Strategies for Handling Missing Data in Detecting Postoperative Surgical Site Infections

Zhen Hu
Institute for Health Informatics
University of Minnesota
Minneapolis, MN, USA
huxxx511@umn.edu

Genevieve B. Melton
Institute for Health Informatics
Department of Surgery
University of Minnesota
Minneapolis, MN, USA
gmelton@umn.edu

Gyorgy J. Simon
Department of Health Sciences
Research
Mayo Clinic
Rochester, MN, USA
gsimon@gmail.com

I. Introduction

Researchers are increasingly interested in the secondary use of EHR data to detect specific outcomes and adverse conditions. Our aim is to develop valid, robust, and practical EHR-derived models for identifying postoperative surgical site infections (SSIs). SSIs can be classified into superficial, deep, and organ/space, and are costly with significant morbidity. Compared with administrative/claims data that previous research heavily relied on, our use of EHR data has the potential to allow for the construction of more informative SSI detection models. Unfortunately, secondary use of EHR data can be challenging due to its often incomplete nature—some specific tests are just ordered to only a subset of patients (e.g., 52% of the surgical patients in our cohort do not have any white blood cell count data within 30 days after the operation). Mostly researchers ignore it by excluding cases or single variables with missing data, or imputing missing values for variables with slight amount of missing data. However, because of the high missingness rate in our data, to simply discard incomplete cases may result in losing important indicators of SSI. In our previous work, we only utilized the complete cases to detect SSIs within 30 days after surgery using the gold standard outcome from a validated national surgical registry—National Surgical Quality Improvement Project (NSQIP)[1]. In the current study, we sought to explore several popular treatments of missing data. The performance of the models after applying different treatments are compared to that of the reference model based on the complete cases.

II. Materials and Methods

A. Data Collection

Patients and their postoperative SSI outcome were identified in NSQIP registry from 2011 to 2013. EHR data of NSQIP patients were extracted from clinical data repository. We collect six types of data: demographics, medications, orders, diagnosis codes, lab results and the vital records.

B. Data Preprocessing

Data preprocessing consists of identifying and removing outliers, and correcting inconsistencies in the data. Many variables are binary, indicating the presence of a medication, order or a diagnosis during postoperative window. Longitudinal results of lab tests and vitals are summarized into features using extreme (highest and lowest) and average values.

C. Missing Data Imputation of the Incomplete Data Set

Eight traditional single and multivariate data imputation methods were applied, resulting in eight completed data sets, one data set for each imputation method. A reference data set consisting only of the complete cases (i.e. cases with no missing values) was also prepared.

D. Model Development

Logistic regression was applied to all of the eight completed data sets as well as the reference data set.

E. Model Evaluation

Data between 2011-2012 were used as training set, and data from 2013 were used a leave-out test set. The detection models were evaluated using two metrics: discrimination (C-statistic) and bias and were compared to the reference models, which are the models constructed on the reference data set. Out primary interest was to better understand the characteristics of the various data imputation methods.

III. Results and Conclusion

Imputation methods improved SSI-detection performance. The C-statistics and bias of the best imputed models versus that of reference model for each SSI type, namely superficial, deep, organ/space and the overall SSI are (0.85, 0.0122) vs (0.85, 0.0600), (0.86, -0.0010) vs (0.70, 0.0032), (0.94, -0.0082) vs (0.86, -0.0136), and (0.93, 0.0072) vs (0.86, -0.0159), respectively. We found imputation to be beneficial for all SSI types except the superficial SSIs and that no single imputation method was uniformly better than all others.

References

[1] Z. Hu, G. Simon, E. Arsoniadis, Y. Wang, M. Kwaan, G. Melton, "Automated Detection of Postoperative Surgical Site Infections Using Supervised Methods with Electronic Health Record Data," Stud Health Technol Inform 2015 (MEDINFO 2015), in press.

This research was supported by University of Minnesota Academic Health Center Faculty Development Award (GM, GS), and NIH/NCRR U54RR026066 (Blazar).

978-1-4673-9548-9/15 $31.00 © 2015 IEEE
DOI 10.1109/ICHI.2015.89