Report on Methods for Complex Linked Data

Li-Chun Zhang*†
James Dawber*

*S3RI and Dept of Social Statistics & Demography, University of Southampton
†Statistics Norway and University of Oslo

July 2019

This work was funded by the Economic & Social Research Council under grant no. ES/T001038/1.

Report on Methods for Complex Linked Data

Li-Chun Zhang and James Dawber
University of Southampton

Executive Summary

The UK's longitudinal study resources have been largely survey-based, but there is potential for increasing the range of variables and coverage of the information through linkage and harmonisation with other datasets. Combining multiple sources in this way creates data with complex structures which require appropriate methodologies for analysis. This report describes the nature of complexities in linked datasets for analysis, and summarises the methodological requirements for

- analysis of partially overlapping repeated measures;
- analysis of networks within longitudinal data;
- secondary analysis of linked data;
- longitudinal population size estimation.

These approaches all share the feature that they have to deal with the potential for errors in the data linkage process, particularly where automated solutions are needed to control costs. A summary of the challenges in providing a scalable linkage methodology which can deal with multiple datasets is included. Secondary analysis of data that cannot be linked without errors is a central topic area in the landscape created by longitudinal data linkage.

Key areas where methodological development seems possible and useful are in the use of structural equation models and related approaches to make the best use of the *all* the available data, and deployment of the entity resolution approach to data linkage to deal with conflicting information in multiple sources.

1 Introduction

The existing core of UK longitudinal study resources consists chiefly of a suite of longitudinal survey programmes, comprising of both household panels and birth cohort studies, most of which can be found listed in Coleman (2015). Davis-Kean et al. (2018) outline two related approaches to enhancing the existing datasets via linkage and harmonisation with other datasets, including relevant administrative, biomarker, geographical and environmental data as the main additional sources.

The combination for research of multiple sources of data of different types requires complementary methodological approaches to allow robust statistical analysis of and inference from complex, multi-level and/or linked data. For integrative usage of multiple sources, it is helpful to maintain a broad view of *integrated data*, which extends beyond a single linked dataset. For instance, in data fusion (e.g. Braverman, 2008) two datasets may be concatenated rather than matched. Moreover, the combined data may be given a network representation instead of as an $n \times p$ matrix, in order to capture the interactions among the n individuals in addition to the p measures at the individual level.

This report aims to describe specific analytic challenges inherent in complex and linked data, in the context of UK longitudinal study resources. In doing so, we shall

- highlight the attributes of complex/big data and linked data, systemise potential data quality issues and their implications for robust statistical analysis;
- evidence opportunities and challenges for future UK longitudinal study resources, which provides a context for integrative usage of disparate data sources;
- outline topics for the analysis and methodology of complex and, in particular, linked or integrated data, with applications to the likely UK future longitudinal study resources.

2 Complex, linked or integrated data

2.1 Complex data and quality issues

In the tradition of longitudinal data analysis, *complex* is taken to refer to the common features of sampling and non-sampling errors, such as stratified multistage selection, compound rotation and attrition patterns over time, which generally complicate the distribution of the observed sample survey data, and call for adaptation of standard statistical methods of analysis, such as regression and hypothesis testing, when applied to such data; see e.g. Skinner et al. (1989) and Lynn (2009). Meanwhile, notwithstanding the lack of a unifying definition, *big data* has become omnipresent in discussions of innovative integrative usage of data from multiple sources. All the major additional datasets envisaged for the UK longitudinal study resources can be big data in one way or another. For instance, administrative and transaction data can be large, fast and varying in quality (e.g. Hand, 2018). Likewise with biomarker or GIS data.

It is thus helpful to extend the traditional scope of complex data features, which may cause analytical or computational complications, including when the data is combined with others. Roughly speaking, the data may be complex unless it is selected according to a design, complete in content and accurate in observation. The following are some typical examples of the features that can make the data complex.

- The observed units form a non-probability sample, which is not obtained according to a probability sampling design. In non-survey sources, the representativeness of the

sample can often be affected by a combination of coverage and selection issues (see also work package 5).

- The measures of interest may be missing or subjected to observational deficiencies more often than is acceptable, because they are not important to the data owner.
- Precise time labelling may be lacking in many administrative datasets, regarding the time of an event and the time of registration or updating, which is a common difficulty for alignment of observations that can change over time, such as place of residence.
- Data may be available in an 'organic' format, such as a text string of an address, a satellite image of a given resolution, etc. Application of Machine Learning techniques is then necessary to extract the content of interest, which inevitably generates its own errors.

One may apply the total survey error framework of Groves et al. (2004), and systematise the challenging features of complex data along the two dimensions of *representation* and *measurement*. Put simply, for a dataset that can be given as an $n \times p$ matrix, the issues related to the rows (of units) refer to the representation dimension, and those related to the columns (of measures) refer to the measurement dimension. For both survey and non-survey sources including big data, the quality issues can generally be caused by

- initial coverage errors, non-probability sample selection or an unknown missing data mechanism, along the representation dimension;
- initial definitional discrepancy, imperfect measurement instrument or processing or learning algorithm, along the measurement dimension.

