1

Introduction

Katie Harron[1], Harvey Goldstein[2,3] and Chris Dibben[4]
[1] *London School of Hygiene and Tropical Medicine, London, UK*
[2] *Institute of Child Health, University College London, London, UK*
[3] *Graduate School of Education, University of Bristol, Bristol, UK*
[4] *University of Edinburgh, Edinburgh, UK*

1.1 Introduction: data linkage as it exists

The increasing availability of large administrative databases for research has led to a dramatic rise in the use of data linkage. The speed and accuracy of linkage have much improved over recent decades with developments such as string comparators, coding systems and blocking, yet the methods still underpinning most of the linkage performed today were proposed in the 1950s and 1960s. Linkage and analysis of data across sources remain problematic due to lack of identifiers that are totally accurate as well as being discriminatory, missing data and regulatory issues, especially concerned with privacy.

In this context, recent developments in data linkage methodology have concentrated on bias in the analysis of linked data, novel approaches to organising relationships between databases and privacy-preserving linkage. *Methodological developments in data linkage* bring together a collection of chapters on cutting-edge developments in data linkage methodology, contributed by members of the international data linkage community.

The first section of the book covers the current state of data linkage, methodological issues that are relevant to linkage systems and analyses today and case studies from the United Kingdom, Canada and Australia. In this introduction, we provide a brief background to the development of data linkage methods and introduce common terms. We highlight the most important issues that have emerged in recent years and describe how the remainder of the book attempts to deal with these issues. Chapter 2 summarises the advances in linkage accuracy and speed that have arisen from the traditional probabilistic methods proposed by Fellegi and Sunter. The first section concludes with a description of the data linkage environment as it is today, with

Methodological Developments in Data Linkage, First Edition. Edited by Katie Harron, Harvey Goldstein and Chris Dibben.

case study examples. Chapter 3 describes the opportunities and challenges provided by data linkage, focussing on legal and security aspects and models for data access and linkage.

The middle section of the book focusses on the immediate future of data linkage, in terms of methods that have been developed and tested and can be put into practice today. It concentrates on analysis of linked data and the difficulties associated with linkage uncertainty, highlighting the problems caused by errors that occur in linkage (false matches and missed matches) and the impact that these errors can have on the reliability of results based on linked data. This section of the book discusses two methods for handling linkage error, the first relating to regression analyses and the second to an extension of the standard multiple imputation framework. Chapter 7 presents an alternative data storage solution compared to relational databases that provides significant benefits for linkage.

The final section of the book tackles an aspect of the potential future of data linkage. Ethical considerations relating to data linkage and research based on linked data are a subject of continued debate. Privacy-preserving data linkage attempts to avoid the controversial release of personal identifiers by providing means of linking and performing analysis on encrypted data. This section of the book describes the debate and provides examples.

The establishment of large-scale linkage systems has provided new opportunities for important and innovative research that, until now, have not been possible but that also present unique methodological and organisational challenges. New linkage methods are now emerging that take a different approach to the traditional methods that have underpinned much of the research performed using linked data in recent years, leading to new possibilities in terms of speed, accuracy and transparency of research.

1.2 Background and issues

A statistical definition of data linkage is ‘a merging that brings together information from two or more sources of data with the object of consolidating facts concerning an individual or an event that are not available in any separate record’ (Organisation for Economic Co-operation and Development (OECD)). Data linkage has many different synonyms (record linkage, record matching, re-identification, entity heterogeneity, merge/purge) within various fields of application (computer science, marketing, fraud detection, censuses, bibliographic data, insurance data) (Elmagarmid, Ipeirotis and Verykios, 2007).

The term ‘record linkage’ was first applied to health research in 1946, when Dunn described linkage of vital records from the same individual (birth and death records) and referred to the process as ‘assembling the book of life’ (Dunn, 1946). Dunn emphasised the importance of such linkage to both the individual and health and other organisations. Since then, data linkage has become increasingly important to the research environment.

The development of computerised data linkage meant that valuable information could be combined efficiently and cost-effectively, avoiding the high cost, time and effort associated with setting up new research studies (Newcombe et al., 1959). This led to a large body of research based on enhanced datasets created through linkage. Internationally, large linkage systems of note are the Western Australia Record Linkage System, which links multiple datasets (over 30) for up to 40 years at a population level, and the Manitoba Population-Based Health Information System (Holman et al., 1999; Roos et al., 1995). In the United Kingdom, several large-scale linkage systems have also been developed, including the Scottish Health Informatics Programme (SHIP), the Secure Anonymised Information Linkage (SAIL) Databank

and the Clinical Practice Research Datalink (CPRD). As data linkage becomes a more established part of research relating to health and society, there has been an increasing interest in methodological issues associated with creating and analysing linked datasets (Maggi, 2008).

