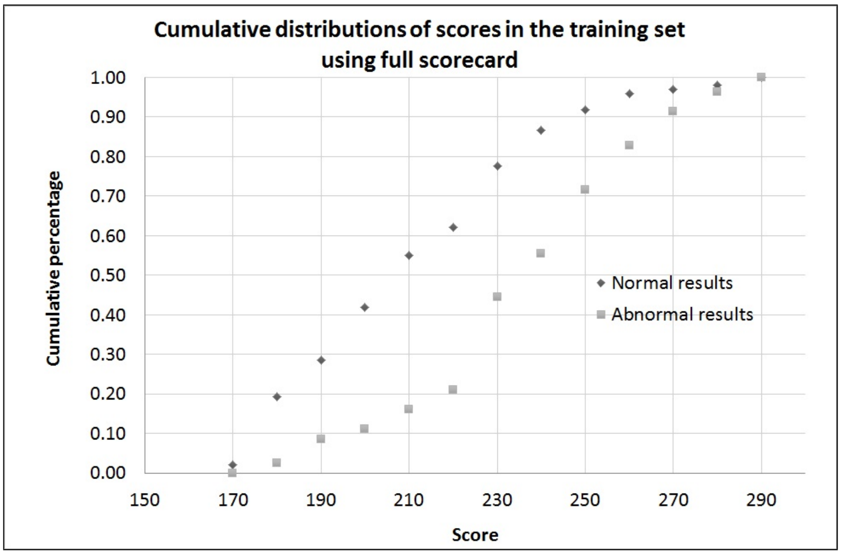


FIGURE 5.7: Cumulative distributions of scores in the original data set using the full scorecard

In both cases, the charts are as expected, with a higher proportion of patients with normal results than patients with abnormal results receiving the lower scores. For example, using the simple scorecard, almost 40% of patients with normal results have scores less than or equal to 10 compared to only 10% of patients with abnormal results. The scores from the full scorecard observed in the original data set do not cover the whole of the possible range (167 to 299); the lowest score here is 170 and the highest is 283.

Next I performed Hosmer-Lemeshow goodness-of-fit tests on bootstrap samples. I repeated the test with 5 different random seeds, and for each of these performed 5000 iterations. I calculated the average p-value and the percentage of the 5000 p-values that were below 0.05, i.e. providing insufficient evidence of a good fit at the 5% significance level. For the simple scorecard, 8 bins were used, corresponding to the 8 possible scores and their related probabilities of having an abnormal result. For the full scorecard 7 bins with different probabilities (p) of having an abnormal result were used: $p > 0.7$, $0.6 < p \leq 0.7$, $0.5 < p \leq 0.6$, $0.4 < p \leq 0.5$, $0.3 < p \leq 0.4$, $0.2 < p \leq 0.3$ and $p \leq 0.2$. With fewer than 5000 iterations, the results were quite unstable due to not having enough observations in each bin.

The results are shown in Tables 5.5 and 5.6. For the simple scorecard, the percentage of p-values that would lead us to reject the model as a poor fit ranges from 19% to 21%. The average p-value is comfortably above 0.05 in each repetition. For the full scorecard, the percentage of p-values that would lead us to reject the model as a poor fit ranges from 28% to 30%, which is worse than for the simple model. The average p-value is larger than 0.05 in each repetition. Although the scorecards do not provide a good fit for all bootstrap samples, in this particular application it is more important to discriminate well between normal and abnormal results than it is to accurately predict the probability of an abnormal result, i.e. calibrate the models. So for our purpose, these goodness-of-fit results are acceptable.

Table 5.5: Hosmer-Lemeshow goodness-of-fit test for the simple scorecard using 5000 iterations for each repetition

Repetition number	% of p-values<0.05	Average p-value
1	19%	0.280
2	18%	0.293
3	21%	0.272
4	19%	0.285
5	20%	0.273

5.2.5 Measures of discrimination

Measures of discrimination for the simple and full scorecards were calculated for the original data set and 1000 bootstrap samples. The results are shown in Table 5.7 and the receiver operating characteristic (ROC) curves for the original data set are shown in Figure 5.8. The ROC curve for the simple scorecard is smoother because there are

Table 5.6: Hosmer-Lemeshow goodness-of-fit test for the full scorecard using 5000 iterations for each repetition

Repetition number	% of p-values < 0.05	Average p-value
1	28%	0.215
2	28%	0.208
3	29%	0.213
4	30%	0.208
5	29%	0.217

fewer possible cut-off values than for the full scorecard. Otherwise, the two curves follow similar trajectories. According to Hosmer et al. (1997), an area under the receiver operating characteristic curve (AUROC) value of between 0.7 and 0.8 is acceptable. Hence both the full and simple scorecards have acceptable AUROC values. Given that Kolmogorov-Smirnov (KS) statistic values are related to AUROC values (Thomas, 2009), these are also acceptable. As a general rule, a KS statistic above 0.2 suggests fair discrimination, and above 0.4 suggests good discrimination, so our full model has good discriminatory power, and the separation ability of the simple model falls comfortably within the fair range (Mays, 2001). This small decrease in separation ability is outweighed by the much more parsimonious structure in the simple model. For the simple scorecard, the KS is achieved at 14 on average for the bootstrap samples. For the full scorecard, it is achieved at a score of 220 points on average for the bootstrap samples.

Table 5.7: Measures of discrimination for the simple and full scorecards

	AUROC	KS statistic	Score where KS is achieved
Simple scorecard			
Original data	0.725	0.350	12
Average of bootstrap samples [95% CI]	0.724 [0.721, 0.726]	0.377 [0.373, 0.381]	14
Full scorecard			
Original data	0.752	0.413	218
Average of bootstrap samples [95% CI]	0.754 [0.751, 0.756]	0.438 [0.434,0.441]	220

5.2.6 Performance measures that depend on choice of cut-off score

It only becomes possible to assess how often a scorecard predicts correctly once a cut-off score has been specified. A cut-off score of S means that patients with scores of

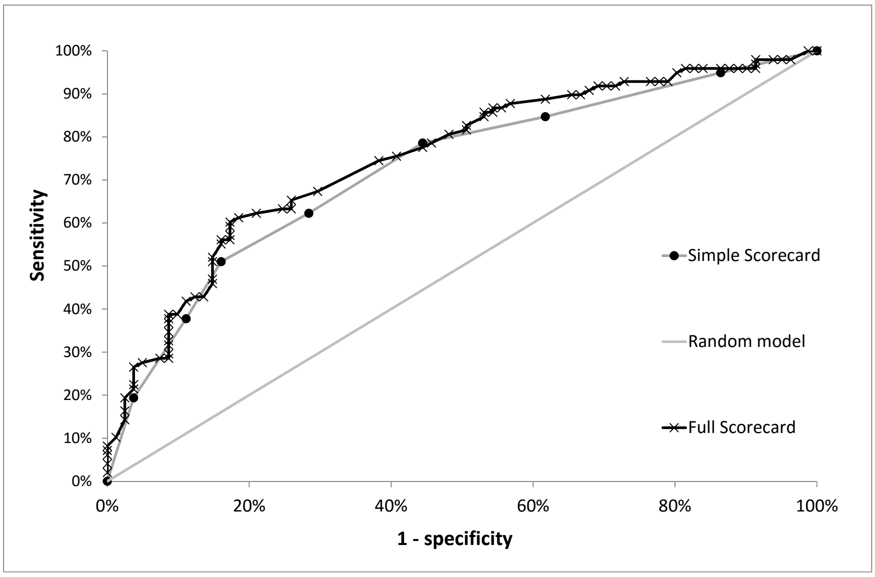


Figure 5.8: ROC curves for the simple and full scorecards compared to a random model

at least S are predicted abnormal results. For the simple scorecard, there are are eight possible scores. Including the scenario predicting all patients will have abnormal results and the scenario predicting all patients will have normal results, there are hence nine possible unique cut-off scores. Table 5.8 shows cut-off specific performance measures for the simple scorecard, based on the original data. For the full scorecard, there are a large number of possible scores, so we present performance measures for a selection of cut-off scores only, in Table 5.9.

Table 5.8: Cut-off specific performance measures for the simple scorecard, assessed on original data. TP=True positives, TN=True negatives, FP=False positives, FN=False negatives

Cut-off	TP	TN	FP	FN	Sensitivity	Specificity	Classification accuracy
26	98	0	81	0	100%	0%	55%
24	93	11	70	5	95%	14%	58%
22	83	31	50	15	85%	38%	64%
17	77	45	36	21	79%	56%	68%
15	61	58	23	37	62%	72%	66%
13	50	68	13	48	51%	84%	66%
11	37	72	9	61	38%	89%	61%
8	19	78	3	79	19%	96%	54%
4	0	81	0	98	0%	100%	45%

In our context, true positives are patients with normal results who were correctly

Table 5.9: Cut-off specific performance measures for the full scorecard, assessed on original data. TP=True positives, TN=True negatives, FP=False positives, FN=False negatives

Cut-off	TP	TN	FP	FN	Sensitivity	Specificity	Classification accuracy
290	98	0	81	0	100%	0%	55%
280	96	3	78	2	98%	4%	55%
270	95	7	74	3	97%	9%	57%
260	94	15	66	4	96%	19%	61%
250	90	24	57	8	92%	30%	64%
240	85	37	44	13	87%	46%	68%
230	76	45	36	22	78%	56%	68%
220	60	66	15	38	61%	81%	70%
210	51	69	12	47	52%	85%	67%
200	38	73	8	60	39%	90%	62%
190	28	74	7	70	29%	91%	57%
180	16	79	2	82	16%	98%	53%
170	0	81	0	98	0%	100%	45%
160	0	81	0	98	0%	100%	45%

classified, and true negatives are patients with abnormal results who were correctly classified. On the other hand false positives are patients who were predicted normal results but who actually have abnormal results. False negatives are patients who were predicted abnormal results but who actually have normal results. Sensitivity is the proportion of normal results that were correctly predicted, and specificity is the proportion of abnormal results that were correctly predicted.

For the simple scorecard, the best classification accuracy is 68% and is achieved when the cut-off is set at 17. Both sensitivity and specificity are greater than 60% when the cut-off is 15. If sensitivity and specificity are equally important, and if this were the only consideration, then 15 would be the best choice of cut-off. For the full scorecard, the best classification accuracy among the cut-off scores considered is 70%, and is achieved when the cut-off score is 220. This cut-off also achieves the best balance between specificity and sensitivity, with both greater than 60%. The best classification accuracy from the full scorecard offers a marginal improvement over the simple scorecard (70% versus 68%), and the cut-off with the best balance between specificity and sensitivity improves the specificity substantially (81% versus 72%) with a small decrease in sensitivity (61% versus 62%).

5.3 Conclusion to predicting abnormal diagnostic results from GP referral information

Two logistic regression scorecard models were fitted to find out which GP referral information, if any, is useful in distinguishing between patients with normal and abnormal results. The two most predictive referral characteristics in our sample (n=179) are *lump* and *age*. The simple scorecard, which contains only these two characteristics, is only slightly worse at separating normal and abnormal results than the full scorecard, which contains seven referral characteristics (the average AUROCs on bootstrap samples are 0.72 and 0.75 respectively). To the best of our knowledge this is the first study providing quantitative evidence of a relationship between GP referral information and the chance of a breast abnormality being diagnosed.

For the simple scorecard, setting the cut-off score at 17 would give the best classification accuracy (68%), and a cut-off score of 15 achieves the best balance between specificity and sensitivity, with both greater than 60%. For the full scorecard, setting the cut-off score at 220 is best both in terms of classification accuracy (70%) and so that both sensitivity and specificity are as high as possible (both above 60%).

Although these findings can help with deciding where to set the cut-off, our context is an example of a situation where the impact on operational measures is more important than a scorecard's predictive performance. We propose a novel approach for selecting the cut-off score by evaluating the operational impact of different cut-offs in a simulation model. Specifically, in Chapter 6 we describe development of a simulation model to decide upon the best risk cut-off in terms of the proportion of patients' time at the clinic that is value-added. However this approach could be used for any operational performance measure or a weighted function of several measures.

Chapter 6

Incorporating GP referral data in breast diagnostic clinic management: discrete-event simulation

In Chapters 4 and 5 we established that GP referral information provided to the Whittington breast clinic is complete and accurate enough to effectively predict the risk of patients having abnormal results in our sample. As explained in Section 4.2.2, we say a patient has an abnormal result if an abnormality was detected from mammogram, ultrasound or biopsy.

In this chapter we describe our novel way of combining Poisson regression, logistic regression scorecards and discrete-event simulation (DES). We test whether using the referral information provided by GPs to plan diagnostic tests in specialist breast clinics could improve operational performance.

In particular we investigate the operational impact of routing patients based on their risk level for different cut-off scores (thresholds between low- and high-risk patients). This is unlike the usual approach of choosing a cut-off score based on predictive performance measures alone. We consider both the simple and full scorecards in turn, for a range of cut-off scores. Our main aim is to optimise the proportion of patients' time at the clinic that is value-added. I developed discrete-event simulation (DES) models to test different scenarios in a virtual environment. This technique was chosen over optimisation for this research because it can easily capture the many sources of variability present in the diagnostic clinic, such as patient punctuality, service times, test requirements and demand. Outpatient clinics are commonly modelled using DES (Hulshof et al., 2012) and as discussed in Section 3.4, a small number of researchers have previously combined classification models and DES. However, our unique approach is to combine these methods for the purpose of recommending cut-off scores in scorecards.

A technical challenge is presented by the full scorecard, since our dataset contains only a small subset of the possible combinations of scorecard attributes (and hence scores) and results. Our review of simulation input modelling techniques in Section 3.3 did not identify a suitable method for this situation. In particular using empirical proportions would not fully capture the likely distribution of scores and results present in the population of patients attending the breast diagnostic clinic as a whole. This leads us to use a more sophisticated technique to estimate the distribution, namely Poisson regression, as explained in Section 6.3.3. This is a new approach to generating characteristics for simulation, and has potential for wide-ranging use as discussed in Chapter 7.1.3.

This chapter proceeds as follows. First in Section 6.1 we introduce our two DES models: the core and extended simulation. It is the extended simulation that we use for testing different cut-off scores.

Next in Section 6.2 we provide an explanation of our chosen operational performance measures. Our close work with staff at the Whittington breast diagnostic clinic helped with choosing appropriate performance measures; a clinician proposed the main measure we consider.

Then in Section 6.3 we describe how patient label values are assigned in the simulations. In particular, at the start of the extended simulation, a set of initial patient labels is assigned to each new patient as explained in Section 6.3.1. One of these labels is the patient's predicted result according to a scorecard. We elaborate on how joint empirical distributions for these labels are assigned when testing the simple scorecard in Section 6.3.2 and how Poisson regression modelling is used instead when testing the full scorecard in Section 6.3.3.

Section 6.4 describes further simulation inputs and assumptions. Next Section 6.5 explains steps to verify and validate the models and Section 6.6 describes sensitivity analysis.

Results for a range of scenarios are presented in Section 6.7, notably we experiment with a range of cut-off scores to evaluate the impact on operational performance.

We conclude the chapter with Section 6.8.

6.1 Overview of the core and extended simulation models

Two simulation models were developed in Simul8 and the difference between them is as follows. The *core simulation* represents the current clinic set-up, so is derived from the process map of the breast diagnostic clinic (see Figure 6.1). A screen-shot of the core simulation model is given in Figure 6.2.

The *extended simulation*, on the other hand, models patients being routed according to their risk of having an abnormal result and the tests required. A process map for this set-up is shown in Figure 6.3 and a screenshot of the extended simulation is shown in Figure 6.4. We use the extended simulation to test the impact of different cut-off scores on operational performance; our novel method for choosing the best cut-off score for each scorecard. Since patient characteristics, including their predicted result according to the full or simple scorecard, affect their routes through the simulation, we need to generate sets of patient characteristics. We introduce our new approach using Poisson regression for situations where not all possible combinations of characteristics are present in the data sample in Section 6.3.3.

According to Gunal (2012) there is a balance to be struck between the level of detail and level of generality in a model, which is decided based on the modelling objectives. Since we are using the simulations to model changes at the Whittington breast diagnostic clinic in particular, enough detail needs to be included to resemble this clinic. This is important so that the model is trusted by stakeholders. Some simplifying assumptions are made due to lack of data, but the effects of these assumptions are tested using sensitivity analysis (see Section 6.6).

6.2 Key operational measures

Here we describe the key operational measures we use to compare scenarios. Our main operational measure is the *clinic efficiency*, which is the average proportion of value-added time per patient on a particular day. When deciding which of several scenarios achieving the same mean clinic efficiency is best, we use the mean *average total time* per patient as a tie-breaker. This means that systems are preferred where on an average day, patients spend less time on average at the diagnostic clinic. Additionally, for each of the final scenarios we provide results for the mean *clinic end time*. This is the time that the last patient finishes their results consultation on average. The purpose of providing this metric is to enable decision makers to assess the feasibility of recommended scenarios in terms of costs. We do not calculate costs in this research because we were not able to obtain the required data.

The reasons for choosing clinic efficiency as our main operational measure are as follows. The idea came from a clinician at the Whittington breast clinic who proposed

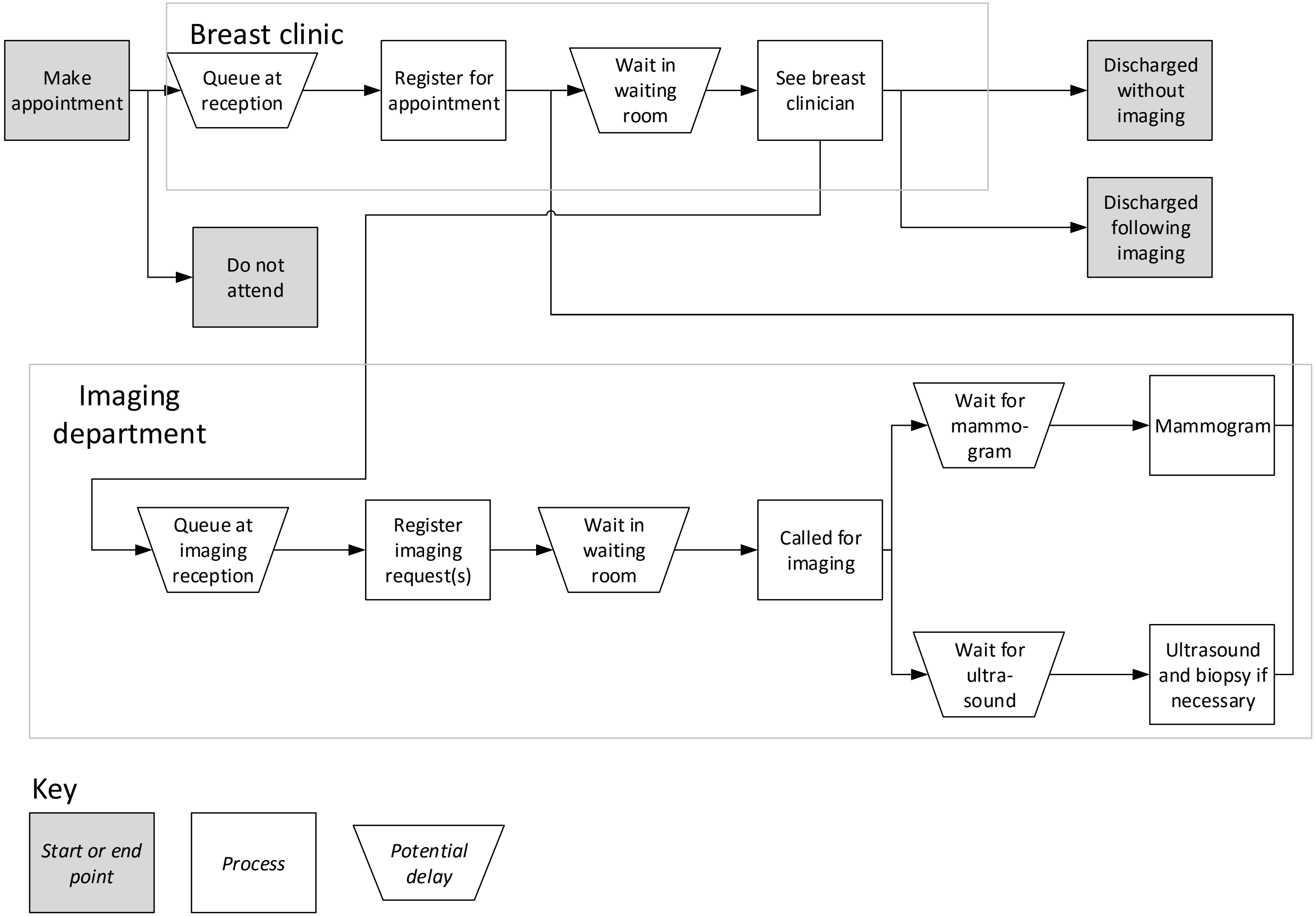


FIGURE 6.1: Process map of breast diagnostic clinic

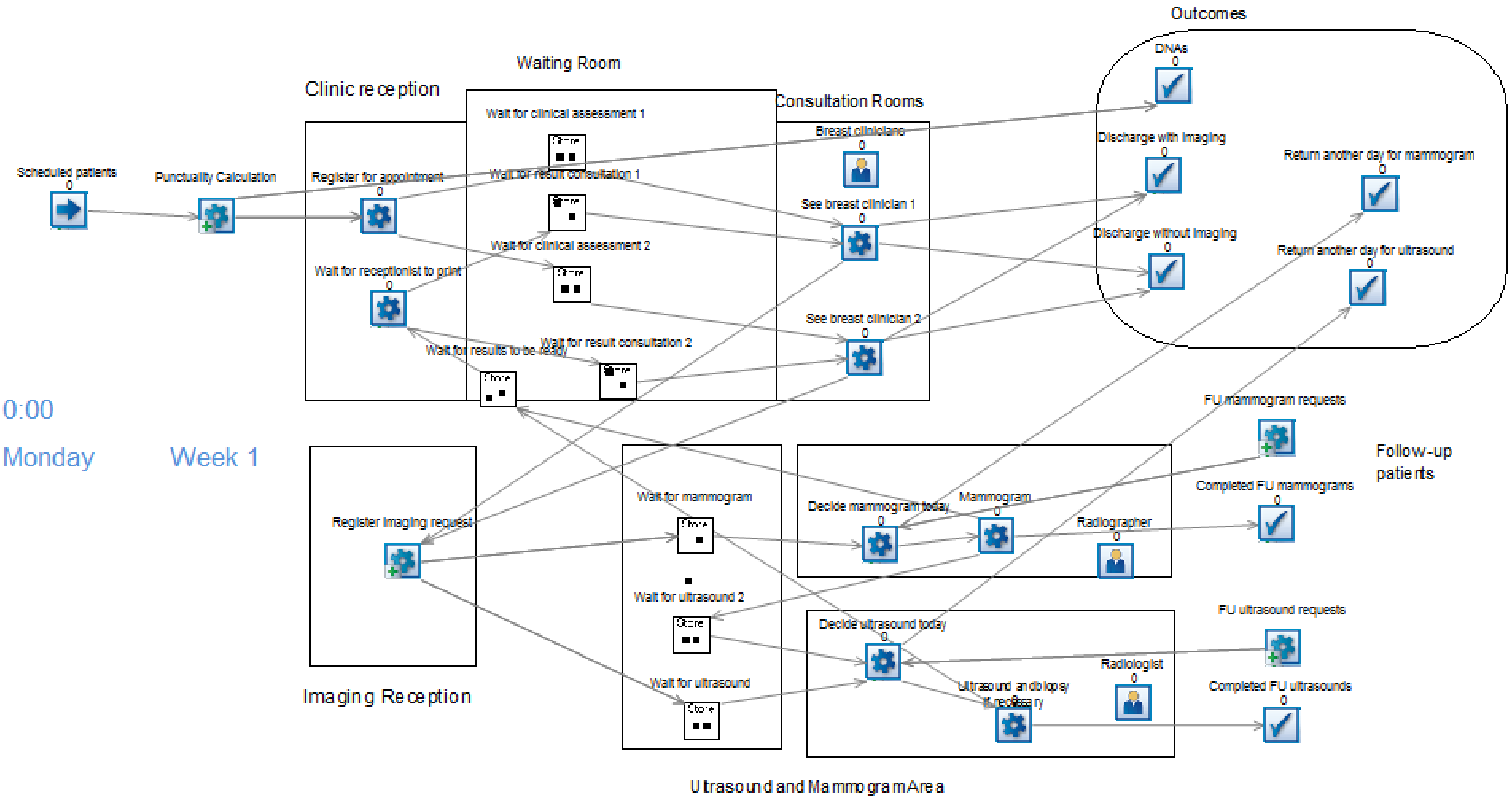


Figure 6.2: Screenshot of the core DES

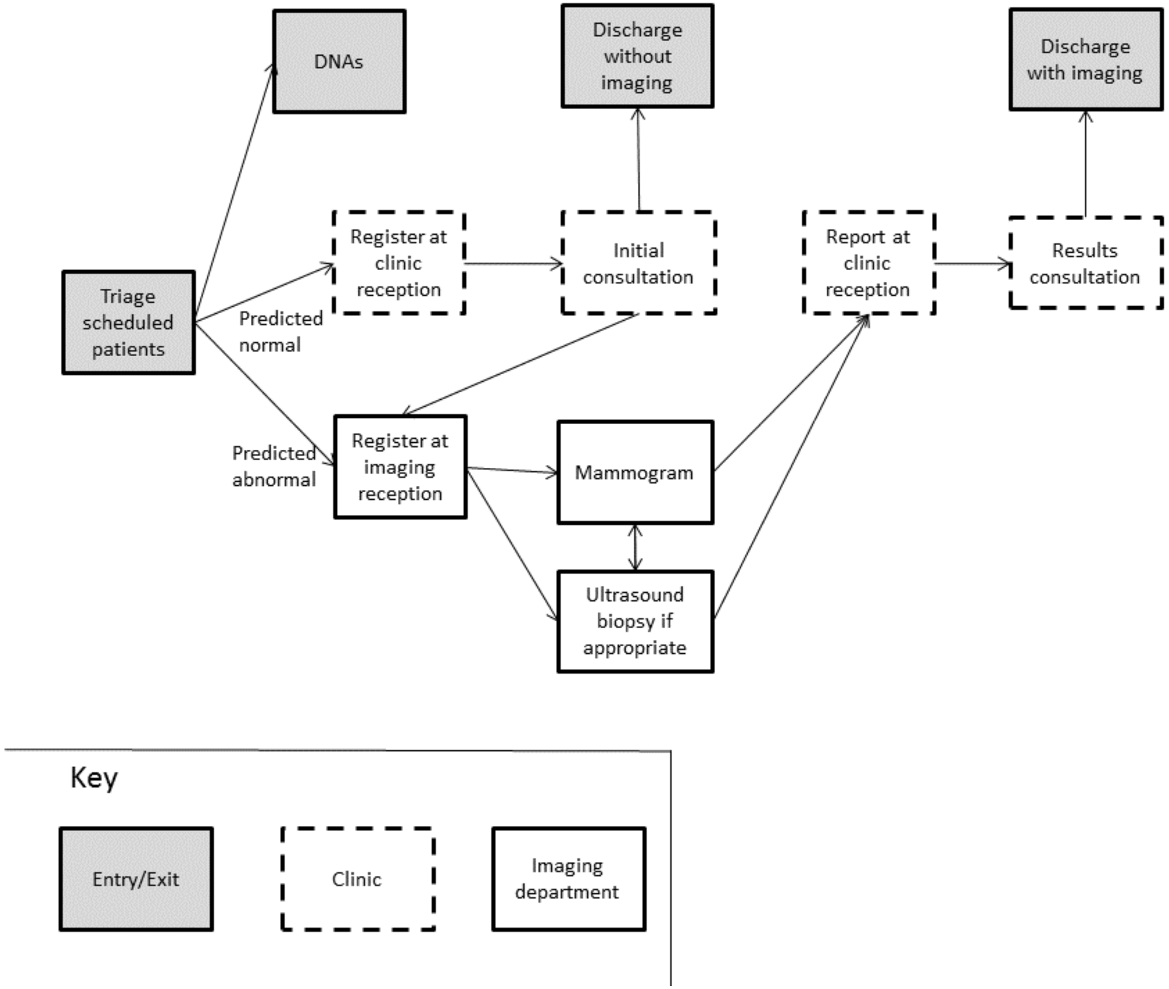

DNAs
Discharge without imaging
Discharge with imaging
Triage scheduled patients
Register at clinic reception
Initial consultation
Report at clinic reception
Results consultation
Predicted normal
Predicted abnormal
Register at imaging reception
Mammogram
Ultrasound biopsy if appropriate
Key
Entry/Exit
Clinic
Imaging department

Figure 6.3: Process map of breast diagnostic clinic where a scorecard is used to triage patients

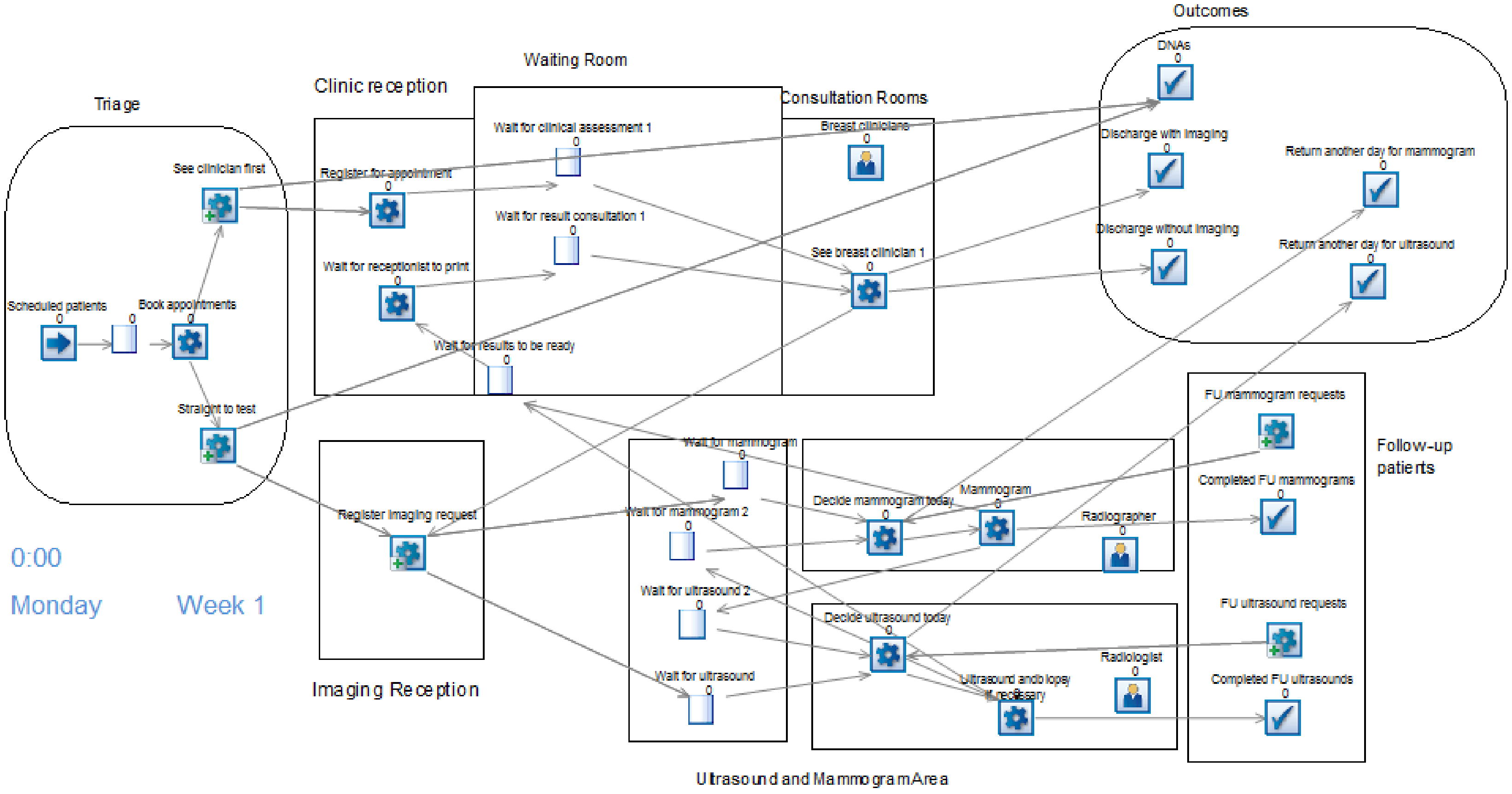


FIGURE 6.4: Screenshot of the extended DES

using a measure of how much of a patient's time at the clinic contributes to their care. Since our research is motivated by the worry and anxiety patients are likely to face when in a cancer clinic waiting for a diagnosis, the focus is on patient in-clinic waiting times. However, since the number of patients following each route through the clinic varies between scenarios, it does not make sense to sum (potentially weighted) waiting times as is done in some other multi-stage applications (Saremi et al., 2015; Klassen and Yoogalingam, 2018) that were reviewed in Section 3.5. This would favour scenarios where patients visit fewer stages, without rewarding the fact that patients visiting more stages receive more care. Instead we use the measure suggested by the clinician. We call this the *efficiency* for a particular patient, and the *clinic efficiency* is defined as the average efficiency across patients on a particular day. These are not new concepts; for example, in lean manufacturing, distinctions are made between value-added and non-value-added time in process flows (Slack et al., 2013). The clinic efficiency measure meets our purpose well, since it takes into account both the amount of value-added time, i.e. time spent in imaging tests or with a clinician, and the amount of non-value-added time, i.e. time spent waiting, queueing and undertaking administrative tasks. Therefore the mean clinic efficiency is the measure that we choose to maximise over when comparing different cut-off scores. Another approach, commonly used in appointment scheduling, would be to optimise a weighted function of waiting times and overtime or idle time (Gupta and Denton, 2008), but this would involve agreeing on suitable weights, for example using the Delphi approach. This is out of scope of our study but would be a possible extension to this research. After agreeing the weights, one could straightforwardly calculate the weighted function for each possible cut-off score.

