frontiers in Epidemiology

SPECIALTY GRAND CHALLENGE
published: 18 November 2021
doi: 10.3389/fepid.2021.786988

Grand Challenge—Crossing Borders to Develop Epidemiologic Methods

Rolf H. H. Groenwold [1,2]*

[1] Department of Clinical Epidemiology, Leiden University Medical Center, Leiden, Netherlands, [2] Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, Netherlands

Keywords: bias (epidemiology), bias analysis, prediction, replicability, electronic health record (EHR), statistical analysis

INTRODUCTION

Over the last decades, the epidemiologic landscape has changed dramatically. The amount of data that are currently being made available for epidemiologic research are unprecedented (1). Some argue that the sheer size of, for instance, electronic health records (EHR) databases, in combination with their representativeness of daily clinical practice (the "real world"), carries enormous potential for clinical research (2, 3). Due to open access initiatives, e.g., by funding agencies and journals, datasets are increasingly made available for new collaborative analyses between researchers.

Although the availability of datasets that are both representative of the target population and large (or huge) holds promise in terms of applicability (or generalizability) of research results, at the same time, it still requires valid and accurate methodology. Merely due to the size of many registry-based epidemiologic studies, confidence intervals around effect estimates tend to be very narrow indicating precision. Often, precision is no longer the limiting factor when it comes to applying research results. However, the role of bias can still be considerable: even modest bias can substantially change a study's conclusions (4). Nevertheless, with large datasets, the potential to investigate multiple independent research questions increases, such as on subgroup effects. Yet by studying subgroups of patients in the database, the apparent value of using large datasets is lost, and common epidemiologic challenges become more prominent. Thus, efficient modeling is needed to adequately and appropriately conduct these analyses.

While datasets have become larger, more representative of daily practice, and more accessible to researchers who did not originally collect the data, epidemiologic methods for data collection and data analysis have been refined and are still being improved. Arguably, the development and innovation of new methods occurs within disciplines that seem to become more and more specialized. Below, I discuss different divisions into disciplines and argue that we should look beyond the boundaries of those disciplines to make the next step in developing methods for contemporary epidemiologic research.

ETIOLOGY vs. PREDICTION

The field of epidemiology has different subfields. While "traditional" epidemiology can be defined as "the investigation of causes of health-related events in populations" (5), which closely links it to health etiology, *clinical epidemiology* appears to be broader by also including research into diagnostic test accuracy, prognosis modeling, and therapeutic interventions (6). Historically, research into methods for etiologic research focused on confounding and other sources of bias, whereas e.g., prediction modeling dealt with optimizing individual risk predictions (7). Although this distinction is artificial and probably too simplistic, it appears to exist in teaching, research, and—importantly—methods development.

OPEN ACCESS

Edited and reviewed by:
David Rich,
University of Rochester, United States

***Correspondence:**
Rolf H. H. Groenwold
r.h.h.groenwold@lumc.nl

Specialty section:
This article was submitted to Research Methods and Advances in Epidemiology, a section of the journal Frontiers in Epidemiology

Received: *30 September 2021*
Accepted: *26 October 2021*
Published: *18 November 2021*

Citation:
Groenwold RHH (2021) Grand Challenge—Crossing Borders to Develop Epidemiologic Methods. Front. Epidemiol. 1:786988. doi: 10.3389/fepid.2021.786988

Nevertheless, the different subfields can learn from each other, and ideas developed in one subfield can spark developments in another. For example, discussions about estimands (8) and target trial emulation (9) started in research areas with a focus on causality, notably the effects of medical treatments. Nevertheless, these ideas appear worthwhile in a prediction modeling context, too. Similar considerations help to articulate which predictions could be made and how different data analytic strategies potentially lead to predictions with a different meaning (10). Likewise, where measurement error was originally a topic that was deemed relevant only for research of causal effects, recent work suggests that the conceptual framework behind measurement error can be used to understand the impact of variations in measurement procedures when working with prediction models (11). The other way around, penalization and shrinkage methods entered the epidemiological toolbox *via* prediction modeling. Yet, they may also be useful in studies of causal effects, for example, when confronted with a large set of potential confounding variables (12).

MACHINE LEARNING vs. TRADITIONAL STATISTICAL METHODS

Another distinction within epidemiology that is often made is that between machine learning and "traditional" statistical methods (13). This distinction focuses on the analytical methods that are applied. Yet, the technical aspects are not distinctive; many texts about machine learning start with a description of, for example, linear regression, which is also at the heart of textbooks on (medical) statistics (14). Nevertheless, the more conceptual approach to data analysis may be different between machine learning and "traditional" statistical methods. Where traditional statistical methods focus on testing (pre-specified) hypotheses, machine learning is more open to unraveling (unexpected?) patterns in a dataset. This distinction makes it obvious to resort to traditional statistical methods for etiological research, while machine learning is more popular in prediction modeling research. Nevertheless, data-driven machine learning methods are increasingly applied and could very well have an added value in research into causal effects too (15).

REPLICATION OF METHODOLOGICAL RESEARCH

Over the past decades, replicability has increasingly received attention in the social and biomedical sciences, including epidemiology. However, methodological studies seem to lag behind in this development, and statistical simulation studies are no exception. Under strict assumptions and for relatively simple methods, it is possible to mathematically derive how statistical methods will behave when applied to real data, e.g., what type-I error rates will be or to what extent a method can identify an association if it truly exists. However, there are many situations in which this is not the case, for instance, when methods are applied in complex high-dimensional data structures, and when simulation studies may offer a solution (16).