1.3 Data linkage methods

Data linkage brings together information relating to the same individual that is recorded in different files. A set of linked records is created by comparing records, or parts of records, in different files and applying a set of linkage criteria or rules to determine whether or not records belong to the same individual. These rules utilise the values on 'linking variables' that are common to each file. The aim of linkage is to determine the true **match status** of each comparison pair: a **match** if records belong to the same individual and a **non-match** if records belong to different individuals.

As the true match status is unknown, linkage criteria are used to assign a **link status** for each comparison pair: a l**ink** if records are classified as belonging to the same individual and a **non-link** if records are classified as belonging to different individuals.

In a perfect linkage, all matches are classified as links, and all non-matches are classified as non-links. If comparison pairs are misclassified (false matches or missed matches), error is introduced. **False matches** occur when records from different individuals link erroneously; **missed matches** occur when records from the same individual fail to link.

1.3.1 Deterministic linkage

In deterministic linkage, a set of predetermined rules are used to classify pairs of records as links and non-links. Typically, deterministic linkage requires exact agreement on a specified set of identifiers or matching variables. For example, two records may be classified as a link if their values of National Insurance number, surname and sex agree exactly. Modifications of strict deterministic linkage include 'stepwise' deterministic linkage, which uses a succession of rules; the '$n-1$' deterministic procedure, which allows a link to be made if all but one of a set of identifiers agree; and ad hoc deterministic procedures, which allow partial identifiers to be combined into a pseudo-identifier (Abrahams and Davy, 2002; Maso, Braga and Franceschi, 2001; Mears et al., 2010). For example, a combination of the first letter of surname, month of birth and postcode area (e.g. H01N19) could form the basis for linkage.

Strict deterministic methods that require identifiers to match exactly often have a high rate of missed matches, as any recording errors or missing values can prevent identifiers from agreeing. Conversely, the rate of false matches is typically low, as the majority of linked pairs are true matches (records are unlikely to agree exactly on a set of identifiers by chance) (Grannis, Overhage and McDonald, 2002). Deterministic linkage is a relatively straightforward and quick linkage method and is useful when records have highly discriminative or unique identifiers that are well completed and accurate. For example, the community health index (CHI) is used for much of the linkage in the Scottish Record Linkage System.

1.3.2 Probabilistic linkage

Newcombe was the first to propose that comparison pairs could be classified using a probabilistic approach (Newcombe et al., 1959). He suggested that a match weight be assigned to each comparison pair, representing the likelihood that two records are a true match, given the

agreement of their identifiers. Each identifier contributes separately to an overall match weight. Identifier agreement contributes positively to the weight, and disagreement contributes a penalty. The size of the contribution depends on the discriminatory power of the identifier, so that agreement on name makes a larger contribution than agreement on sex (Zhu et al., 2009). Fellegi and Sunter formalised Newcombe's proposals into the statistical theory underpinning probabilistic linkage today (Fellegi and Sunter, 1969). Chapter 2 provides details on the match calculation.

In probabilistic linkage, link status is determined by comparing match weights to a threshold or cut-off match weight in order to classify as a match or non-match. In addition, manual review of record pairs is often performed to aid choice of threshold and to deal with uncertain links (Krewski et al., 2005). If linkage error rates are known, thresholds can be selected to minimise the total number of errors, so that the number of false matches and missed matches cancels out. However, error rates are usually unknown. The subjective process of choosing probabilistic thresholds is a limitation of probabilistic linkage, as different linkers may choose different thresholds. This can result in multiple possible versions of the linked data.

There are certain problems with the standard probabilistic procedure. The first is the assumption of independence for the probabilities associated with the individual matching variables. For example, observing an individual in any given ethnic group category may be associated with certain surname structures, and hence, the joint probability of agreeing across matching variables may not simply be the product of the separate probabilities. Ways of dealing with this are suggested in Chapters 2 and 6. A second typical problem is that records with match weights that do not reach the threshold are excluded from data analysis, reducing efficiency and introducing bias if this is associated with the characteristics of the variables to be analysed. Chapter 6 suggests a way of dealing with this using missing data methods. A third problem occurs when the errors in one or more matching variables are associated with the values of the secondary data file variables to be transferred for analysis. This non-random linkage error can lead to biases in the estimates from subsequent analyses, and this is discussed in Chapters 4–6. Chapter 4 reviews the literature and sets out the situations where linkage bias of any kind can arise, including the important case when individual consent to linkage may be withheld so leading to incomplete administrative registers. Chapter 5 looks explicitly at regression modelling of linked data files when different kinds of errors are present, and Chapter 6 proposes a Bayesian procedure for handling incomplete linkages.