To define clinic efficiency and total time mathematically, first we define a patient's *start time* as the time at which the clinic's performance for that patient begins to be measured. For a patient arriving on time, the *start time* is the scheduled appointment time, which is the same as the registration time. For a late patient, the *start time* is the registration time, since the delay between the scheduled time and the registration time is caused by the patient not the clinic. For early patients there are two possibilities. If an early patient is seen early, their *start time* is the actual appointment time. If an early patient is seen late, the *start time* is the scheduled appointment time. This corresponds to how waiting times for unpunctual patients are dealt with in the literature (see for example Santibáñez et al. (2009)). This allows us to define a patient's *total time* as the period from the *start time* until the *end time*, when the patient leaves the clinic.

$$\text{Total time} = \text{End time} \; - \; \text{Start time} \tag{6.1}$$

For each day (run) of the simulation, we calculate the *average total time*. We report the mean *average total time* across runs along with the 95% confidence interval. The mean *average total time* is the tie-breaker when clinic efficiency is equal for several scenarios.

We define the *value-added time* as the time during which a patient is in a consultation or having tests done.

$$\begin{split}\text{Value-added time} = \text{Time in consultations } + \text{ time in mammogram room} \\ + \text{ time in ultrasound room}\end{split} \tag{6.2}$$

Hence *efficiency* is the proportion of time at the clinic during which a patient is in a consultation or having tests done.

$$\text{Efficiency} = \frac{\text{Value-added time}}{\text{Total time}} \tag{6.3}$$

The overall *clinic efficiency* is the average *efficiency* over all patients, so can be used as a performance measure on a particular day.

$$\text{Clinic efficiency} = \frac{\sum_{patients} \text{Efficiency}}{\text{Number patients}} \tag{6.4}$$

We report the mean *clinic efficiency* across all runs (days) of the simulation, along with 95% confidence intervals.

6.3 Assigning patient label values

A range of labels are attached to patients in the simulations, for example to determine patient pathways, as well as tracking waiting and service times. Details of all the labels for new patients assigned in the core simulation are included in Table D.3, and those for follow-up patients assigned in both simulations are listed in Table D.4. The extended simulation contains some extra new patient labels (see Table D.5) and the patient labels from the core simulation are assigned in some additional places due to the changes in how patients are routed (see Table D.6).

Several label values are assigned probabilistically, for example, the imaging tests that a patient has. This is achieved by using the inversion method as described by Devroye (1986). Briefly, since the probabilities of mutually exclusive label values add to 1, these values can be mapped to the (0,1) interval. When assigning such a label value to a patient, a random number is sampled from the standard uniform distribution, and then the corresponding label value is assigned. An illustrative example is provided in Figure 6.5.

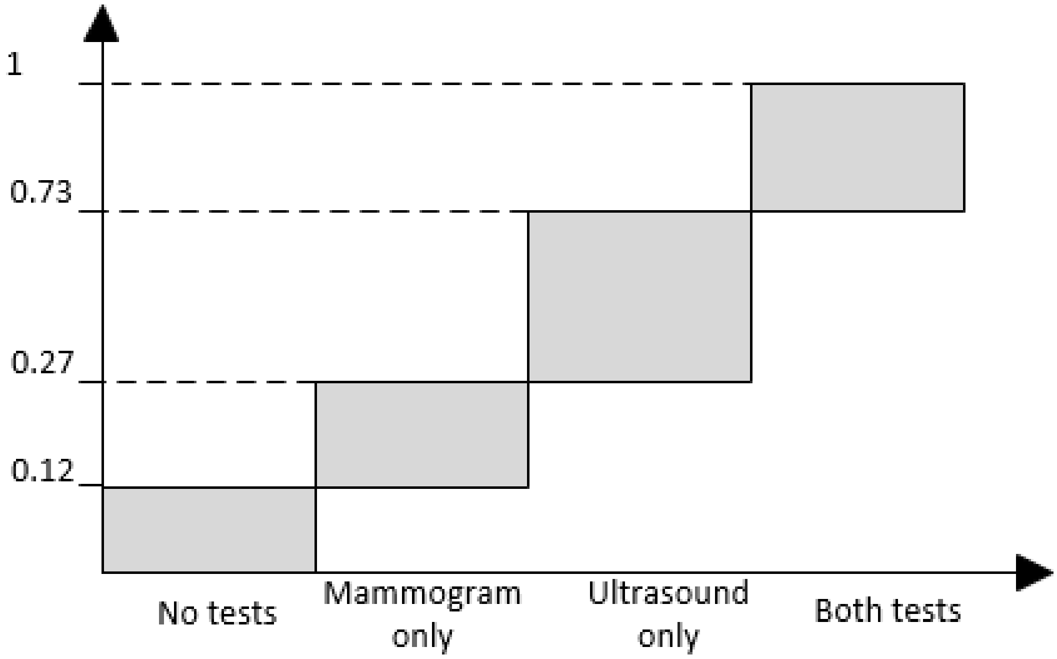


FIGURE 6.5: Illustrative example of the inversion method for sampling

6.3.1 Extra patient labels in the extended simulation

At the start of the extended simulation, a set of initial patient labels is assigned to each new patient. These patient labels are *age group*, *predicted result* and *actual result*. These all influence progress through the simulation of breast diagnostic services. In particular, our novel approach of testing the operational impact of alternative cut-off scores involves changing the proportions of patients with abnormal and normal predicted results between scenarios.

The *age group* label takes one of two values: below 35 years old or at least 35 years old. This label affects which imaging tests are required; according to our data the probability of needing an ultrasound, mammogram, both tests, or neither, depends partly on this. More details are provided in Section 6.4.7.

The *predicted result* (either normal or abnormal) is calculated from a scorecard. When using the simple scorecard, only two referral characteristics are needed to calculate the *predicted result*, but for the full scorecard, seven referral characteristics are needed. In the simulation, patients are routed to imaging first, if an abnormal result is predicted, or to a clinician first, if a normal result is predicted.

The *actual result* label indicates whether a patient's tests show a normal or abnormal result. Hence in reality this label value is only known to patients and staff after tests are done. However it is assigned at the start of the simulation with the other labels since it affects patients' progress through the simulation in the following ways. In order to be consistent with reality, only patients with actual abnormal results may have a biopsy or be diagnosed with cancer in the simulation, and only patients with actual normal results may be discharged without tests.

Since the *predicted result* depends on the cut-off score, and we want to test different cut-offs, we need to know how patients are likely to be distributed between combinations of risk scores, *actual result* and *age group*. For the simple scorecard, each possible combination of these is present in our dataset. Therefore we assume that the empirical joint distribution is a reasonable representation of the true distribution in the population of patients visiting the clinic (see Section 6.3.2). However, for the full scorecard, given the large number of different scores, only a small subset of the possible combinations are present in our data sample. Thus using empirical distributions in this case would not fully capture the likely underlying distribution. Our novel approach for generating characteristics in this case involves fitting a Poisson regression model and is described in Section 6.3.3.

6.3.2 Simple scorecard: Empirical distributions of patient label combinations

In order to test the impact of using the simple scorecard to predict abnormal results and route patients accordingly, I generated the patient labels (*age group*, *actual result* and *predicted result*) in the following way. For the 179 patients in our sample, I calculated the *age group* and *actual result* labels. Then from the simple scorecard referral characteristics (*lump* and *age*), I calculated each patient's risk score.

The following scores are possible: 5, 10, 12, 14, 16, 21, 23, 25. There are seven ways of dividing the scores into two groups at higher and lower risk, for example using the cut-off scores 24, 22, 17, 15, 13, 11 and 6. It is also possible to predict all patients will have a normal result (for example with cut-off 26), or that all patients will have an abnormal result (for example with cut-off 4).

I aggregated the data to find the proportions of our sample with each combination of label values (*age group*, *actual result* and *predicted result*). Since the *predicted result* depends on the cut-off score, for each cut-off I calculated the empirical proportions with each label combination. In the simulation, combinations of patient labels are sampled from the empirical joint distributions, given in Table 6.1.

Table 6.1: Empirical joint distributions of age divide, actual result and predicted result for different cut-offs with the simple scorecard

			Cut-off score								
Actual result	Predicted result	Age divide	4	6	11	13	15	17	22	24	26
Abnormal	Abnormal	Age<35	20%	19%	19%	17%	17%	11%	11%	0%	0%
Abnormal	Normal	Age<35	0%	1%	1%	3%	3%	8%	8%	20%	20%
Normal	Abnormal	Age<35	25%	18%	18%	11%	11%	6%	6%	0%	0%
Normal	Normal	Age<35	0%	7%	7%	14%	14%	20%	20%	25%	25%
Abnormal	Abnormal	Age≥35	26%	25%	21%	21%	16%	14%	6%	6%	0%
Abnormal	Normal	Age≥35	0%	1%	4%	4%	10%	12%	20%	20%	26%
Normal	Abnormal	Age≥35	30%	26%	16%	16%	9%	6%	3%	3%	0%
Normal	Normal	Age≥35	0%	4%	14%	14%	20%	23%	27%	27%	30%
Total predicted abnormal (*imaging first*)			100%	88%	74%	65%	53%	37%	26%	9%	0%

Table 6.2: Factors used in Poisson loglinear models

Factor	Shorthand	Levels	Number of levels
Age	A	Age<29, 29<=Age<35, 35<=Age<42, 42<=Age<47, 47<=Age<52, 52<=Age	6
Family History	F	No or NR, Y	2
Lump	L	No or NR, Y	2
Unilateral Pain	P	No or NR, Y	2
Other Symptom	O	No or NR, Y	2
Duration of symptoms	D	Less than 2 weeks, 2 weeks-2 months, 2-5 months, Over 5 months, NA or NR	5
Urgency	U	Suspected cancer, Symptomatic, Other or NR	3
Result	R	Abnormal, normal	2

6.3.3 Full scorecard: Poisson loglinear model for distributions of patient label combinations

For testing the impact of using the full scorecard, we also need a joint distribution of the *age group*, *actual result* and *predicted result* for different cut-off scores. In this case, the risk score is calculated from the referral characteristics *age*, *family history of cancer*, *lump*, *unilateral pain*, *other symptom*, *duration of symptoms* and *urgency*. By amending the *age* attributes used in the scorecard, they can also be used to generate the *age group* label value. Thus we need the joint distribution of the seven (amended) scorecard referral characteristics plus the *actual result*.

Table 6.2 shows abbreviations and levels for each of these eight factors. There are 2880 possible combinations of levels, with only 166 unique combinations present in our data sample. Since our dataset contains only a small subset of the possible combinations, using empirical proportions would not fully capture the likely joint distribution present in the population of patients attending the breast diagnostic clinic.

Therefore we need a method that can generate combinations of categorical, potentially dependent variable values, including combinations that are not present in our sample. When reviewing the literature in Section 3.3 we did not find a suitable existing method. We address this research gap by introducing a new approach for this situation.

Our new approach consists of firstly fitting a Poisson regression to generate the probabilities of each combination of variables. The Poisson loglinear model is a generalized linear model, which can be used to model count data, as detailed by Agresti (2013). A log function links the expected count to a linear combination of explanatory variables. In our case we want to predict counts in a contingency table with eight factors.

After identifying a Poisson regression model that fits well, the probabilities of each factor combination are generated by dividing the expected counts of each combination by the sample size. As for the simple scorecard, the probabilities of each factor combination are aggregated to obtain the probabilities of each label combination. This aggregation is repeated for each cut-off score.

Illustrative example of Poisson loglinear model for predicting contingency table counts

We introduce some notation to represent Poisson loglinear models for ease of exposition. A purely illustrative two-way contingency table for the factors *lump* (L) and *urgency* (U) is shown in Table 6.3. Here μ_{ij} is the expected count for the cell in row i and column j of the contingency table. For instance, μ_{00} is the expected number of patients with no lump who are symptomatic. The Poisson loglinear model that includes row effects, column effects and the row-column interaction can be summarised as (L,U,LU) and is written in full as shown in Equation (6.5).

Table 6.3: Illustrative two-way contingency table

	Column j	0	1	2
Row i	Urgency Lump	Symptomatic	Suspected Cancer	Other
0	No lump	μ_{00}	μ_{01}	μ_{02}
1	Lump present	μ_{10}	μ_{11}	μ_{12}

$$\ln(\mu_{ij}) = \lambda + \lambda_{11}X_{11} + \lambda_{21}X_{21} + \lambda_{22}X_{22} + \lambda_{311}X_{11}X_{21} + \lambda_{312}X_{11}X_{22} \tag{6.5}$$

In the above equation, λ is the offset parameter and dummy variables are used to code the factor levels. Since *lump* has two levels, *lump present* and *no lump*, this is coded using one dummy variable, X_{11}. *Urgency* has three levels, *symptomatic*, *suspected cancer*, and *other*, so is coded using two dummy variables X_{21} and X_{22}. The dummy variable codings are given in Equations 6.6 to 6.8. The parameter λ_{11} represents the effect of the level *lump present*, compared to *no lump*, on the expected count. Similarly λ_{21} and λ_{22} estimate the effects of *suspected cancer* and *other* urgency compared to the reference level, *symptomatic*. The dependence between *lump* and *urgency* is captured by the row-column interaction effects, λ_{3ij} (for $i = 1$, and $j = 1$ or 2).

$$X_{11} = \begin{cases} 0 & \text{if } i = 0, \\ 1 & \text{if } i = 1. \end{cases} \tag{6.6}$$

$$X_{21} = \begin{cases} 0 & \text{if } j = 0 \text{ or } 2, \\ 1 & \text{if } j = 1. \end{cases} \tag{6.7}$$

$$X_{22} = \begin{cases} 0 & \text{if } j = 0 \text{ or } 1, \\ 1 & \text{if } j = 2. \end{cases} \tag{6.8}$$

Alternative Poisson loglinear models for the same dataset differ in which interaction effects they include, and consequently also how many parameters must be estimated. Using dummy variables means that the number of parameters for each single variable effect is equal to the number of levels minus one. For each two-way interaction included, the number of parameters to estimate is equal to the product of the number of dummy variables for the two factors. Inclusion of higher-order interactions is also possible, but the feasibility of estimating the associated parameters depends on the size of the dataset (Agresti, 2013).

Poisson loglinear models for our contingency table

I fitted two alternative models to predict counts in our 8-way contingency table using the *glm()* command in the *stats* package in R. This command performs maximum likelihood estimation using iteratively reweighted least squares (Quick-R, 2017). The R code is provided in Appendix B. I fitted firstly Model 1 which contained single variable effects only, (A, F, L, P, O, D, U, R), and secondly Model 2 which contained single variable effects and as well as all two-way interactions, (A, F, L, P, O, D, U, R, AF, AL, AP, AO, AD, AU, AR, FL, FP, FO, FD, FU, FR, LP, LO, LD, LU, LR, PO, PD, PU, OD, OU, OR, DU, DR, UR). Fitting Model 1 using dummy variables involved estimating 17 parameters (for single effects) while Model 2 had 120 parameters (for both single effects and interactions).

The Pearson statistic is commonly used to measure the goodness-of-fit of alternative models. It is calculated as

$$\chi^2 = \sum_k \frac{(n_k - \mu_k)^2}{\mu_k} \tag{6.9}$$

where n_k is the observed count in cell k and μ_k is the fitted (expected) count in cell k. The Pearson statistic converges to a chi-squared distribution. However its distribution is not necessarily near chi-squared if there are many empty cells, i.e. missing factor combinations, as in our case. To overcome this, I used small sample statistics as proposed by Hirji (2005). For this, I first calculated the Pearson goodness-of-fit statistic for the sample. Then I simulated Pearson statistics for 3000 samples from the fitted Poisson model distribution. Finally I calculated the simulated p-value, which is the proportion of simulated statistics that were larger than the statistic for the data. The large number of iterations used gives stable results. Another commonly-used goodness-of-fit statistic is the likelihood ratio statistic, also known as the deviance. We cannot use it here though, because there are zero counts for some combinations, since they would evaluate as *not a number* when logs are taken in the deviance formula.

The simulated p-values are 0.39 for Model 1 and 0.59 for Model 2. This means that at the 0.05 level of significance, we do not reject the null hypotheses that the models fit the data. We want a model that provides good estimates of expected counts. Given that both models fit well, Model 2 was chosen because the variables are not considered independent; for example, older women are more likely to have a lump.

The probabilities of each factor combination are generated by dividing the 2880 expected counts by 179 (the sample size). The same aggregation procedure as for the simple scorecard is followed to obtain probabilities of each label combination, but the underlying probabilities are from the Poisson model rather than directly from the data.

Table 6.4 shows the Model 2 joint distribution of label combinations for the full scorecard with some different cut-off scores.

TABLE 6.4: Model 2 joint distribution of age divide, actual result and predicted result for different cut-offs with the full scorecard

			Cut-off					
Actual result	Predicted result	Age divide	220	225	230	235	240	245
Abnormal	Abnormal	Age<35	15%	14%	11%	10%	9%	9%
Abnormal	Normal	Age<35	5%	6%	9%	10%	10%	11%
Normal	Abnormal	Age<35	10%	9%	6%	5%	5%	4%
Normal	Normal	Age<35	15%	16%	19%	20%	20%	21%
Abnormal	Abnormal	Age≥35	20%	19%	16%	12%	11%	8%
Abnormal	Normal	Age≥35	6%	7%	10%	14%	15%	17%
Normal	Abnormal	Age≥35	12%	10%	7%	4%	3%	2%
Normal	Normal	Age≥35	18%	19%	22%	25%	26%	28%
Total predicted abnormal (*imaging first*)			56%	53%	40%	32%	28%	23%

6.4 Other inputs and assumptions

Following on from the explanation of how initial patient labels are assigned in the extended simulation, further simulation inputs and assumptions are described in this section. First we make some general points about how we fitted statistical distributions. Then for each area of the simulation we discuss the core simulation first, and then provide any differences in the extended simulation.

6.4.1 Fitting distributions

Stochastic inputs to the simulation models include demand, punctuality, service times, report times and imaging registration times. The collection and descriptive analysis of underlying data samples for each of these is discussed in Chapter 4. In the simulation models these quantities are represented by statistical distributions.

When empirical distributions are used, only events that happened in the data can occur in the simulation. Since it is more realistic that different events can occur (for example a service time that is shorter, in between, or larger than service times in the data), Santibáñez et al. (2009) recommend fitting theoretical distributions where possible.

Therefore for most stochastic inputs (unless otherwise stated) I compared the fit of different theoretical distributions using the Auto-Fit feature in the Stat::Fit plug-in for Simul8 Professional. This estimates distribution parameters using maximum likelihood. Distributions are given an overall rank based on Kolmogorov-Smirnov and Anderson-Darling goodness of fit tests, as well as on an empirical test for over-fitting the tails of the distributions (Geer Mountain Software Corporation, 2016). It is unclear from the documentation of this software which of these elements are given more priority. I assessed the appropriateness of distributions suggested by Stat::Fit, by considering the characteristics of the quantities being described and distributions used for similar quantities in the literature.

The statistical distributions used in the simulations are summarised in tables. Table 6.5 shows details of the distributions for new and follow-up patient demand. The service time distributions for initial consultations, mammograms, ultrasounds and results consultations are summarised in Table 6.6, while distributions for punctuality, report times and imaging registration times are in Table 6.7.

TABLE 6.5: Demand distributions

Description	Distribution	Parameters: Mean [Range]	Data source	Reason for choice of distribution
New patient appointments per day	Empirical (separate for Mondays, Tuesdays and Wednesdays)	25 [22,27] on Monday, 26 [22,31] on Tuesday, 23 [16,27] on Wednesday	Medway appointment system	Number of appointments not completely random
Follow-up patient ultrasounds per day	Empirical	4 [1,7]	RIS	Multi-modal distribution, so pragmatic decision to use empirical distribution
Follow-up patient mammograms per day	Empirical	5 [1,10]	RIS	As above

Since it is not feasible to have a negative service time, a lower bound of 0 minutes is fixed for all service time distributions. As can be seen in Table 6.6, data sources for service time distributions vary. Owing to the artificial peaks at 5 minute intervals in the questionnaire data for service times (see Section 4.1.1), in most cases I decided to fit distributions to the data I collected. In particular, my data was used where the summary statistics for questionnaire data and my data were similar. For the cases where there were differences, extra data or staff opinions were collected to help with fitting distributions (see Section 4.1.1). As can be seen in the table, questionnaire data were used in addition to my data for the ultrasound and biopsy times. Questionnaire data were also used for the imaging reception queue times (see Table 6.7).

Table 6.6: Service time distributions

Service	Distribution	Parameters (minutes)	Data source	Reason for distribution choice
Initial consultation	Lognormal	mean=12.6, SD=4.34	Observed timings	Best fit according to Stat::Fit
Mammogram	Lognormal	mean=8.94, SD=2.57	Observed timings	Good fit according to Stat::Fit and used in literature
Ultrasound (no biopsy)	Weibull	min=0, α=3.67, β=8.74	Times shorter than 12 minutes from observed timings	Best fit according to Stat::Fit
Ultrasound (with biopsy)	Empirical	mean=22.3, SD=6.5	Times longer than 12 minutes from observed timings and patient questionnaires with outliers removed	Sensitivity analysis showed no significant difference when some alternative distributions were applied
Results consultation (no cancer)	Lognormal	mean=3.33, SD=2.3	Observed timings	Good fit according to Stat::Fit and used in literature
Results consultation (cancer)	Triangular	lower=15, mode=20, upper=30	Assumption provided by clinic nurses	Sensitivity analysis showed no significant difference to results using some alternative distributions

Table 6.7: Other distributions

Description	Distribution	Parameters (minutes)	Data source	Reason for distribution choice
Punctuality	Chi-squared	min=-99, ν=95.1	Medway information system	Best fit according to Stat::Fit
Imaging registration and queue time	Pearson V	α=2.18, β=7.3	Patient question-naires	Best fit according to Stat::Fit. Sensitivity analysis showed no significant difference to results using some alternative distributions
Time for report to be ready following mammogram	Beta	min=0, max=57, p=1.53, q=3.98	RIS	It is realistic that there should be upper and lower bounds, so manually fitted Beta distribution. Showed no significant difference to results using Pearson V distribution suggested by Stat:Fit
Time for report to be ready following ultrasound	Beta	min=0, max=50, p=1.73, q=7.69	RIS	As above

6.4.2 New patient demand for appointments

Core simulation

The possible appointment times to see a clinician are based on schedules. Examples of typical appointment schedules are shown in Table D.2. The number of patients scheduled on each day is sampled from empirical distributions dependent on the day of the week (Monday, Tuesday or Wednesday). I chose to use empirical distributions because the number of appointments is not purely random; it is partly based on demand and partly a supply decision. When allocating appointments in the DES, the most commonly-booked appointment slots are filled first. A fixed no-show rate of 10% is assumed, based on the average daily rate from these data.

Extended simulation

The total number of new patients scheduled on one day and the overall no-show rate, which is independent of patient labels, are the same as in the core simulation. In the extended simulation, patients with normal predicted results are sent to a clinician first, whereas patients with abnormal predicted results are sent to imaging first. Each patient is given an appointment time for when they should arrive at the hospital (clinic or imaging department). Hence two separate appointment schedules are specified: one for patients visiting a clinician first, and one for patients visiting imaging first.

The appointment times for patients visiting a clinician first are based on real schedules but with appointments spread out more to allow space for the extra patients arriving from imaging (see Table D.7, and compare with Table D.2). Specifically the appointment times are 15 minutes apart unless there are many patients sent to clinician first, in which case, patients are booked into slots in the gaps between appointments.

The appointment times for patients visiting imaging first are set 10 minutes apart at the start of the morning then spread out to 20 minutes apart for the rest of the day. If there are not enough appointments, patients are booked into the gaps in the order shown in Table D.8. The rationale for this is that at the start of the morning there will be a delay before the first patient arrives from seeing a clinician.

Further experiments with different appointment times are described in Section 6.7.2.

6.4.3 Follow-up patient demand for imaging

Core simulation

The focus of this research is on diagnostics of new patients attending the clinic, but there are also follow-up patients attending who share ultrasound and mammogram resources with the new patients. For this reason the daily numbers of follow-up patients having ultrasounds and mammograms are model inputs. In particular, the numbers of ultrasounds and mammograms performed on follow-up patients each day are sampled from empirical distributions. Since the data seems to follow multi-modal distributions in both cases, I decided to use empirical distributions for pragmatic reasons. In the absence of evidence otherwise, it is assumed that the follow-up patients have their mammograms and ultrasounds at random times throughout the day. This is achieved by uniformly distributing arrivals over the opening hours of the imaging department.

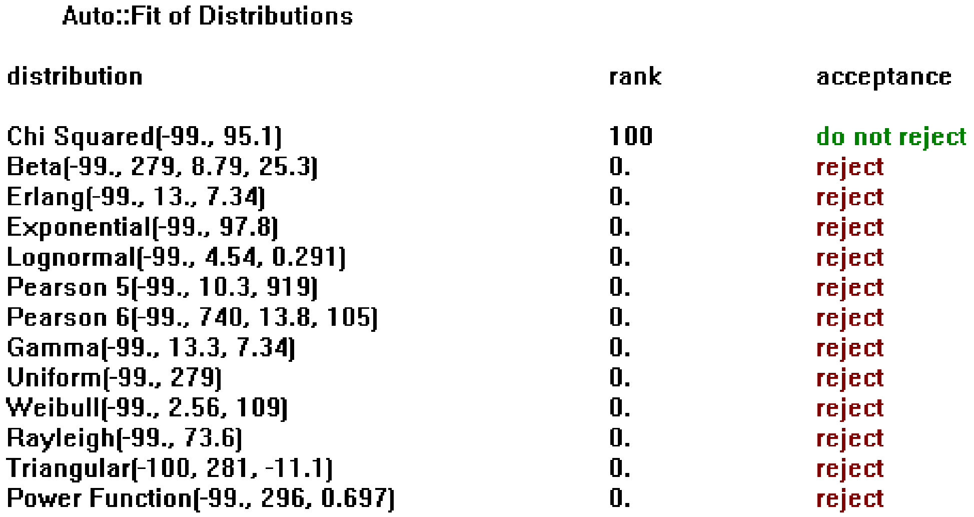
Extended simulation

The same assumptions regarding follow-up patients are made as in the core simulation. In particular, the unavailability of imaging resources for new patients when they are in use by follow-up patients is modelled.

6.4.4 Patient punctuality

Core simulation

When fitting distributions to the punctuality data, the lower bound is set as the earliest time in the data, since there is a limit to how early patients will be for their appointment. Liang et al. (2015) fit normal distributions for the punctuality of patients attending an oncology clinic. However our data are skewed so a normal distribution is not appropriate (see Figure 6.7). On the other hand, Williams et al. (2014) use beta distributions to model patients' punctuality, enabling them to match their observed minimum and maximum, as well as two shape parameters. However in our case, a beta distribution is rejected by the goodness of fit tests and the only distribution accepted is the chi-squared distribution (see Figure 6.6). The beta distribution is unable to capture both the very steep shape in the centre and the long tails (see Figure 6.7). The chi-squared distribution is implemented in Simul8 using a Gamma distribution with a fixed offset.