2.2 Linked or integrated data and quality issues

Data from combining two or more sources can be represented as a single linked dataset, when the units residing in the separate datasets overlap each other. Contextual or multilevel variables can be obtained, when separate datasets are organised around different units that are nested in each other. For example, let A be a dataset organised around persons, and let B be one organised around postcodes. The linked dataset can be organised around persons, which are nested in postcodes, such that the measures at postcode level become now contextual variables associated with the persons.

When the units residing in the separate datasets do not overlap each other, they can be concatenated to form the units of a single *fused* dataset. For integrative usage of data from multiple sources, we include also fused data. There are no joint observations of the measures in the separate datasets. The aim of data fusion, statistical matching (D'Orazio et al., 2006) or ecological inference (King, 1997) and so on, is to construct or make

inference about the joint distribution based on observations of the marginal distributions. A schematic representation of linked vs. fused dataset can be given as

$$[y\ z] \overset{linkage}{=} \begin{bmatrix} y_{A\setminus B} & z^*_{A\setminus B} \\ y^*_{B\setminus A} & z_{B\setminus A} \\ y_{A\cap B} & z_{A\cap B} \end{bmatrix} \quad vs. \quad [y\ z] \overset{fusion}{=} \begin{bmatrix} y_A & z^*_A \\ y^*_B & z_B \\ (y_\emptyset) & (z_\emptyset) \end{bmatrix}$$

where the y-variable is observed in dataset A, and z in B. For emphasis the nonexistent part $[(y_\emptyset)\ (z_\emptyset)]$ is included to contrast the observed data $[y_{A\cap B}\ z_{A\cap B}]$. The superscript * is used to indicate existent but unobserved variables.

It may be more appropriate to adopt a network representation of the combined data. For instance, given n persons, the kinships among them can be given by the $n\times n$ adjacency matrix, where the (i,j)-th element is 1 if i and j are kins of each other, and 0 otherwise. All the diagonal elements of this adjacency matrix will be 0. One can associate it with an $n \times n$ value matrix, where the i-th diagonal element is a measure of person i, and the (i,j)-th element is a measure of the type of kinship between i and j, provided they are adjacent to each other, which can take on a value different to the (j,i)-th element. Or, one can use an $n \times p$ matrix for all the individual measures. In any case, a network is then a valued graph (Frank, 2011), where the graph representing the network structure is defined by the adjacency matrix, and the values associated with the nodes and edges of the graph by the value matrices. Instead of kinship, one can use other socio-economic interactions derived from additional data sources, which may be relevant to the analysis of individuals, or such interactions may even be the primary target of analysis.

For integration of data from multiple sources, Zhang (2012) outlines a total error framework at the integration phase. Di Zio et al. (2017) identify various statistical tasks pertaining to (i) transformation of input data objects and attributes to relevant statistical units and measurements for integrative uses, and (ii) micro- and macro-level integration of separate datasets, often with overlapping units and measurements. The following are some typical examples of the problems one needs to deal with in data integration.

- Different datasets may be organised around different objects. For instance, cellphone data containing location information may be available by time, cell tower and phone number, which can cause ambiguity regarding the matched individuals.

- In the absence of a common identifier, linkage error is the case if a pair of linked records do not refer to the same entity, or one may fail to link the records of the same entity.

- Different datasets may contain similar measures that are not fully compatible with each other. For instance, educational attainment in the different sources may not agree with each other, and none of sources is deemed the gold standard *a priori*.

- The statistical unit of interest may not exist in any of the sources, and thus need to be created, which inevitably generates its own errors. A typical example is creating

living households from individuals (Zhang, 2011), based on information about family relationships, various addresses of residence, work and other activities.

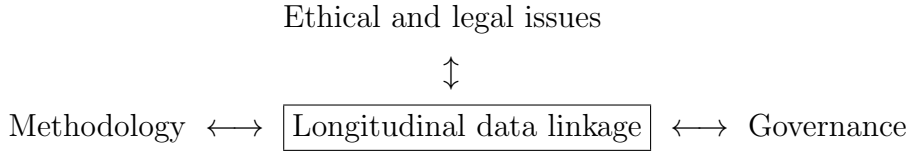
3 Opportunities and challenges

Longitudinal data linkage is the ability to link longitudinal survey data to a range of other (often also longitudinal) administrative data, such as health, tax, welfare and educational records, open or free data, as well as to big data such as digital footprints (Davis-Kean et al., 2017). In epidemiology and medical studies, longitudinal data linkage is extensively used to enhance data on clinical performance and health outcomes; see e.g. Guenel et al. (1990) and Harron et al. (2016) for an early and a more recent example, respectively. Calderwood and Lessof (2009) review the practice, benefits and challenges of linking longitudinal surveys to administrative data. In addition to the obvious potential for enhancing substantive survey data, they stress the possibility for enhancing the quality of the survey process itself. An example is the ability to use administrative records to track respondents and to understand and adjust for attrition to longitudinal samples, and these are important benefits which can ultimately improve the quality of longitudinal surveys and longitudinal data; see e.g. Cameron et al. (2017) for a recent study.

