

Machine learning to deal with missing disability status: Ascertainment and imputation of outcomes should be distinguished

David M Kent  and Ewout W Steyerberg

Multiple Sclerosis Journal—
Experimental, Translational
and Clinical

October–December 2022,
1–2

DOI: 10.1177/
20552173221128874

© The Author(s), 2022.
Article reuse guidelines:
sagepub.com/journals-
permissions

Dear Editors,

Using natural language processing (NLP) to ascertain missing outcomes from unstructured notes has great potential for enhancing real world data to assess the care of patients with multiple sclerosis and other diseases. The novel approach of Alves et al.¹ uses the full information available in unstructured notes to address missingness in the expanded disability status scale (EDSS). This blurs the important distinction between extraction of information that is plainly in the medical records and single imputation of missing information using prediction based on correlations with other variables. This novel approach raises new and interesting questions. We focus on two that we believe warrant further consideration when dealing with missing patient outcomes.

First, how can algorithm performance be evaluated on the notes where EDSS is missing? Performance is readily measured in the group with EDSS assessments, but matters for those without. The excellent performance on the small test set of notes where EDSS is recorded may not reflect, even approximately, that on the large set of notes where EDSS is missing, since the former set might differ from the latter in many ways. Patient- and practice-level differences between the note samples with and without EDSS should be described. But beyond Table 1 characteristics, the notes without EDSS may differ from those with EDSS in the degree to which disability is assessed and documented, which may have important effects on algorithm performance.

Second, can these imputed outcomes be used for unbiased treatment effect estimation? This is a complex issue that is understandably not the focus of the paper. But consider that any feature from the unstructured

notes that correlate with outcome in the small development sample, comprising 5% of the population, may be used to project eEDSS outcomes in the full sample. Any of these features also correlated to treatment in an observational study is a potential confounder, but these features may not be known or available for deconfounding during effect estimation. Even small biases in outcome prediction across treatment groups can cause clinically important biases in effect estimation, since prediction errors in each treatment group may compound and be magnified in the effect estimate.^{2,3}

Given the above and other issues, we suggest maintaining a clear distinction between data (particularly outcome data) that has been extracted from the medical record and is verifiable at the individual patient level (e.g. by expert review of the medical record) and imputed outcomes that are not. We also wondered, since the binarized EDSS depends only on the ability to ambulate without an assistive device, whether a simple rule-based NLP may provide high-integrity, verifiable outcome data on many patients, avoiding some of the issues above and preserving transparency and analytically important information about where this outcome is actually missing in accord with emerging guidance.⁴

Sincerely,



David M Kent, MD, MS

Director, Predictive Analytics and Comparative Effectiveness (PACE) Center, Institute for Clinical Research and Health Policy Studies (ICRHPS), Tufts Medical Center.

Director, Clinical and Translational Science Program, Graduate School of Biomedical Sciences.

Correspondence to:

David M Kent,
Predictive Analytics and
Comparative Effectiveness
(PACE) Center, Institute for
Clinical Research and Health
Policy Studies, Tufts Medical
Center, 800 Washington St,
Box 63, Boston, MA 02111,
USA.
dkent1@tuftsmedicalcenter.org

David M Kent,
Predictive Analytics and
Comparative Effectiveness
Center, Tufts Medical Center,
Boston, MA, USA

Ewout W Steyerberg,
Department of Biomedical
Data Sciences, Leiden
University Medical Centre,
Leiden, the Netherlands



Professor of Medicine, Neurology, and Clinical and Translational Science, Tufts University.



Ewout W Steyerberg, PhD
Professor of Clinical Biostatistics and Medical Decision Making.
Chair, Department of Biomedical Data Sciences, Leiden University Medical Center.


Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

David M Kent  <https://orcid.org/0000-0002-9205-5070>

References

1. Alves P, Green E, Leavy M, et al. Validation of a machine learning approach to estimate expanded disability status scale scores for multiple sclerosis. *Mult Scler J Exp Transl Clin* 2022; 8: 20552173221108635.
2. Kent DM, van Klaveren D, Paulus JK, et al. The predictive approaches to treatment effect heterogeneity (PATH) statement: explanation and elaboration. *Ann Intern Med* 2020; 172: W1–W25.
3. van Klaveren D, Balan TA, Steyerberg EW, et al. Models with interactions overestimated heterogeneity of treatment effects and were prone to treatment mistargeting. *J Clin Epidemiol* 2019; 114: 72–83.
4. Lee KJ, Tilling KM, Cornish RP, et al. Framework for the treatment and reporting of missing data in observational studies: the treatment and reporting of missing data in observational studies framework. *J Clin Epidemiol* 2021; 134: 79–88.