Machine Learning Can Unlock Insights Into Mortality

Jessica C. Young, PhD, MSPH, Cory Pack, BS, Teresa B. Gibson, PhD, Frank Yoon, PhD, Debra E. Irwin, PhD, MSPH, Shalu Shiv, PhD, MPH, Toska Cooper, MPH, and Nabarun Dasgupta, PhD

ABOUT THE AUTHORS

Jessica C. Young is with the Cecil. G. Sheps Center for Health Services Research, University of North Carolina, Chapel Hill. Cory Pack, Teresa B. Gibson, Frank Yoon, Debra E. Irwin, and Shalu Shiv are with IBM Watson Health, Bethesda, MD. Toska Cooper and Nabarun Dasgupta are with the Injury Prevention Research Center, University of North Carolina, Chapel Hill.

The study of mortality is fundamental in public health research, but our ability to derive detailed insights is often limited by the practical constraints of the available data. Although the National Vital Statistics System maintains a national record of death certificate data enabling basic research on mortality trends and life expectancy, clinical information on the context of a given death in national death records is limited. When the necessary information is available, the ability to examine clinical details surrounding death and health indicators in the period before death enables research that can meaningfully inform public health strategies and interventions.

Insurance claims data can fill gaps in mortality research by providing details about diagnosed health conditions and key health care services a person receives before death, such as surgical procedures, laboratory tests, and drug prescriptions. This level of granularity composed of data that are continuously and prospectively collected for each patient offers insights that are not available with vital statistics alone and opens the door to uncovering how medications, health conditions, and health encounters may be associated with mortality.

A major limitation in mortality research is that data sets that have rich longitudinal health information (claims) and those that have recorded death dates (vital statistics) are often separate, and linkage may be prohibitively expensive or prohibited because of data privacy restrictions. We discuss the research implications of having disparate streams of health and mortality data; introduce how machine learning can help overcome these limitations; highlight important considerations for machine learning, including the risk of algorithmic bias; and briefly discuss best practices for applying machine learning to enhance public health research.

RESEARCH IMPLICATIONS

Studies using detailed longitudinal health data to better understand risks of mortality are limited to subpopulations whose claims data can be linked to death records (e.g., from Medicare claims); consequently, these studies present limited generalizability (e.g., to commercially insured populations). In some instances, inpatient claims data might contain discharge status reflecting death in the hospital. However, the Centers for Disease Control and Prevention estimates that the majority of all deaths occur outside the hospital (72% in 2019), and therefore deaths that occur in a hospital are just a small part of total mortality.[1] Other than data on cases for which hospital discharge status may indicate deaths that occur in a hospital, claims data do not regularly include death events. Individuals who die show a “disenrolled” status in their enrollment record, and researchers are largely unable to separate those who died from those who ended coverage with the insurance plan or employer.

Incomplete death data and the inability to differentiate between disenrollment because of death and disenrollment because of changes in insurance coverage present a significant missing data problem in health outcomes research. In epidemiologic terms, death is a competing risk. Incorrectly assuming that disenrollment is uninformative (i.e., independent of clinical disposition) and failing to account for death as a competing risk (when risk of death is nonnegligible) will induce bias in risk estimates for primary outcomes. Berry et al. demonstrate this bias empirically in a study of second hip fracture among elderly patients, showing that incidence of second hip fracture was 21% when not accounting for death as a competing risk and 12% when properly incorporating death information.[2] More recently, a study examining patients hospitalized for severe COVID-19 estimated the percentage of patients with clinical improvement after treatment with remdesivir.[3] By failing to account for the 13% of patients who died after receiving remdesivir, the authors overestimated the percentage of patients with clinical improvement by 10 percentage points.

MACHINE LEARNING AS A SOLUTION

Although custom linkage between claims and death data is challenging because of privacy concerns, innovations in privacy-preserving methods such as differential privacy may improve data availability and usability. When differential privacy allows claims-based linkage between clinical data elements and death status, machine learning can be trained in these linked data. For instance, vendors of administrative claims databases often have internal access to death data that can be analyzed to create algorithms differentiating disenrollment because of death and disenrollment because of changes in insurance coverage in claims-based studies. Effectively, this translates to distinguishing death from what may be considered uninformative administrative censoring.

In contrast to traditional statistical methods, which are better suited to testing prespecified hypotheses, machine learning focuses on empirical prediction of an outcome, irrespective of the model's parametric form (i.e., explanatory power).[4,5] Machine-learning methods can efficiently analyze thousands of potential predictors using data-adaptive identification of complex patterns, including nonlinear relationships and high-order interactions, optimizing predictor selection for a final algorithm.[6] By optimizing predictive performance, a machine-learning algorithm of mortality in claims data can be used to effectively impute missing or incomplete death status.

Reps et al. illustrated the potential for this type of work using an administrative claims database that had death records up to 2013 (sourced from the Death Master File) to develop and test a machine-learning algorithm for death status at the end of observation.[7] In the spirit of this work, new predictive algorithms based on machine learning can be developed in different data sources or those with more recent death data. Such algorithms can be disseminated for use in claims-based studies to address mortality as a primary outcome or competing risk, dramatically mitigating potential bias and broadening the scope and utility of health outcomes research.