One may expect that such an outlook to longitudinal data linkage will continue to be the likely scenario for enhancing future UK longitudinal study resources. The key advantages are easily understood. To put it shortly, more and better data is expected to yield more and better analysis and inference, which in turn leads to more and better knowledge and insight. The relevant data that can be obtained via linkage can be broadly classified into two categories: (I) missing or additional study variables, (II) auxiliary information for survey processing, estimation and analysis.

Davis-Kean et al. (2017) describe in addition an ambitious outlook, which is more concerned with the longitudinal population, than the longitudinal measures that have been the focus of the traditional perspective. In particular, this aims at standardising the designs of the various longitudinal surveys so that they all use the same *longitudinal population register* (i.e. a population spine) as their sampling frame, and with all ESRC research-related linkage of different administrative and survey data sources harmonised to this spine. As an example of such a constructed longitudinal population spine, in countries that do not have a population register to start with, one may refer to the Integrated Data Infrastructure (IDI) at Statistics New Zealand (SNZ, 2018). A similar but currently more limited undertaking in the UK context is the Statistical Population Dataset (SPD) at the Office for National Statistics (ONS, 2019).

The various challenges to the opportunities raised above can be summarised below:



Data access is obviously the central challenge to legal and ethical issues. In addition, the fact that the data may be linked to many other sources can be expected to affect survey consent and response, which in turn can cause coverage and selection errors.

As examples of challenges related to governance, one may mention privacy preserving linkage infrastructure, access management to aggregated outputs vs. de-identified micro-data, etc. Establishing approved standards for linkage and dissemination protocols is of critical importance in this respect; see e.g. GUILD by Gilbert et al. (2017).

When it comes to the methodological issues *at the stage of estimation and analysis*, it seems helpful to distinguish between the situations where the key challenges all arise from the features that make the (integrated) data complex, because data linkage or integration does not pose serious problems of its own, and those situations where data linkage or integration is the root cause of the analytic challenges. The following is a classification of the main issues in accordance with this distinction.

- Regarding complex data features:
  - weighting adjustment for coverage and sample selection errors
  - imputation and weighting for missing or incomplete observations
  - adjustment for measurement errors or analysis of latent variables
  - estimation of dynamic populations and analysis of dynamic units
  - Machine Learning techniques for content extraction (in order to construct a measure)
  - estimation and analysis of network data or network dynamics over time
- Regarding integrated data features:
  - generic record linkage methods that are scalable to multiple large datasets
  - secondary analysis of linked data (without access to the separate input datasets)
  - privacy preserving linkage and analysis methods
  - data fusion and inference from fused data

It is useful to keep in mind that, while many of the challenges apply to any type of complex or integrated data, the longitudinal set-up can add its own extra twist. For instance, one almost always needs to handle multivariate observations in a longitudinal study, whereas it is often possible to focus on scalar observations in theoretical and methodological developments for cross-sectional studies.

A particular challenge above that is specific to longitudinal studies arises from the fact that the population and units of interest are generally dynamic (rather than static) over the period of study. The population of individuals is dynamic due to the natural demographic events of birth, death, internal and external migration. The unit that can be relevant in longitudinal studies is dynamic as long as it can consist of more than one individual, such as family, household, siblings, school class, etc.

4 Analysis and methodology

Many of the complex data features outlined above can be dealt with by novel applications and extensions of existing methods, some of which may well have been developed outside the longitudinal data setting. However, Davis-Kean et al. (2017) draw special attention to the linked data: "The creation of longitudinal survey databases with links to various administrative databases is half the story. The other half is what is done with these linked datasets", "This is in sharp contrast to the sophisticated methodologies, and associated software implementation, that have been developed to account for sampling and attrition biases when analysing data collected from respondents to longitudinal surveys."

We believe it is useful to extend the scope, from linked data to integrated data, to include in particular fused data and network data as well. Below we motivate and highlight some topic areas of methodology for integrated data, in the context of longitudinal data linkage envisaged for the UK future longitudinal study resources.

4.1 Analysis of partially overlapping repeated measures

Analysis of repeated measures is obviously a standard topic in longitudinal studies. Complications due to missing data have received ample attention in the past (e.g. Little and Rubin, 1987). Similarly with applications of latent variable/class analysis to the estimation of gross flows over time in the presence of measurement errors (e.g. Vermunt, 1996). Longitudinal data linkage can easily create a situation where multiple conflicting measures become available over time, albeit for different individuals over the study period, in which case one is faced with *partially overlapping repeated measures*.