Simulation studies have an important role in guiding the planning of studies and data analysis of applied biomedical research. A striking example is the simulation study by Peduzzi et al. on the sample size required for logistic regression analysis, one of the most used statistical models in the biomedical sciences (17). Since its publication in 1996, this simulation study has had a major impact, has been cited > 6,000 times, and even led to a widely used one-in-ten rule about the number of variables that can be included in a multivariable logistic regression model. However, a replication study by Van Smeden et al. could not confirm these results and questions whether this "one-in-ten rule" is appropriate (18). Undoubtedly, simulation studies are a powerful tool for methodological research, yet this example illustrates that they may have limitations too and that their results are clearly not definitive. Instead of considering the results of simulation studies to be set in stone, simulation studies need to be replicated, just as empirical studies need replication (19).

CROSSING BORDERS

In parallel with spectacular developments in terms of the dimension of datasets available for epidemiological research, we have witnessed developments regarding the methods being applied to analyse those data. Even though the abovementioned distinctions may appear a bit artificial, they also illustrate that looking for appropriate methodology outside of one's own field may prove worthwhile. When doing so, we need to keep an eye out on the appropriateness of those methods and the fundamental questions about whether (and how) the new methods still provide correct and interpretable answers to the research questions that are being asked. Testing and refining methods and discussing their suitability within a new field of application is obviously essential. Often innovations come from adjacent fields, and why would epidemiology be an exception?

The section *Research Methods and Advances in Epidemiology* aims to contribute to the development of new and existing epidemiologic methods and to help researchers cross the borders that have emerged between epidemiological fields. This section of Frontiers in Epidemiology seeks not only original research describing new methods, but will also seek challenges of the use of existing methods for modern epidemiology research questions, as well as replication of already developed analytical methods for such research questions.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

ACKNOWLEDGMENTS

M. S. Jansen, B. B. L. Penning de Vries, L. Nab, and T. Kurth provided comments on an earlier version of this manuscript.

REFERENCES

1. Raghupathi W, Raghupathi V. Big data analytics in healthcare: promise and potential. *Health Inf Sci Syst.* (2014) 2:3. doi: 10.1186/2047-2501-2-3
2. Beam AL, Kohane IS. Big data and machine learning in health care. *J Am Med Assoc.* (2018) 319:1317–8. doi: 10.1001/jama.2017.18391
3. McKinstry B. All watched over by machines of loving grace: an optimistic view of big data. *BMJ.* (2017) 358:j3967. doi: 10.1136/bmj.j3967
4. Kaplan RM, Chambers DA, Glasgow RE. Big data and large sample size: a cautionary note on the potential for bias. *Clin Transl Sci.* (2014) 7:342–6. doi: 10.1111/cts.12178
5. Morabia A. *A History of Epidemiologic Methods and Concepts*. Berlin/Heidelberg: Springer Science & Business Media (2005). doi: 10.1007/978-3-0348-7603-2
6. Grobbee DE, Hoes AW. *Clinical Epidemiology: Principles, Methods, and Applications for Clinical Research*. Burlington, MA: Jones & Bartlett Publishers (2014).
7. Shmueli G. To explain or to predict? *Statist Sci.* (2010) 25:289–310. doi: 10.1214/10-STS330
8. Lipkovich I, Ratitch B, Mallinckrodt CH. Causal inference and estimands in clinical trials. *Statist Biopharmaceut Res.* (2020) 12:54–67. doi: 10.1080/19466315.2019.1697739
9. Hernán MA, Robins JM. Using big data to emulate a target trial when a randomized trial is not available. *Am J Epidemiol.* (2016) 183:758–64. doi: 10.1093/aje/kwv254
10. van Geloven N, Swanson SA, Ramspek CL, Luijken K, van Diepen M, Morris TP, et al. Prediction meets causal inference: the role of treatment in clinical prediction models. *Eur J Epidemiol.* (2020) 35:619–30. doi: 10.1007/s10654-020-00636-1
11. Luijken K, Wynants L, van Smeden M, Van Calster B, Steyerberg EW, Groenwold RHH. Changing predictor measurement procedures affected the performance of prediction models in clinical examples. *J Clin Epidemiol.* (2020) 119:7–18. doi: 10.1016/j.jclinepi.2019.11.001
12. Greenland S. Invited commentary: variable selection vs. shrinkage in the control of multiple confounders. *Am J Epidemiol.* (2008) 167:523–9. doi: 10.1093/aje/kwm355
13. Breiman L. Statistical modeling: the two cultures. *Statist Sci.* (2001) 16:199–215. doi: 10.1214/ss/1009213726
14. Friedman J, Hastie T, Tibshirani R. *The Elements of Statistical Learning*. New York, NY: Springer Series in Statistics (2001).
15. Hernán MA, Hsu J, Healy B. A second chance to get causal inference right: a classification of data science tasks. *Chance.* (2019) 32:42–9. doi: 10.1080/09332480.2019.1579578
16. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. *Stat Med.* (2019) 38:2074–102. doi: 10.1002/sim.8086
17. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. *J Clin Epidemiol.* (1996) 49:1373–9. doi: 10.1016/S0895-4356(96)00236-3
18. van Smeden M, de Groot JA, Moons KG, Collins GS, Altman DG, Eijkemans MJ, et al. No rationale for 1 variable per 10 events criterion for binary logistic regression analysis. *BMC Med Res Methodol.* (2016) 16:163. doi: 10.1186/s12874-016-0267-3
19. Lohmann A, Astivia OLO, Morris TP, Groenwold RH. *It's Time! 10+ 1 Reasons We Should Start Replicating Simulation Studies*. Available online at: https://psyarxiv.com/agsnt (accessed September 16, 2021).

Conflict of Interest: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.