Auto::Fit of Distributions

distribution	rank	acceptance
Chi Squared(-99., 95.1)	100	do not reject
Beta(-99., 279, 8.79, 25.3)	0.	reject
Erlang(-99., 13., 7.34)	0.	reject
Exponential(-99., 97.8)	0.	reject
Lognormal(-99., 4.54, 0.291)	0.	reject
Pearson 5(-99., 10.3, 919)	0.	reject
Pearson 6(-99., 740, 13.8, 105)	0.	reject
Gamma(-99., 13.3, 7.34)	0.	reject
Uniform(-99., 279)	0.	reject
Weibull(-99., 2.56, 109)	0.	reject
Rayleigh(-99., 73.6)	0.	reject
Triangular(-100, 281, -11.1)	0.	reject
Power Function(-99., 296, 0.697)	0.	reject

FIGURE 6.6: Punctuality distribution: Stat::Fit Output

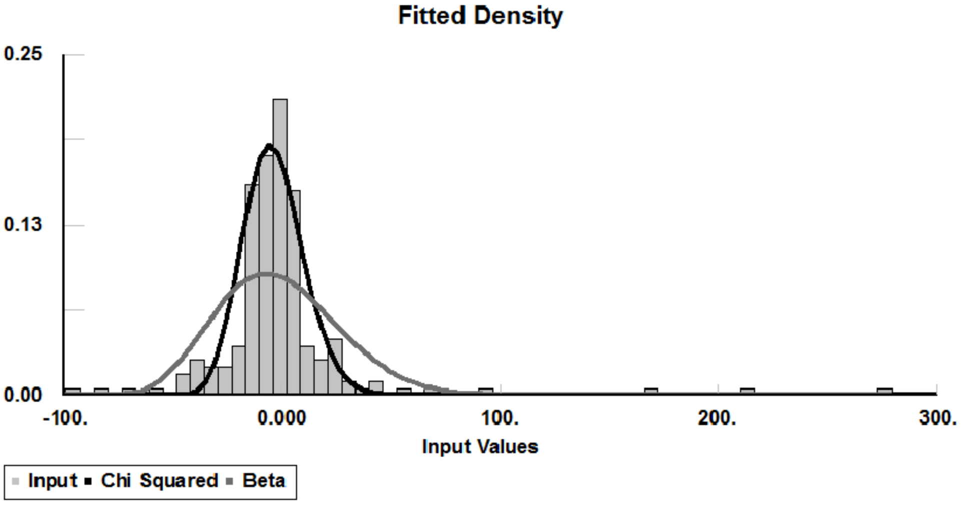


Figure 6.7: Fit of chi-Squared and beta distributions to punctuality data

Extended simulation

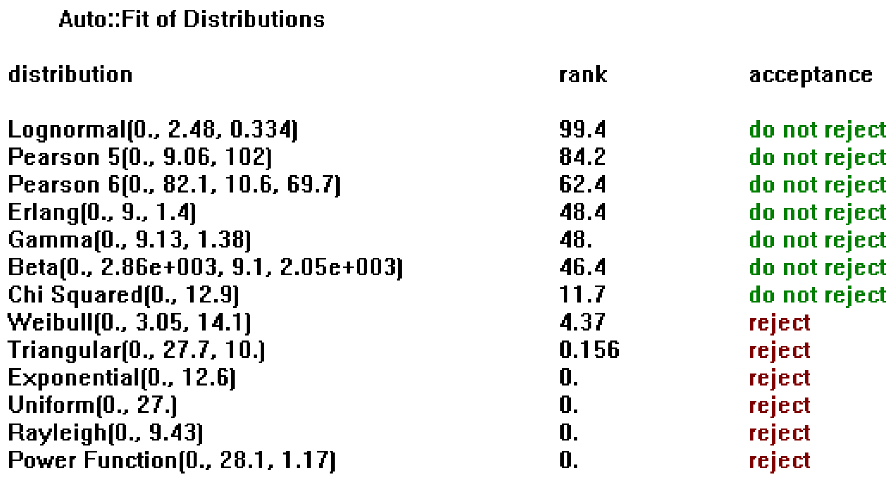
It is assumed that being asked to attend imaging rather than clinic first does not affect patient punctuality. That is, the same punctuality distributions as in the core simulation are assumed for both patients visiting imaging first and those visiting a clinician first.

6.4.5 Initial consultation

Core simulation

It is assumed that patients are called into their initial consultation even if they have not finished their new patient questionnaire. There are separate lists (queues) for each clinician, consisting of both patients waiting for initial consultations and those waiting for results consultations. By observing the nurses and talking to them, it was found that there are no standard rules on how to prioritise which patient is seen next. Thus in the base scenario, it is assumed arbitrarily that patients are seen in order of their waiting time (longest first) and it is possible to be seen before the scheduled appointment time. Other prioritisation behaviours were also tested in the simulation (see Section 6.7). The average turnaround time between patients is 2 minutes, as calculated from data I collected.

Huang et al. (2013) report that authors use lognormal, gamma or exponential distributions to model physician's service times. In our case, lognormal is ranked as the best fit for initial consultation service times, with gamma a close second (see Figure 6.8). Exponential distributions do not fit the data very well since they peak at the lower bound (see Figure 6.9). I therefore chose to use a lognormal distribution.



Auto::Fit of Distributions

distribution	rank	acceptance
Lognormal(0., 2.48, 0.334)	99.4	do not reject
Pearson 5(0., 9.06, 102)	84.2	do not reject
Pearson 6(0., 82.1, 10.6, 69.7)	62.4	do not reject
Erlang(0., 9., 1.4)	48.4	do not reject
Gamma(0., 9.13, 1.38)	48.	do not reject
Beta(0., 2.86e+003, 9.1, 2.05e+003)	46.4	do not reject
Chi Squared(0., 12.9)	11.7	do not reject
Weibull(0., 3.05, 14.1)	4.37	reject
Triangular(0., 27.7, 10.)	0.156	reject
Exponential(0., 12.6)	0.	reject
Uniform(0., 27.)	0.	reject
Rayleigh(0., 9.43)	0.	reject
Power Function(0., 28.1, 1.17)	0.	reject

Figure 6.8: Initial consultation service time distribution: Stat::Fit Output

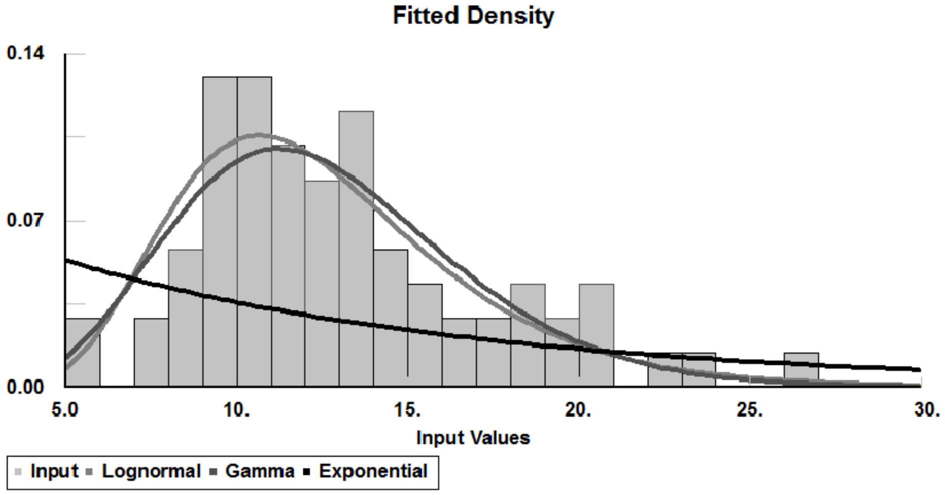


Figure 6.9: Fit of lognormal, gamma and exponential distributions to initial consultation time data

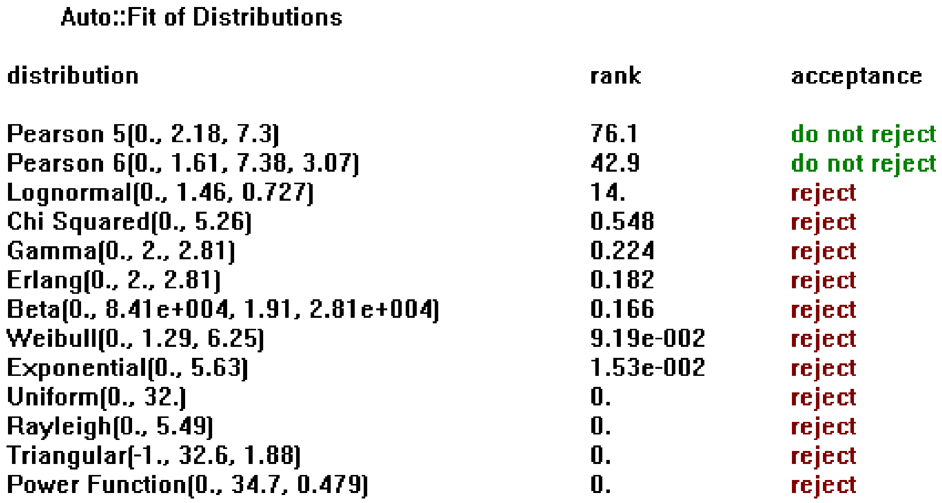
Extended simulation

The same assumptions hold as in the core simulation.

6.4.6 Registering for imaging

Core simulation

It is assumed that it takes 5 minutes to travel between the clinic and the imaging department, as well as 5 minutes to travel from the imaging reception to the mammogram and ultrasound waiting areas. The time from joining the queue at the imaging reception desk until leaving the desk was calculated from the patient questionnaire data. Unfortunately it is not possible to validate these data since many patients from other departments also queue at the same desk and may have different service times to the patients we are interested in. The distribution that best fit these data is a Pearson V distribution (see Figure 6.10). This distribution has a fixed minimum and no upper bound. This makes sense because there is a minimum time that all patients spend at the desk, which is the time to register, and the total time is unbounded since the queue is unbounded. Sensitivity analysis was performed to see the effect of changing the chosen distribution and is discussed in Section 6.6.



Auto::Fit of Distributions

distribution	rank	acceptance
Pearson 5(0., 2.18, 7.3)	76.1	do not reject
Pearson 6(0., 1.61, 7.38, 3.07)	42.9	do not reject
Lognormal(0., 1.46, 0.727)	14.	reject
Chi Squared(0., 5.26)	0.548	reject
Gamma(0., 2., 2.81)	0.224	reject
Erlang(0., 2., 2.81)	0.182	reject
Beta(0., 8.41e+004, 1.91, 2.81e+004)	0.166	reject
Weibull(0., 1.29, 6.25)	9.19e-002	reject
Exponential(0., 5.63)	1.53e-002	reject
Uniform(0., 32.)	0.	reject
Rayleigh(0., 5.49)	0.	reject
Triangular(-1., 32.6, 1.88)	0.	reject
Power Function(0., 34.7, 0.479)	0.	reject

FIGURE 6.10: Queue and imaging registration time distribution: Stat::Fit Output

Extended simulation

The same assumptions regarding registering for imaging as in the core simulation hold for patients who visit a clinician first. For patients who attend imaging first, their arrival time (calculated using their appointment time and punctuality) is when they arrive at the imaging reception desk queue. It is assumed that the duration of time from joining the queue until leaving the desk follows the same distribution as for patients arriving from clinic. That is, we do not model a dedicated reception desk, and we assume it takes the same length of time to register them as those patients who have arrived from clinic, i.e. their details are not entered into the imaging system in advance.

6.4.7 Imaging

Core simulation

The percentage of new patients having imaging tests is based on data and sampled probabilistically (see Table 6.8). The probability of new ultrasound patients having a

TABLE 6.8: Imaging tests probabilities (core simulation)

Imaging tests	Probability
Ultrasound only	0.46
Mammogram only	0.15
Both ultrasound and mammogram	0.27
Neither ultrasound nor mammogram	0.12

biopsy is 27% and of follow-up ultrasound patients having a biopsy is 9% as in our data. Where applicable, it is assumed patients have their mammogram before their ultrasound.

In the core simulation, patients are prioritised for mammograms in order of decreasing waiting times (following registration at the Imaging Reception). Based on observations, it is assumed that there is zero turnaround time between mammograms. Coelli et al. (2007) model mammogram service times with a normal distribution, but this is not appropriate in our case since the data are right skewed (see Figure 4.2). So instead, as for clinician times, lognormal, gamma and exponential distributions were considered. Of these, a lognormal distribution best fits the data for mammogram service times (see Figure 6.12).

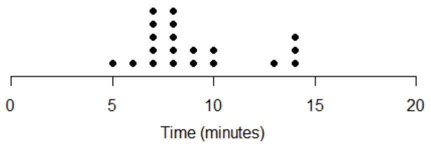
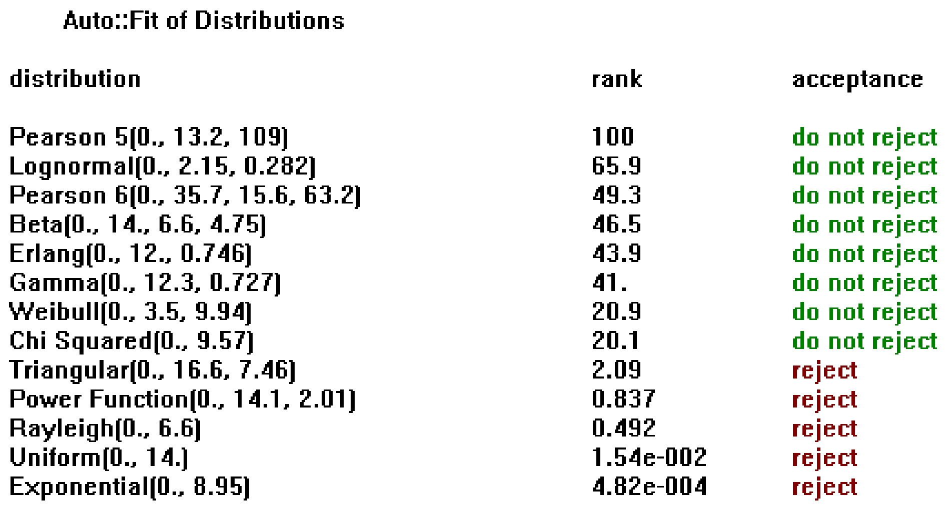


FIGURE 6.11: Empirical mammogram service time distribution (from observed timings)

In the core simulation, patients who have had a mammogram are prioritised for ultrasounds and seen in order of waiting time. There is an average 3.5 minute turnaround time between ultrasounds according to our data, which includes time for the radiologist to report both ultrasounds and mammograms, so this is implemented in the simulation. The ultrasound service time dotplot in Figure 6.13 suggests that the data can be split into long times and short times. The split is less clear in the patient-collected data because of the artificial peak at 15 minutes. Separate distributions are fitted to the short times, assumed to correspond to ultrasounds only, and to the long times, assumed to correspond to ultrasounds and biopsies.

The distribution for ultrasound times (no biopsy) was chosen as follows. I fitted distributions to the ultrasound times shorter than 12 minutes in the sample of observed



```
        Auto::Fit of Distributions

distribution                          rank          acceptance

Pearson 5(0., 13.2, 109)              100           do not reject
Lognormal(0., 2.15, 0.282)            65.9          do not reject
Pearson 6(0., 35.7, 15.6, 63.2)       49.3          do not reject
Beta(0., 14., 6.6, 4.75)              46.5          do not reject
Erlang(0., 12., 0.746)                43.9          do not reject
Gamma(0., 12.3, 0.727)                41.           do not reject
Weibull(0., 3.5, 9.94)                20.9          do not reject
Chi Squared(0., 9.57)                 20.1          do not reject
Triangular(0., 16.6, 7.46)            2.09          reject
Power Function(0., 14.1, 2.01)        0.837         reject
Rayleigh(0., 6.6)                     0.492         reject
Uniform(0., 14.)                      1.54e-002     reject
Exponential(0., 8.95)                 4.82e-004     reject
```

FIGURE 6.12: Mammogram service time distribution: Stat::Fit Output

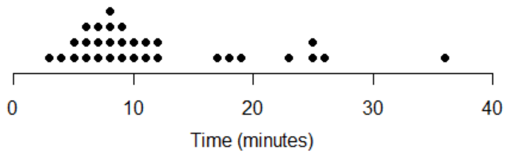
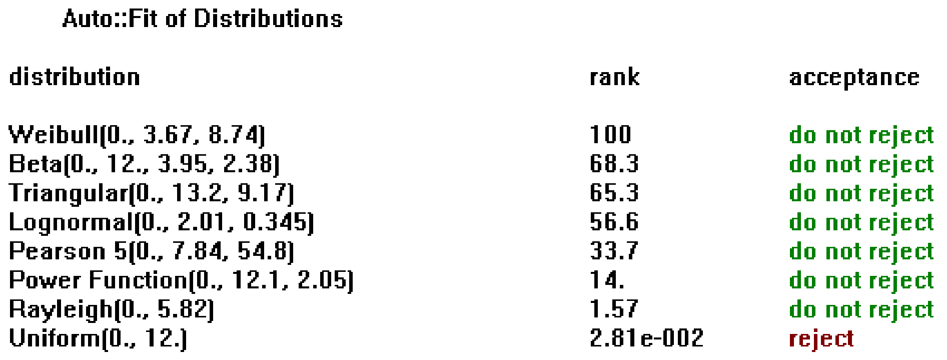


FIGURE 6.13: Empirical ultrasound service time distribution from observed timings

timings. The top-ranking distribution is Weibull, followed by Beta and Triangular (see Figure 6.14). Beta and Triangular distributions both have upper bounds. Although the data are bounded at 12 minutes by construction of the dataset, it seems possible that an ultrasound, even without a biopsy, could take longer than 12 minutes. For this reason the Weibull distribution, which is not bounded above, is chosen.

The service time distribution for ultrasound with a biopsy was chosen as follows. As described in Section 4.1.1, after collecting the initial data sample of ultrasound service times, further times were collected in an attempt to capture more long times. Times longer than 12 minutes from both questionnaires and observed timings (after removing artificial peaks and outliers) were used as an empirical distribution. I also experimented with some alternative distributions as part of sensitivity analysis (see Section 6.6).



Auto::Fit of Distributions

distribution	rank	acceptance
Weibull(0., 3.67, 8.74)	100	do not reject
Beta(0., 12., 3.95, 2.38)	68.3	do not reject
Triangular(0., 13.2, 9.17)	65.3	do not reject
Lognormal(0., 2.01, 0.345)	56.6	do not reject
Pearson 5(0., 7.84, 54.8)	33.7	do not reject
Power Function(0., 12.1, 2.05)	14.	do not reject
Rayleigh(0., 5.82)	1.57	do not reject
Uniform(0., 12.)	2.81e-002	reject

FIGURE 6.14: Ultrasound (no biopsy) service time distribution: Stat::Fit Output

Extended simulation

Assumptions relating to durations of tests, turnaround times and follow-up patients are unchanged from the core simulation. It is assumed that no new patients with actual normal results have a biopsy. Of those new patients with actual abnormal results who have an ultrasound, it is assumed that 44% also have a biopsy (as in the dataset).

According to imaging staff, the order of tests for a particular patient usually depends on their age. It was not possible to model this complexity in the core simulation, where patients did not have *age group* labels. In the extended simulation patients under 35-years-old are directed to have an ultrasound first, followed by a mammogram if needed. On the other hand, it is assumed that patients who are at least 35-years-old and need a mammogram have this first, followed by an ultrasound if needed.

It is assumed that patients who are predicted an abnormal result and so sent to imaging first receive the same imaging tests as those patients with actual abnormal results in our data. The percentage of these patients having ultrasound, mammogram, both and neither are summarised in Table 6.9. In particular, it is assumed that all under 35-year-olds who have an ultrasound have a 6% chance of also having a mammogram, where this decision is made by a radiologist. This agrees with Harvey et al. (2014) who state that ultrasound is commonly the only imaging test done for younger patients, although they consider under 40-year-olds in this group.

For patients with actual abnormal results that are 35-years-old and over, it is assumed that those few for whom a mammogram is not appropriate are identified from their GP referral or from a brief discussion in the mammogram room. Some reasons why an ultrasound and no mammogram were performed on patients aged 35 and over can be seen from GP referral information: two have implants, one had a previous ultrasound but no symptoms so needed another for comparison and one was on hormone replacement therapy (which makes breasts look younger so a mammogram is less useful). There are other conditions where a mammogram is not advised, for example sebaceous cysts, but which are not necessarily recorded on the GP referral information (Harvey et al., 2014). It is assumed that those who do not have an ultrasound are identified by a radiologist based on their mammogram report, and do not enter the ultrasound room.

Table 6.9: Imaging tests for patients with actual abnormal results

	Both	Neither	Mammogram only	Ultrasound only
All ages	34 (42%)	0	3 (4%)	44 (54%)
Age<35	2 (6%)	0	0	33 (94%)
Age≥35	32 (70%)	0	3 (6.5%)	11 (23.5%)

For the patients who are predicted normal results and so visit a clinician first, it is assumed that clinician behaviour in requesting tests remains unchanged from current behaviour. That is, we sample probabilistically using empirical proportions dependent on *age group* and *actual result* (see Tables 6.9 and 6.10). This reflects how the extra knowledge clinicians gain during the initial consultation informs the decision of whether to request tests. By definition, patients who do not have tests have normal results.

Table 6.10: Imaging tests for patients with actual normal results

	Both	Neither	Mammogram only	Ultrasound only
All ages	15 (15%)	21 (21%)	23 (23%)	39 (40%)
Age<35	0	16 (36%)	0	29 (64%)
Age>=35	15 (28%)	5 (9%)	23 (43%)	10 (19%)

Patients are prioritised for tests in the following way. For mammograms, patients who have had an ultrasound are prioritised in order of waiting time. Then other patients are seen in order of their waiting time, regardless of whether they have come from a consultation or straight for imaging. Similarly, for ultrasound, patients who have had a mammogram are prioritised in order of waiting time. Then other patients are seen in order of their waiting time, again regardless of whether they visited a clinician or imaging first.

6.4.8 Waiting for report

Core simulation

A results consultation can only begin once both the patient has returned from the imaging department and the report containing the patient's results has been printed (see the process diagram provided previously in Figure 4.6). If the report is ready when a patient arrives back from clinic, it is assumed that it is printed immediately. However if it is not ready, it is assumed that there is a 5 minute delay from when it is

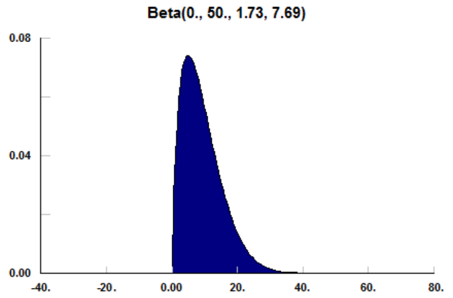


FIGURE 6.15: Time from ultrasound until report ready: Beta distribution

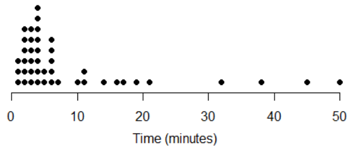


FIGURE 6.16: Time from ultrasound until report ready: Empirical distribution

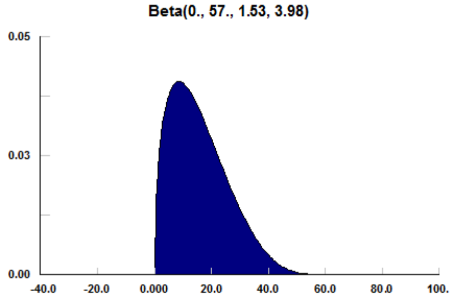


FIGURE 6.17: Time from mammogram until report ready: Beta distribution

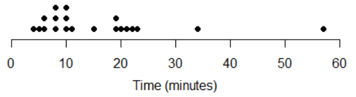


FIGURE 6.18: Time from mammogram until report ready: Empirical distribution

ready until it is printed and the patient can be called to see a clinician. The delay for nurses to collect the print-out is not modelled.

The time for a report to be ready for printing consists of the delay to start reporting (from when a patient's last imaging test finishes) and the time to dictate the report. We have data on the total time for a report to be ready. Since there is a minimum time to dictate a report and it is assumed there is a maximum time by which the radiologist will ensure the report is dictated (although there is no standard rule for this), a beta distribution is appropriate since it has both lower and upper bounds. Since Stat::Fit does not allow specification of an upper bound in its Autofit feature, manual fits were performed instead (see Figure 6.4.8 and Figure 6.4.8). Sensitivity analysis was also performed to see the effect of using Pearson V distributions in the simulation instead (see Section 6.6), since these were suggested as the best fit by Stat::Fit.

Extended simulation

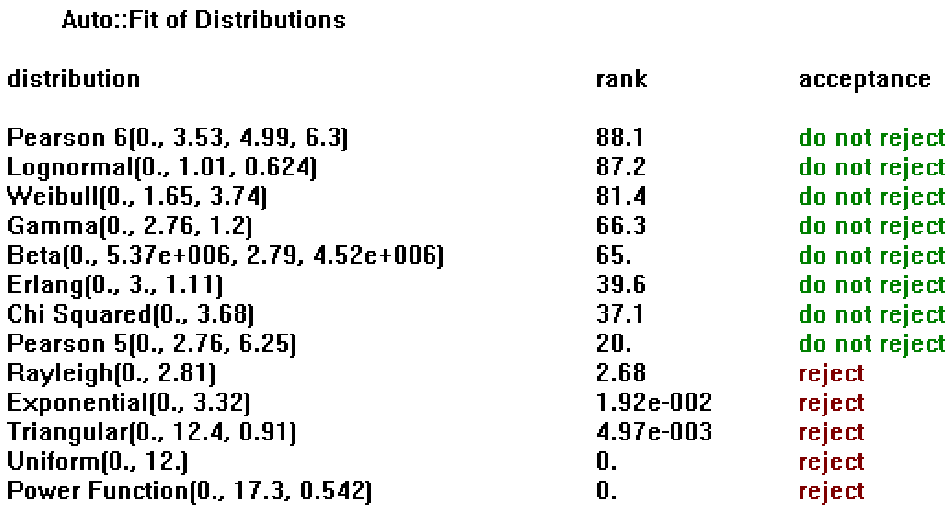
The same assumptions regarding reporting results and printing reports hold as in the core simulation.

6.4.9 Results consultation

Core simulation

As described previously in Section 4.1.1, patients diagnosed with cancer spend a longer time in results consultations than other patients. Based on nurses' opinions and the small number of data points available from patients, the service time distribution for cancer results consultations is assumed to be Triangular taking values between 15 and 30 minutes and a mode of 20 minutes. The probability that a patient is diagnosed with cancer in the simulation is 4%, based on data.

For patients with shorter results consultation times, which we assume are those without cancer, there are sufficient data points in our observed timings sample to fit a service time distribution. All of these times are included since the maximum time is 12 minutes, so shorter than for a cancer results consultation. As for initial consultations, lognormal, gamma and exponential distributions are considered. A lognormal distribution fits the data well (see Figure 6.19) so is used in the simulation.



Auto::Fit of Distributions

distribution	rank	acceptance
Pearson 6(0., 3.53, 4.99, 6.3)	88.1	do not reject
Lognormal(0., 1.01, 0.624)	87.2	do not reject
Weibull(0., 1.65, 3.74)	81.4	do not reject
Gamma(0., 2.76, 1.2)	66.3	do not reject
Beta(0., 5.37e+006, 2.79, 4.52e+006)	65.	do not reject
Erlang(0., 3., 1.11)	39.6	do not reject
Chi Squared(0., 3.68)	37.1	do not reject
Pearson 5(0., 2.76, 6.25)	20.	do not reject
Rayleigh(0., 2.81)	2.68	reject
Exponential(0., 3.32)	1.92e-002	reject
Triangular(0., 12.4, 0.91)	4.97e-003	reject
Uniform(0., 12.)	0.	reject
Power Function(0., 17.3, 0.542)	0.	reject

FIGURE 6.19: Results (no cancer) service time distribution: Stat::Fit Output

Extended simulation

As in the core simulation, the duration of the results consultation varies depending whether patients have been diagnosed with cancer or not. The probability that patients with actual abnormal results are diagnosed with cancer in the simulation is 8.6% (as in our dataset). Patients with actual normal results cannot be diagnosed with cancer. I performed sensitivity analysis to see the effect of longer results consultations for those patients who attend imaging first so only see a clinician for their results.

6.4.10 Staff

Core simulation

The staff included in the core simulation model are listed in Table 6.11 along with their working hours (based on observations). We assume staff will not begin seeing any more patients after their work end time, but will complete the current consultation or test. Other staff working in the clinic, such as nurses and radiography assistants, are out of scope of the simulation model.

Table 6.11: Resource assumptions

		Start time			End time
Location	Staff member	Mon	Tue	Wed	
Clinic 4A	Breast clinician 1	9:45am	9am	9:30am	When last patient has been seen
Clinic 4A	Breast clinician 2	9:45am	9am	N/A	When last patient has been seen
Imaging department	Radiographer	9am	9am	9am	3pm
Imaging department	Radiologist	9:45am	9am	9am	4pm

Extended simulation

The same staff types are included as in the core simulation. For the validation we assume staff working hours are also the same as in the core simulation. However in the sensitivity and scenario analyses the end times for radiographer and radiologist were changed so that they also work until they have seen the last patient.

6.5 Verification and validation

When implementing the models in Simul8, steps were taken to verify that they were working as expected. The models were also validated to check that they represent the real clinic well enough.

6.5.1 Verification

Extra labels were assigned to patients to check they are behaving as expected, for example following the correct paths through the simulations. Detailed results spreadsheets were produced to check the label values and identify problems. To solve problems, the relevant *Visual Logic* code section was run line by line in debugging mode.

6.5.2 Validation

Core simulation

The core simulation was validated through visual checks of face-validity, asking clinicians to check the simulation at various of stages of development and comparing model outputs to historic data. For the latter, the average clinic performance for patients on each day of the week for December 2015 was approximated from patient questionnaires (see Chapter 4 for details about these data). In particular the following measures were considered.

1. Average wait for mammogram - from leaving clinician until entering mammogram room
2. Average wait for ultrasound - from leaving clinician until entering ultrasound room
3. Average wait between mammogram and ultrasound
4. Average wait for results - from returning to clinic waiting room until seeing clinician
5. Average time at clinic - from start appointment

Following the method suggested by Banks et al. (2010b), the simulation model was run many times to obtain a 95% confidence interval for the mean value of each measure. The *trial run calculator* feature of Simul8 was used to find the number of runs required for the 95% confidence limits to be within 10% of the estimate of the mean. This is known as the *precision*. It was assumed that the average of the empirical data is the true mean value for each measure; I checked whether these values lie within an acceptable error of the confidence intervals. The results for Mondays, Tuesdays and Wednesdays are presented in Tables 6.12, 6.13 and 6.14 respectively.

In the results for Mondays, the average waits for ultrasounds both for patient who do and do not have a mammogram first, are shorter in the data than the mean of the simulation output, but sample sizes are small so may not be representative. The average time at clinic is shorter for the sample than the mean of simulation runs because a high proportion of patients in the sample did not have imaging. The other measures are within acceptable limits. Most Tuesday measures are within acceptable limits, aside from the average wait for ultrasound after seeing a clinician which is longer

Table 6.12: Simulation validation - Mondays. a): Sample size. For simulation it is number of days and for data it is number of patients. b): For simulation it is mean across days [95% confidence interval] and for data it is average over patients.

	Average wait for mammogram		Average wait for ultrasound (no mammogram)		Average wait for ultrasound (after mammogram)		Average wait for results		Average time at clinic from start appointment	
Precision	10%		10%		10%		10%		10%	
	a)	b)	a)	b)	a)	b)	a)	b)	a)	b)
Simulation	47	24, [21,26]	101	58, [52,63]	138	53, [48,58]	35	30, [27,33]	15	127, [115,140]
Data	11	26	12	57	8	36	20	30	29	103

Table 6.13: Simulation validation - Tuesdays. a): Sample size. For simulation it is number of days and for data it is number of patients. b): For simulation it is mean across days [95% confidence interval] and for data it is average over patients.

	Average wait for mammogram		Average wait for ultrasound (no mammogram)		Average wait for ultrasound (after mammogram)		Average wait for results		Average time at clinic from start appointment	
Precision	10%		10%		10%		10%		10%	
	a)	b)	a)	b)	a)	b)	a)	b)	a)	b)
Simulation	127	34, [30,37]	127	59, [53,64]	123	36, [32,39]	39	22, [20,24]	22	119, [108,131]
Data	16	29	7	75	12	40	22	26	30	122

Table 6.14: Simulation validation - Wednesdays. a): Sample size. For simulation it is number of days and for data it is number of patients. b): For simulation it is mean across days [95% confidence interval] and for data it is average over patients.