POTENTIAL PITFALLS OF MACHINE LEARNING

Although machine-learning methods have the potential to enable research dealing with mortality, these methods have limitations to be considered. Predictive algorithms developed through machine learning may be subject to less human bias (e.g., model misspecification) given the data-driven nature of these methods; however, the data themselves can be inherently biased. If the input data used to train algorithms lack diversity or reflect structural biases, output models will not generalize across populations. Use of these algorithms can result in algorithmic bias, perpetuating existing inequities.[8] Obermeyer et al. show that for a given level of health, structural inequalities in access to care in the United States result in lower health care costs generated by Black patients than by White patients.[9] Subsequently, the use of an algorithm to identify patients who are most likely to benefit from additional resources based on predicted health care costs exacerbates systemic racial biases by preferentially identifying White patients to be more likely to benefit from additional resources.

For death specifically, racial disparities in the accuracy of death records (race and cause of death) can cause disparities in algorithm performance in certain subgroups. Previous work has found that race was more often misclassified on National Death Index records for American Indians and Alaska Natives,[10] and research conducted by the National Center for Health Statistics found that linkage rates of participants of Hispanic or Asian/Pacific Islander descent with National Death Index records were considerably lower than were those of non-Hispanic White patients.[11] These findings imply that what is often considered the "gold standard" for death recording has differential accuracy across race. Another study comparing race as recorded on death certificates to race reported by next of kin found that cause of death affected race reporting on death records, suggesting that racial information in vital statistics may be influenced by racial stereotypes.[12] This is particularly problematic for algorithms aiming to predict cause of death.

CONCLUSIONS

Death is a critical outcome for many research questions and a significant competing risk for many others. Machine learning can be used in large claims data to unlock insights into mortality, facilitating new public health research. With the growing availability of electronic health data, along with the gaining momentum of machine learning, large claims data are a frontier for mainstream public health research. These data-driven methods are flexible and, in the case of defining death in insurance claims data, can help examine data points from millions of patients, evaluate many predictors, identify the most important factors associated with death, while evaluating complex interactions at a scale not previously possible. In addition to predicting death in claims data, machine-learning methods have been applied to vital statistics data sets themselves, making mortality data more representative of minority populations[13] and providing more granular time estimates of often misclassified deaths such as suicide.[14]

Researchers must be aware that machine-learning tools are not impervious to human bias. If there are biases in whose experience is recorded, machine-learning tools can entrench existing race-based health disparities.[8,9] In an evaluation of approaches to reduce bias in machine-learning models, Park et al. illustrated several methods for evaluating algorithmic bias and found that a reweighting method was most successful in reducing bias.[15] When using machine learning, the population represented by the input data for algorithm creation must be considered, with an understanding that algorithms may not generalize to other populations. An algorithm developed in one population cannot necessarily be applied to a different population. External validation in appropriate populations is an important component of any machine-learning algorithm. Additional important aspects include a deep understanding of the data and potential biases in the underlying mechanisms of data generation, evaluation of algorithm performance across diverse subgroups, and transparency in the dissemination of algorithmic inputs, parameters, and outputs.[8]

When implemented rigorously using these best analytic practices, machine-learning algorithms can also address existing biases in health care. In a recent study examining racial disparities in pain, Pierson et al. found that compared with standard radiologic measures of pain severity that were developed in White patients, a newer machine-learning algorithm trained on racially and socioeconomically diverse data better captured pain in underserved populations.[16] Machine learning is a powerful tool for analyzing large amounts of data for clinical prediction. When applied to claims data linked to vital statistics, machine learning presents an opportunity to create algorithms to predict death, thus unlocking new possibilities into mortality research and reducing bias in estimates of other health outcomes of interest. AJPH

CORRESPONDENCE

Correspondence should be sent to Jessica C. Young, Cecil G. Sheps Center for Health Services Research, The University of North Carolina at Chapel Hill, 725 Martin Luther King Jr Blvd, Chapel Hill, NC 27599 (e-mail: Jessica.young@unc.edu). Reprints can be ordered at http://www.ajph.org by clicking the "Reprints" link.

PUBLICATION INFORMATION

Full Citation: Young JC, Pack C, Gibson TB, et al. Machine learning can unlock insights into mortality. *Am J Public Health*. 2021;111(S2):S65–S68.

Acceptance Date: May 17, 2021.

DOI: https://doi.org/10.2105/AJPH.2021.306418

CONTRIBUTORS

J. C. Young led the structuring of the article and authored the introduction, potential pitfalls, and conclusions sections and portions of the machine learning and research implications sections and was responsible for soliciting feedback and integrating the text and author contributions. C. Pack led the authoring of the machine-learning methods text. C. Pack, T. B. Gibson, F. Yoon, D. E. Irwin, S. Shiv, T. Cooper, and N. Dasgupta reviewed multiple versions of the text and provided comments. T. B. Gibson led the authoring of the implications on research text. F. Yoon was heavily involved in machine-learning efforts and assisted with the text describing these methods. D. E. Irwin provided general project direction and contributed expertise on claims data and mortality. S. Shiv provided oversight for the study team, was involved in discussing methods and clarifying the text, and reviewed the text before submission. T. Cooper managed the larger project and coordinated the study team. N. Dasgupta was responsible for the synthesis of the original project and team, detailed the direction for the article and the vision for the overall message, made substantial additions to the article, and provided feedback.

