Journal of Medical Systems (2019) 43: 142
https://doi.org/10.1007/s10916-019-1274-9

SYSTEMS-LEVEL QUALITY IMPROVEMENT

The Trade-off of Applying Simple vs. Advanced Imputation Techniques in Prediction Modeling

Uri Kartoun[1]

Received: 12 March 2019 / Accepted: 3 April 2019 / Published online: 13 April 2019

With reference to health care data, when values are missing for a certain patient's covariate, these values must be estimated accurately as a preceding step to application of the statistical data analysis and machine learning methods. To accurately assess missing values, a variety of techniques have been introduced over the past few decades, especially within the context of advanced mechanisms such as propensity score analysis. In "A Comparison of Different Methods to Handle Missing Data in the Context of Propensity Score Analysis" (January 2019) [1] Choi et al. emphasized the importance of estimating missing confounders; the authors deserve credit for bringing this topic to the attention of the *European Journal of Epidemiology*'s readers.

In terms of the development of new medical risk scores, which are presumably more accurate than well-established scores, a major component in the success of a new risk score is the ability of its users to use it quickly. The model for end-stage liver disease (MELD) risk score, for example, is one of the most important and widely used risk prediction scores in medicine. Unlike other scores, a patient's MELD score may trigger a major clinical event—his or her rank on the organ allocation waiting list. A crucial factor that contributed to MELD's success is its simplicity; being composed of only three commonly available laboratory values.

Regarding more comprehensive, as well as more accurate, scores such as MELD-Plus, which is composed of 9 variables [2], it could be that not all patients will have the complete set of variables available. While creatinine, INR, and total bilirubin (MELD's components) are likely to be available for most patients, other components of MELD-Plus, such as total cholesterol, may be out of reach for many. Intuitively, the application of the most advanced imputation techniques to estimate missing total cholesterol values would be expected to achieve the highest possible accuracy. However, the users of scores such as MELD-Plus may wish to apply the score to a given population or specific patient immediately, but they cannot allocate the resources of testing complex techniques that may likely require applying time-consuming analyses (rather than immediately calculate the score using Excel or an on-line calculator). The trade-off for the user would thus be to apply the simplest of imputation techniques, such as imputing by the mean laboratory values representing the original population on which the original score was based.

MELD-Plus is an example of the feasibility of imputing by mean for laboratory measurements. The score has consistently been found superior to others in independent electronic medical record repositories (Brigham and Women's Hospital, Massachusetts General Hospital, and IBM's Explorys). Another example of adopting a simple approach to imputing by mean includes noting the discovery that progressive liver disease is associated with increased risk of cardiovascular disease, a finding that was confirmed by researchers from Cleveland Clinic [3]. Alternative approaches to imputation by mean have been explored; however, the high standard deviations of errors reported among different approaches [4] as well as the fact that imputation accuracy may not strictly correlate with predictive performance [5], calls into question their true usefulness in the development of new risk scores. It remains to be seen whether using the most advanced imputation methods could truly result in a significant improvement in the accuracy of associations between covariates and outcomes in prediction modeling.

This article is part of the Topical Collection on *Systems-Level Quality Improvement*

✉ Uri Kartoun
uri.kartoun@ibm.com

[1] Center for Computational Health, IBM Research, Cambridge, MA, USA

Compliance with Ethical Standards

Conflict of Interest IBM neither provided author U.K. salaries related to the study nor played any role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. U.K. received honoraria and travel funding from The American Association for the Study of Liver Diseases (2017), and received travel funding from Merck & Co., Inc. (2017).

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

1. Choi, J., Dekkers, O. M., and le Cessie, S., A comparison of different methods to handle missing data in the context of propensity score analysis. Eur J Epidemiol. 34(1):23–36, 2019 Jan. https://doi.org/10.1007/s10654-018-0447-z Epub 2018 Oct 19.
2. Kartoun, U., Corey, K. E., Simon, T. G., Zheng, H., Aggarwal, R., Ng, K., and Shaw, S. Y., The MELD-Plus: A generalizable prediction risk score in cirrhosis. PLOS ONE 12(10):e0186301, 2017.
3. Mehta, N., Singh, T., Lopez, R., and Alkhouri, N., The heart age is increased in patients with nonalcoholic fatty liver disease and correlates with fibrosis and hepatocyte ballooning. Am J Gastroenterol 111(12):1853–1854, 2016.
4. Waljee, A. K., Mukherjee, A., Singal, A. G., Zhang, Y., Warren, J., Balis, U., Marrero, J., Zhu, J., and Higgins, P. D., Comparison of imputation methods for missing laboratory data in medicine. BMJ Open 3(8), 2013.
5. Beaulieu-Jones, B. K., and Moore, J. H., Missing data imputation in the electronic health record using deeply learned autoencoders. Pac Symp Biocomput 22:207–218, 2017.