	Average wait for mammogram		Average wait for ultrasound (no mammogram)		Average wait for ultrasound (after mammogram)		Average wait for results		Average time at clinic from start appointment	
Precision	10%		10%		10%		10%		10%	
	a)	b)	a)	b)	a)	b)	a)	b)	a)	b)
Simulation	31	18 [16,20] mins	157	29 [26,32] mins	263	19 [17,21] mins	53	77 [69,84] mins	21	131 [118,144] mins
Data	19	24	10	87	18	32	25	30	21	139

in the data, but the sample size is small. Similarly in the results for Wednesdays, the average wait for ultrasound after seeing a clinician is much longer in the data but the sample size is small. On the other hand, the average wait for results is much shorter in the data. This could be because in the simulation model there is only one clinician working on a Wednesday, but in reality perhaps a second clinician helps with the results consultations. The other Wednesday measures are within acceptable limits.

Extended simulation

In order to validate the extended simulation, its outputs were compared to the core simulation and data samples. As explained in Section 6.1, key differences to the core simulation are routing patients based on their *age group*, *predicted result* and *actual result* including more detailed modelling of which tests patients receive.

In particular two high-level outputs, clinic efficiency and average time at clinic from start appointment, were compared for Mondays, Tuesdays and Wednesdays. For comparability, in the extended model, all patients were sent to the clinician first. The findings are displayed in Table 6.15. The mean clinic efficiencies in the extended model are very similar or the same to those in the core model for all days of the week. For the average time at clinic from start appointment, the extended model matches the data more closely than the core model does.

6.6 Sensitivity analysis

In order to check the robustness of the simulation models, sensitivity analyses were performed, as described in this section. In particular, I tested the effects of changing inputs that were based on assumptions or small sample sizes.

6.6.1 Core simulation

For the core simulation, we compared alternative scenarios against a base scenario which models the current clinic set-up on a Tuesday, using the inputs described previously in Section 6.4. We decided to use Tuesday because the validation work showed this day had the best match between the model and data. A range of reasonable values or distributions for each input of interest were tested and compared to this base scenario.

We find that using alternative realistic distributions in the following areas does not significantly affect the mean clinic efficiency or mean average waiting time measures: cancer results consultation service time (see Table D.9), ultrasound and biopsy service

Table 6.15: Validation of extended simulation compared to core simulation and data samples

			Clinic efficiency	Average time at clinic from start appointment
Mon	Extended simulation	Precision	5%	10%
		Number runs	34	16
		Mean [95% confidence limits]	0.22 [0.21, 0.24]	126 [114,138]
	Core simulation	Number runs	16	15
		Mean [95% confidence limits]	0.23 [0.22, 0.24]	127 [115,140]
	Data	Average	N/A	122
Tue	Extended simulation	Number runs	36	6
		Mean [95% confidence limits]	0.27 [0.26,0.29]	113 [107,118]
	Core simulation	Number runs	36	22
		Mean [95% confidence limits]	0.27 [0.26,0.28]	119 [108,131]
	Data	Average	N/A	103
Wed	Extended simulation	Number runs	51	25
		Mean [95% confidence limits]	0.2, [0.19, 0.21]	138 [125,151]
	Core simulation	Number runs	59	21
		Mean [95% confidence limits]	0.2, [0.19, 0.21]	131 [118,144]
	Data	Average	N/A	139

time (see Table D.10), mammogram report time (see Table D.11), ultrasound report time (see Table D.12) and time queuing and registering at imaging reception (see Table D.13). On the other hand, removing follow-up patients does significantly affect the new patient mean clinic efficiency and mean average waiting time measures for new patients (see Table D.14). This justifies that follow-up patients need to be included in the model. Changing the clinic start time also significantly affects the mean clinic efficiency and the mean average total time (see Table D.15). The effect of clinic starting late on Mondays is investigated as part of the scenario testing (see Section 6.7).

6.6.2 Extended simulation

The base scenario for the extended simulation uses the simple scorecard with a cut-off score of 15. The parameters used in this scenario are summarised in Table 6.16. Sensitivity analyses were performed to test how responsive the mean clinic efficiency and mean average total time are to changes.

Table 6.16: Parameter summary of base scenario used for sensitivity analyses (Extended simulation)

Day of week	Appointment times	Cut-off score	Imaging opening hours	Clinicians	Punctuality of patients attending imaging first	Non-cancer results appointment length
Tuesday	As described in Section 6.4.2, separate appointment lists for patients attending imaging first and seeing a clinician first	15	Open until work finished	2	As for patients seeing a clinician first	Same for all patients

Firstly, the effect of the punctuality of patients attending imaging first was investigated. The base scenario assumes that patients attending imaging first have the same punctuality distribution as those visiting a clinician first. The alternative scenario assumes that patients attending imaging first are punctual. The results are shown in Table D.16. There is little difference in mean clinic efficiency. On an average day, the average patient spends slightly longer at hospital if all patients are punctual.

Secondly, the effect of the length of results consultations for patients attending imaging first was investigated. The base scenario assumes that all non-cancer patients' results consultation lengths follow the same distribution. The alternative scenario assumes that the results consultation lengths for patients attending imaging first who do not have cancer follow the same distribution as initial consultations. The outputs are shown in Table D.17. The mean clinic efficiency is higher in the alternative scenario because patients attending imaging first have more value-added time (the extra time they spend with the clinician). The mean average total time is also longer for this reason. Since the size of the impact will depend on how many patients are sent to imaging first, this sensitivity analysis was repeated after deciding upon the best cut-off score (see Section 6.7.2).

6.7 Scenario testing

6.7.1 Core simulation

Some scenario analyses using the core simulation model were carried out that do not directly contribute to our research objectives but provided insights for clinic staff. I performed three different analyses to find the effects of: prioritisation rules, late-starting clinic and discharging patients with normal results from the imaging department.

Test 1: Prioritisation rules

It was observed that nurses have different behaviours when it comes to prioritising which patient the clinician should see next (see Subsection 6.4). The potential impacts of three different prioritisation rules, which we call A, B and C, on mean average waiting times for consultations were investigated in the core simulation model for Tuesday. The results are shown in Table 6.17. It was found that the choice of prioritisation rule has little effect on the mean average wait for the initial consultation. However, the mean average waiting time for a results consultation is about 12 minutes shorter if results consultations are prioritised (rule B) than if initial consultations are prioritised (rule C). This is because results consultations for non-cancer patients, which form the vast majority of patients, are short compared to initial consultations (on average 3 minutes compared to 13 minutes). The waiting time results for rule A are in between those for the other two rules.

Table 6.17: Test 1: Effect of prioritisation rules on average waiting times to see clinician (Tuesday)

	Average wait for initial consultation		Average wait for results consultation	
Precision	5%		5%	
	Number runs	Mean [95% confidence interval]	Number runs	Mean [95% confidence interval]
A. Base: Prioritise longest waiting	299	11.3 [10.8, 11.9] mins	163	22.2 [21.1,23.3] mins
B. Prioritise results consultations	315	12.4 [11.8,13.0] mins	165	14.7 [14, 15.5] mins
C. Prioritise initial consultations	301	10.9 [10.3,11.4] mins	206	27.2 [25.8, 28.5] mins

Test 2: Late-starting clinic

The multidisciplinary breast team meeting takes place on Monday mornings before the start of clinic. If there are many patients to discuss, the meeting may overrun, meaning that the clinicians and radiologist are late to the clinic. The first scheduled appointment on a Monday is at 9:45am. Table 6.18 shows the mean average total time that patients spend at the clinic, depending on the clinic start time. We investigate three different start times, 9:45am, 10:15am and 10:45am. We find that the later clinic starts, the longer average patients spend at the clinic on an average day, with a delay of 30 minutes in beginning the clinic causing average patients to wait about an extra 20 minutes each on an average day.

Table 6.18: Test 2: Effect of clinic start time on average total time

Clinic start time (Monday)	9:45 AM	10:15 AM	10:45 AM
Number runs for 10% precision	13	13	10
Mean average total time [95% confidence interval] (minutes)	147, [133,161]	168, [152, 184]	191, [172,209]

Test 3: Discharging patients with normal results from the imaging department

Currently all those patients who have imaging tests return to see a clinician for their results. However patient feedback and observations have highlighted that some patients with normal results would prefer to leave the clinic straight after their tests to avoid more time spent at the clinic when they have already been told their result in the imaging department. A what-if analysis was run in the simulation to see the potential impact of discharging patients with normal results directly from the imaging department, compared to the as-is situation.

The core simulation model was run for Tuesday with the probability of patients having normal results based on data: 88% of patients who had only a mammogram had normal results and 40% of patients that had an ultrasound had normal results. According to the model, discharging patients with normal results from imaging would cause the mean average total time for patients with normal results to reduce by about 36 minutes (see Table 6.19). This is because patients with normal results would no longer need to wait for and attend a results consultation. Patients with abnormal imaging results would benefit marginally too, since they would wait on average 4 minutes less for their results consultation. The clinic would end on average 20 minutes earlier on a Tuesday.

Table 6.19: Test 3: Effect of discharging patients with normal results from imaging department (Tuesday)

	Average total time for patients with normal results		Average wait for results consultation		Clinic end time	
Precision	10%		10%		0.10%	
	Number runs	Mean [95% confidence interval]	Number runs	Mean [95% confidence interval]	Number runs	Mean [95% confidence interval]
Discharge patients with normal results from imaging	36	109 [98,120] mins	35	18 [16,20] mins	217	16:07, [16:01, 16:13]
As-is	21	145 [131,160]	31	22 [20, 24] mins	373	16:27 [16:21, 16:33]

This idea has been proposed to staff. A locum radiologist confirmed that this is already done in at least one other breast diagnostic clinic elsewhere. In that clinic, the clinician records that they are happy for particular patients to be discharged directly from imaging, thus still being accountable for the decision. Previously at the Whittington hospital there used to be a radiologist-led breast diagnostic clinic for younger women, where all discharge decisions were made by an experienced radiologist (Leonard, 2014).

6.7.2 Extended simulation

The extended simulation was used to test the potential operational impacts of using the simple and full scorecards to route patients. Unlike the usual method of selecting a cut-off score between low and high risk patients, where only predictive performance of different cut-off scores is considered, our novel method involves choosing the cut-off score based on operational criteria.

For the simple scorecard, the full range of possible cut-off scores (equivalent to sending different proportions of higher-risk patients to imaging first) were tested. Due to the results of those experiments, it was decided to next test resource and appointment changes in conjunction with using the simple scorecard. Finally a subset of cut-off scores for the full scorecard were tested.

Simple scorecard with current resources and appointment times

In this test, the same inputs and assumptions were used as in the base scenario for the sensitivity analyses (see Table 6.16), except that the cut-off scores (in the simple scorecard) were varied. All possible ways of dividing patients into higher and lower risk groups were tested, by using the following cut-off scores: 24, 22, 17, 15, 13, 11 and 6. I also tested the cut-off score 26 which corresponds to sending all patients to a clinician first, as today, as well as the cut-off score 4 which corresponds to sending all patients to imaging first.

The mean clinic efficiency and mean average total time for each cut-off are shown in Table 6.20. Unexpectedly, using the simple scorecard in the simulation (for the current set-up on a Tuesday) does not improve the mean clinic efficiency, our main measure, nor the mean average total time, our tie-breaker. The best results are for cut-off score 26 (as today). This is surprising because in all other scenarios, some patients are sent straight to imaging, meaning that one stage of their visit (the initial consultation) is removed. However, by observing the simulation, the ultrasound test appears to be the bottleneck process where queues build up. This agrees with the reports of long waits for ultrasounds that staff have observed (see Section 2.3.3). Using the simple scorecard to identify patients to send straight to imaging means that some patients have tests who otherwise would not have done. In particular, under-35-year-olds with actual normal results who are sent straight to imaging have an 100% chance of having an ultrasound in the simulation compared to only a 64% chance if they see a clinician first (see Section 6.4.7). Similarly, women who are 35-years-old and above with an actual normal result have a 93.5% chance of having an ultrasound if sent straight to imaging, compared to a 47.5% if they see a clinician first. These extra ultrasound patients would add to the pressure on already overburdened resources. Thus when using the simple scorecard in the current clinic set-up on Tuesdays, the benefit of fewer initial consultations appears to be outweighed by more unnecessary ultrasounds and so longer waits.

Table 6.20: Simple scorecard with current resources and appointment times

Precision	Clinic efficiency 5%		Average total time 10%	
Cut-off score	Number runs	Mean [95% confidence interval]	Number runs	Mean [95% confidence interval]
4: All to imaging first	39	0.17 [0.15,0.19]	56	146 [132,161] mins
6	36	0.18 [0.16,0.12]	33	144 [130, 158] mins
11	11	0.18 [0.16,0.2]	38	150 [136,165] mins
13	22	0.18 [0.16,0.12]	31	159 [144,175] mins
15	20	0.18 [0.17, 0.2]	26	156 [141,171] mins
17	17	0.2 [0.18, 0.21]	18	154 [139, 169] mins
22	20	0.22 [0.19,0.24]	21	147 [133,162] mins
24	16	0.25 [0.22,0.27]	19	133 [120, 146] mins
26: All to clinician first	8	0.27 [0.24,0.3]	15	125 [113,137] mins

Simple scorecard with one clinician and different appointment times

Next, I visually inspected the interaction between initial consultation work-flow and ultrasound work-flow in the simulation. With 2 clinicians working simultaneously and at full utilisation (i.e. assuming that there is always a patient waiting to be seen), patients arrive at the ultrasound area at a faster rate than they can be processed. This prompted us to investigate whether having one clinician working at a time would better match the ultrasound flow rate; a further slow run of the simulation showed that this might be the case.

In order to test this idea more rigorously, I conducted a range of simulation experiments with one clinician working at a time. I tested appointments spaced at 10-, 15- and alternating 10- and 15-minute intervals. For these scenarios there is a shared appointment list for both patients who attend imaging first and those who visit a clinician first. In other respects, the inputs and assumptions are the same as for the base scenario.

The results are shown in Table 6.21. The best mean clinic efficiency is 0.27 and is achieved with appointments spaced 15 minutes apart with cut-off scores 15, 17 and 22. Among these options, the shortest mean average total time is 107 minutes, when the cut-off score is 15. Therefore cut-off 15 appears to be the best option according to the patient-focussed measures. Another benefit of using 15 as the cut-off is that it simplifies use of the scorecard; it is equivalent to sending patients with a lump recorded straight to test, and patients without a lump recorded to a clinician first.

The potential impact of the proposed changes (using the scorecard with cut-off 15, one clinician and 15-minute appointment gaps) on all three clinic days is shown in Table 6.22. Now that the appointment gaps and number of clinicians are the same each day, the difference between days of the week is only in the demand for appointments and the start times. Although the cut-off score was chosen based on Tuesday's model, the simulation shows improved mean clinic efficiency and mean average total times on all three days. Applying the changes would likely cause Wednesday's clinic to finish slightly earlier on average, Tuesday's to finish about the same time but Monday's to finish much later. To combat this, it might be possible to redistribute patients across days of the week, although perhaps Monday appointments are most popular with patients. There are potential cost savings in terms of clinician time, since on Mondays and Tuesdays there are currently two clinicians working. However these cost savings would need to be balanced against the increased costs of longer opening hours of the imaging department on Mondays.

As mentioned in Section 6.6, it is unknown how long patients who have been sent to imaging first would spend seeing the clinician afterwards. In the above analysis it was assumed that these patients spend the same length of time in results consultations as those patients who have already seen the clinician once. If instead the results consultations for patients who attend imaging first are the same length as initial consultations, the potential impact of the changes is as in Table 6.23. Although the cut-off score was not optimised for this situation, there are still improvements in mean clinic efficiency and mean average patient total time for Mondays and Wednesdays; however Tuesday results are worse than for the current situation.

Table 6.21: Simple scorecard with one clinician and different appointment times

			Clinic efficiency		Average total time (mins)
Precision			5%		10%
Appointment times	Cut-off score	Number of runs	Mean [95% confidence interval]	Number of runs	Mean [95% confidence interval]
10-minute slots	26	39	0.16 [0.15,0.16]	16	231 [210,253]
	24	45	0.17 [0.16,0.17]	15	209 [189,230]
	22	38	0.18 [0.17,0.19]	22	175 [157,192]
	17	53	0.19 [0.18,0.2]	31	167 [151,183]
	15	67	0.18 [0.17,0.19]	32	158 [143,174]
	13	97	0.18 [0.17,0.19]	38	158 [142,173]
	11	81	0.16 [0.16,0.17]	40	162 [146,178]
	6	84	0.15 [0.14,0.16]	33	168 [152,185]
	4	102	0.13 [0.13,0.14]	34	172 [155,189]
15-minute slots	26	48	0.24 [0.23,0.25]	18	143 [129,156]
	24	60	0.26 [0.24,0.27]	29	134 [121,147]
	22	59	**0.27 [0.25,0.28]**	35	124 [111,136]
	17	69	**0.27 [0.26,0.29]**	42	116 [105,128]
	15	106	**0.27 [0.26,0.28]**	47	**107 [97,118]**
	13	112	0.26 [0.24,0.27]	46	109 [98,119]
	11	132	0.25 [0.24,0.26]	52	110 [99,121]
	6	150	0.23 [0.22,0.24]	65	116 [104,127]
	4	166	0.2 [0.19,0.22]	61	123 [110,135]
Alternate 10- and 15-minute slots	26	41	0.19 [0.18,0.2]	20	187 [169,204]
	24	53	0.21 [0.2,0.22]	24	170 [153,187]
	22	57	0.22 [0.21,0.24]	26	142 [128,155]
	17	63	0.23 [0.22,0.24]	31	133 [120,146]
	15	113	0.22 [0.21,0.23]	42	136 [123,150]
	13	132	0.22 [0.21,0.23]	46	129 [116,141]
	11	142	0.21 [0.2,0.22]	53	134 [121,147]
	6	128	0.19 [0.18,0.20]	43	137 [123,150]
	4	156	0.17 [0.16,0.17]	52	148 [133,162]

Full scorecard with one clinician and 15-minute gap between appointments

Next we investigate whether by using the full scorecard we could further improve mean clinic efficiency. Experiments for cut-off scores at intervals of 5 were carried out. As before, there is one clinician and 15-minute appointment gaps.

The results are shown in Table 6.24. The highest mean clinic efficiency is 0.28 and is achieved with a cut-off score of 235, which corresponds to sending 32% of patients straight to imaging. Hence using the full scorecard has made negligible improvements to the mean clinic efficiency. The mean average total time for this cut-off score is 113 minutes, which is slightly longer than for the simple scorecard with cut-off 15 (107

Table 6.22: Potential impact of using the simple scorecard with cut-off 15, one clinician and 15-minute appointment slots

	Clinic efficiency		Average total time		Clinic end time	
Precision	5%		10%		0.50%	
Scenario	Runs	Mean [95% confidence interval]	Runs	Mean [95% confidence interval]	Runs	Mean [95% confidence interval]
Monday as today	53	0.21 [0.2, 0.22]	20	165 [148, 181] mins	1001	16:13 [16:08, 16:17]
Monday with changes	104	0.27 [0.26,0.28]	52	116 [105, 127] mins	636	18:06 [18:01, 18:11]
Tuesday as today	30	0.26 [0.28,0.29]	17	127 [114,139] mins	59	17:29 [17:17, 17:41]
Tuesday with changes	106	0.27 [0.26,0.28]	47	107 [97,118] mins	119	17:26 [17:14, 17:39]
Wednesday as today	51	0.2 [0.19, 0.21]	20	174 [156,191] mins	59	17:06 [16:46, 17:25]
Wednesday with changes	101	0.26 [0.25,0.27]	41	115 [104, 126] mins	71	16:47 [16:28, 17:07]

minutes). Since the full scorecard is also much more complicated to use in practice, involving assessing seven attributes per patient rather than two, we do not recommend using it in place of the simple scorecard.

6.8 Conclusion to incorporating GP referral data in breast diagnostic clinic management

In this chapter, the development of two DES models was described: a core and an extended model. To capture the variability of quantities such as service times and patient punctuality, I fitted a range of statistical distributions to data from various sources. For simulation inputs where insufficient suitable data were available, assumptions were made and the sensitivity of results to these inputs was analysed. The simulation models were verified and validated including by comparing average outputs against averages from data samples.

TABLE 6.23: Potential impact of using the simple scorecard with cut-off 15, one clinician and 15-minute appointment slots. Assumes longer results consultations for patients visiting imaging first.

Precision	Clinic efficiency 5%		Average total time 10%		Clinic end time 0.50%	
Scenario	Runs	Mean [95% confidence interval]	Runs	Mean [95% confidence interval]	Runs	Mean [95% confidence interval]
Monday as today	53	0.21 [0.2, 0.22]	20	165 [148, 181] mins	1001	16:13 [16:08, 16:17]
Monday with changes	60	0.27 [0.26, 0.28]	31	135 [122, 148] mins	546	18:25 [18:20, 18:31]
Tuesday as today	30	0.26 [0.28,0.29]	17	127 [114,139] mins	59	17:29 [17:17, 17:41]
Tuesday with changes	73	0.25 [0.24, 0.26]	36	140 [126, 153] mins	125	17:54 [17:42, 18:07]
Wednesday as today	51	0.2 [0.19, 0.21]	20	174 [156,191]mins	59	17:06 [16:46, 17:25]
Wednesday with changes	62	0.25 [0.24, 0.26]	32	141 [127, 155] mins	68	17:10 [16:51, 17:30]

The core simulation model, which represents the current diagnostic clinic set-up, was used for initial scenario analyses, as discussed in Section 6.7. These provided insights to clinic staff but do not contribute directly to our research objectives.

The extended simulation was used to test more fundamental changes to clinic processes; in particular using data from GP referrals to plan pathways. The patient labels *age group*, *predicted result* and *actual result* all affect patients' progress through the extended simulation. These label values were generated from empirical joint distributions for the simple scorecard but for the full scorecard not all label combinations appear in the data sample. Therefore I instead applied our novel method for generating the probabilities of each combination of label values, which uses Poisson regression.

Another methodological contribution is our approach for selecting scorecard cut-offs based on operational rather than predictive performance. I tested the potential operational impact of using the scorecards (developed in Chapter 5), to identify higher risk patients to send straight to imaging.

Table 6.24: Full scorecard with one clinician and 15-minute gap between appointments

Precision	Clinic efficiency 5%		Average total time (mins) 10%	
Cut-off score	Number runs	Mean [95% confidence interval]	Number runs	Mean [95% confidence interval]
165	166	0.20 [0.19,0.22]	61	123 [110,135]
170	162	0.21 [0.20,0.22]	61	123 [111,135]
175	158	0.21 [0.20,0.22]	61	123 [111,135]
180	147	0.22 [0.21,0.23]	66	117 [105,128]
185	140	0.23 [0.22,0.24]	59	115 [104,127]
190	141	0.24 [0.23,0.25]	66	113 [102,124]
195	132	0.25 [0.24,0.26]	57	113 [102,125]
200	123	0.25 [0.24,0.26]	47	105 [95,116]
205	120	0.26 [0.24,0.27]	47	107 [96,117]
210	116	0.26 [0.25,0.27]	46	106 [95,116]
215	112	0.26 [0.25,0.27]	46	108 [97,118]
220	112	0.26 [0.25,0.28]	52	109 [98,119]
225	83	0.27 [0.25,0.28]	43	110 [99,121]
230	71	0.27 [0.26,0.29]	35	114 [103,126]
235	63	**0.28 [0.26,0.29]**	41	113 [102,124]
240	64	0.27 [0.26,0.29]	37	120 [108,131]
245	54	0.27 [0.26,0.28]	34	125 [112,137]
250	51	0.26 [0.25,0.28]	33	131 [118,144]
255	52	0.26 [0.25,0.28]	24	127 [115,139]
260	62	0.26 [0.25,0.27]	24	132 [119,144]
265	45	0.25 [0.24,0.27]	24	134 [121,147]
270	49	0.25 [0.24,0.27]	20	138 [124,151]
275	48	0.25 [0.23,0.26]	19	141 [127,155]
280	43	0.24 [0.23,0.25]	16	142 [128,156]
285	44	0.24 [0.23,0.25]	18	143 [129,156]
290	44	0.24 [0.23,0.25]	18	143 [129,156]
295	48	0.24 [0.23,0.25]	18	143 [129,156]
300	48	0.24 [0.23,0.25]	18	143 [129,156]

It was found that in the current set-up on Tuesdays, using the simple scorecard in this way would not improve mean clinic efficiency. However, when coupled with new appointment times (15 minutes apart) and changing to having one clinician working at a time, using the simple scorecard in the simulation did improve mean clinic efficiency. In particular, the best results for Tuesday (using mean average total time as a tie-breaker) were obtained for a cut-off score of 15; conveniently corresponding to sending patients with a lump recorded straight to test, and patients without a lump recorded to a clinician first. This is also the cut-off score that achieves the best balance between specificity and sensitivity, as found in Chapter 5. These changes also improve mean clinic efficiency and mean average total times in the simulation on both Mondays and Wednesdays.

However another factor that should be taken into account from a cost and feasibility perspective is the time at which the clinicians finish working; the results for this indicator differ by day of the week. It may be possible to change the numbers of appointments offered each day to better balance finishing times across the days of the week. In any case, I have shown that it may be possible to see the same number of patients in fewer hours of clinician time.

Using the full scorecard rather than the simple scorecard in the simulation yielded only a minor improvement to the mean clinic efficiency, achieved for a cut-off score of 235. Additionally, the full scorecard is more complicated to apply in practice, since it involves considering seven referral characteristics rather than two. For these reasons, we recommend the simple scorecard as more promising for use in practical pathway planning.

Chapter 7

Conclusion

In this chapter we conclude the thesis by discussing the achievement of each of our research questions in turn, then highlighting limitations of the research and suggesting extensions. The research questions, restated from Section 1.2, are as follows.

1. For the Whittington breast diagnostic clinic, is GP referral information accurate and complete enough to be used to predict the risk of an abnormal result (defined later), and so to identify patients who could be sent straight for imaging tests?
2. For the Whittington breast diagnostic clinic, could introducing risk-based pathways increase the proportion of patients' time at the clinic that adds value?
3. What general insights, practical and methodological, can be drawn from the models developed for this case study?

7.1 Achievement of research questions

7.1.1 Research question 1

For the first objective, as described in Chapter 4, I collected data from 179 eligible consenting patients' records to form a unique dataset linking GP referral information with clinic tests and results. From analysing the data it became apparent that GP referral information is complete enough to be considered feasible for use in predicting abnormal results (i.e. predicting for which patients an abnormality is likely to be detected from mammogram, ultrasound or biopsy) in the sample. Therefore I fitted two logistic regression scorecard models that predict abnormal results from a selection of GP referral characteristics. A major strength of the logistic regression scorecard methodology is that resulting models are easy to use, while using WoE (weights of evidence) enables non-linear relationships to be captured effectively. It was found that the most predictive referral characteristics were a patient's age and whether or not a lump was recorded. The full scorecard which contains seven referral characteristics was only slightly better at discriminating between normal and abnormal results than the more parsimonious simple scorecard which contains these two most predictive referral characteristics. Since the simple and full scorecards achieved average AUROC on bootstrap samples of 0.72 and 0.75 respectively, we can say that GP referral information provided to the Whittington breast clinic is accurate enough to effectively predict the risk of patients having abnormal results in our sample.

7.1.2 Research question 2

In order to understand the current performance of the clinic, I created process maps, as well as collecting and analysing data from a range of sources. This led to development, verification and validation of two simulation models. The core model was used to investigate relatively minor changes at the clinic, while the extended model, which includes more detailed modelling of patient characteristics, was used to model patients routed according to their risk of an abnormal result.

The impact of using the scorecards to route high risk patients straight to imaging was tested for a range of cut-off scores for both scorecards in the extended simulation for Tuesday. Scenarios were compared in terms of their mean *clinic efficiency* (daily average proportion of a patient's time at the clinic that is value-added), and ties were resolved by comparing the mean *average total time* (daily average of patients' total times at the clinic).

It was found that in order to improve mean clinic efficiency, other improvements were necessary in conjunction with using a scorecard, namely changing the gaps between appointments to 15 minutes, and changing to one clinician working at a time. In this adapted clinic set-up, the best cut-off score (in terms of mean clinic efficiency and using mean average total time as a tie-breaker) for the simple scorecard was 15. Although optimised for Tuesdays, these changes also improved mean clinic efficiency and mean average total times in the simulation on both Mondays and Wednesdays. The simulation results show that the same number of patients could be seen in fewer hours of clinician time by making these changes, since only one clinician would work at a time, although any resulting cost savings would need to be balanced against the cost of longer imaging opening hours on Mondays.

When using the full scorecard instead of the simple scorecard, only a slight improvement in mean clinic efficiency was achieved. Therefore we recommend the simple scorecard as the more promising option for practical use, since it involves assessing only two rather than seven characteristics. As a result of these simulation experiments, we can say that there is potential for improving mean clinic efficiency by introducing a scorecard based triaging system when implemented together with other changes.

7.1.3 Research question 3

From a methodological perspective, we combined logistic regression, Poisson regression and discrete-event simulation in a novel way to test whether using diagnostic information provided by non-specialists to plan diagnostic tests could improve operational criteria. Although our case study is for a particular breast diagnostic clinic, the methods we use are generalisable to other diagnostic clinics and more widely in health and other application areas.

From a practical perspective, we have provided the first evidence that we are aware of that links GP referral information with abnormal results of breast diagnostic tests. This implies that referral information could help to identify those patients most likely to require imaging tests. Although we cannot say that the scorecard models and findings are directly generalisable, they do offer proof of concept.

We contribute a novel method for generating combinations of dependent categorical characteristics, where unseen combinations of characteristics are possible, for use in DES modelling. In our literature review of how patient characteristic inputs are generated for pathway DES models in Section 3.3 and search of the related theoretical literature, we did not find a suitable method for this situation. Our approach involves fitting a Poisson loglinear model to estimate the distribution of counts of characteristic combinations. This is a statistically sound approach that avoids the need to make assumptions about the relationship between characteristics when data samples are small. The benefit of this approach over using empirical distributions is that all combinations are possible, not just those present in the sample.

Our unique approach for generating characteristics for DES is not only applicable to health services but also to other areas such as manufacturing, call centres and marketing. This approach is of particular benefit in applications where data are scarce. Some reasons for having missing (but possible) combinations of characteristics in a data sample are limited time or money to collect data, as well as extremely rare events or a very large number of possible combinations. Potential applications of our approach could include the following. Firstly, an IT call centre may wish to simulate resource use for different problems characterised by urgency, customer type, software and type of problem. Some combinations of these problem characteristics may be rare (so do not appear in a data sample) but important to capture, for example if they are more expensive or time-consuming to deal with. Secondly, a simulation of organ transplants has many possible combinations of characteristics that must match between donor and recipient including age, genetics and blood type. Thirdly, a simulation of a new production line, following a smaller-scale trial, may use the characteristics of individual products, including different faults that can occur which need to be handled differently depending on the product type.

Another methodological contribution of this study is the combination of logistic regression with simulation in order to determine the best cut-off from an operational perspective. Although classification and DES models have been combined before, as described in Section 3.4, this is not common and we are the first to use simulation to decide on the cut-off score. For example, Huang and Hanauer (2016) use DES to evaluate the cost per patient for appointment overbooking policies based on logistic regression models with increasing numbers of parameters. In that paper, the risk cut-off for each model was chosen (without using the simulation) so that the number of misclassified patients was minimised. This differs from our approach which uses the output of the simulation to decide upon the best cut-off. Our approach has the advantage that a wider range of performance measures can be considered than solely predictive accuracy. Thus it could be applied to risk classification problems in any area where operational measures such as waiting times or throughput are of importance in the choice of cut-off.