ACKNOWLEDGMENTS

This work is supported by the following sources: US Food and Drug Administration (grant HHSF223201810183C); the National Institutes of Health, National Center for Advancing Translational Sciences (grants UL1TR001998, NIH NCATS UL1TR002489); and a National Research Service Award Pre-Doctoral/Post-Doctoral Traineeship from the Agency for Healthcare Research and Quality sponsored by the Cecil G. Sheps Center for Health Services Research, University of North Carolina at Chapel Hill (grant T32-HS000032).

We thank the editors and reviewers at *AJPH* for their feedback on this article.

Note. The views expressed in this article do not necessarily reflect those of the funders.

CONFLICTS OF INTEREST

J. C. Young is a consultant to CERobs Consulting, LLC, which had no role in this work. N. Dasgupta is a methods consultant to the RADARS System of Denver Health and Hospitals Authority, a political subdivision of the State of Colorado, which had no role in this work. N. Dasgupta does not accept compensation from any pharmaceutical company or distributor.

REFERENCES

1. Centers for Disease Control and Prevention, National Center for Health Statistics. About underlying cause of death 1999–2019. Available at: http://wonder.cdc.gov/ucd-icd10.html. Accessed April 1, 2021.
2. Berry SD, Ngo L, Samelson EJ, Kiel DP. Competing risk of death: an important consideration in studies of older adults. *J Am Geriatr Soc.* 2010;58(4): 783–787. https://doi.org/10.1111/j.1532-5415.2010.02767.x
3. Grein J, Ohmagari N, Shin D, et al. Compassionate use of remdesivir for patients with severe COVID-19. *N Engl J Med.* 2020;382(24):2327–2336. https://doi.org/10.1056/NEJMoa2007016
4. Beam AL, Kohane IS. Big data and machine learning in health care. *JAMA.* 2018;319(13):1317–1318. https://doi.org/10.1001/jama.2017.18391
5. Shmueli G. To explain or to predict? *Stat Sci.* 2010; 25(3):289–310. https://doi.org/10.1214/10-STS330
6. Bzdok D, Krzywinski M, Altman N. Points of significance: machine learning: a primer. *Nat Methods.* 2017;14(12):1119–1120. https://doi.org/10.1038/nmeth.4526
7. Reps JM, Rijnbeek PR, Ryan PB. Identifying the DEAD: development and validation of a patient-level model to predict death status in population-level claims data. *Drug Saf.* 2019;42(11):1377–1386. https://doi.org/10.1007/s40264-019-00827-0
8. Panch T, Mattie H, Atun R. Artificial intelligence and algorithmic bias: implications for health systems. *J Glob Health.* 2019;9(2):010318. https://doi.org/10.7189/jogh.09.020318
9. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science.* 2019;

366(6464):447–453. https://doi.org/10.1126/science.aax2342

10. Espey DK, Jim MA, Richards TB, Begay C, Haverkamp D, Roberts D. Methods for improving the quality and completeness of mortality data for American Indians and Alaska Natives. *Am J Public Health.* 2014;104(suppl S3):S286–S294. https://doi.org/10.2105/AJPH.2013.301716
11. Miller EA, McCarty FA, Parker JD. Racial and ethnic differences in a linkage with the National Death Index. *Ethn Dis.* 2017;27(2):77–84. https://doi.org/10.18865/ed.27.2.77
12. Noymer A, Penner AM, Saperstein A. Cause of death affects racial classification on death certificates. *PLoS One.* 2011;6(1):e15812. https://doi.org/10.1371/journal.pone.0015812
13. Koivu A, Sairanen M, Airola A, Pahikkala T. Synthetic minority oversampling of vital statistics data with generative adversarial networks. *J Am Med Inform Assoc.* 2020;27(11):1667–1674. https://doi.org/10.1093/jamia/ocaa127
14. Choi D, Sumner SA, Holland KM, et al. Development of a machine learning model using multiple, heterogeneous data sources to estimate weekly US suicide fatalities. *JAMA Netw Open.* 2020;3(12):e2030932. https://doi.org/10.1001/jamanetworkopen.2020.30932
15. Park Y, Hu J, Singh M, et al. Comparison of methods to reduce bias from clinical prediction models of postpartum depression. *JAMA Netw Open.* 2021;4(4):e213909. https://doi.org/10.1001/jamanetworkopen.2021.3909
16. Pierson E, Cutler DM, Leskovec J, Mullainathan S, Obermeyer Z. An algorithmic approach to reducing unexplained pain disparities in underserved populations. *Nat Med.* 2021;27(1):136–140. https://doi.org/10.1038/s41591-020-01192-7