博士論文（要約）

Performance Comparison of Probabilistic Linkage and Deterministic Linkage: A Simulation Study and a Real-Life Example of Linking Registry to Medical Insurance Database

（確率的リンケージと決定的リンケージの性能比較：
シミュレーション実験とアメリカ医療保険データベースを
用いた実例による検討）

シュ イン

ZHU YING

博士論文の要約

Performance Comparison of Probabilistic Linkage and Deterministic Linkage: A Simulation Study and a Real-Life Example of Linking Registry to Medical Insurance Database

（確率的リンケージと決定的リンケージの性能比較：
シミュレーション実験とアメリカ医療保険データベースを用いた実例による検討）

シュ　イン
ZHU YING

Introduction: Record linkage is the means through which person- or event-specific data from one source is combined with additional data from another source. Linking existing databases with complementary information will make the data more complete and useful for a variety of public health researches including health services and outcomes, comparative effectiveness and pharmacoepidemiology. Studies using linked data sources may serve as an alternative to prospective studies and clinical trials as they are more cost-effective and representative of real world scenarios. Robust linkage methods are crucial to the validity of linked data and studies that use such data. Deterministic linkage and probabilistic linkage are two common methods. Only a few studies have compared the performance of the two linkage methods and findings are equivocal. In the following three studies, key factors of database and linkage methodology that affect linkage outcomes were explored, including the rate of missing and error of linkage variables, the discriminative power of linkage rules, the file sizes of databases in linkage project, and cutoff selection methods in probabilistic linkage. The performances of deterministic linkage and probabilistic linkage were compared using both real-life databases and simulated data. The study data were from the Centers for Medicare and Medicaid Services (CMS) implantable cardioverter-defibrillator (ICD) registry and administrative Medicare inpatient claims data. The studies also aimed to address the challenges of record linkage in the context of public health researches.

Study 1

Introduction: Record linkage is expected to be most accurate if a unique record identifier is common to, accurately recorded in, and not missing in both sources. However, registries often do not collect unique identifiers such as Social Security Number and Medicare beneficiary identification number, or even names or addresses that are not unique but commonly used and often sufficient for linkage. Even if collected, the information is not usually released to researchers for ethical or technical reasons. Nonetheless, hospitalization records from registries and administrative databases can be linked using multiple non-unique identifiers, such as date of birth, sex, admission date, and provider information such as hospital ID. Although previous studies have demonstrated the feasibility of record linkage between registries and claims data using multiple non-unique identifiers, the validity of this method relative to linkage using unique identifiers, generally considered the gold standard, has not been assessed.

Methods: We compared the validity of 6 deterministic linkage rules with multiple non-unique identifiers by using data from CMS ICD registry and administrative Medicare inpatient claims data between 2005 and 2008, and validated against a gold standard using a combination of both unique and non-unique identifiers.

Results: Linkage rules using 2 or 3 non-unique, patient-level identifiers (i.e. date of birth, sex, admission date) and hospital ID produced linkages with high sensitivity (≥95%) and positive predictive value (≥98%). Linkage rules with higher discriminative power generated more accurate linkage outcomes. When linking hospitalization-level records in the absence of unique identifiers, provider information was necessary for successful linkage.

Discussion: Our results are likely generalizable to attempts that link hospitalization-level records. However, in outpatient records, completeness and accuracy of physician IDs may be more compromised because of a more ambiguous definition of providers in outpatient settings where physicians work in a group practice. Detailed methodological work and separate validation are needed to understand the best practice in linking outpatient records. Our results may not be applicable in settings in which databases have high error rates in linkage variables. When error rates are high, a deterministic linkage method using multiple identifiers will produce a large number of false-negative links. A probabilistic linkage method should be considered to overcome this limitation.

Study 2

Introduction: Probabilistic linkage allows imperfection such as missing value and error in linkage variables and thereby allows more true matches to be identified. As the true matching status of record pairs is unknown without a gold standard, there is always a trade-off between sensitivity and positive predictive value. A valid probabilistic linkage requires selection of appropriate cutoff of the weights above which record pairs are considered matches. Traditionally, the cutoff is chosen by visual inspection of the weight histogram. This process is usually arbitrary and varies greatly among different inspectors. Many studies have discussed other issues of record linkage such as the empirical selection of linkage variables and the differences between deterministic and probabilistic linkages, however, very few have examined the issue of cutoff selection.