Take employment status as an example. Suppose a panel of individuals are linked to the Labour Force Survey (LFS) over time, from which one obtains the LFS status in addition to the status originally observed in the panel survey. However, the observations may be in conflict with each other, which indicates the presence of measurement error. Moreover, due to the rotating design of the LFS, the LFS status will only be available for some but not all the time points covered by the panel study. In short, at any given time point, one would observe both (i.e. overlapping) measures for some of the panel members, while for the others one would observe only the status in the panel survey. This creates a setting of partially overlapping measures at the given moment, the pattern of which

varies over time. Hence, a situation of partially overlapping repeated measures.

The data can be further enhanced, given the ability to link to the relevant administrative registers on employment and social benefits. Suppose as a result one can obtain a fully overlapping series of register-based employment status for all the panel members over the entire study period. This not only strengthens the analysis of partially overlapping repeated measures using the same approach, but it can offer additional possibilities via data fusion. As shown in Zhang (2015b), data fusion techniques can be effective for the estimation of gross flows from one time point to another, for which the LFS status is only observed for *different* individuals at the two time points, i.e. in the absence of any repeated LFS measures, when the marginal LFS observations are combined with the proxy repeated measures derived from the relevant administrative data.

In short, longitudinal data linkage can create partially overlapping repeated measures, which require more advanced methods in order to make use of *all* the available data. For instance, it seems possible and useful to extend methods based on structural equation modelling (e.g. Kline, 2016), including latent class models, possibly in combination with data fusion techniques, in order to make better use of *all* the available data.

4.2 Analysis of networks in longitudinal populations over time

Any study unit that consists of more than one individual (e.g. a household) is dynamic if it can change over time. Lavallée (2007) applies the generalised weight share method (GWSM), by which the measure of a dynamic unit at a later time point is transferred to the initial sampling units, which can lead to the observation of the given dynamic unit. This allows one to base estimation and inference on the known initial sampling design, and avoid the complications associated with the unknown sampling design for the dynamic units (and the population) at the later time points.

The GWSM is in fact a variation of multiplicity estimation, originally discovered by Birnbaum and Sirken (1965). A related method is adaptive cluster sampling (Thompson, 1990), where the observation procedure leading from the initial sampling units is such that the inclusion probability cannot be calculated for all the units observed by the end of the survey. Zhang and Patone (2017) show that all these sampling methods are special cases of sampling in finite graphs, under a unified definition of graph sampling.

The graph sampling approach can thus be used to treat many problems associated with dynamic units over time, which occur in longitudinal studies. Moreover, graph (or network) sampling methods can naturally be expected to gain increasing importance, when the integrated data from longitudinal data linkage is given a network representation. Finally, one can draw on an extensive literature of model-based analysis of networks; see e.g. Goldberg et al. (2009) for a survey on statistical network models.

In summary, an emerging rich topic area is the application of design- or model-based methods for network data resulting from longitudinal data linkage. Not only can they

provide solutions to some perennial problems in longitudinal studies, they can greatly extend the scope of analysis, from isolated units to network characteristics that capture the various interactions among the units over time.

4.3 Generic scalable linkage methodology

A pair of records from two separate files (A and B) can be linked deterministically, if they match exactly and uniquely on a chosen set of linkage key variables, such as name, birthdate, address, etc. The main shortcoming of deterministic linkage is that it will miss many true matches. In probabilistic record linkage one recognises that the key variables may be subject to measurement errors, and allows for links to be made despite the records not matching completely. Fellegi and Sunter (1969) partition the Cartesian product set $A \times B$ into $M \cup U$, where M contains the true matched pairs and U the true unmatched pairs, and set up the Likelihood Ratio Test (LRT) for each pair of records $(ab) \in A \times B$:

$$H_0 : \ (ab) \in M, \ vs. \ H_1 : \ (ab) \in U$$

based on the so-called m-probability, which is the probability of observing how the key variables of a and b compare to each other under H_0, and the u-probability of that under H_1. The pairs whose LRT statistics are above a threshold value are accepted as links, those below a threshold value are rejected as non-links, and the rest to be determined by additional processing including clerical review.

The FS-methodology has proven to be very useful in practice, enabling industrial-strength applications to large datasets of the size of the general population, obtained from census, tax, health and other sources; see e.g. Zhang and Campell (2009), Owen et al. (2015). Nevertheless, the formulation does have certain theoretical issues.

- Applying the LRT to all the pairs in $A \times B$ creates a multiple comparison problem. The acceptable pairs require deduplication, e.g. when both (ab) and (ab') are above the acceptance threshold. It is difficult to apply the approach to link multiple files in a transitive manner, e.g. if (ab) in $A \times B$ and (bc) in $B \times C$ are links, it does not necessarily follow that (ac) is an acceptable link when looking at $A \times C$.