One of the features of traditional probabilistic methods is that once weights have been computed, the full pattern of similarities that give rise to these weights, based upon the matching variables, is either discarded or stored in a form that requires any future linkage to repeat the whole process. In Chapter 7, a graphical approach to data storage and retrieval is proposed that would give the data linker efficient access to such patterns from a graph database. In particular, it would give the linker the possibility to readily modify her algorithm or update files as further information becomes available. Chapter 7 discusses implementation details.

1.3.3 Data preparation

Quality of data linkage ultimately depends on the quality of the underlying data. If datasets to be linked contained sufficiently accurate, complete and discriminative information, data linkage would be a straightforward database merging process. Unfortunately, many administrative datasets contain messy, inconsistent and missing data. Datasets also vary in structure,

format and content. The way in which data are entered can influence data quality. For example, errors may be more likely to occur in identifiers that are copied from handwritten forms, scanned or transcribed from a conversation. These issues mean that techniques to handle complex and imperfect data are required. Although data preparation is an important concern when embarking on a linkage project, we do not attempt to cover this in the current volume. A good overview can be found in Christen (2012a).

1.4 Linkage error

Linkage error occurs when record pairs are misclassified as links or non-links. Errors tend to occur when there is no unique identifier (such as NHS number or National Insurance number) or when available unique identifiers are prone to missing values or errors. This means that linkage relies on partial identifiers such as sex, date of birth or surname (Sariyar, Borg and Pommerening, 2012).

False matches, where records from different individuals link erroneously, occur when different individuals have similar identifiers. These errors occur more often when there is a lack of discriminative identifiers and file sizes are large (e.g. different people sharing the same sex, date of birth and postcode). For records that have more than the expected number of candidate records in the linking file, the candidate(s) with the most agreeing identifiers or with the highest match weight is typically accepted as a link. This may not always be the correct link.

Missed matches, where records from the same individual fail to link, occur where there are errors in identifiers. This could be due to misreporting (e.g. typographical errors), changes over time (e.g. married women's surnames) or missing/invalid data that prevent records from agreeing.

Many linkage studies report the proportion of records that were linked (match rate). Other frequently report measures of linkage quality are sensitivity and specificity (Christen and Goiser, 2005). These measures are directly related to the probability of false matches and missed matches. However, interpretation of these measures is not always straightforward. For example, match rate is only relevant if all records are expected to be matched. Furthermore, such measures of linkage error can be difficult to relate to potential bias in results.

Derivation of measures of linkage error can also be difficult, as estimation requires that either the error rate is known or that the true match status of comparison pairs is known. A common method for measuring linkage error is the use of a gold-standard dataset. Gold-standard data may be an external data source or a subset of data with additional identifiers available (Fonseca et al., 2010; Monga and Patrick, 2001). Many linkage projects create a gold-standard dataset by taking a sample of comparison pairs and submitting the records to manual review (Newgard, 2006; Waien, 1997). The aim of the manual review is to determine the true match status of each pair (Belin and Rubin, 1995; Gill, 1997; Morris et al., 1997; Potz et al., 2010). Once a gold-standard dataset has been obtained, it is used to calculate sensitivity, specificity and other measures of linkage error by comparing the true match status of each comparison pair (in the gold-standard data) with the link status of each pair (Wiklund and Eklund, 1986; Zingmond et al., 2004). These estimates are assumed to apply to the entire linked dataset (the gold-standard data are assumed to be representative). This is a reasonable assumption if the gold-standard data were a random sample; otherwise, potential biases might be introduced.

Manual review is convenient but can take a substantial amount of time, particularly for large files (Qayad and Zhang, 2009; Sauleau, Paumier and Buemi, 2005). It also may not always be completely accurate. If samples are only taken from linked pairs – which is often the case due to the smaller number of links compared to non-links – the rate of missed matches would not be estimated. If the sample of pairs reviewed is not representative, estimates of linkage error may be biased.