Methods: We linked the ICD registry to Medicare inpatient files of one-year 2008 with anonymous non-unique identifiers, provider location, provider ID, admission date, date of birth and gender. We assessed the validity of three methods of cutoff selection, i.e. histogram inspection, duplicate method and formula method, against an internally derived gold standard with unique identifiers. The aim was to find a valid method that would minimize false positives, maximize positive predictive value and not compromise sensitivity. We also aimed to discuss real-world challenges in working with administrative databases.

Results: Of the 64 890 registry records with an expected linkage rate of 55% to 65%, 55% were linked at cutoffs associated with positive predictive values of ≥90%. Histogram inspection suggested an approximate range of

optimal cutoffs, however the associated PPV could not be directly determined without a gold standard. The duplicate method made accurate estimations of cutoff and positive predictive value against the gold standard if the method's assumption was met --- that is if there were less than 2 links for each linkable record in the smaller dataset (ICD) which required very high discriminative power of linkage variables. The odds formula method overestimated the cutoff. It was overly simplistic, not taking data quality into consideration and performed poorly with very large file size which is not uncommon in real-life linkage projects.

Discussion: Probabilistic linkage without unique identifiers generated valid linkages when an optimal cutoff was chosen. Cutoff selection remains challenging, however, the duplicate method can be used in conjunction with histogram inspection when a gold standard is not available. The linkage rate and validity measures of deterministic linkage in our first study and probabilistic linkage in our second study were comparative. This is likely due to the high quality of data and choice of highly discriminative linkage variables.

Study 3

Introduction: The performance of both deterministic linkage and probabilistic linkage depend greatly on the completeness and accuracy of data. Error rate and missing rate are particular to the files involved in the linkage project, which are highly variable. Some previous studies reported error rates ranging from 4% to 15% and missing rates of 0% to 9% in surnames, given names, zip code and date of birth in disease registries and medical administrative databases. Theoretically, when these rates are high, deterministic linkage will produce a large number of false negative links and probabilistic linkage will capture more true matches by allowing imperfection in linkage variables. However in practice, studies have shown inconclusive results. Furthermore, unique identifiers such Social Security Number are rarely available, non-unique linkage variables must be chosen carefully such that there is sufficient discriminative power to identify a unique record.

Methods: The third study aimed to understand how different linkage methods perform with data of varying quality and to describe the scenario in which one method outperforms the other. We created a series of scenarios that represent real linkage situations by using variables commonly found in large medical and administrative databases, i.e. the first two studies. The linkage variables cover a range of distribution (discriminative power), rate of missing and error, and file size. We then assessed the difference in performance between deterministic linkage and probabilistic linkage in each of these scenarios, in terms of Type I error, Type II error, sensitivity, positive predictive value and computation time.

Results: Type I error of probabilistic linkage and Type II error of deterministic linkage, the limiting factor of respective method, were mainly affected by the rate of missing and error. Probabilistic linkage uniformly outperformed deterministic linkage. However, with very high quality data, deterministic linkage performed not significantly worse. The implementation of deterministic linkage in SAS took less than 1 minute, and probabilistic linkage took 2 minutes to 5 hours depending on file size.

Discussion: Our simulation study demonstrated that knowing the intrinsic error rate (quality) and missing rate (completeness) of linkage variables are key to choosing between the two linkage methods. As the quality of databases worsened, there were increasingly more undiscernible records in the larger dataset. In general, probabilistic linkage is suggested. It improved sensitivity compared to deterministic linkage and maintained a high PPV with variables of high discriminative power. For databases of very good quality, both linkage methods performed comparably well and deterministic linkage was a more cost-effective choice. Future research is needed to circumvent information loss from missing data, such as multiple imputation. Also needed is validation of these linkage methods in real-life databases of more complex structures and lesser quality.