- The joint distribution of all the $n_A n_B$ comparison scores is ill-defined, if one treats e.g. the comparison scores for (ab) and (ab') as if they were independent of each other. The so-called maximum likelihood estimator (MLE) of the parameters of the m- and u-probabilities (Jaro, 1989) is biased in reality; see e.g. Fortini and Tuoto (2019).

It should be noticed that these issues are usually not critical when linking two files because, to put it this way, what is then important is whether (ab) is more likely to be a match than (ab') but not *how much* more likely. However, one would no longer be able to ignore these problems, if biased estimates of the linkage errors are used as necessary inputs to

the analysis of linked data, or when longitudinal data linkage requires linking multiple large datasets, e.g. in order to create the latent population spine.

Entity resolution provides a theoretically more attractive formulation (e.g. Christen, 2012), where the set of (unique) entities underlying the separate datasets are envisaged as a latent spine of unknown size, and each record (in any dataset) is attached to one and only one latent entity on the spine. In this way, the records in different files are linked to each other or deduplicated, provided they are attached to the same latent entity, in a transitive manner regardless of the number of datasets involved. There have been a few applications of the entity resolution perspective under the Bayesian paradigm of computation (e.g. Tancredi and Liseo, 2011; Stoerts et al., 2017), although there is nothing intrinsically Bayesian about this perspective to record linkage. Lack of scalability has been a central challenge to the proposed methodology so far, which is not yet feasible e.g. to link the population census file with the patient register.

In conclusion, generic scalable linkage methodology is of key interest to the vision of longitudinal data linkage. Scalable methods for the linkage of multiple population-size datasets are clearly important to the creation of a latent population register. Moreover, to enhance a survey dataset by linkage, the generic ability to link multiple files in a transitive manner can be expected to improve the quality of statistical information harnessed in the linked dataset. Replacing the FS-paradigm of record linkage by the entity resolution perspective can provide the angle for innovative approaches. Finally, it is essential that the methodology allows for consistent estimation of the different types of linkage error probabilities, which are necessary for subsequent analysis of the linked data.

4.4 Secondary analysis of linked data

A secondary analyst of linked data can be assumed to have access only to non-sensitive summary information about the performance of the data linkage method (Chambers and Da Silva, 2019), but not the unlinked records in the different files, nor all variables in the linkage key, nor the details of the linkage procedure (Zhang, 2019). By contrast, a primary analyst would have access to all of these information. Secondary analysis of linked data is the default setting for users of the UK longitudinal study resources.

It requires a bit more notation than so far in this report, to properly appreciate the challenges in secondary analysis of linked data, because the underlying entity ambiguity problem is rather different to the selection or measurement error problems common in the statistical literature. Let A and B denote two separate datasets, of respective size n_A and n_B. Denote by AB the true matched entities, and by A_u and B_u the unmatched entities in A and B, respectively, such that $\omega = (AB, A_u, B_u)$ forms a tripartition of the set of underlying entities, with unknown size $\min(n_A, n_B) \leq |A_u| + |AB| + |B_u| \leq n_A + n_B$, in the absence of duplicated records in either dataset. Let (X_A, K_A) denote the statistical and linkage key variables associated with A, and (Y_B, K_B) those of B, all of which are

observed. What is unobserved is the partition ω. Denote the complete data by (Z, ω), where $Z = (X_A, K_A, Y_B, K_B)$ is observed. A joint model of (Z, ω) can e.g. be given as

$$f(X_A, Y_B|\omega; \psi, \eta, \theta) = \prod_{A_u} f(x_a; \psi) \cdot \prod_{B_u} f(y_b; \eta) \cdot \prod_{AB} f(x_a, y_b; \theta)$$
$$f(K_A, K_B|\omega, X_A, Y_B) = \prod_{A_u} f(k_a|x_a) \cdot \prod_{B_u} f(k_b|y_b) \cdot \prod_{AB} f(k_a, k_b|x_a, y_b)$$

DeGroot and Goel (1980) consider inference of the correlation coefficient from a broken sample from a bivariate normal distribution, where $A_u = B_u = \emptyset$ and ω is a permutation matrix. The MLE is found to behave erratically, whether one treats ω as part of the parameter to be maximised or the missing data to be integrated out. Zhang (2019) shows that for linear regression, where the dependent and independent variables come separately from two datasets, the MLE of the regression coefficients is inconsistent, if it is obtained via the EM algorithm by treating ω as missing observations to be integrated out at the E-step, even when the non-informative model of the key variables $f(K_A, K_B|\omega)$ is known.

Thus, at present, it is unclear whether or how exact likelihood-based inference of linked data can be correctly formulated, unlike inference in the presence of missing observations or measurement errors. Instead, adjustment methods need to be tailored for different problems, of which linear regression has so far received the most extensive treatment in the literature; see e.g. Lahiri and Larsen (2005), Chambers (2009), Kim and Chambers (2012a, 2012b), Han and Lahiri (2018), Chambers and Da Silva (2019). However, as outlined below, these existing approaches under the so-called linkage model do have some serious restrictions, which point to directions for future development.