Some advantages of our unique approach combining logistic regression and simulation to select a cut-off score are as follows. This approach allows decision makers to consider the wider implications of their choice of cut-off. It is more versatile than considering solely predictive performance measures; any measure that can be simulated can be used to compare cut-off scores. This means that performance measures can be customised to particular contexts. The practical impact of the cut-off score may be more important than the predictive accuracy, particularly where the classification model is being used to sequence or prioritise services rather than deciding whether or not to offer a service.

On the other hand, disadvantages of combining logistic regression and simulation to select a cut-off score are as follows. The optimal cut-off is likely to depend on the specic setting, depending on, for example, the service times and number of patients, while using a predictive performance measure may be more transferrable between settings. Also, it may be that several different cut-offs generate quite similar operational performance results. However in this case the best scores provide a range from which decision makers can choose.

Although the models were developed for a particular clinic, with some modifications they could be applied elsewhere. The scorecard points should not be used elsewhere, but the methods and possibly characteristics could be used. As discussed in Section 2.2, GP referral forms differ between regions and sometimes even hospitals, however they have similar information recorded: symptoms, family history, age, duration of symptoms. Two different referral forms were used to provide the data for our study.

Thus it is likely that the same characteristics can be considered for inclusion in scorecard models for other clinics.

The simulation models could be used for other clinics by changing the parameters such as number of appointments, test proportions and punctuality profiles where needed. In particular, the base simulation is suitable for clinics which send all patients to see a clinician first, while the extended simulation can be used for clinics where some patients are sent to test first. The joint distributions of patient characteristics should also be based on local data where possible. For the full scorecard, the R code for using the Poisson regression method of fitting a distribution is provided in Appendix B. With some simulation expertise, the performance measures over which to optimise the cut-off could also be changed in the simulation.

7.2 Limitations and future work

Some research limitations and possible extensions are as follows.

One limitation of the scorecards developed is that the sample used was relatively small due to time constraints on data collection; nonetheless, the fact that development of scorecards is feasible in a small-scale setting is promising as a precursor to larger scale studies. In particular, the scorecard development approach followed is generalisable to other breast diagnostic clinics and more broadly to other types of diagnostic clinics. However, the resulting models themselves should not be used in other clinics, due to potential differences in referral forms, patient populations and referral behaviours. Automatic text recognition software and text mining techniques could help with developing scorecards based on scanned-in referrals including GPs' letters.

The second limitation of our scorecards is that we do not know whether our sample is representative; it may be that certain types of patients were more likely to consent to their data being included. Therefore we recommend validating the scorecard on an independent test set at the Whittington breast clinic, preferably as part of a service evaluation by staff, in which case consent would not be required. In addition, a feasibility and impact assessment of using the scorecard in practice, for example through a pilot study, could be carried out.

When simulating the impacts of using scorecards to triage patients, we limited our attention to the following operational performance measures: clinic efficiency, average total time and clinic end time. Thus a limitation of our modelling is that we did not consider other performance measures such as costs. The research could be extended by performing a cost-effectiveness analysis and by considering additional performance measures, perhaps in a weighted function. Another extension could be looking at the range in performance across different patients and across different days, for example by calculating percentiles rather than average measures. For the Whittington clinic in particular, we recommend investigating whether the numbers of appointments offered each day could be better balanced over the three clinic days.

In this research we did not attempt to compare alternative ways of modelling patient characteristics in the simulation; when testing the simple scorecard we used empirical distributions and when testing the full scorecard we used Poisson regression to model the distributions. Future work could compare recommendations obtained from using a Poisson model distribution to the empirical distribution, for a series of case studies. We suspect that situations where the rarer combinations of characteristics correspond to higher service use will particularly benefit from the Poisson approach.

The problem under study can be generalised to other contexts outside diagnostic clinics, and even outside healthcare. The general problem is as follows. Is information provided by a non-specialist (or a patient or customer themselves) complete and accurate enough to make decisions related to the patient or customer (e.g. assign resources or allow access), without first performing a specialist assessment? One example of this problem is in IT support: Is the information collected by a front-line technician sufficiently detailed and accurate to resolve the issue or does it require follow-up by a second-line technician? A second example is in security: Does an airline passenger's face match their passport photograph closely enough to automatically open the gate or should they queue for a further human check? A third example is the problem of detecting false identities created by thieves to borrow money that is purposefully not repaid. In which situations should credit be granted online and in which should the customer be invited into a bank branch? The classification-DES approach allows different operational measures to be considered depending on the context.

Appendices

Appendix A

Formulating scorecards

A diagnostic result, Y, may be normal ($Y = 1$) or abnormal ($Y = 0$). A binary logistic regression model predicts the probability, p, of a normal result from patient-specific variables, $X_1, X_2, ..., X_n$, obtained from GP referral information, as given in equation A.1. The parameters $\beta_1, \beta_2, ..., \beta_n$ show the relative importance of each characteristic in the prediction and β_0 is the intercept. These parameters are obtained from maximum likelihood estimates. In reality there is some error, ϵ, that is not captured by the model.

$$p := Prob(Y = 1|X_1, X_2, ..., X_n) = \frac{1}{1 + e^{-(\beta_0 + \sum_{i=1}^{n} \beta_i X_i)}} \tag{A.1}$$

One way of coding the variables is to use weights of evidence. The weight of evidence, W, of a particular grouped attribute j, for example "young", of a characteristic i, for example "age", is the strength of evidence that patients who have the attribute will have an abnormal result. Let the number of normal (abnormal) results with attribute j be n_j (a_j) and the total number of normal (abnormal) results be n_{total} (a_{total}). Then the weight of evidence variable, W_{ij}, is given by the following.

$$W_{ij} = \ln \left(\frac{a_j}{a_{total}} \div \frac{n_j}{n_{total}} \right) \tag{A.2}$$

The scaled score, S, which is calculated from a scorecard, is related to the unscaled score, $\beta_0 + \sum_{i=1}^{n} \beta_i X_i$, which appears in the logistic regression, as follows.

$$S = (\beta_0 + \sum_{i=1}^{n} \beta_i X_i) \cdot \text{factor} + \text{offset} \tag{A.3}$$

The factor and offset are the solution to the following system of linear equations. This follows since the unscaled score is equal to the log odds. The pair $(Score, Odds)$ is the alignment point, e.g. it is assumed that the score 300 corresponds to odds of 12(:1) of having an abnormal result. The PDO is the specified number of points to double the odds, e.g. if PDO is 20, then the odds double for every increase in 20 points.

$$\text{Score} = \ln(\text{Odds}) \cdot \text{factor} + \text{offset} \tag{A.4}$$

$$\text{Score} + \text{PDO} = \ln(2 \cdot \text{Odds}) \cdot \text{factor} + \text{offset} \tag{A.5}$$

Using the weights of evidence codings as the X variables, the scaled score for a particular patient becomes

$$S = (\beta_0 + \sum_{i=1}^{n} \beta_i W_{ij}) \cdot \text{factor} + \text{offset} \tag{A.6}$$

where n is the number of variables and j is the attribute that this patient has for characteristic i.

The points of the scaled score can be split between the characteristics by dividing the parts of the score that are not characteristic-specific between them.

$$\text{Points for characteristic } i = (\frac{\beta_0}{n} + \beta_i W_{ij}) \cdot \text{factor} + \frac{\text{offset}}{n} \tag{A.7}$$

Finally, to find the total score for a particular patient, the points for each of the patient's attributes are added up.

Appendix B

R code used for fitting Poisson loglinear models

```
PearsonP <- function(NumCombos, ProbVector, FittedVector,
PearsonOrig, NumSamples){
x<-1:NumCombos

#Load plyr package for "count"
library(plyr)

#Sample, count frequencies of each combination, and add to
observed combinations vector
sample<-sample(x,size=179,replace=TRUE,prob=ProbVector)
freq<-count(sample)
names(freq)[2]<-1
observed<-merge(x,freq,by="x",all=TRUE)

y<-NumSamples
for (i in 2:y) {
sample<-sample(x,size=179,replace=TRUE,prob=ProbVector)
freq<-count(sample)
names(freq)[2]<-i
observed<-merge(observed,freq,by="x",all=TRUE)
}

#Replace NAs with 0s
observed[is.na(observed)] <- 0

#Initialise Pearson test stat vector
```

```
PearsonStat<-vector(mode = "numeric", length = y)

#Calculate Pearson test stat for each sample
for (i in 1:y) {
PearsonStat[i]<-sum((observed[,i+1]-FittedVector)^2/FittedVector)
}

#Count how often simulated Pearson test stats are larger
#than original data's Pearson test stat
Count<-0
for (i in 1:y) {
if (PearsonStat[i]>=PearsonOrig)
{Count<-Count+1}
}

#Get simulated p value (proportion of test statistic values
#that are at least as large as observed value.)
Proportion<-Count/y
return(Proportion)
}

allmydata<-data.frame(expand.grid(
Age=factor(c("1","2a","2b","3","4","5"),
labels=c("Age < 29","29 <= Age < 35","35<= Age < 42","42 <= Age < 47",
"47 <= Age < 52","52<= Age")),
Duration=factor(c("1","2","3","4","5"),
labels=c("Less than 2 weeks","2 weeks - 2 months","2 - 5 months","Over 5 months",
"NA or NR")),
Urgency=factor(c("1","2","3"),
labels=c("Suspected cancer","Symptomatic","Other or NR")),
OtherSymptom=factor(c(1,2),
labels=c("No or NR","Y")),
Lump=factor(c("No or NR","Y"),
labels=c("No or NR","Y")),
UnilateralPain=factor(c("1","2"),
labels=c("No or NR","Y")),
FamilyHistory=factor(c("1","2"),
labels=c("No or NR","Y")),
NormalResult=factor(c(0,1),
labels=c("abnormal","normal"))),
count=c(#String of 2880 integers that are mainly zeroes
))

#Fit model 1. This model assumes independence.
IndepGenLin<-glm(count ~ Age + Duration + Urgency + OtherSymptom + Lump +
UnilateralPain + FamilyHistory + NormalResult,
data=allmydata, family=poisson())
summary(IndepGenLin)
Fit1<- c(fitted(IndepGenLin))
Prob1<- Fit1/179
ChiSq1<-sum((allmydata["count"]-Fit1)^2/Fit1)
Model1_p<-PearsonP(2880,Prob1,Fit1,ChiSq1,3000)
Model1_p

#Fit model 2 with all two-way interactions
M_Complex<- glm(count ~ Age + Duration + Urgency + OtherSymptom + Lump +
UnilateralPain + FamilyHistory + NormalResult +

Age:Duration + Age:Urgency + Age:OtherSymptom + Age:Lump +
Age:UnilateralPain + Age:NormalResult + Age:FamilyHistory +

Duration:Urgency + Duration:OtherSymptom + Duration:Lump +
Duration:UnilateralPain + Duration:FamilyHistory +  Duration:NormalResult +

Urgency:OtherSymptom + Urgency:Lump + Urgency:UnilateralPain +
Urgency:FamilyHistory + Urgency:NormalResult +

OtherSymptom:Lump + OtherSymptom:UnilateralPain +
OtherSymptom:FamilyHistory + OtherSymptom:NormalResult +

Lump:UnilateralPain + Lump:FamilyHistory + Lump:NormalResult +

UnilateralPain:FamilyHistory + UnilateralPain:NormalResult

+ FamilyHistory:NormalResult,
data=allmydata, family=poisson())

Fit2<-fitted(M_Complex)
data2<-data.frame(allmydata[1:8], expected = c(Fit2))
summary(M_Complex)
Prob2<- Fit2/179
ChiSq2<-sum((allmydata["count"]-Fit2)^2/Fit2)
```

```
Model2_p<-PearsonP(2880,Prob2,Fit2,ChiSq2,3000)
Model2_p
```

Appendix C

Figures

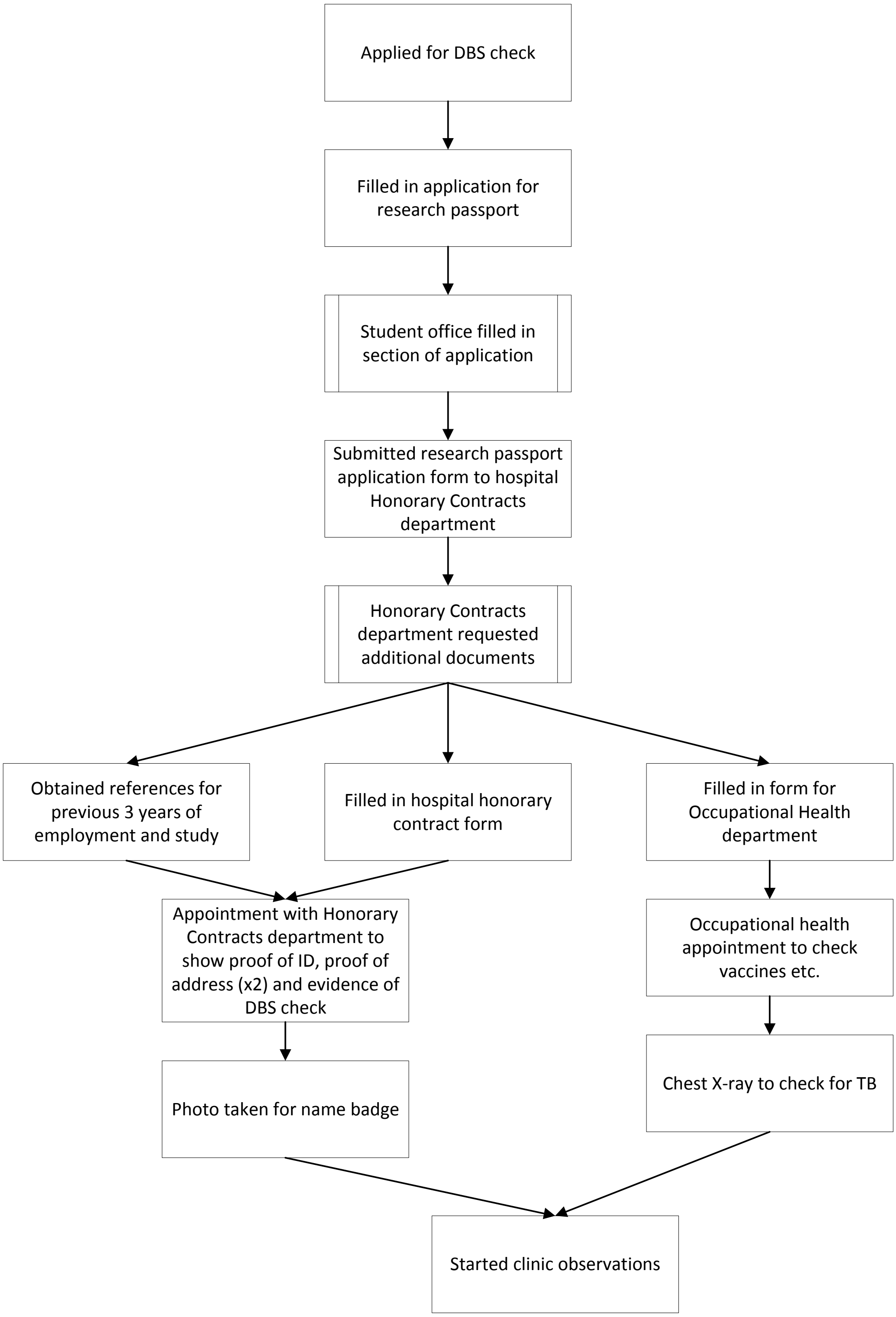

Applied for DBS check
Filled in application for research passport
Student office filled in section of application
Submitted research passport application form to hospital Honorary Contracts department
Honorary Contracts department requested additional documents
Obtained references for previous 3 years of employment and study
Filled in hospital honorary contract form
Filled in form for Occupational Health department
Appointment with Honorary Contracts department to show proof of ID, proof of address (x2) and evidence of DBS check
Occupational health appointment to check vaccines etc.
Photo taken for name badge
Chest X-ray to check for TB
Started clinic observations

Figure C.1: Process to obtain honorary contract

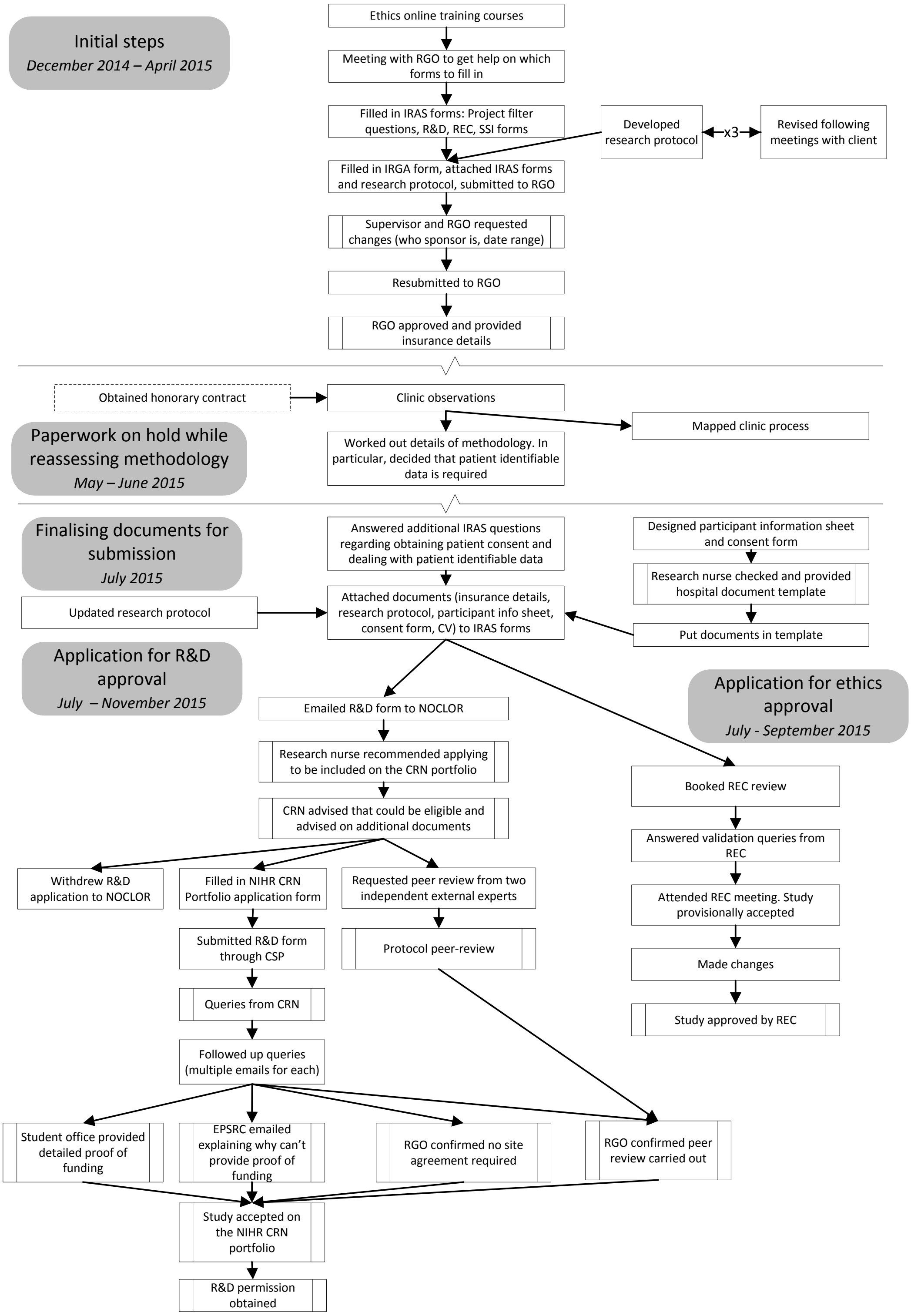
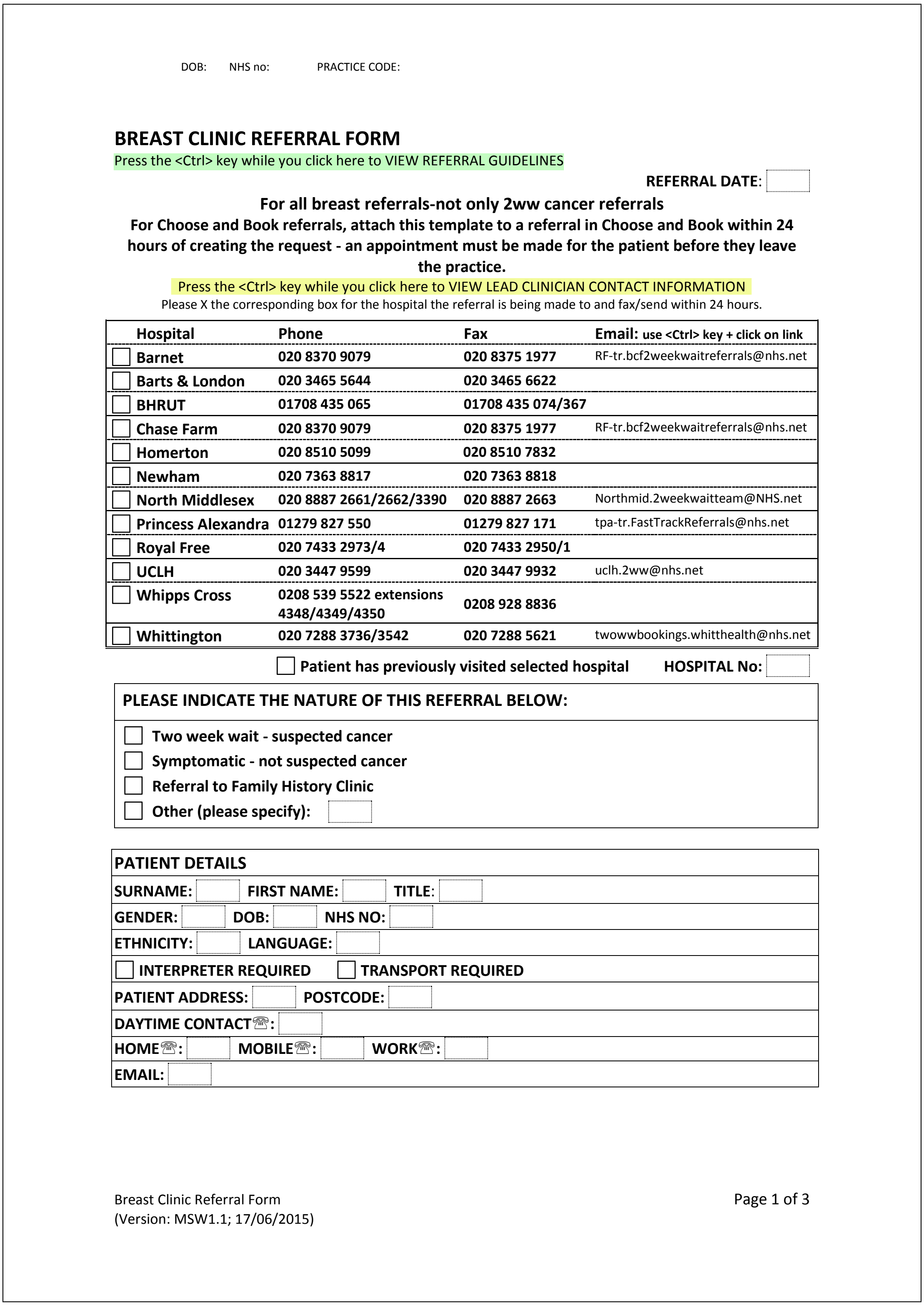

Initial steps
December 2014 – April 2015
Ethics online training courses
Meeting with RGO to get help on which forms to fill in
Filled in IRAS forms: Project filter questions, R&D, REC, SSI forms
Developed research protocol
x3
Revised following meetings with client
Filled in IRGA form, attached IRAS forms and research protocol, submitted to RGO
Supervisor and RGO requested changes (who sponsor is, date range)
Resubmitted to RGO
RGO approved and provided insurance details
Obtained honorary contract
Clinic observations
Mapped clinic process
Paperwork on hold while reassessing methodology
May – June 2015
Worked out details of methodology. In particular, decided that patient identifiable data is required
Finalising documents for submission
July 2015
Answered additional IRAS questions regarding obtaining patient consent and dealing with patient identifiable data
Designed participant information sheet and consent form
Research nurse checked and provided hospital document template
Updated research protocol
Attached documents (insurance details, research protocol, participant info sheet, consent form, CV) to IRAS forms
Put documents in template
Application for R&D approval
July – November 2015
Application for ethics approval
July - September 2015
Emailed R&D form to NOCLOR
Research nurse recommended applying to be included on the CRN portfolio
CRN advised that could be eligible and advised on additional documents
Booked REC review
Answered validation queries from REC
Attended REC meeting. Study provisionally accepted
Made changes
Study approved by REC
Withdrew R&D application to NOCLOR
Filled in NIHR CRN Portfolio application form
Requested peer review from two independent external experts
Submitted R&D form through CSP
Protocol peer-review
Queries from CRN
Followed up queries (multiple emails for each)
Student office provided detailed proof of funding
EPSRC emailed explaining why can't provide proof of funding
RGO confirmed no site agreement required
RGO confirmed peer review carried out
Study accepted on the NIHR CRN portfolio
R&D permission obtained

Figure C.2: Process to obtain research approvals



DOB: NHS no: PRACTICE CODE:

BREAST CLINIC REFERRAL FORM

Press the <Ctrl> key while you click here to VIEW REFERRAL GUIDELINES

REFERRAL DATE:

For all breast referrals-not only 2ww cancer referrals

For Choose and Book referrals, attach this template to a referral in Choose and Book within 24 hours of creating the request - an appointment must be made for the patient before they leave the practice.

Press the <Ctrl> key while you click here to VIEW LEAD CLINICIAN CONTACT INFORMATION

Please X the corresponding box for the hospital the referral is being made to and fax/send within 24 hours.

	Hospital	Phone	Fax	Email: use <Ctrl> key + click on link
□	Barnet	020 8370 9079	020 8375 1977	RF-tr.bcf2weekwaitreferrals@nhs.net
□	Barts & London	020 3465 5644	020 3465 6622	
□	BHRUT	01708 435 065	01708 435 074/367	
□	Chase Farm	020 8370 9079	020 8375 1977	RF-tr.bcf2weekwaitreferrals@nhs.net
□	Homerton	020 8510 5099	020 8510 7832	
□	Newham	020 7363 8817	020 7363 8818	
□	North Middlesex	020 8887 2661/2662/3390	020 8887 2663	Northmid.2weekwaitteam@NHS.net
□	Princess Alexandra	01279 827 550	01279 827 171	tpa-tr.FastTrackReferrals@nhs.net
□	Royal Free	020 7433 2973/4	020 7433 2950/1	
□	UCLH	020 3447 9599	020 3447 9932	uclh.2ww@nhs.net
□	Whipps Cross	0208 539 5522 extensions 4348/4349/4350	0208 928 8836	
□	Whittington	020 7288 3736/3542	020 7288 5621	twowwbookings.whitthealth@nhs.net

□ **Patient has previously visited selected hospital** **HOSPITAL No:**

PLEASE INDICATE THE NATURE OF THIS REFERRAL BELOW:

- □ **Two week wait - suspected cancer**
- □ **Symptomatic - not suspected cancer**
- □ **Referral to Family History Clinic**
- □ **Other (please specify):**

PATIENT DETAILS

SURNAME: **FIRST NAME:** **TITLE**:

GENDER: **DOB:** **NHS NO:**

ETHNICITY: **LANGUAGE:**

□ **INTERPRETER REQUIRED** □ **TRANSPORT REQUIRED**

PATIENT ADDRESS: **POSTCODE:**

DAYTIME CONTACT☎:

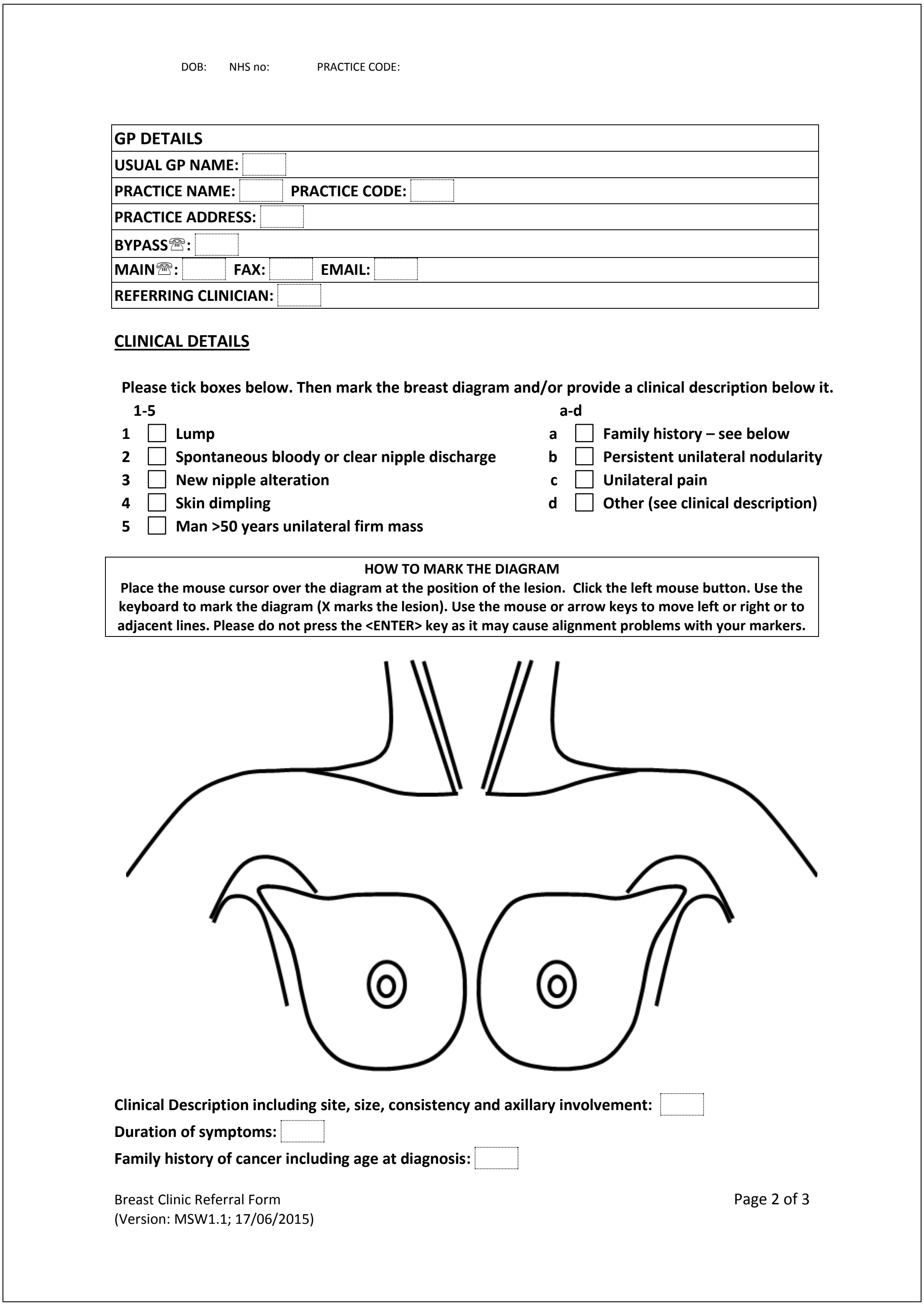
HOME☎: **MOBILE☎:** **WORK☎:**

EMAIL:

Breast Clinic Referral Form
(Version: MSW1.1; 17/06/2015)

Page 1 of 3

Figure C.3: Standard referral form page 1



DOB: NHS no: PRACTICE CODE:

GP DETAILS
USUAL GP NAME:
PRACTICE NAME: **PRACTICE CODE:**
PRACTICE ADDRESS:
BYPASS☎:
MAIN☎: **FAX:** **EMAIL:**
REFERRING CLINICIAN:

CLINICAL DETAILS

Please tick boxes below. Then mark the breast diagram and/or provide a clinical description below it.

