Chapter 12
New Technology and Implications for Healthcare and Public Health:
The Case of Probabilistic Record Linkage

Gulzar H. Shah
National Association of County and City Health Officials, USA

Kaveepan Lertwachara
California Polytechnic State University, USA

Anteneh Ayanso
Brock University, Canada

ABSTRACT

In this chapter, the authors provide a review of recent developments in probabilistic record linkage and their implications in healthcare research and public health policies. Their primary objective is to pique the interest of researchers and practitioners in the healthcare and public health communities to take full advantage of record linkage technologies in completing a health care scenario where different pieces of patient records are collected and managed by different agencies. A brief overview of probabilistic record linkage, software available for such record linkage, and type of functions provided by probabilistic record linkage software is provided. Specific cases where probabilistic linkage has been used to bridge information gaps in informing public health policy and enhancing decision-making in healthcare delivery are described in this chapter. Issues and challenges of integrating medical records across distributed databases are also outlined, including technical considerations as well as concerns about patient privacy and confidentiality.

INTRODUCTION

The necessity for comparing and matching data records from multiple sources in order to determine which sets of records belong to the same person, object, or event can arise in many contexts. Scholarly interests in this area span several academic disciplines (e.g., statistics, information systems, management sciences) as well as communities

DOI: 10.4018/978-1-61520-733-6.ch012

of practitioners (e.g., in electronic commerce, public health, vital records, welfare fraud detection, e-government). As computerized databases have now become ubiquitous in business sectors, the possibility of extensive analysis using these databases relies on the ability to integrate heterogeneous databases across organizations and functional units. Such data integration requires the presence of an error-free unique identifier or *key* attribute common among the data sets being matched. Unfortunately, in most real-world situations, this common key attribute across data sets is rarely available. Consequently, instead of relying upon a deterministic approach using unique identifiers, past research studies have proposed probabilistic algorithms to achieve the goal of record matching across heterogeneous databases. Among these early studies, seminal work by Newcombe, Kennedy, Axford, and James (1959) and Fellegi and Sunter (1969) provide theoretical frameworks for computer-aided record linkage operations. Other more recent scholarly studies on this topic include Dey, Sarkar, and De (1998); Bell and Sethi (2001); Sarathy and Muralidhar (2006); and Jiang, Sarkar, De, and Dey (2007). Although the algorithmic procedures to match data records suggested in these studies may vary, they share a common objective of linking records that belong to the same entity while minimizing the likelihood of erroneous matching (i.e., ensuring sensitivity and specificity).

Statistical theory used in record linkage was developed in the 1950s and was further refined in the 1970s and 1980s (Jaro, 1989; Newcombe et al., 1959). Until the early1980s, no commercial record linkage software was marketed, and those with a need for record linkage had to develop their own software (e.g., the Generalized Record Linkage System (GRLS) developed at Statistics Canada). They often faced the choice of using less accurate methods or expending considerable staff time to create proprietary systems. For example, in the late 1970s, the U.S. National Agricultural Statistics Service spent what is conservatively estimated as 50 staff-years to develop a state-of-the-art system (Day, 1997).

OVERVIEW OF RECORD LINKAGE CONCEPTS AND TECHNIQUES

Record linkage is a computer-based process to match records from different and often heterogeneous sources of data that refer to the same entities such as persons, events, or other objects of interest. However record linkage is sometimes performed within a single data set when multiple records are present in a single database for a person or other entity (e.g., records for multiple hospitalizations in a hospital discharge data set for a 12-month period). Record linkage within a single data set is also performed to remove duplicate records, referred to as "deduplication" (Winkler, 1999).

There are many applications of record linkage in both public and private sectors and its use has become even more significant with advances in the underlying techniques and the implementation tools. Detailed technical descriptions of record linkage are available elsewhere (see, for example, Fair, 1995, 1997; Newcombe, 1994). In addition to applications in health care and public health, record linkage is widely employed in other fields. For example, Probert, Semenciw, Mao, and Gentleman (1997) described how record linkage was used to integrate immigration and mortality databases in Canada. Quass and Starkey (2003), White (1997), and NeSmith (1997) provided examples on how record linkage could be used to consolidate genealogical data. Other examples of record linkage applications in public sectors include consolidating tax return records (Czajka, 1997; Harville & Moore, 1999; Steel & Konschnik 1997; Wahl, 1997*);* verifying housing mortgage applications (Herzog & Eilerman, 1997; Herzog, Scheuren, & Winkler, 2007); tracking academic progress in public schools (Miller, 1997); and public safety (Scheetz, Zhang, Kolassa, Allen, & Allen, 2008; Utter, 1997). In the private sector,