- Take the case of two datasets, it is commonly assumed either that $|A_u| = |B_u| = \emptyset$ or $|A_u| = \emptyset$, which may be referred to as the complete match space assumption. However, this is unrealistic in many practical situations where there exist unmatched entities in each separate file, such as when linking tax and health data.

- For secondary analysis, one has to adopt greatly simplifying assumptions, such as the exchangeable linkage error (ELE) model (Chambers, 2009), where there exists a constant false linkage probability and completely random mismatching in the case of false linkage. While the ELE assumption is practically appealing, heterogeneous linkage error is generally the case, where the probability of false linkage is not constant across the different entities. Moreover, in situations with an incomplete match space, the unmatched entities in either dataset cannot possibly be correctly linked, so that the linkage errors for these unmatched entities can never be exchangeable with the other matched entities that do have a chance to be correctly linked.

- The linkage model is based on the probability of two records being linked, averaged over all the possible key variable errors that can cause linkage errors. However, in

any given application of record linkage, one only faces a particular set of key variable errors, meaning the linkage result is better in some applications than others, even if the underlying error mechanism is the same in all these applications. The existing methods do no allow for conditional inference, given the actual precision achieved.

- In one form or another, assumptions of non-informative linkage errors are required in all the methods mentioned above.

In summary, secondary analysis of data that cannot be linked without errors is a central topic area in the landscape created by longitudinal data linkage. The current lack of exact likelihood-based inference is a fundament challenge that needs to be addressed. To improve the various tailored solutions to different problems, it is important to overcome the restrictive assumption of a complete match space, to accommodate heterogeneous linkage errors, to potentially allow for conditional inference given the achieved linkage precision, and to be able to test the underlying non-informative linkage error assumption and develop suitable adjustments in the case of informative linkage error.

4.5 Longitudinal population size estimation

The experience with SPD (ONS, 2019) so far suggests that a longitudinal population register is likely to suffer from non-negligible coverage errors, when based purely on linkage of relevant administrative datasets, even if one disregards all the false links committed. The union of the administrative datasets will most likely have appreciable over-coverage, in the sense that it contains erroneous enumeration of out-of-scope individuals. Under-coverage will also be unavoidable, in terms of missing individuals who do not interact with any of the administrative authorities; see e.g. Zhang and Dunne (2017) for an example of such under-coverage of the Irish Person Activity Register.

Longitudinal population size estimation is a necessary enabler, in order to support the bold outlook to the UK future longitudinal study resources, as outlined by Davis-Kean et al. (2017). One would need to resolve some difficulties, before the traditional capture-recapture models (Fienberg, 1972) can be extended to the longitudinal setting.

- The log-linear models presume a single closed population, with one structural zero cell consisting of the individuals missed by all the enumeration lists. In the longitudinal context, one has at least two populations U_t and U_{t+1}. Setting $U = U_t \cup U_{t+1}$ to be the population implies a peculiar kind of association between U_t and U_{t+1}, considered as two additional enumeration lists. Moreover, it would create more than one structural zero cell, of the individuals missed by all the original enumeration lists.

- The log-linear models are traditionally only applied to deal with under-coverage. However, erroneous enumeration is unavoidable and possibly non-negligible in the administrative datasets. Modelling approaches that can simultaneously handle both coverage

errors are being developed; see e.g. Di Cecco (2019), Zhang (2015a, 2019). A related problem is sub-population domain misclassification (e.g. Van der Heijden et al., 2019), which causes domain over-coverage and under-coverage at the same time.

- It is typically assumed that different enumeration lists can be matched without error, in order to identify the cross-classified cells, which will be violated as long as linkage error is inevitable; see Di Consiglio et al. (2019) for a recent review.

In summary, longitudinal population size estimation from multiple administrative datasets, possibly in combination with additional survey sampling data, is an integral component of the UK's future longitudinal study resources, in accordance with the bold outlook of Davis-Kean et al. (2017). Methodological developments and systematic applications are required to deal with both over- and under-coverage of the relevant administrative datasets, possibly with extra complications due to linkage error. All in all, the benefits, of having a time series of plausible, accepted and internally consistent longitudinal domain population sizes, will likely outweigh the known shortcomings of whichever methodology is chosen in the end.

References

[1] Birnbaum, Z.W. and Sirken, M.G. (1965). *Design of Sample Surveys to Estimate the Prevalence of IRareDiseases: Three Unbiased Estimates*. Vital and Health Statistics, Ser. 2, No.11. Washington:Government Printing Office.

[2] Boyd, J. H., Randall, S. M., Ferrante, A. M., Bauer, J. K., Brown, A. P. and Semmnes, J. B. (2014). Technical challenges of providing record linkage services for research. *BMC Med Inform Decis Mak*, **14**, 23.