1-5

1 □ **Lump**
2 □ **Spontaneous bloody or clear nipple discharge**
3 □ **New nipple alteration**
4 □ **Skin dimpling**
5 □ **Man >50 years unilateral firm mass**

a-d

a □ **Family history – see below**
b □ **Persistent unilateral nodularity**
c □ **Unilateral pain**
d □ **Other (see clinical description)**

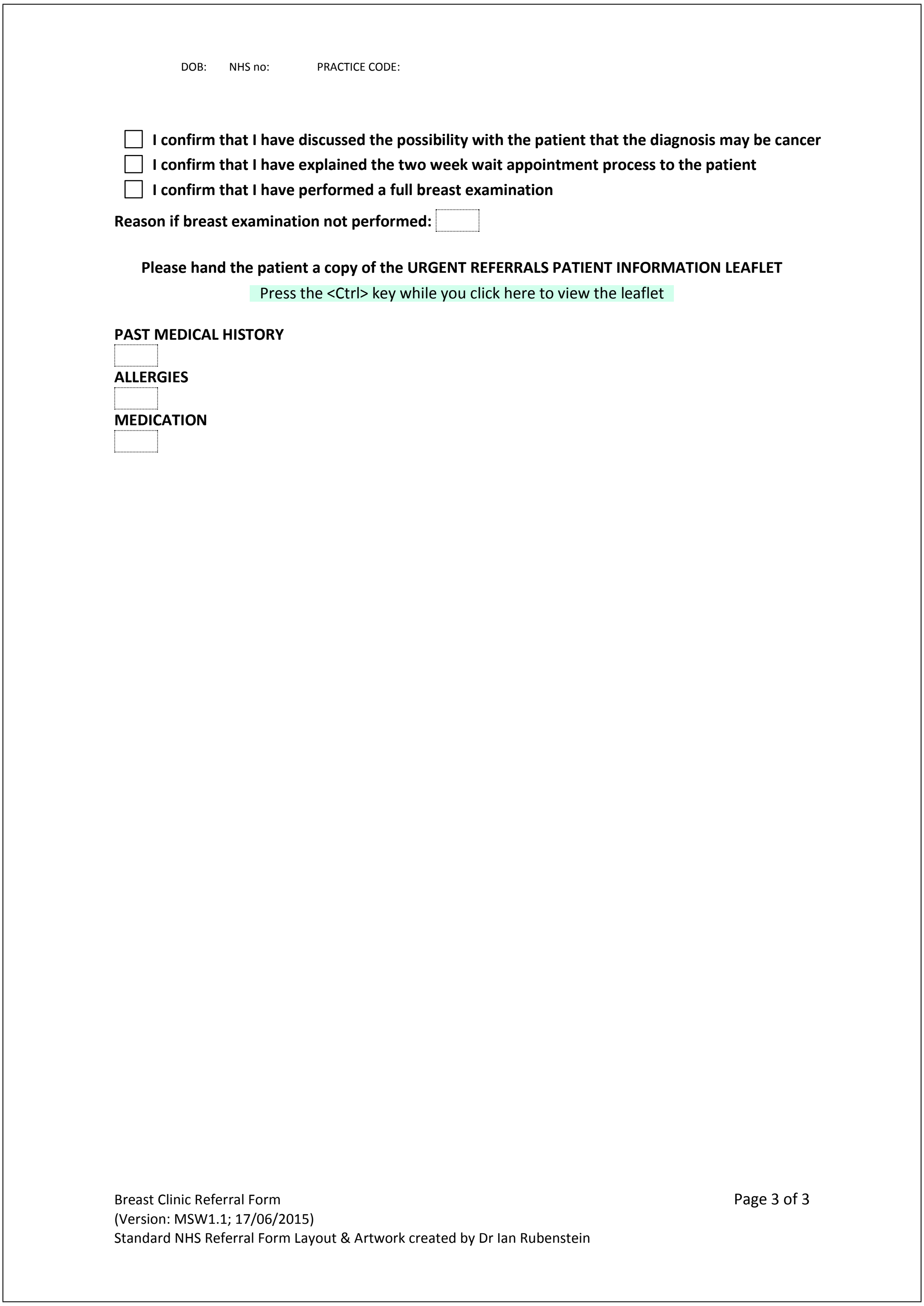
HOW TO MARK THE DIAGRAM
Place the mouse cursor over the diagram at the position of the lesion. Click the left mouse button. Use the keyboard to mark the diagram (X marks the lesion). Use the mouse or arrow keys to move left or right or to adjacent lines. Please do not press the <ENTER> key as it may cause alignment problems with your markers.

Clinical Description including site, size, consistency and axillary involvement:
Duration of symptoms:
Family history of cancer including age at diagnosis:

Breast Clinic Referral Form
(Version: MSW1.1; 17/06/2015)

Page 2 of 3

Figure C.4: Standard referral form page 2



DOB: NHS no: PRACTICE CODE:

□ **I confirm that I have discussed the possibility with the patient that the diagnosis may be cancer**
□ **I confirm that I have explained the two week wait appointment process to the patient**
□ **I confirm that I have performed a full breast examination**

Reason if breast examination not performed:

Please hand the patient a copy of the URGENT REFERRALS PATIENT INFORMATION LEAFLET

Press the <Ctrl> key while you click here to view the leaflet

PAST MEDICAL HISTORY

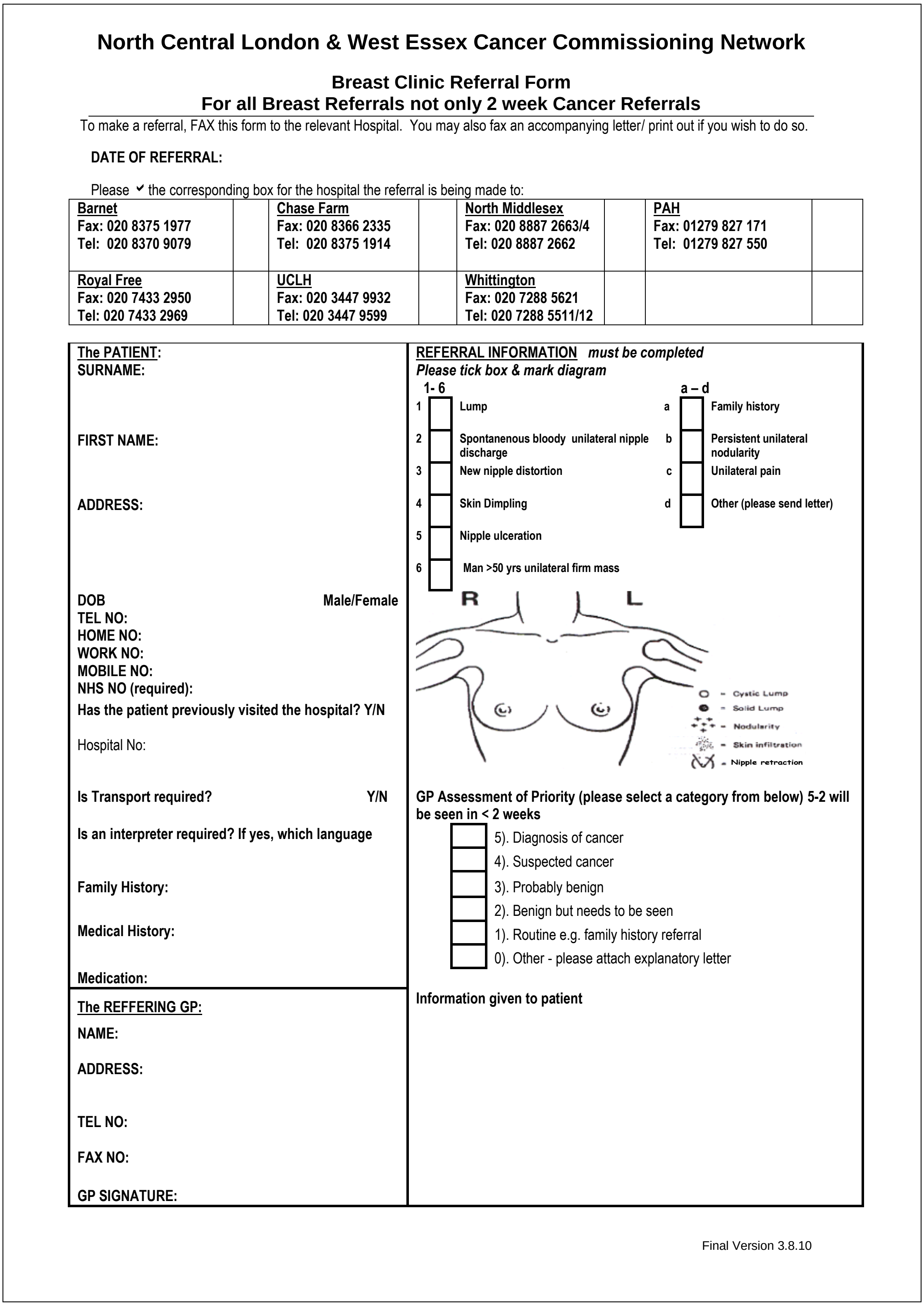
ALLERGIES

MEDICATION

Breast Clinic Referral Form
(Version: MSW1.1; 17/06/2015)
Standard NHS Referral Form Layout & Artwork created by Dr Ian Rubenstein

Page 3 of 3

Figure C.5: Standard referral form page 3



North Central London & West Essex Cancer Commissioning Network

Breast Clinic Referral Form
For all Breast Referrals not only 2 week Cancer Referrals

To make a referral, FAX this form to the relevant Hospital. You may also fax an accompanying letter/ print out if you wish to do so.

DATE OF REFERRAL:

Please ✓ the corresponding box for the hospital the referral is being made to:

Barnet Fax: 020 8375 1977 Tel: 020 8370 9079	□	**Chase Farm** Fax: 020 8366 2335 Tel: 020 8375 1914	□	**North Middlesex** Fax: 020 8887 2663/4 Tel: 020 8887 2662	□	**PAH** Fax: 01279 827 171 Tel: 01279 827 550	□
Royal Free Fax: 020 7433 2950 Tel: 020 7433 2969	□	**UCLH** Fax: 020 3447 9932 Tel: 020 3447 9599	□	**Whittington** Fax: 020 7288 5621 Tel: 020 7288 5511/12	□		□

The PATIENT:

SURNAME:

FIRST NAME:

ADDRESS:

DOB **Male/Female**

TEL NO:

HOME NO:

WORK NO:

MOBILE NO:

NHS NO (required):

Has the patient previously visited the hospital? Y/N

Hospital No:

Is Transport required? Y/N

Is an interpreter required? If yes, which language

Family History:

Medical History:

Medication:

The REFFERING GP:

NAME:

ADDRESS:

TEL NO:

FAX NO:

GP SIGNATURE:

REFERRAL INFORMATION ***must be completed***

Please tick box & mark diagram

1- 6			a – d		
1	□	Lump	a	□	Family history
2	□	Spontanenous bloody unilateral nipple discharge	b	□	Persistent unilateral nodularity
3	□	New nipple distortion	c	□	Unilateral pain
4	□	Skin Dimpling	d	□	Other (please send letter)
5	□	Nipple ulceration			
6	□	Man >50 yrs unilateral firm mass			

R L

○ - Cystic Lump

● - Solid Lump

+ - Nodularity

- Skin infiltration

- Nipple retraction

GP Assessment of Priority (please select a category from below) 5-2 will be seen in < 2 weeks

□ 5). Diagnosis of cancer

□ 4). Suspected cancer

□ 3). Probably benign

□ 2). Benign but needs to be seen

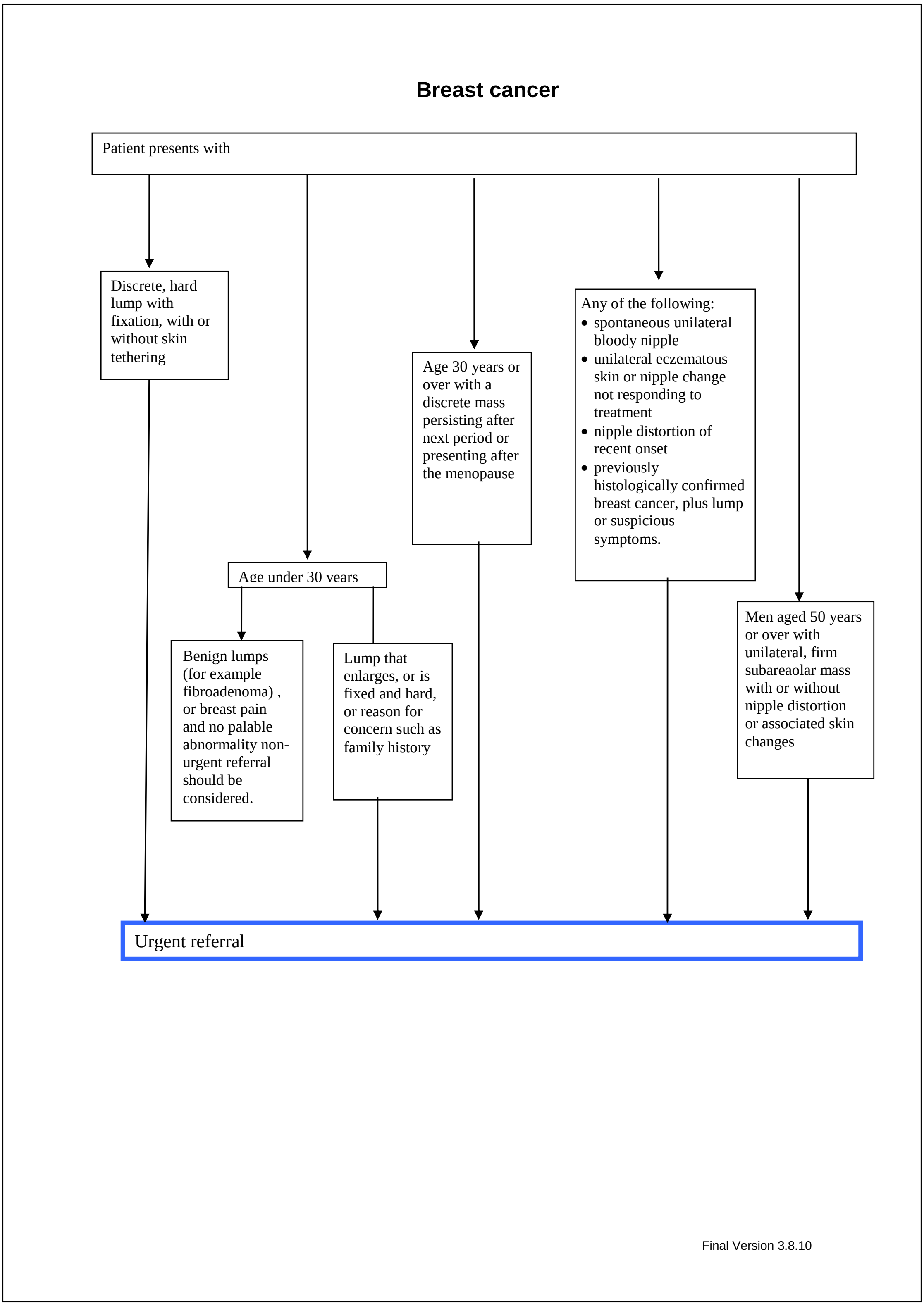
□ 1). Routine e.g. family history referral

□ 0). Other - please attach explanatory letter

Information given to patient

Final Version 3.8.10

Figure C.6: Outdated referral form page 1



Breast cancer

Patient presents with

Discrete, hard lump with fixation, with or without skin tethering

Age under 30 years

Benign lumps (for example fibroadenoma) , or breast pain and no palable abnormality non-urgent referral should be considered.

Lump that enlarges, or is fixed and hard, or reason for concern such as family history

Age 30 years or over with a discrete mass persisting after next period or presenting after the menopause

Any of the following:

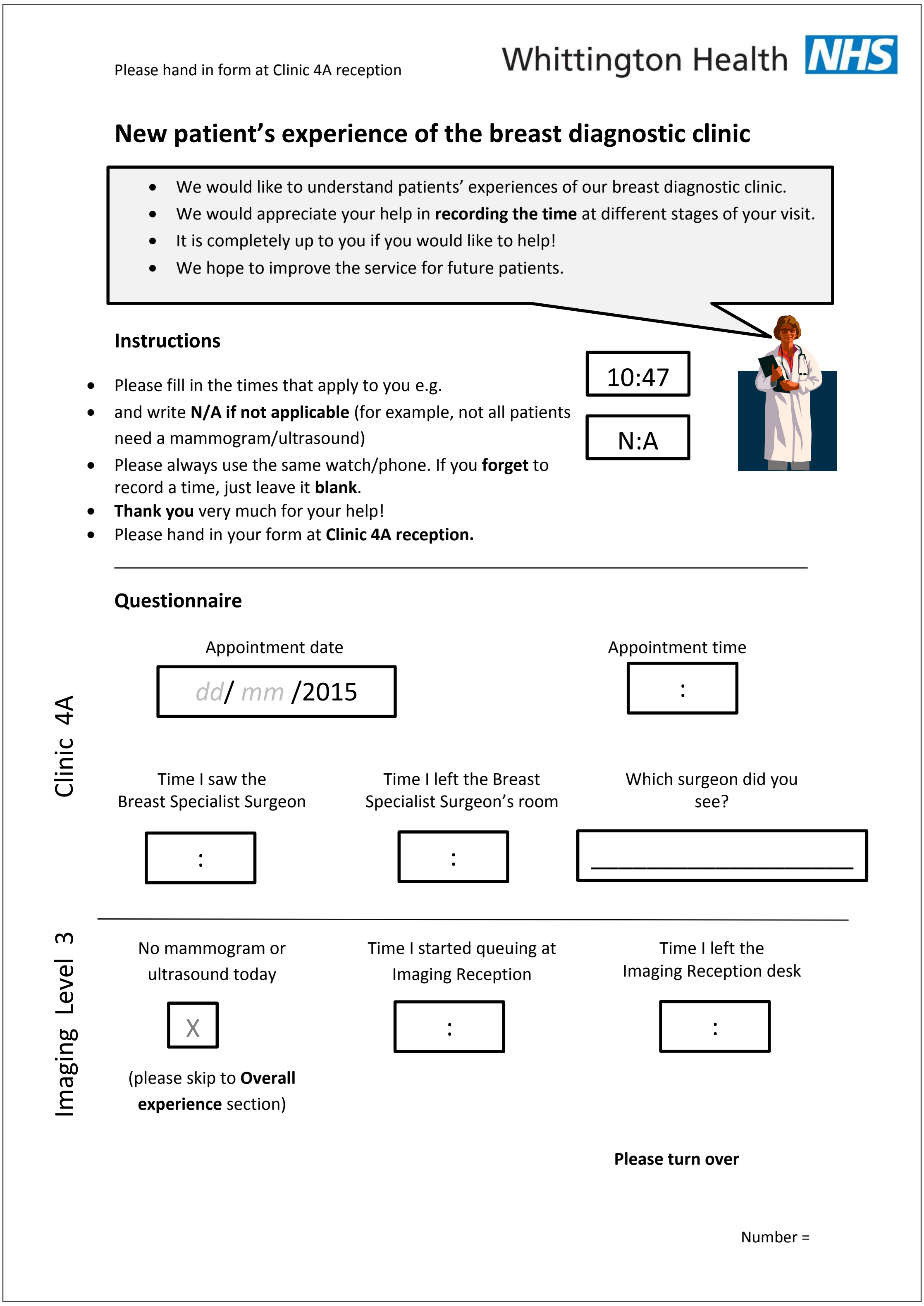
- spontaneous unilateral bloody nipple
- unilateral eczematous skin or nipple change not responding to treatment
- nipple distortion of recent onset
- previously histologically confirmed breast cancer, plus lump or suspicious symptoms.

Men aged 50 years or over with unilateral, firm subareaolar mass with or without nipple distortion or associated skin changes

Urgent referral

Final Version 3.8.10

FIGURE C.7: Outdated referral form page 2



Please hand in form at Clinic 4A reception

Whittington Health NHS

New patient's experience of the breast diagnostic clinic

- We would like to understand patients' experiences of our breast diagnostic clinic.
- We would appreciate your help in **recording the time** at different stages of your visit.
- It is completely up to you if you would like to help!
- We hope to improve the service for future patients.

Instructions

- Please fill in the times that apply to you e.g. 10:47
- and write **N/A if not applicable** (for example, not all patients need a mammogram/ultrasound) N:A
- Please always use the same watch/phone. If you **forget** to record a time, just leave it **blank**.
- **Thank you** very much for your help!
- Please hand in your form at **Clinic 4A reception.**

Questionnaire

Clinic 4A

Appointment date: dd/ mm /2015

Appointment time: :

Time I saw the Breast Specialist Surgeon: :

Time I left the Breast Specialist Surgeon's room: :

Which surgeon did you see? ______________________

Imaging Level 3

No mammogram or ultrasound today: ☒

(please skip to **Overall experience** section)

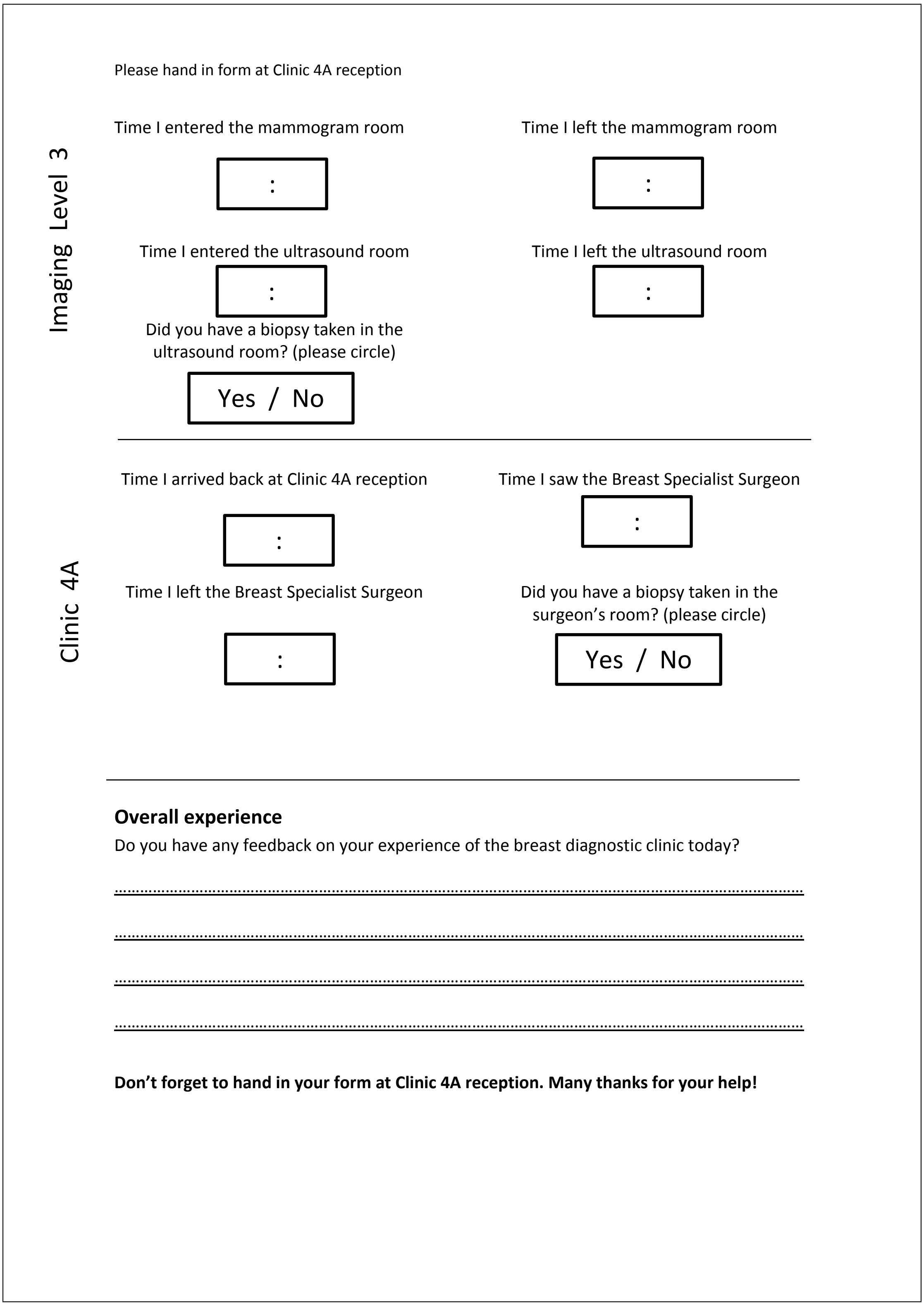
Time I started queuing at Imaging Reception: :

Time I left the Imaging Reception desk: :

Please turn over

Number =

FIGURE C.8: Questionnaire page 1



Please hand in form at Clinic 4A reception

Imaging Level 3

Time I entered the mammogram room

:

Time I left the mammogram room

:

Time I entered the ultrasound room

:

Time I left the ultrasound room

:

Did you have a biopsy taken in the ultrasound room? (please circle)

Yes / No

Clinic 4A

Time I arrived back at Clinic 4A reception

:

Time I saw the Breast Specialist Surgeon

:

Time I left the Breast Specialist Surgeon

:

Did you have a biopsy taken in the surgeon's room? (please circle)

Yes / No

Overall experience

Do you have any feedback on your experience of the breast diagnostic clinic today?

……

……

……

……

Don't forget to hand in your form at Clinic 4A reception. Many thanks for your help!

Figure C.9: Questionnaire page 2

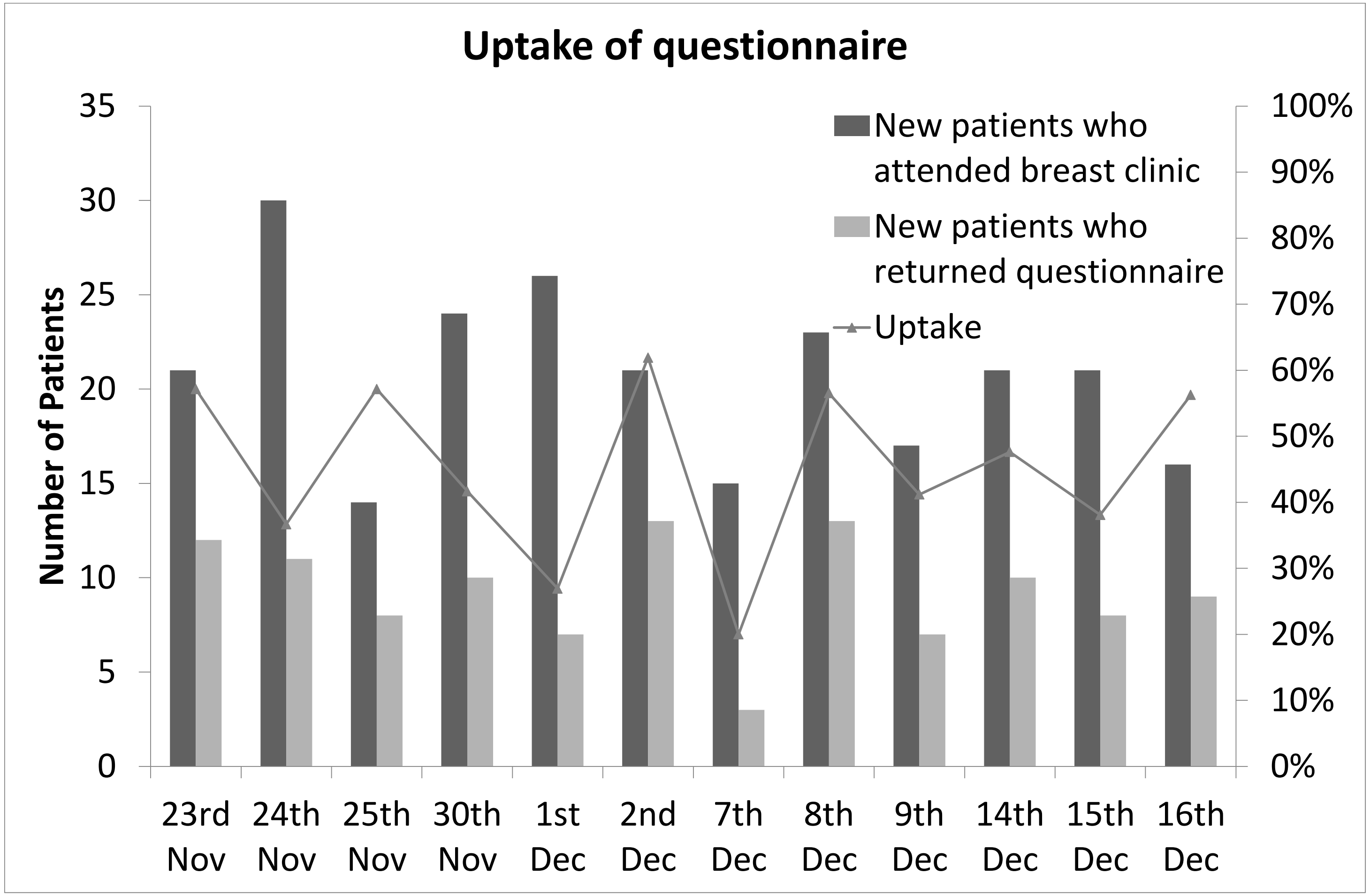


Figure C.10: Uptake of patient questionnaire (2015)

Appendix D

Tables

Table D.1: Sample sizes and date ranges

Source	Dates	Preprocessing	Variables (Sample size)
Patient questionnaires	23rd November to 16th December	Removed outliers, missing and invalid entries	Initial consultation service time (99), Mammogram service time (54), Ultrasound service time (68), Results consultation (73)
Observed timings	23rd November to 8th December 2015 and supplementary sample on 25th April 2016	Removed outliers	Initial consultation service time (69), Mammogram service time (20), Ultrasound service time (31), Results consultation (41), Clinician turnaround times (95), Mammogram turnaround times (12), Ultrasound turnaround times (16)
Medway appointment system	23rd December 2015- 23rd March 2016	None	Number new patients per day (91 days), No-show rate (91 days)
Medway appointment system	23rd November and 9th December 2015	None	Punctuality (193)
RIS	March 2016	None	number follow-up patients mammograms per day (31 days, 58 patients), follow-up patients ultrasounds per day (31 days, 69 patients)
PAS and RIS	January - March 2016	Removed outliers	Mammogram report times (20), Ultrasound report times (43)
RIS matched to GP referral information	January - March 2016	Standardised coding of characteristics and grouped into attributes	New patient ultrasounds, mammograms, biopsies and cancer diagnoses linked to characteristics (179 patients)

Table D.2: Typical initial consultation schedules

Monday		Tuesday		Wednesday	
Time	Slots	Time	Slots	Time	Slots
09:45	2	09:00	1	09:30	1
09:55	2	09:10	2	09:40	1
10:05	2	09:20	2	09:50	1
10:15	2	09:30	2	10:00	1
10:25	2	09:40	2	10:10	1
10:35	1	09:50	2	10:20	1
10:45	1	10:00	1	10:30	1
10:55	2	10:10	2	10:40	1
11:05	2	10:20	2	10:50	1
11:15	2	10:30	2	11:00	1
11:25	2	13:50	1	11:10	1
11:35	2	14:00	1	11:20	1
11:45	2	14:10	1	11:30	1
		14:20	1	11:40	1
		14:30	1	11:50	1
		14:40	1	12:00	1
		14:50	1	12:10	1
		15:00	1	12:20	1
				13:30	1
				13:40	1
				13:50	1
				14:00	1
				14:10	1
				14:20	1
				14:30	1

Table D.3: New patient labels in core simulation

Label name	Type	Where assigned	Description
lblVisit	Number	Register for appointment on work complete, Mammogram on work complete, Ultrasound on work complete	Shows whether first or second time to see clinician on this day.
lblRoute	Number	Assign appointment time route in after, See breast clinician (1 and 2) work complete, See breast clinician (1 and 2) route in after, Queue and register imaging request(s) on work complete, Decide mammogram today route in after, Decide ultrasound today route in after, Mammogram route in after, Ultrasound and biopsy if necessary route in after	Sets which route a patient should follow.
lblShelfLife	Number	Assign appointment time route in after, Assign ultrasound time on exit, Assign mammogram time on exit	Time from beginning of day until new patient arrives at reception. Also time from beginning of day until follow-up patient arrives at ultrasound/mammogram room.
lblArrivalTime	Number	Assign appointment time route in after	Sets a patient's registration time using their lblScheduledTime and lblLateness.
lblClinician	Number	Assign appointment time route in after	Assigned to clinician 1 or 2.

