審査の結果の要旨

シュ　イン
ZHU YING

In the three studies of this thesis, the application of record linkage methodologies was discussed in the context of public health research. The performances of two main linkage methods, deterministic linkage and probabilistic linkage, were validated in cardiovascular device registry and the U.S. Medicare inpatient claims databases. Three key parameters that affect linkage methodologies were examined, including the rate of missing and error of linkage variables, the discriminative power of linkage rules, and the file sizes of databases. The followings are the main study findings.

1. In linking device registry to Medicare claims data, deterministic linkage without unique identifiers required linkage rules with sufficiently high discriminative power to control for Type II error. When linking hospitalization-level records in the absence of unique identifiers, provider information is necessary for successful linkage.

2. In linking device registry to Medicare claims data, probabilistic linkage without unique identifiers required an optimal method of cutoff weight selection to control for and balance the Type I and Type II errors. When a gold standard by unique identifier is not available, the duplicate method can be used in conjunction with histogram inspection if the data conformed to the method's assumption, that is sufficient discriminative power is maintained by having not more than 2 links for every linkage record.

3. Linkage outcomes by deterministic linkage and probabilistic linkage were comparable in the linkage of device registry to Medicare claims data. This is likely due to the high quality of data and choice of highly discriminative linkage variables.

4. The simulation study demonstrated that steadily worsening data quality (missing and error) resulted in increasingly more undiscernible records and poorer linkage outcomes by both linkage methods. In general, probabilistic linkage generated more

accurate linkage outcomes (higher sensitivity and positive predictive value) than deterministic linkage, but was also more time-consuming as file sizes increased.

5. The simulation study also demonstrated that for databases of very good quality (<5% missing and error), both deterministic linkage and probabilistic linkage performed comparably well and deterministic linkage was a more cost-effective choice as it was less time-consuming and easier to implement.

This thesis discussed the issues and challenges of record linkage in public health research, and demonstrated the applicability of linkage methods to medical databases and the comparative effectiveness of deterministic linkage and probabilistic linkage. Based on these important findings, this thesis fulfills the requirement for the degree of doctor of philosophy.