Record Linkage in Healthcare: Applications, Opportunities, and Challenges for Public Health

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

Recent years have witnessed the development of new record linkage technologies that are increasingly being used for data integration in various application settings. The authors’ objective in this article is to provide a review of recent developments in medical record linkage and their implications in healthcare research and public health policies. In particular, the authors assess the key advantages and possible limitations of record linkage techniques and technologies in various health care scenarios where different pieces of patient records are collected and managed by different agencies. First, the authors provide a brief overview of deterministic, probabilistic, and unsupervised record linkage techniques and their advantages and limitations. Then, the authors describe current probabilistic record linkage software and their functionalities, and present specific cases where probabilistic linkage has been successfully used to enhance decision-making in healthcare delivery as well as in healthcare-related public policy making. Finally, the authors outline some of the critical issues and challenges of integrating medical records across distributed databases, including technical considerations as well as concerns about patient privacy and confidentiality.

Keywords: Electronic Health Records, Healthcare Information Systems, Patient Confidentiality, Probabilistic Record Linkage, Public Health Data, Record Linkage, Unique Patient Identifier

INTRODUCTION

As computerized databases have now become ubiquitous in many business and public sectors, the amount of data collected by these organizations have multiplied exponentially during the past decade. However, these data collection efforts are often implemented by various and uncoordinated constituents. As a result, the information derived from these heterogeneous databases tend to be fragmented and incomplete. Thus, the ability to compare and match data records from multiple sources in order to determine which sets of records belong to the same person, object, or event has become a critical task for many organizations. However, 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 be-
ing 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); Dey, Sarkar, and De (2002); Verykios, Moustakides, and Elfeky (2002); 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 early 1980s, 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 a considerable amount of resources 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). In addition to the past studies mentioned above, scholarly work in this area span several other academic disciplines (e.g., statistics, information systems, management sciences) as well as communities of practitioners (e.g., in electronic commerce, public health, vital records, welfare fraud detection, e-government).

In this article, we present a review of recent development in record linkage technologies relevant to healthcare research and public health policies. The remaining of the article is organized as follows. The next section summarizes the existing literature on record linkage and the importance of record linkage in healthcare and public health. A brief introduction to different record linkage techniques is presented. Examples of successful applications of record linkage in healthcare and public health are also offered. We then discuss potential opportunities and challenges in using record linkage. The last section concludes our discussion on this topic.

PAST RESEARCH IN RECORD LINKAGE

Record linkage can be applicable both within and across data sources. Typically, record linkage is defined as a computer-based process of matching two or more 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 (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
A Strategic Approach to e-health Interoperability Using e-government Frameworks
www.igi-global.com/chapter/strategic-approach-health-interoperability-using/47128?camid=4v1a