Continued on next page

Label name	Type	Where assigned	Description
lblID	Number	Assign appointment time route in after	Appointment slot number on a particular day.
lblScheduledTime	Number	Assign appointment time route in after	Scheduled appointment time.
lblLateness	Number	Register for appointment route in after	Number of minutes late for appointment (earliness is negative). Used to calculate lblArrivaltime.
lblConsultation	Text	Queue for first consult (1 and 2) on entry and Queue for results (1 and 2) on entry	Type of consultation with breast clinician (FirstConsult, CancerResults or NonCancerResults). Used to set clinician service time.
lblImaging	Text	See breast clinician (1 and 2) route in after	Imaging tests (N=None, U=Ultrasound only, M=Mammogram only or B=Both)
lblRand	Number	See breast clinician (1 and 2) route in after	A random number used to decide what imaging if any this patient will have. (Random numbers are also used when deciding which patients to biopsy, and which have cancer)
lblTimeSeenBySurgeon	Number	See breast clinician (1 and 2) route in after	Time arrived in breast clinician's room for clinical assessment. Used to calculate lblConsultDur and so value-added time.

Continued on next page

Label name	Type	Where assigned	Description
lblTimeSeenBySurgeon2	Number	See breast clinician (1 and 2) route in after	Time arrived in breast clinician's room for results. Used to calculate lblResDur and so value-added time
lblConsultDur	Number	See breast clinician (1 and 2) work complete	Duration of clinical assessment. Used to calculate value-added time.
lblEndTime	Number	See breast clinician (1 and 2) work complete	Time left clinic. Used to calculate total time and value-added time.
lblMammDur	Number	See breast clinician (1 and 2) work complete	Duration of mammogram. Used to calculate value-added time.
lblProductiveTime	Number	See breast clinician (1 and 2) work complete	Value-added time.
lblResDur	Number	See breast clinician (1 and 2) work complete	Duration of results consultation. Used to calculate value-added time.
lblStartTime	Number	See breast clinician (1 and 2) work complete	See definition of start time on page \pagerefdef:start.
lblTimeLeftSurgeon	Number	See breast clinician (1 and 2) work complete	End time of clinical assessment. Used to calculate lblConsultDur and so value-added time.
lblTimeLeftSurgeon2	Number	See breast clinician (1 and 2) work complete	End time of results consultation. Used to calculate lblResDur and so value-added time.

Continued on next page

Label name	Type	Where assigned	Description
lblTotalTime	Number	See breast clinician (1 and 2) work complete	Total time at clinic. Used to calculate total time spent at clinic for all patients on one day.
lblUltDur	Number	See breast clinician (1 and 2) work complete	Duration of ultrasound. Used to calculate value-added time.
lblStartQTime	Number	Queue at imaging reception on entry	Time started queuing at imaging reception. For debugging.
lblEndQTime	Number	Register imaging request on work complete	Time left imaging reception desk. For debugging.
lblDecisionHour	Number	Decide mammogram today before exit and Decide ultrasound today before exit	Hour of the day when deciding whether to mammogram/ultrasound today.
lblMammStart	Number	Mammogram route in after	Time entered mammogram room.
lblMammEnd	Number	Mammogram on work complete	Time left mammogram room.
lblUltrasoundType	Text	Wait for ultrasound on entry	Ultrasound type (Ult = ultrasound only or UltBiop= ultrasound and biopsy). Used to assign appropriate ultrasound process time distribution.
lblUltStart	Number	Ultrasound route in after	Time entered ultrasound room. Used to calculate lblUltDur.
lblUltEnd	Number	Ultrasound on work complete	Time left ultrasound room. Used for lblUltDur.

Continued on next page

Label name	Type	Where assigned	Description
lblWaitForPrint	Text	Mammogram work complete, Ultrasound work complete	Label-based distribution giving length of time "wait for receptionist to print". If journey to clinic takes longer than report time, report is ready (ReportReady), otherwise need to wait until report is ready plus 5 minutes for printing (ReportNotReady).
lblWaitResults	Number	Mammogram work complete, Ultrasound work complete	Sets waiting time for results to be ready = report time - travel time (5), or zero if this is negative
lblReturnTime	Number	Wait for results to be ready on entry	Time returned to clinic waiting room following imaging.
lblResultsReadyTime	Number	Wait for result consultation (1 and 2) on entry	Time results have been reported.

Table D.4: Labels used for follow-up patients in base simulation

Label name	Type	Where assigned	Description
lblType	Text	FU ultrasound, On entry, FU mammogram, On entry	Used to identify follow-up patients (F) so that they can be routed to appropriate exit.
lblMammTime	Number	Assign mammogram time, actions	Assigned time for follow-up patient to start queuing for mammogram.
lblUltTime	Number	Assign ultrasound time, actions	Assigned time for follow-up patient to start queuing for ultrasound.
lblUltrasoundType	Text	Assign ultrasound time, route in after	Ultrasound type (Ult = ultrasound only or UltBiop= ultrasound and biopsy). Used to assign appropriate ultrasound service time distribution.

Table D.5: Extra labels that are assigned in the extended simulation

Label name	Type	Where assigned	Description
lblCombo	Number	Triage patients, after loading work	A number referring to the combination of label values (lblNormRes, lblPredNorm, lblAgeUnder35) for this patient
lblNormRes	Number	Triage patients, after loading work	Actual result. 1 if have a normal result and 0 if have an abnormal result.
lblPredNorm	Number	Triage patients, after loading work	Predicted result. 2 if scorecard predicts a normal result and 0 if it predicts an abnormal result.
lblAgeUnder35	Number	Triage patients, after loading work	Age divide. 1 if aged under 35, 0 if aged 35 or over. This is used for specifying which tests and the order of tests.

Table D.6: Extra places that labels from core simulation are assigned in extended simulation

Label name	Type	Extra places assigned in extended simulation	Description
lblRoute	Number	Triage patients after loading work, assign imaging appointment time route in after	Sets which route a patient should follow.
lblRand	Number	Triage patients after loading work, assign imaging appointment time route in after	A random number used to decide which combination of label values, lblCombo, to assign to this patient. A new random number is generated to decide what imaging if any this patient will have.
lblShelfLife	Number	Assign imaging appointment time route in after	Time from beginning of day until "imaging first" patient joins imaging reception queue.
lblArrivalTime	Number	Assign imaging appointment time route in after	Sets a patient's arrival time at imaging reception queue using their lblScheduledTime and lblLateness.
lblClinician	Number	Assign imaging appointment time route in after	Assigned to clinician 1 or 2 ("imaging first" patients)
lblScheduledTime	Number	Assign imaging appointment time route in after	Scheduled appointment time ("imaging first" patients)
lblLateness	Number	Assign imaging appointment time route in after	Number of minutes late for appointment (earliness is negative). Used to calculate lblArrivaltime.
lblImaging	Text	Assign imaging appointment time route in after	Imaging tests (N=None, U=Ultrasound only, M=Mammogram only or B=Both)

Table D.7: Appointment times for patients visiting a clinician first. The times are listed in the order in which they are booked

Monday		Tuesday		Wednesday	
Time	Slots	Time	Slots	Time	Slots
09:45	2	09:30	2	09:30	1
10:00	2	09:45	2	09:45	1
10:15	2	10:00	2	10:00	1
10:30	2	10:15	2	10:15	1
10:45	2	10:30	2	10:30	1
11:00	2	13:45	1	10:45	1
11:15	2	14:00	1	11:00	1
11:30	2	14:15	1	11:15	1
11:45	2	14:30	1	11:30	1
		14:45	1	11:45	1
09:50	2	15:00	1	12:00	1
10:05	2			12:15	1
10:20	2	09:00	2	13:30	1
10:35	2	09:15	2	13:45	1
10:50	1	10:45	2	14:00	1
		11:00	2	14:15	1
		09:05	2	14:30	1
		09:20	2	14:45	1
		09:35	2		
		09:50	1	09:35	1
				09:50	1
				10:05	1
				10:20	1
				10:35	1
				10:50	1
				11:05	1
				11:20	1
				11:35	1

Table D.8: Appointment times for patients visiting imaging first. The times are listed in the order in which they are booked

Monday		Tuesday		Wednesday	
Time	Slots	Time	Slots	Time	Slots
09:00	1	09:00	1	09:00	1
09:10	1	09:10	1	09:10	1
09:20	1	09:20	1	09:20	1
09:30	1	09:30	1	09:30	1
09:40	1	09:40	1	09:40	1
09:50	1	10:00	1	10:00	1
10:00	1	10:20	1	10:20	1
10:20	1	10:40	1	10:40	1
10:40	1	11:00	1	11:00	1
11:00	1	11:20	1	11:20	1
11:20	1	11:40	1	11:40	1
11:40	1	12:00	1	12:00	1
		12:20	1	12:20	1
10:10	1	12:40	1	12:40	1
10:30	1	13:00	1	13:00	1
10:50	1	13:20	1	13:20	1
11:10	1	13:40	1	13:40	1
11:30	1	14:00	1	14:00	1
11:50	1	14:20	1	14:20	1
		14:40	1	14:40	1
09:05	1	15:00	1		
09:15	1			09:50	1
09:25	1	09:50	1	10:10	1
09:35	1	10:10	1	10:30	1
09:45	1	10:30	1	10:50	1
09:55	1	10:50	1	11:10	1
10:05	1	11:10	1	11:30	1
10:15	1	11:30	1	11:50	1
10:25	1	11:50	1		
		12:10	1		
		12:30	1		
		12:50	1		

Table D.9: Sensitivity analysis: Cancer results consultation duration

	Average wait for results consultation		Clinic efficiency	
Precision	10%		5%	
	Number runs	Mean [95% confidence interval]	Number runs	Mean [95% confidence interval]
A) Base: Triangular (min=15, mode=20, max=30)	39	22 [20,24] mins	36	0.27, [0.26, 0.28]
B) Exponential (mean=20)	32	23 [20, 25] mins	36	0.27, [0.25, 0.28]
C) Lognormal (mean=20, SD=3)	32	22 [20,24] mins	36	0.27, [0.25, 0.28]

Table D.10: Sensitivity analysis: Ultrasound and biopsy duration

	Average wait for ultrasound (no mammogram)		Average wait for ultrasound after mammogram		Clinic efficiency	
Precision	10%		10%		5%	
	Number runs	Mean [95% confidence interval]	Number runs	Mean [95% confidence interval]	Number runs	Mean [95% confidence interval]
A) Base: Empirical	127	59, [53,64] mins	123	36, [32,39] mins	36	0.27, [0.26, 0.28]
B) Uniform (min=17, max=35)	121	64, [58, 71] mins	137	39, [35,43] mins	44	0.26, [0.25,0.28]
C) Exponential (mean=26)	176	59, [53,65] mins	340	48, [43,52] mins	46	0.26 [0.25,0.27]

TABLE D.11: Sensitivity analysis: Time until mammogram report is ready

	Average wait for results consultation		Clinic efficiency	
Precision	10%		5%	
	Number runs	Mean [95% confidence interval]	Number runs	Mean [95% confidence interval]
A) Base: Beta (min= 0, max= 57, p=1.53, q=3.98)	39	22, [20,24] mins	36	0.27, [0.26, 0.28]
B) Pearson V (α=2.01, β=21.5)	30	22, [20,24] mins	36	0.27, [0.25, 0.28]

TABLE D.12: Sensitivity analysis: Time until ultrasound report is ready

	Average wait for results consultation		Clinic efficiency	
Precision	10%		5%	
	Number runs	Mean [95% confidence interval]	Number runs	Mean [95% confidence interval]
A) Base: Beta (min=0, max=50, p=1.73, q=7.69)	39	22 [20,24] mins	36	0.27, [0.26, 0.28]
B) Pearson V (α=1.38, β=4.91)	70	22 [20,24] mins	42	0.27, [0.26, 0.29]

Table D.13: Sensitivity analysis: Queue and service time at imaging reception

	Average wait for mammogram		Average wait for ultrasound (no mammogram)		Clinic efficiency	
Precision	10%		10%		5%	
	Number runs	Mean [95% confidence interval]	Number runs	Mean [95% confidence interval]	Number runs	Mean [95% confidence interval]
A) Base: Pearson V (α=2.18, β=7.3)	127	34, [30,37] mins	39	22, [20,24] mins	36	0.27, [0.26, 0.28]
B) Pearson V (α=2.47, β=14.7)	100	35, [32,39] mins	124	57, [52, 63] mins	35	0.27, [0.26, 0.28]
C) Pearson V (α=2.38, β=8.56)	125	34, [30, 40] mins	130	58, [52, 64] mins	38	0.27, [0.26, 0.28]

Table D.14: Sensitivity analysis: Follow-up patients

	Average wait for mammogram		Average wait for ultrasound (no mammogram)		Average wait for ultrasound (after mammogram)		Clinic efficiency	
Precision	10%		10%		10%		5%	
	Number runs	Mean [95% confidence interval]	Number runs	Mean [95% confidence interval]	Number runs	Mean [95% confidence interval]	Number runs	Mean [95% confidence interval]
A) Base: Imaging resources are shared with follow-up patients	127	34, [30,37] mins	39	22, [20,24] mins	123	36, [32,39] mins	36	0.27, [0.26, 0.28]
B) Imaging resources are dedicated to new patients	20	18, [17,20] mins	129	41, [36,45] mins	162	28, [25,31] mins	32	0.31, [0.29, 0.32]

Table D.15: Sensitivity analysis: Start time of clinic (Tuesday)

	Average total time at clinic		Clinic efficiency	
Precision	10%		5%	
	Number runs	Mean [95% confidence interval]	Number runs	Mean [95% confidence interval]
A) Base: Clinic starts at 9am	18	130, [118, 143] mins	36	0.27, [0.26, 0.28]
B) Clinic starts at 9:30am	16	149, [135, 163] mins	28	0.24, [0.23,0.25]
C) Clinic starts at 10am	13	175, [158, 191] mins	30	0.2, [0.19, 0.21]

Table D.16: Sensitivity analysis: Punctuality of patients attending imaging first

	Clinic efficiency		Average total time for patients seeing a clinician first		Average total time for patients attending imaging first		Average total time for all patients	
Precision	10%		10%		10%		10%	
	Number runs	Mean [95% confidence interval]	Number runs	Mean [95% confidence interval]	Number runs	Mean [95% confidence interval]	Number runs	Mean [95% confidence interval]
Alternative	16	0.17 [0.16,0.19]	20	151 [136,165] mins	31	176 [159,193] mins	19	165 [149, 181] mins
Base	20	0.18 [0.17, 0.2]	22	148 [134,162] mins	35	167 [150,183] mins	26	156 [141,171] mins

TABLE D.17: Sensitivity analysis: Results consultation durations for patients attending imaging first

	Clinic efficiency		Average total time	
Precision	10%		10%	
	Number runs	Mean [95% confidence interval]	Number runs	Mean [95% confidence interval]
Alternative	19	0.2 [0.18,0.22]	25	160 [145,176] mins
Base	20	0.18 [0.17, 0.2]	26	156 [141,171] mins

Glossary

- Abnormal result: Overall test result is 2 or higher, i.e. an abnormality was detected from mammogram, ultrasound or biopsy.

- Actual result: Indicates whether a patient's tests show a normal or abnormal result. It is a label in the extended simulation that takes values normal or abnormal.

- Adjuvant treatment: Treatment to lower the risk of cancer returning.

- Attribute: Possible categorical values of characteristics in the scorecards, e.g. *young* could be an attribute of the characteristic *age*.

- Average total time: The daily average of patients' total times at the clinic.

- Biopsy: This involves removing a sample of cells for analysis by a pathologist. Image-led biopsy uses ultrasound to locate the area from which to remove cells.

- Bootstrap: Sampling with replacement.

- Cancer: Disease that develops when abnormal cells keep dividing to form a lump, called a tumour, which grows into neighbouring tissue. Also known as malignancy.

- Chemotherapy: Treatment that uses drugs that are designed to destroy cancer cells.

- Classification: Statistical and data mining techniques that are used to predict class membership; unlike linear regression they predict categorical rather than continuous values. Examples are logistic regression and decision trees.

- Clinic efficiency: The average proportion of a patient's time at the clinic that is value-added on a particular day.

- Clinic end time: The time that the last patient finishes their results consultation.

- Clinician first: A possible route through clinic. Patients see a clinician who decides whether imaging is required.

- Core simulation: The simulation model that represents the current clinic set-up.
- Cut-off score: The threshold between low- and high-risk patients chosen for a scorecard. Patients with a score below the cut-off are predicted a normal result and those with a score at least as high as the cut-off score are predicted an abnormal result.
- Extended simulation: The simulation that models patients being routed according to their risk of having an abnormal result and the tests required.
- Follow-up patient: Patient returning for repeat appointment with a clinician or for imaging, for example patients undergoing treatment for breast diseases and cancer.
- Full scorecard: Predicts risk of abnormal result from seven referral characteristics.
- Imaging first: A possible route through clinic. Patients go straight for imaging tests then see a clinician afterwards for their results.
- Imaging tests: Mammogram, ultrasound.
- Initial consultation: First visit to breast clinician. During this consultation, the clinician discusses the new patient questionnaire, performs a clinical examination and decides which imaging tests, if any, to request.
- Interactive Grouping: Feature in SAS Enterprise Miner for generating scorecard attributes for continuous characteristics.
- Malignant: Cancerous.
- Mammogram: Low dose X-ray.
- New patient: Patient visiting the breast diagnostic clinic.
- Normal result: Overall test result is 1, i.e. normal results for all mammogram, ultrasound and biopsy tests that were performed, or no tests were done.
- One-stop clinic: Clinic that offers patients a series of diagnostic tests on a single day to confirm or exclude a cancer diagnosis.
- Overall test result: The highest (worst) score of mammogram, ultrasound and biopsy (for those tests that were done). A score of 1 means “normal”, 2 “benign”, 3 “unusual/uncertain but probably benign”, 4 “suspicious” and 5 “malignant”.
- Palliative treatment: Treatments for relieving symptoms (rather than curing cancer).

- Pathway: Sequence of services that a patient receives, i.e. an operational pathway, unless otherwise specified.
- Precision: When deciding on the number of simulation runs, this percentage specifies the desired distance of confidence intervals from the estimate of the mean.
- Predicted result: Indicates whether the scorecard predicts a normal or abnormal result. It is a label in the extended simulation that takes values normal or abnormal. In the simulation, patients are routed to imaging first, if an abnormal result is predicted, or to a clinician first, if a normal result is predicted.
- Radiographer: Staff member who performs mammograms.
- Radiologist: Staff member who performs ultrasounds as well as reporting both ultrasounds and mammograms.
- Radiotherapy: Treatment using radiation to destroy cancer cells. May be internal or external.
- Recurrence: Return of cancer.
- Referral characteristic: Patient characteristics obtained from GP referral information. Used in the scorecards to predict risk of an abnormal result.
- Report: To interpret mammogram and ultrasound images and dictate the findings.
- Results consultation: Visit to clinician after imaging tests to discuss the imaging results.
- Sensitivity: also known as the true positive rate. a) the proportion of cancers that are classified correctly e.g. by screening test or classification model. b) For our logistic regression scorecards, the proportion of normal results that were correctly predicted.
- Simple scorecard: Predicts risk of abnormal result from two referral characteristics.
- Specificity: also known as the true negative rate. a) the proportion of non-cancers that are classified correctly e.g. by screening test or classification model. b) For our logistic regression scorecards, the proportion of abnormal results that were correctly predicted.
- Staging: tests for determining the size of the cancer and how much it has spread.
- Total time: Total time that patient spends at clinic from registration until leaving.

- Ultrasound: imaging scan using ultrasonic waves.
- Unilateral pain: One-sided pain.
- Value-added time: Time that contributes to a patient's care, i.e. time spent in imaging tests or with a clinician as opposed to time spent waiting, queueing and undertaking administrative tasks.
- Weights of evidence (WoE): The strength of evidence that patients with a particular attribute will have an abnormal result.

Acronyms

- ANN: Artificial neural network
- AUROC: Area under a receiver operating characteristic curve
- DES: Discrete-event simulation
- IMRT: Intensity modulated radiotherapy
- KS: Kolmogorov-Smirnov
- LDA: Linear discriminant analysis
- MSM: Multisurface separation
- QDA: Quadratic discriminant analysis
- PACS: Picture Archiving and Communications System
- RIS: Radiology Information System
- ROC: Receiver operating characteristic curve
- SVM: Support vector machine
- WoE: Weights of evidence

References

Abbass, H. A. (2002). An evolutionary artificial neural networks approach for breast cancer diagnosis. *Artificial Intelligence in Medicine*, 25(3):265–281.

Adeyemi, S., Demir, E., and Chaussalet, T. (2013). Towards an evidence-based decision making healthcare system management: Modelling patient pathways to improve clinical outcomes. *Decision Support Systems*, 55(1):117–125.

Agresti, A. (2013). *Categorical data analysis*. Wiley series in probability and statistics. John Wiley & Sons, Hoboken, New Jersey, 3rd edition.

Ahn, S. K., Han, W., Moon, H. G., Kim, M. K., Noh, D.-Y., Jung, B.-w., et al. (2018). Management of benign papilloma without atypia diagnosed at ultrasound-guided core needle biopsy: Scoring system for predicting malignancy. *European Journal of Surgical Oncology*, 44(1):53–58.

Alagoz, O., Ayer, T., and Erenay, F. S. (2011). Operations Research models for cancer screening. In Cochran, J. J., Cox, L. A. J., Keskinocak, P., Kharoufeh, J. P., and Smith, J. C., editors, *Wiley Encyclopedia of Operations Research and Management Science*. John Wiley & Sons, Hoboken, NJ.

Alagoz, O., Chhatwal, J., and Burnside, E. S. (2013). Optimal policies for reducing unnecessary follow-up mammography exams in breast cancer diagnosis. *Decision Analysis*, 10(3):200–224.

Alam, M. S., Hossain, M. A., Algoul, S., Majumader, M. A. A., Al-Mamun, M. A., Sexton, G., et al. (2013). Multi-objective multi-drug scheduling schemes for cell cycle specific cancer treatment. *Computers and Chemical Engineering*, 58:14–32.

Arrospide, A., Rue, M., van Ravesteyn, N. T., Comas, M., Larrañaga, N., Sarriugarte, G., et al. (2015). Evaluation of health benefits and harms of the breast cancer screening programme in the Basque Country using discrete event simulation. *BMC Cancer*, 15(1):671.

Ayer, T. (2015). Inverse optimization for assessing emerging technologies in breast cancer screening. *Annals of Operations Research*, 230(1):57–85.

Ayer, T., Alagoz, O., and Stout, N. K. (2012). OR Forum–A POMDP approach to personalize mammography screening decisions. *Operations Research*, 60(5):1019–1034.

Ayer, T., Alagoz, O., Stout, N. K., and Burnside, E. S. (2016). Heterogeneity in women's adherence and its role in optimal breast cancer screening policies. *Management Science*, 62(5):1339–1362.

Baesler, F. F. and Sepúlveda, J. A. (2001). *Multi-objective simulation optimization for a cancer treatment center.* In: Peters, BA, Smith, JA, Medeiros, DA And Rohrer, MW (Eds), Proceedings of the 2001 Winter Simulation Conference. Piscataway: NJ pp 1405-1411.

Bailey, N. (1952). A study of queues and appointment systems in hospital outpatient departments with special reference to waiting times. *Journal of the Royal Statistical Society*, 14:185–199.

Baker, R. D. and Atherill, P. L. (2002). Improving appointment scheduling for medical screening. *IMA Journal of Management Mathematics*, 13(4):225–243.

Banks, J., Carson II, J. S., Nelson, B. L., and Nicol, D. M. (2010a). Input modeling. In Fabrycky, W. J. and Mize, J. H., editors, *Discrete-event system simulation*, chapter 9, pages 331–384. Pearson Prentice Hall, New Jersey, 5th edition.

Banks, J., Carson II, J. S., Nelson, B. L., and Nicol, D. M. (2010b). Verification, calibration and validation of simulation models. In Fabrycky, W. J. and Mize, J. H., editors, *Discrete-event system simulation*, chapter 10, pages 385–414. Pearson Prentice Hall, New Jersey, 5th edition.

Bayer, S., Petsoulas, C., Cox, B., Honeyman, A., and Barlow, J. (2010). Facilitating stroke care planning through simulation modelling. *Health Informatics Journal*, 16(2):129–143.

Berg, B., Murr, M., Chermak, D., Woodall, J., Pignone, M., Sandler, R., et al. (2013). Estimating the cost of no-shows and evaluating the effects of mitigation strategies. *Medical Decision Making*, 33(8):976–85.

Bhattacharjee, P. and Ray, P. K. (2014). Patient flow modelling and performance analysis of healthcare delivery processes in hospitals: A review and reflections. *Computers and Industrial Engineering*, 78:299–312.

Bhattacharjee, P. and Ray, P. K. (2016). Simulation modelling and analysis of appointment system performance for multiple classes of patients in a hospital: A case study. *Operations Research for Health Care*, 8:71–84.

Bikker, I. A., Kortbeek, N., van Os, R. M., and Boucherie, R. J. (2015). Reducing access times for radiation treatment by aligning the doctor's schemes. *Operations Research for Health Care*, 7:111–121.

Bortfeld, T. (2006). IMRT: A review and preview. *Physics in Medicine and Biology*, 51(13):R363–R379.

Brailsford, S. and Vissers, J. (2011). OR in healthcare: A European perspective. *European Journal of Operational Research*, 212(2):223–234.

Brailsford, S. C., Harper, P. R., Patel, B., and Pitt, M. (2009). An analysis of the academic literature on simulation and modelling in health care. *Journal of Simulation*, 3:130–140.

Brailsford, S. C., Harper, P. R., and Sykes, J. (2012). Incorporating human behaviour in simulation models of screening for breast cancer. *European Journal of Operational Research*, 219(3):491–507.

Brighton and Sussex University Hospitals NHS Trust (2017). *Your first visit to the Park Centre for Breast Care*. https://www.bsuh.nhs.uk/wp-content/uploads/sites/5/2016/09/Your-first-visit-to-the-Park-Centre-for-Breast-Care.pdf (accessed May 31, 2018).

Burr, J. M., Botello-Pinzon, P., Takwoingi, Y., Hernandez, R., Vazquez-Montes, M., Elders, A., et al. (2012). Surveillance for ocular hypertension: An evidence synthesis and economic evaluation. *Health Technology Assessment*, 16(29):1–271.

Cabrera, G. G., Ehrgott, M., Mason, A., and Philpott, A. (2014). Multi-objective optimisation of positively homogeneous functions and an application in radiation therapy. *Operations Research Letters*, 42(4):268–272.

Çakir, A. and Demirel, B. (2011). A software tool for determination of breast cancer treatment methods using data mining approach. *Journal of Medical Systems*, 35(6):1503–1511.

Cancer Research UK (2011). *People fear cancer more than other serious illness*. Press release http://www.cancerresearchuk.org/about-us/cancer-news/press-release/2011-08-15-people-fear-cancer-more-than-other-serious-illness (accessed October 11, 2016).

Cancer Research UK (2016a). *About cancer*. http://www.cancerresearchuk.org/about-cancer/ (accessed November 12, 2016).

Cancer Research UK (2016b). *Breast cancer diagnosis and treatment statistics*. http://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/breast-cancer/diagnosis-and-treatment#heading-Zero (accessed October 11, 2016).

Cancer Research UK (2016c). *Breast cancer statistics.* http://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/breast-cancer#heading-Zero (accessed October 11, 2016).

Cancer Research UK (2016d). *Testing times to come? An evaluation of pathology capacity across the UK.* https://www.cancerresearchuk.org/sites/default/files/testing_times_to_come_nov_16_cruk.pdf (accessed May 31, 2018).

Cannon, J. W., Mueller, U. A., Hornbuckle, J., Larson, A., Simmer, K., Newnham, J. P., et al. (2013). Economic implications of poor access to antenatal care in rural and remote Western Australian Aboriginal communities: An individual sampling model of pregnancy. *European Journal of Operational Research*, 226(2):313–324.

Cant, P. J. and Yu, D. S. L. (2000). Impact of the '2 week wait' directive for suspected cancer on service provision in a symptomatic breast clinic. *British Journal of Surgery*, 87(8):1082–1086.

Cario, M. C. and Nelson, B. L. (1997). Modeling and generating random vectors with arbitrary marginal distributions and correlation matrix. Technical report, Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, Illinois.

Castro, E. and Petrovic, S. (2012). Combined mathematical programming and heuristics for a radiotherapy pre-treatment scheduling problem. *Journal of Scheduling*, 15(3):333–346.

Cayirli, T. and Veral, E. (2003). Outpatient scheduling in health care: A review of literature. *Production and Operations Management*, 12(4):519–549.

Ceglowski, R., Churilov, L., and Wasserthiel, J. (2006). Combining data mining and discrete event simulation for a value-added view of a hospital emergency department. *Journal of the Operational Research Society*, 58(2):246–254.

Censor, Y. and Unkelbach, J. (2012). From analytic inversion to contemporary IMRT optimization: Radiation therapy planning revisited from a mathematical perspective. *Physica Medica*, 28(2):109–118.

Chan, T. C., Mahmoudzadeh, H., and Purdie, T. G. (2014). A robust-CVaR optimization approach with application to breast cancer therapy. *European Journal of Operational Research*, 238(3):876–885.