[3] Braverman, A. (2008) Data fusion. In the *Encyclopedia of Quantitative Risk Analysis and Assessment*, Volume 2. Wiley, New York.

[4] Calderwood, L. and and C. Lessof (2009). Enhancing Longitudinal Surveys by Linking to Administrative Data. In *Methodology of Longitudinal Surveys, ed. P. Lynn*, 55-72.

[5] Cameron, C.M., Osborne, J.M., Spinks, A.B., Davey, T.M., Sipe, N. and R.J. McClure (2017). Impact of participant attrition on child injury outcome estimates: a longitudinal birth cohort study in Australia. *BMJ Open*, **7**, nr. 6, doi: `10.1136/bmjopen-2016-015584`

[6] Chambers R. (2009). Regression analysis of probability-linked data. *Official Statistics Research Series, Vol. 4.*

[7] Chambers, R.L. and A.D. da Silva (2019). Improved secondary analysis of linked data: a framework and an illustration. *Journal of the Royal Statistical Society: Series A*. doi: `10.1111/rssa.12477`

[8] Christen, P. (2012). A survey of indexing techniques for scalable record linkage and deduplication. *ISEE Transactions on Knowledge and Data Engineering*, **24**.

[9] Coleman, N. (2015). *Summary of Longitudinal Surveys*. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/444794/DFE-RR458_Summary_of_longitudinal_surveys.pdf

[10] Davis-Kean, P., Chambers, R.L., Kleinert, C., Rem, Q. and S. Tang (2017). *Longitudinal Studies Strategic Review 2017: Report to the Economic and Social Research Council*. https://esrc.ukri.org/files/news-events-and-publications/publications/longitudinal-studies-strategic-review-2017/

[11] DeGroot, M.H. and Goel, P.K. (1980). Estimation of the correlation coefficient from a broken random sample. *The Annals of Statistics*, **8**, 264–278.

[12] Di Cecco, D. (2019). Estimating population size in multiple record systems with uncertainty of state identification. In *Analysis of Integrated Data*, eds. L.-C. Zhang and R.L. Chambers. Chapter 8, pp. 169-196. Chapman & Hall/CRC.

[13] Di Consiglio L., Tuoto, T. and Zhang, L.-C. (2019) Capture-recapture methods in the presence of linkage errors. In *Analysis of Integrated Data*, eds. L.-C. Zhang and R.L. Chambers. Chapter 3, pp. 39-72. Chapman & Hall/CRC.

[14] D'Orazio, M., Di Zio, M. and Scanu, M. (2006). *Statistical Matching: Theory and Practice*. Chichester: Wiley.

[15] Di Zio, M., Zhang, L.-C. and T. De Waal (2017). Statistical methods for combining multiple sources of administrative and survey data. *The Survey Statistician*, **76**, 17-26.

[16] Fellegi, I.P. and Sunter, A.B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, **64**, 1183-121 0.

[17] Fienberg, S.E. (1972). The multiple recapture census for closed populations and incomplete 2^k contingency tables. *Biometrika* **59** 409–439.

[18] Fortini, M. and Tuoto, T. (2019). Use of record linkage in Official Statistics and feedbacks on research. *Presented at ITACOSM 2019, Florence*.

[19] Frank, O. (2011). Survey sampling in networks. *The SAGE Handbook of Social Network Analysis*, 389-403.

[20] Gilbert, R., Lafferty, R., Hagger-Johnson, G., Harron, K., Zhang, L-C., Smith, P., ... H. Goldstein (2017). GUILD: GUidance for Information about Linking Data sets. *Journal of Public Health*. DOI: 10.1093/pubmed/fdx037

[21] Goldenberg, A., Zheng, A.X., Fienberg, S.E. and E.M. Airoldi (2009). A survey of statistical network models. https://arxiv.org/abs/0912.5410

[22] Groves, R. M., F. J. Fowler Jr., M. Couper, J. M. Lepkowski, E. Singer and R. Tourangeau (2004). *Survey methodology*. Wiley, New York.

[23] Guenel, P., Engholm, G. and E. Lynge (1990). Laryngeal cancer in Denmark: a nationwide longitudinal study based on register linkage data. *Occupational and Environmental Medicine*, **47**, 473-479. doi: 10.1136/oem.47.7.473

[24] Han, Y. and Lahiri, P. (2018). Statistical analysis with linked data. *International Statistical Review*. https://doi.org/10.1111/insr.12295

[25] Hand, D. (2018). Statistical challenges of administrative and transaction data (with discussion). *Journal of Royal Statistical Society, Series A*, **181**, 555-605.