1.5 Impact of linkage error on analysis of linked data

Although a large body of literature exists on methods and applications of data linkage, there has been relatively little methodological research into the impact of linkage error on analysis of linked data. The issue is not a new one – Neter, Maynes and Ramanathan (1965) recognised that even relatively small errors could result in substantially biased estimates (Neter, Maynes and Ramanathan, 1965). The lack of comprehensive linkage evaluation seems to be due to a lack of awareness of the implications of linkage error, possibly resulting from a lack of communication between data linkers and data users. However, the relevance and reliability of research based on linked data are called into question in the presence of linkage error (Chambers, 2009).

Data custodians (organisations that hold personally identifiable data that could be used for research) have a responsibility to protect privacy by adhering to legislation and guidelines, avoiding unauthorised copies of data being made and distributed and ensuring data are used only for agreed purposes. For these reasons, data custodians can be unwilling or unable to release identifiable data for linkage. To overcome this issue, many linkage projects adhere to the 'separation principle'. This means that the people performing the linkage (the data linkers – sometimes a trusted third party) do not have access to 'payload' data and people performing the analysis (the data users) do not have access to any personal identifiers. This protects confidentiality and means that linked datasets can be used for a range of purposes (Goeken et al., 2011). Approaches to privacy and security are discussed in detail in Chapter 3.

While the potential knowledge benefits from data linkage can be very great, these have to be balanced against the need to ensure the protection of individuals' personal information. Working with data that is non-personal (i.e. truly anonymous) guarantees such protection but is rarely practicable in a data linkage context. Instead, what is required is the construction of a data linkage environment in which the process of re-identification is made so difficult that the data can be judged as practicably anonymous. This type of environment is created both through the governance processes operating across the environment and the data linkage and analysis models that structure the operational processes. Chapter 3 reviews the main models that are used and their governance processes. Some examples from across the world are presented as case studies.

The separation principle is recognised as good practice but means that researchers often lack the information needed to assess the impact of linkage error on results and are unable to report useful evaluations of linkage (Baldi et al., 2010; Harron et al., 2012; Herman et al., 1997; Kelman, Bass and Holman, 2002). Separation typically means that any uncertainty in linkage is not carried through to analysis. The examples of linkage evaluation that do appear in the literature are often extreme cases of bias due to linkage error. However, as there is a lack of consistent evaluation of linkage, it is difficult to identify the true extent of the problem.

Reported measures of linkage error are important, as they offer a simple representation of linkage quality. However, in isolation, measures of sensitivity and specificity cannot always provide interpretation of the validity of results (Leiss, 2007). Although it is useful to quantify linkage error, it is most important to understand the impact of these errors on results.

The impact of linkage error on analysis of linked data depends on the structure of the data, the distribution of error and the analysis to be performed (Fett, 1984; Krewski et al., 2005). In some studies, it may be important to capture all true matches, and so a more specific approach could be used. For example, if linkage was being used to detect fraud, it may be important that all possible links were captured. In other studies, it might be more important that linked records are true matches, and missed matches are less important. For example, if linked healthcare records were being used to obtain medical history, it might be more important to avoid false matches (German, 2000). For these reasons, it is important that the impact of linkage error is understood for a particular purpose, with linkage criteria ideally tailored to that purpose. Bias due to linkage error is explored in detail in Chapter 4.

1.6 Data linkage: the future

Methods for data linkage have evolved over recent years to address the dynamic, error-prone, anonymised or incomplete nature of administrative data. However, as the size and complexity of these datasets increase, current techniques cannot eliminate linkage error entirely. Manual review is not feasible for the linkage of millions of records at a time. With human involvement in the creation of these data sources, recording errors will always be an issue and lead to uncertainty in linkage. Furthermore, as opportunities for linkage of data between organisations and across sectors arise, new challenges will emerge.

Chapter 9 looks at record linking approaches that are used to support censuses and population registers. There is increasing interest in this with the growing availability of large-scale administrative datasets in health, social care, education, policing, etc. It looks at issues of data security and, like Chapter 3, addresses the balance between individual privacy protection and knowledge and explores technical solutions that can be implemented, especially those that can operate on so-called ‘hashed’ data or data that is ‘pseudonymised at source’ where the linker only has access to linking variables where the original information has been transformed (similarly but irreversibly) into non-disclosive pseudonyms.

The second decade of the twenty-first century is an exciting and important era for data linkage, with increasing amounts of resources being applied and a broad range of different disciplinary expertise being applied. Our hope is that the present volume, by setting out some of these developments, will encourage further work and interest.