Addressing errors in a retrospective observational ICU database

Fiona M. Callaghan, PhD[1], Dina Demner-Fushman, MD, PhD[1], Swapna Abhyankar, MD[1], Clement J. McDonald, MD[1]
[1]Lister Hill Center, National Institutes of Health, Bethesda, MD

Abstract

Real-world, observational hospital databases are often error-prone, with incorrect values, mislabeled entries and missing data problems, which present problems for the medical researcher. In this poster, we present some of the challenges in performing a retrospective observational analysis of obesity and mortality using the Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II) database of ICU patients maintained by the Laboratory for Computational Physiology Department at MIT.

Introduction

The Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II) database is maintained by the Laboratory for Computational Physiology department at MIT and is comprised of all patients that were admitted to the ICU at Beth Israel Deaconess Medical Center from 2001 to 2008. The database contains information on approximately 19,000 adult patients. This database has the potential to form the basis for many retrospective observational studies, but like many real-world, hospital databases the data presents problems for analysis including missing data, and confusing or misstated values in hand-entered fields. In this study, we focus on the missing data problem as it relates to a specific observational study of body mass index (BMI) and mortality.

Missing Data

All predictor variables required some error checking, but by far the biggest challenge to the study of BMI and mortality was the large amount of missing information on the height of the patients. BMI is calculated using a patient's weight (kg) divided by their height (m) squared. Of the 19,000 patients in the data base (over age 18), height chart measurements were only available for approximately 9,500 patients. Further values were derived from body surface area (BSA) measurements and echocardiogram reports brought the number of heights to approximately 12,500. An analysis of the risk factors for missing versus non-missing values tended to show that the patients with one or more missing covariates were over-represented for traditional mortality risk factors: for example, they weighed more, had more medical and surgical admissions versus cardiac surgery admission, and were more likely to be males than females. Therefore, it was important to perform an analysis that attempted to account for these differences in order to see if the results remained robust.

Several methods were applied to account for the missing data. The first approach involved data imputation. Height values were generated using the medians taken from age/gender groups from the non-missing height values, and the addition of a random normal covariate with mean 0 and standard deviation equal to the estimate of the sample standard deviation to preserve the estimates of the variance. An implicit assumption was made that the reasons for the missing height values were not related to height and the values could be inferred from the observed data (namely, age and gender) under a missing at random (MAR) model[1]. With imputation, the sample size was increased to approximately 17,000. A Cox proportional hazards model was fitted in order to estimate the quantity of interest, the hazard rate. The largest percent change in estimates of risk from a model was a change of 11% in the estimate of the hazard rate for one BMI grouping. No predictors changed the direction of the risk. Other missing data methods were applied with similar results. Further methods are needed in order to attempt to quantify the amount of bias caused by similar missing data situations.

Conclusion

Data from observational studies must always be analyzed carefully to minimize bias, and this is especially true of hospital databases. By applying various methods for missing data some of this bias can at least be assessed, if not corrected for.

References

1. Little RJ, Rubin DB. Statistical analysis with missing data. 2nd ed. Wiley-Interscience, 2002