[26] Harron, K, Gilbert, K., Cromwell, D. and van der Meulen, J. (2016). Linking data for mothers and babies in de-identified electronic health data. *PLoS ONE 11(10): e0164667.* doi: 10.1371/journal.pone.0164667

[27] Jaro, M.A. (1989). Advances in record-linkage methodology as applied to matching the 1985 Census of Tampa, Florida. *Journal of the American Statistical Association*, **84**, 414-420. doi: 10.1080/01621459.1989.10478785.

[28] Kim, G. and Chambers, R.C. (2012a). Regression analysis under incomplete linkage. *Comutational Statistics and Data Analysis*, **56**, 2756–2770.

[29] Kim, G. and Chambers, R.C. (2012b). Regression analysis under probabilistic multi-linkage. *Statistica Neerlandica*, **66**, 64–79.

[30] King, G. (1997). *A Solution to the Ecological Inference Problem.* Princeton University Press.

[31] Kline, R. B. (2016) *Principles and Practice of Structural Equation Modeling (4th Edition).* The Guilford Press.

[32] Lahiri, P. and Larsen, M.D. (2005). Regression analysis with linked data. *Journal of the American Statistical Association*, **100**, 222–230.

[33] Lavallée, P. (2007). *Indirect Sampling.* Springer.

[34] Little, R.J.A and D.B. Rubin (1987). *Statistical Analysis with Missing Data.* Wiley: New York.

[35] Lynn, P. (2009). *Methodology of Longitudinal Surveys.* Wiley: Chichester.

[36] ONS (2019). *Methodology of Statistical Population Dataset V2.0.* https://www.ons.gov.uk/census/censustransformationprogramme/administrativedatacensusproject/methodology/methodologyofstatisticalpopulationdatasetv20

[37] Owen, A., Jones, P. and Ralphs, M. (2015). Large-scale linkage for total populations in official statistics. In *Methodological Developments in Data Linkage (eds. K. Harron, H. Goldstein and C. Dibben)*, Chapter 8.

[38] Skinner, C.J., Holt, D. and T. M. F. Smith (1989). *Analysis of Complex Surveys.* Wiley: New York.

[39] SNZ (2018). *Integrated Data Infrastructure.* https://www.stats.govt.nz/integrated-data/integrated-data-infrastructure

[40] Stoerts, R., Hall, R. and Fienberg, S. (2017). A Bayesian approach to graphical record linkage and de-duplication. *Journal of the American Statistical Association*, **111**, 1660–1672

[41] Tancredi, A. and Liseo, B. (2011). A hierarchical Bayesian approach to record linkage and population size problems. *The Annals of Applied Statistics*, **5**, 1553–1585.

[42] Thompson, S.K. (1990). Adaptive cluster sampling. *Journal of the American Statistical Association*, 85:1050–1059.

[43] Van der Heijden, P.G.M., Smith, P.A., Whittaker, J., Cruyff, M. and B. Bakker (2019). Dual and multiple system estimation with fully and partially observed covariates. In *Analysis of Integrated Data*, eds. L.-C. Zhang and R.L. Chambers. Chapter 7, pp. 137-168. Chapman & Hall/CRC.

[44] Vermunt, J.K. (1996). *Log-linear event history analysis: A general approach with missing data, latent variables and unobserved heterogeneity.* Tilburg: Tilburg University Press.

[45] Zhang, G. and Campbell, P. (2012) Data Survey: Developing the Statistical Longitudinal Census Dataset and identifying its potential uses. *Australian Economic Review*, **45**, 125–133.

[46] Zhang, L.-C. (2019). On secondary analysis of datasets that cannot be linked without errors. In *Analysis of Integrated Data*, eds. L.-C. Zhang and R.L. Chambers. Chapter 2, pp. 13-38. Chapman & Hall/CRC.

[47] Zhang, L.-C. (2019). Log-linear models of erroneous list data. In *Analysis of Integrated Data*, eds. L.-C. Zhang and R.L. Chambers. Chapter 9, pp. 197-218. Chapman & Hall/CRC.

[48] Zhang, L.-C. (2015a). On modelling register coverage errors. *Journal of Official Statistics*, **31**, 381-396.

[49] Zhang, L.-C. (2015b). On proxy variables and categorical data fusion. *Journal of Official Statistics*, **31**, 783-807.

[50] Zhang, L.-C. (2012). Topics of statistical theory for register-based statistics and data integration. *Statistica Neerlandica*, **66**, 41-63.

[51] Zhang, L.-C. (2011). A unit-error theory for register-based household statistics. *Journal of Official Statistics*, **27**, 415-432.

[52] Zhang, L.-C. and Patone, M. (2017). Graph sampling. *Metron*, **75**, 277-299. doi: 10.1007/s40300-017-0126-y

[53] Zhang, L.-C. and Dunne, J. (2017). Trimmed dual system estimation. In *Capture-Recapture Methods for the Social and Medical Sciences*, eds. D. Böhning, J. Bunge and P. v. d. Heijden, Chapter 17, pp. 239-259. Chapman & Hall/CRC.