Chemweno, P., Thijs, V., Pintelon, L., and van Horenbeek, A. (2014). Discrete event simulation case study: Diagnostic path for stroke patients in a stroke unit. *Simulation Modelling Practice and Theory*, 48:45–57.

Chhatwal, J., Alagoz, O., and Burnside, E. S. (2010). Optimal breast biopsy decision-making based on mammographic features and demographic factors. *Operations Research*, 58(6):1577–1591.

Coelli, F. C., Ferreira, R. B., Almeida, R. M. V. R., and Pereira, W. C. A. (2007). Computer simulation and discrete-event models in the analysis of a mammography clinic patient flow. *Computer Methods and Programs in Biomedicine*, 87(3):201–207.

Colditz, G. A. and Rosner, B. (2000). Cumulative risk of breast cancer to age 70 years according to risk factor status: Data from the Nurses' Health Study. *American Journal of Epidemiology*, 152(10):950–964.

Conforti, D., Guerriero, F., and Guido, R. (2010). Non-block scheduling with priority for radiotherapy treatments. *European Journal of Operational Research*, 201(1):289–296.

Cooper, K., Davies, R., Roderick, P., Chase, D., and Raftery, J. (2002). The development of a simulation model of the treatment of coronary heart disease. *Health Care Management Science*, 5(4):259–267.

Cotteels, C., Peeters, D., Coucke, P. A., and Thomas, I. (2012). Localisation des centres de radiothérapie : Une analyse géographique exploratoire pour la Belgique. *Cancer Radiothérapie*, 16(7):604–612.

Cramér, H. (1946). *Mathematical Methods of Statistics*. Princeton University Press, Princeton.

Crane, G. J., Kymes, S. M., Hiller, J. E., Casson, R., and Karnon, J. D. (2013). Development and calibration of a constrained resource health outcomes simulation model of hospital-based glaucoma services. *Health Systems*, 2(3):181–197.

Crawford, E. A., Parikh, P. J., Kong, N., and Thakar, C. V. (2014). Analyzing discharge strategies during acute care: A discrete-event simulation study. *Medical Decision Making*, 34(2):231–241.

De Boeck, L., Beliën, J., and Egyed, W. (2014). Dose optimization in high-dose-rate brachytherapy: A literature review of quantitative models from 1990 to 2010. *Operations Research for Health Care*, 3(2):80–90.

Delen, D. (2009). Analysis of cancer data: A data mining approach. *Expert Systems*, 26(1):100–112.

Delen, D., Walker, G., and Kadam, A. (2005). Predicting breast cancer survivability: A comparison of three data mining methods. *Artificial Intelligence in Medicine*, 34(2):113–127.

Devroye, L. (1986). *Non-uniform random variate generation*. Springer-Verlag, New York, NY.

Dias, J., Rocha, H., Ferreira, B., and do Carmo Lopes, M. (2014). A genetic algorithm with neural network fitness function evaluation for IMRT beam angle optimization. *Central European Journal of Operations Research*, 22(3):431–455.

Eatock, J., Clarke, M., Picton, C., and Young, T. (2011). Meeting the four-hour deadline in an A&E department. *Journal of Health Organization and Management*, 25(6):606–624.

Ehrgott, M., Güler, Ç., Hamacher, H. W., and Shao, L. (2010). Mathematical optimization in intensity modulated radiation therapy. *Annals of Operations Research*, 175(1):309–365.

Ekaette, E., Lee, R. C., Kelly, K.-L., and Dunscombe, P. (2006). A Monte Carlo simulation approach to the characterization of uncertainties in cancer staging and radiation treatment decisions. *Journal of the Operational Research Society*, 58(2):177–185.

Elliott, J., Staetsky, L., Smith, P. W. F., Foster, C. L., Maher, E. J., and Corner, J. (2011). The health and well-being of cancer survivors in the UK: Findings from a population-based survey. *British Journal of Cancer*, 105:S11–S20.

Elliss-Brookes, L., McPhail, S., Ives, A., Greenslade, M., Shelton, J., Hiom, S., et al. (2012). Routes to diagnosis for cancer - Determining the patient journey using multiple routine data sets. *British Journal of Cancer*, 107(8):1220–1226.

European Pathway association (2016). *Care pathways.* http://e-p-a.org/care-pathways/ (accessed July 22, 2016).

García-Laencina, P. J., Abreu, P. H., Abreu, M. H., and Afonoso, N. (2015). Missing data imputation on the 5-year survival prediction of breast cancer patients with unknown discrete values. *Computers in Biology and Medicine*, 59:125–133.

Geer Mountain Software Corporation (2016). *Stat::Fit.* Connecticut. http://www.geerms.com/files/114225421.pdf (accessed May 23, 2018).

Gillespie, J., McClean, S., Garg, L., Barton, M., Scotney, B., and Fullerton, K. (2016). A multi-phase DES modelling framework for patient-centred care. *Journal of the Operational Research Society*, 67(10):1239–1249.

Glasziou, P., Moynihan, R., Richards, T., and Godlee, F. (2013). Too much medicine; too little care. *British Medical Journal*, 347:1–2.

Gu, W., Wang, X., and McGregor, S. E. (2010). Optimization of preventive health care facility locations. *International Journal of Health Geographics*, 9(1):17.

Gunal, M. M. (2012). A guide for building hospital simulation models. *Health Systems*, 1(1):17–25.

Gupta, D. and Denton, B. (2008). Appointment scheduling in health care: Challenges and opportunities. *IIE Transactions*, 40(9):800–819.

Guy's and St Thomas' NHS Foundation Trust (2011). *The breast clinic.* https://guysandstthomas.nhs.uk/resources/patient-information/radiology/Breastclinic.pdf (accessed May 31, 2018).

Haase, K. and Müller, S. (2015). Insights into clients' choice in preventive health care facility location planning. *OR Spectrum*, 37(1):273–291.

Hahn-Goldberg, S., Carter, M. W., Beck, J. C., Trudeau, M., Sousa, P., and Beattie, K. (2014). Dynamic optimization of chemotherapy outpatient scheduling with uncertainty. *Health Care Management Science*, 17(4):379–392.

Harper, P. R. (2002). A framework for operational modelling of hospital resources. *Health Care Management Science*, 5(3):165–173.

Harper, P. R. and Gamlin, H. M. (2003). Reduced outpatient waiting times with improved appointment scheduling: A simulation modelling approach. *OR Spectrum*, 25(2):207–222.

Harper, P. R., Sayyad, M. G., De Senna, V., Shahani, A. K., Yajnik, C. S., and Shelgikar, K. M. (2003). A systems modelling approach for the prevention and treatment of diabetic retinopathy. *European Journal of Operational Research*, 150(1):81–91.

Harvey, J., Down, S., Bright-Thomas, R., Winstanley, J., and Bishop, H. (2014). *Breast disease management: A multidisciplinary manual.* Oxford Care Manuals. Oxford University Press, Oxford.

Haybittle, J. L., Blamey, R. W., Elston, C. W., Johnson, J., Doyle, P. J., Campbell, F. C., et al. (1982). A prognostic index in primary breast cancer. *British Journal of Cancer*, 45(3):361–366.

Healthy London Partnership (2018). *Pan London suspected cancer referral forms version 2.* https://www.healthylondon.org/suspected-cancer-referrals/ (accessed May 31, 2018).

Heidenberger, K. (1996). Strategic investment in preventive health care: Quantitative modelling for programme selection and resource allocation. *OR Spektrum*, 18(1):1–14.

Herbert, R. D. (2014). Cohort studies of aetiology and prognosis: They're different. *Journal of Physiotherapy*, 60(4):241–244.

Hippisley-Cox, J. and Coupland, C. (2013). Symptoms and risk factors to identify women with suspected cancer in primary care: Derivation and validation of an algorithm. *British Journal of General Practice*, 63(606):11–21.

Hippisley-Cox, J. and Coupland, C. (2015). Development and validation of risk prediction algorithms to estimate future risk of common cancers in men and women: Prospective cohort study. *BMJ Open*, 5(3):e007825.

Hirji, K. F. (2005). *Exact analysis of discrete data.* Chapman and Hall/CRC, New York.

Holder, A. and LLagostera, D. (2008). Optimal treatments for photodynamic therapy. *4or*, 6(2):167–182.

Hosmer, D. W., Hosmer, T., Le Cessie, S., and Lemeshow, S. (1997). A comparison of goodness-of-fit tests for the logistic regression model. *Statistics in Medicine*, 16(9):965–980.

Hosmer, D. W., Lemeshow, S., and Sturdivant, R. X. (2013). *Applied Logistic Regression.* Wiley, New York, 3rd edition.

Huang, Y.-L. and Hanauer, D. A. (2016). Time dependent patient no-show predictive modelling development. *International Journal of Health Care Quality Assurance*, 29(4):475–488.

Huang, Y.-L., Zuniga, P., and Marcak, J. (2013). A cost-effective urgent care policy to improve patient access in a dynamic scheduled clinic setting. *Journal of the Operational Research Society*, 65(5):763–776.

Hulshof, P. J. H., Kortbeek, N., Boucherie, R. J., Hans, E. W., and Bakker, P. J. M. (2012). Taxonomic classification of planning decisions in health care: A structured review of the state of the art in OR/MS. *Health Systems*, 1(2):129–175.

Hung, M. S., Shanker, M., and Hu, M. Y. (2002). Estimating breast cancer risks using neural networks. *Journal of the Operational Research Society*, 53(2):222–231.

Isken, M. W. and Rajagopalan, B. (2002). Data mining to support simulation modeling of patient flow in hospitals. *Journal of Medical Systems*, 26(2):179–197.

Jalalimanesh, A., Haghighi, H. S., Ahmadi, A., and Soltani, M. (2017). Simulation-based optimization of radiotherapy: Agent-based modeling and reinforcement learning. *Mathematics and Computers in Simulation*, 133:235–248.

Jin, S. Y., Won, J. K., Lee, H., and Choi, H. J. (2012). Construction of an automated screening system to predict breast cancer diagnosis and prognosis. *Basic and Applied Pathology*, 5(1):15–18.

Jonsdottir, T., Hvannberg, E. T., Sigurdsson, H., and Sigurdsson, S. (2008). The feasibility of constructing a Predictive Outcome Model for breast cancer using the tools of data mining. *Expert Systems with Applications*, 34(1):108–118.

Keogh, B. (2009). *Letter regarding* Operational Standards for the Cancer Waiting Times Commitments. Available from: http://webarchive.nationalarchives.gov.uk/20130107105354/http://www.dh.gov.uk/prod_consum_dh/groups/dh_digitalassets/documents/digitalasset/dh_103431.pdf (accessed December 4, 2016).

Khanna, S., Sier, D., Boyle, J., and Zeitz, K. (2016). Discharge timeliness and its impact on hospital crowding and emergency department flow performance. *Emergency Medicine Australasia*, 28(2):164–170.

Klassen, K. J. and Rohleder, T. R. (1996). Scheduling outpatient appointments in a dynamic environment. *Journal of Operations Management*, 14(2):83–101.

Klassen, K. J. and Yoogalingam, R. (2018). *Appointment scheduling in multi-stage outpatient clinics*. https://doi.org/10.1007/s10729-018-9434-x.

Knudsen, A. B., McMahon, P. M., and Gazelle, G. S. (2007). Use of modeling to evaluate the cost-effectiveness of cancer screening programs. *Journal of Clinical Oncology*, 25(2):203–208.

Koleva-Kolarova, R. G., Zhan, Z., Greuter, M. J., Feenstra, T. L., and De Bock, G. H. (2015). Simulation models in population breast cancer screening: A systematic review. *The Breast*, 24(4):354–363.

Koo, M. M., von Wagner, C., Abel, G. A., McPhail, S., Rubin, G. P., and Lyratzopoulos, G. (2017). Typical and atypical presenting symptoms of breast cancer and their associations with diagnostic intervals: Evidence from a national audit of cancer diagnosis. *Cancer Epidemiology*, 48:140–146.

Laganga, L. R. and Lawrence, S. R. (2007). Clinic overbooking to improve patient access and increase provider productivity. *Decision Sciences*, 38(2):251–276.

Law, A. M. and Kelton, W. D. (2000). *Simulation modeling and analysis*. Mcgraw-Hill Higher Education, Boston, 3rd edition.

Legrain, A., Fortin, M.-A., Lahrichi, N., and Rousseau, L.-M. (2015). Online stochastic optimization of radiotherapy patient scheduling. *Health Care Management Science*, 18(2):110–123.

Lehaney, B., Clarke, S. A., and Paul, R. J. (1999). A case of intervention in an outpatients department. *Journal of the Operational Research Society*, 50(9):877–891.

Leonard, P. (2014). *Personal communication with Dr Pauline Leonard, Lead Cancer Clinician at the Whittington Hospital.*

Li, J., Dong, M., Ren, Y., and Yin, K. (2015). How patient compliance impacts the recommendations for colorectal cancer screening. *Journal of Combinatorial Optimization*, 30(4):920–937.

Liang, B., Turkcan, A., Ceyhan, M. E., and Stuart, K. (2015). Improvement of chemotherapy patient flow and scheduling in an outpatient oncology clinic. *International Journal of Production Research*, 53(24):7177–7190.

Lichman, M. (2016). *UCI Machine Learning Repository.* http://archive.ics.uci.edu/ml/ (accessed July 22, 2016).

Lim, G. J., Mobasher, A., Bard, J. F., and Najjarbashi, A. (2016). Nurse scheduling with lunch break assignments in operating suites. *Operations Research for Health Care*, 10:35–48.

Lord, J., Willis, S., Eatock, J., Tappenden, P., Trapero-Bertran, M., Miners, A., et al. (2013). Economic modelling of diagnostic and treatment pathways in National Institute for Health and Care Excellence clinical guidelines: The Modelling Algorithm Pathways in Guidelines (MAPGuide) project. *Health Technology Assessment*, 17(58):1–150.

Lundin, M., Lundin, J., Burke, H. B., Toikkanen, S., and Pylkkänen, L. (1999). Artificial neural networks applied to survival prediction in breast cancer. *Oncology*, 57:281–286.

Macmillan Cancer Support (2013). *Mortality trends 2013 executive summary final.* http://www.macmillan.org.uk/documents/aboutus/newsroom/mortality-trends-2013-executive-summary-final.pdf (accessed November 12, 2016).

Macmillan Cancer Support (2016). *Information and support.* http://www.macmillan.org.uk/information-and-support/index.html (accessed November 12, 2016).

Macmillan Cancer Support (July 2014). The rich picture - people with cancer. Technical report. http://www.macmillan.org.uk/documents/aboutus/research/richpictures/update/rp-people-with-cancer.pdf.

Madadi, M., Zhang, S., and Henderson, L. M. (2015). Evaluation of breast cancer mammography screening policies considering adherence behavior. *European Journal of Operational Research*, 247(2):630–640.

Mahmoudzadeh, H., Purdie, T. G., and Chan, T. C. Y. (2016). Constraint generation methods for robust optimization in radiation therapy. *Operations Research for Health Care*, 8:85–90.

Mangasarian, O. L., Street, W. N., and Wolberg, H. (1995). Breast cancer diagnosis and prognosis via linear programming. *Operations Research*, 43(4):570–577.

Masselink, I. H., van der Mijden, T. L., Litvak, N., and Vanberkel, P. T. (2012). Preparation of chemotherapy drugs: Planning policy for reduced waiting times. *Omega*, 40(2):181–187.

Matta, M. E. and Patterson, S. S. (2007). Evaluating multiple performance measures across several dimensions at a multi-facility outpatient center. *Health Care Management Science*, 10(2):173–194.

Mays, E. (2001). *Handbook of credit scoring.* Glenlake Publishing Company, Chicago.

Medway NHS Foundation Trust (2013). *Breast care unit.* https://www.medway.nhs.uk/downloads/services/breast-care/PIL00000924-1%20Symptomatic%20Breast%20Clinic.pdf (accessed May 31, 2018).

Mimouni, M., Lecuru, F., Rouzier, R., Lotersztajn, N., Heitz, D., Cohen, J., et al. (2015). Reexcision for positive margins in breast cancer: A predictive score of residual disease. *Surgical Oncology*, 24(3):129–135.

Mobasher, A., Lim, G., Bard, J. F., and Jordan, V. (2011). Daily scheduling of nurses in operating suites. *IIE Transactions on Healthcare Systems Engineering*, 1(4):232–246.

Monks, T., Worthington, D., Allen, M., Pitt, M., Stein, K., and James, M. A. (2016). A modelling tool for capacity planning in acute and community stroke services. *BMC Health Services Research*, 16:1–8.

Mutlu, S., Benneyan, J., Terrell, J., Jordan, V., and Turkcan, A. (2015). A co-availability scheduling model for coordinating multi-disciplinary care teams. *International Journal of Production Research*, 53(24):7226–7237.

National Breast Cancer Foundation (2016). *Lymph node removal & lymphedema.* www.nationalbreastcancer.org/breast-cancer-lymph-node-removal (accessed November 12, 2016).

National Cancer Institute (2016a). *Cancer treatment.* www.cancer.gov/about-cancer/treatment (accessed November 12, 2016).

National Cancer Institute (2016b). *NCI dictionary of cancer terms.* www.cancer.gov/publications/dictionaries/cancer-terms (accessed November 12, 2016).

National Health Service (2017). *Be clear on cancer.* https://www.nhs.uk/be-clear-on-cancer/symptoms/breast-cancer (accessed October 20, 2017).

National Institute for Health and Care Excellence (2015). Suspected cancer: Recognition and referral (NG12).

NHS (2016). *Cancer screening programmes.* www.cancerscreening.nhs.uk (accessed November 12, 2016).

NHS England (2016). *Provider-based cancer waiting times.* https://www.england.nhs.uk/statistics/statistical-work-areas/cancer-waiting-times/ (accessed July 5, 2016).

NHS Kent and Medway Cancer Collaborative (2018). *GP referral proformas.* http://www.kmcc.nhs.uk/tumour-sites/gp-referral-proformas/ (accessed May 31, 2018).

Nuffield Trust (2016). *NHS spending on the top three disease categories in England.* http://www.nuffieldtrust.org.uk/data-and-charts/nhs-spending-top-three-disease-categories-england (accessed October 11, 2016).

O'Mahony, J. F., van Rosmalen, J., Mushkudiani, N. A., Goudsmit, F.-W., Eijkemans, M. J. C., Heijnsdijk, E. A. M., et al. (2015). The influence of disease risk on the optimal time interval between screens for the early detection of cancer: A mathematical approach. *Medical Decision Making*, 35(2):183–195.

Pendharkar, P. C., Rodger, J. A., Yaverbaum, G. J., Herman, N., and Benner, M. (1999). Association, statistical, mathematical and neural approaches for mining breast cancer patterns. *Expert Systems with Applications*, 17(3):223–232.

Petrovic, D., Morshed, M., and Petrovic, S. (2011). Multi-objective genetic algorithms for scheduling of radiotherapy treatments for categorised cancer patients. *Expert Systems with Applications*, 38(6):6994–7002.

Pierskalla, W. P. and Brailer, D. J. (1994). Applications of operations research in health care delivery. *Handbooks in Operations Research and Management Science*, 6:469–505.

Pilgrim, H., Tappenden, P., Chilcott, J., Bending, M., Trueman, P., Shorthouse, A., et al. (2008). The costs and benefits of bowel cancer service developments using discrete event simulation. *Journal of the Operational Research Society*, 60(10):1305–1314.

Quick-R (2017). *Generalized Linear Models.* https://www.statmethods.net/advstats/glm.html (accessed April 01, 2017).

Ravdin, P. M., Siminoff, L. A., Davis, G. J., Mercer, M. B., Hewlett, J., Gerson, N., et al. (2001). Computer program to assist in making decisions about adjuvant therapy for women with early breast cancer. *Journal of Clinical Oncology*, 19(4):980–991.

Razavi, A. R., Gill, H., Åhlfeldt, H., and Shahsavar, N. (2007). Predicting metastasis in breast cancer: Comparing a decision tree with domain experts. *Journal of Medical Systems*, 31(4):263–273.

Referral support service (2018). *2WW forms.* http://www.valeofyorkccg.nhs.uk/rss/index.php?id=2ww-referral-forms (accessed May 31, 2018).

Revankar, N., Ward, A. J., Pelligra, C. G., Kongnakorn, T., Fan, W., and LaPensee, K. T. (2014). Modeling economic implications of alternative treatment strategies for acute bacterial skin and skin structure infections. *Journal of Medical Economics*, 17(10):730–740.

Roberts, S., Wang, L., Klein, R., Ness, R., and Dittus, R. (2007). Development of a simulation model of colorectal cancer. *ACM Transactions on Modeling and Computer Simulation*, 18(1):4.

Rue, M., Carles, M., Vilaprinyo, E., Pla, R., Forne, C., Roso, A., et al. (2012). *How to optimize population screening programs for breast cancer using mathematical models.* In: Uchiyama, N and Do Nascscimento, MZ (Eds), Mammography - Recent Advances. InTech: Chap 3 pp 47-71.

Ryu, J. M., Lee, S. K., Kim, J. Y., Yu, J., Kim, S. W., Lee, J. E., et al. (2017). Predictive factors for nonsentinel lymph node metastasis in patients with positive sentinel lymph nodes after neoadjuvant chemotherapy: nomogram for predicting nonsentinel lymph node metastasis. *Clinical Breast Cancer*, 17(7):550–558.

Ryu, Y. U., Chandrasekaran, R., and Jacob, V. S. (2007). Breast cancer prediction using the isotonic separation technique. *European Journal of Operational Research*, 181(2):842–854.

Santibáñez, P., Aristizabal, R., Puterman, M. L., Chow, V. S., Huang, W., Kollmannsberger, C., et al. (2012). Operations Research methods improve chemotherapy patient appointment scheduling. *Joint Commission Journal on Quality and Patient Safety*, 38(12):541–553.

Santibáñez, P., Chow, V. S., French, J., Puterman, M. L., and Tyldesley, S. (2009). Reducing patient wait times and improving resource utilization at British Columbia Cancer Agency's ambulatory care unit through simulation. *Health Care Management Science*, 12(4):392–407.

Santos, S. P., Belton, V., and Howick, S. (2007). Enhanced performance measurement using OR: A case study. *Journal of the Operational Research Society*, 59(6):762–775.

Saremi, A., Jula, P., Elmekkawy, T., and Wang, G. G. (2013). Appointment scheduling of outpatient surgical services in a multistage operating room department. *International Journal of Production Economics*, 141(2):646–658.

Saremi, A., Jula, P., ElMekkawy, T., and Wang, G. G. (2015). Bi-criteria appointment scheduling of patients with heterogeneous service sequences. *Expert Systems with Applications*, 42(8):4029–4041.

SAS (2009). Building credit scorecards using credit scoring for SAS Enterprise Miner, a SAS best practices paper. Technical report, SAS Press, Cary, NC.

Sauré, A., Patrick, J., Tyldesley, S., and Puterman, M. L. (2012). Dynamic multi-appointment patient scheduling for radiation therapy. *European Journal of Operational Research*, 223(2):573–584.

Sauven, P. (2001). Specialists, not GPs, may be best qualified to assess urgency. *British Medical Journal*, 323(7317):864–865.

Sense about science (2017). *Making sense of screening.* http://senseaboutscience.org/r (accessed October 20, 2017).

Shi, J., Alagoz, O., Erenay, F. S., and Su, Q. (2014). A survey of optimization models on cancer chemotherapy treatment planning. *Annals of Operations Research*, 221(1):331–356.

Slack, N., Brandon-Jones, A., and Johnston, R. (2013). *Operations management.* Pearson Education Ltd, Harlow, UK, 7th edition.

Soerjomataram, I., de Vries, E., Pukkala, E., and Coebergh, J. W. (2007). Excess of cancers in Europe: A study of eleven major cancers amenable to lifestyle change. *International Journal of Cancer*, 120(6):1336–1343.

Sree, S. V., Ng, E. Y. K., and Acharya, U. R. (2010). Data mining approach to evaluating the use of skin surface electropotentials for breast cancer detection. *Technology in Cancer Research and Treatment*, 9(1):95–106.

Stevenson, C. E. (1995). Statistical models for cancer screening. *Statistical Methods in Medical Research*, 4(1):18–32.

Štrumbelj, E., Bosnić, Z., Kononenko, I., Zakotnik, B., and Kuhar, C. G. (2010). Explanation and reliability of prediction models: The case of breast cancer recurrence. *Knowledge and Information Systems*, 24(2):305–324.

Swisher, J. R., Jacobson, S. H., Jun, J. B., and Balci, O. (2000). Modeling and analyzing a physician clinic environment using discrete-event (visual) simulation. *Computers and Operations Research*, 28(2):105–125.

Tafazzoli, A., Roberts, S., Klein, R., Ness, R., and Dittus, R. (2009). Probabilistic cost-effectiveness comparison of screening strategies for colorectal cancer. *ACM Transactions on Modeling and Computer Simulation*, 19(2):6.

Takin, Z. C. and Cevik, M. (2013). Combinatorial Benders cuts for decomposing IMRT fluence maps using rectangular apertures. *Computers and Operations Research*, 40(9):2178–2186.

Tejada, J. J., Ivy, J. S., King, R. E., Wilson, J. R., Ballan, M. J., Kay, M. G., et al. (2014). Combined DES/SD model of breast cancer screening for older women, II: Screening-and-treatment simulation. *IIE transactions*, 46(7):707–727.

Tejada, J. J., Ivy, J. S., Wilson, J. R., Ballan, M. J., Diehl, K. M., and Yankaskas, B. C. (2015). Combined DES/SD model of breast cancer screening for older women, I: Natural-history simulation. *IIE Transactions*, 47(6):600–619.

Thomas, L. C. (2009). Using logistic regression to build scorecards. In *Consumer credit models: Pricing, profit and portfolios*, chapter 1.9, pages 79–84. Oxford University Press, Oxford, 1st edition.

Thrush, S., Sayer, G., Scott-Coombes, D., and Roberts, J. V. (2002). Grading referrals to specialist breast unit may be ineffective. *British Medical Journal*, 324(7348):1279.

Tourassi, G. D., Markey, M. K., Lo, J. Y., and Floyd, C. E. (2001). A neural network approach to breast cancer diagnosis as a constraint satisfaction problem. *Medical Physics*, 28(5):804–811.

Tran-Duy, A., Boonen, A., Kievit, W., van Riel, P. L. C. M., van de Laar, M. A. F. J., and Severens, J. L. (2014). Modelling outcomes of complex treatment strategies following a clinical guideline for treatment decisions in patients with rheumatoid arthritis. *PharmacoEconomics*, 32(10):1015–1028.

University Hospital Southampton NHS Foundation Trust (2014). *Imaging for patients with breast symptoms.* http://www.uhs.nhs.uk/OurServices/Breastservice/Breastradiology/BreastImaging/Imagingforpatientswithbreastsymptoms.aspx (accessed May 31, 2018).

University of Southampton (2016). *Delphis.* http://library.soton.ac.uk/delphis (accessed November 12, 2016).

Vanberkel, P. T., Boucherie, R. J., Hans, E. W., Hurink, J. L., van Lent, W. A. M., and van Harten, W. H. (2011). An exact approach for relating recovering surgical patient workload to the master surgical schedule. *Journal of the Operational Research Society*, 62(10):1851–1860.

Vanden Bosch, P. M. and Dietz, D. C. (2000). Minimizing expected waiting in a medical appointment system. *IIE Transactions*, 32(9):841–848.

Vataire, A.-L., Aballea, S., Antonanzas, F., Hakkaart-van Roijen, L., Lam, R. W., McCrone, P., et al. (2014). Core discrete event simulation model for the evaluation of health care technologies in major depressive disorder. *Value in Health*, 17(2):183–195.

Verjee, A. (2015). *Personal communication with Azmina Verjee, Clinical Research Practitioner (Oncology) at the Whittington Hospital.*

Vidal, L.-A., Sahin, E., Martelli, N., Berhoune, M., and Bonan, B. (2010). Applying AHP to select drugs to be produced by anticipation in a chemotherapy compounding unit. *Expert Systems with Applications*, 37(2):1528–1534.

Vieira, B., Hans, E. W., van Vliet-Vroegindeweij, C., van de Kamer, J., and van Harten, W. (2016). Operations research for resource planning and -use in radiotherapy: A literature review. *BMC Medical Informatics and Decision Making*, 16(1):149.

Vieira, I. T., de Senna, V., Harper, P. R., and Shahani, A. K. (2011). Tumour doubling times and the length bias in breast cancer screening programmes. *Health Care Management Science*, 14(2):203–211.

Wang, F., Zhang, S., and Henderson, L. M. (2018). Adaptive decision-making of breast cancer mammography screening: A heuristic-based regression model. *Omega*, 76:70–84.

Wang, H.-I., Smith, A., Aas, E., Roman, E., Crouch, S., Burton, C., et al. (2017). Treatment cost and life expectancy of diffuse large B-cell lymphoma (DLBCL): A discrete event simulation model on a UK population-based observational cohort. *European Journal of Health Economics*, 18(2):255–267.

Wang, K.-J., Makond, B., and Wang, K.-M. (2013). An improved survivability prognosis of breast cancer by using sampling and feature selection technique to solve imbalanced patient classification data. *BMC Medical Informatics and Decision Making*, 13(124):2–14.

Werker, G., Sauré, A., French, J., and Shechter, S. (2009). The use of discrete-event simulation modelling to improve radiation therapy planning processes. *Radiotherapy and Oncology*, 92(1):76–82.

West, D., Mangiameli, P., Rampal, R., and West, V. (2005). Ensemble strategies for a medical diagnostic decision support system: A breast cancer diagnosis application. *European Journal of Operational Research*, 162(2):532–551.

Whittington Health NHS (2015a). *About us.* http://www.whittington.nhs.uk/default.asp?c=3920 (accessed June 12, 2015).

Whittington Health NHS (2015b). *Breast cancer.* https://www.whittington.nhs.uk/default.asp?c=4737 (accessed June 16, 2015).

Willett, A. M., Michell, M. J., and Lee, M. J. R. (2010). Best practice diagnostic guidelines for patients presenting with breast symptoms. Technical report, CRS Breast Cancer Working Group.

Williams, K. A., Chambers, C. G., Dada, M., McLeod, J. C., and Ulatowski, J. A. (2014). Patient punctuality and clinic performance: Observations from an academic-based private practice pain centre: A prospective quality improvement study. *BMJ Open*, 4(5):e004679.

Winkler, S. M., Affenzeller, M., Kronberger, G., Kommenda, M., Wagner, S., Dorfer, V., et al. (2013). On the use of estimated tumour marker classifications in tumour diagnosis prediction - A case study for breast cancer. *International Journal of Simulation and Process Modelling*, 8(1):29–41.

Wishart, G. C., Azzato, E. M., Greenberg, D. C., Rashbass, J., Kearins, O., Lawrence, G., et al. (2010). PREDICT: A new UK prognostic model that predicts survival following surgery for invasive breast cancer. *Breast Cancer Research*, 12(1):R1.

Witten, I. H., Frank, E., and Hall, M. A. (2011). *Data mining: Practical machine learning tools and techniques.* Morgan Kaufmann, Burlington, 3rd edition.

Woodall, J. C., Gosselin, T., Boswell, A., Murr, M., and Denton, B. T. (2013). Improving patient access to chemotherapy treatment at Duke Cancer Institute. *Interfaces*, 43(5):449–461.

Zhang, Y., Berman, O., and Verter, V. (2012). The impact of client choice on preventive healthcare facility network design. *OR Spectrum*, 34(2):349–370.