ELSEVIER

Journal of Clinical Epidemiology 114 (2019) v−vi

Journal of Clinical Epidemiology

EDITORIAL

Confounding obscures our view, effect modification is part of reality

In 1983, one of us (AK) attended a course given by one of the world's most prominent scholars in epidemiological methodology, Olli Miettinen [1], and heard him say that in the face of the Lord there is no confounding, but there is effect modification. This may well be the clearest available definition of the difference between two concepts which are still too often confused and not appropriately distinguished, theoretically, and in data-analytic practice dealing with 'covariables.'

Miettinen's compact expression makes seemingly complex things immediately understandable. If we could fully know and see reality, there would be no confounding to struggle with. But with that same sharp eye we would see reality in all its variety and diversity, including differences between subgroups and individuals that modify the effects of determinants.

At the same time, this presentation makes clear that we as researchers, not having all-seeing eyes, must develop and appropriately apply methods to avoid or eliminate confounding as much as possible, by design (e.g., by randomization) and analysis (e.g., by adjusting for it). It also makes clear that for effect modification the challenge for researchers is nothing of the sort. On the contrary, we should not eliminate it but get the sharpest possible picture of it, by design (e.g., by studying effect differences between subgroups, or using N of 1 designs [2,3]); analysis (e.g., by evaluating interactions); and meta-analysis (e.g., by synthesizing results of multiple studies conducted in diverse populations, in order to identify heterogeneïty and its determinants).

These considerations apply to the study of the whole spectrum of etiology, pathogenesis, diagnosis, prognosis, prevention and treatment, although there may be differences in objectives and consequences. For example, in studying etiology and effectiveness, the goal is clarifying causal relationships, while in studying diagnosis and prognosis, identifying reproducible, consistently predictive - but not necessarily causal - determinants may be sufficient.

Ultimately, for patient care, we strive to use all available knowledge, tailored as much as possible in the context of an individualized approach. This implies effect modification focused on $N = 1$, thoughtfully using as many relevant - obviously valid - data as are applicable [4−7].

Against this background, the work of Van Klaveren et al. is highly interesting. They argue that the main goal of predictive analyses of heterogeneous treatment effects is to develop models to predict which of a number treatment options will be better for a particular individual. In this context, the authors compared different regression modeling approaches, with and without interactions, for the prediction of heterogeneous treatment effects. Trial samples from a population with 12 binary risk predictors of treatment benefit, both without and with true treatment interactions, were simulated. The group assessed a "risk model" (with the treatment effect being constant) and three "effect models" (including interactions of risk predictors with treatment), while evaluating three novel performance measures focused on, respectively, calibration, discrimination, and prediction error for benefit. The authors conclude that, consistently, a risk modeling approach yields models well calibrated for benefit, whereas in the presence of a true interaction effect, modeling may improve discrimination for benefit but is prone to overfitting. According to the investigators, effect models should only be considered when treatment interactions are plausible and should always be fitted with penalized regression. The findings of these authors underline that evaluation of interaction is not primarily a statistical endeavor, but needs to start from thorough biomedical and clinical prior knowledge.

In order to match study results with those to whom these results can be considered applicable, an optimal description of the included subjects is a basic requirement, together with a transparently described process of recruitment and selection. Epidemiologic and clinical research publications generally describe the study sample in the first table. Hayes-Larson c.s. review how to make 'Table 1' as useful as possible. They found little appropriate guidance for designing a Table 1, especially for complex study designs and analyses. Therefore, the authors synthesized and developed reporting recommendations for Table 1, driven by study design, focusing on transparency regarding threats to internal and external validity. With respect to possible effect modification and confounding, they recommend to show the distributions of all variables according to strata of both the exposure and the modifier, and the distributions of the exposure and modifier in the total sample. They also highlight some analytic complexities common in epidemiologic research and possible related Table 1 variations. At the same time, the authors address the balance between comprehensiveness and reader-friendliness. While concentrating on validity of studies on causal effects, their considerations may be also relevant for other study types.

https://doi.org/10.1016/j.jclinepi.2019.08.011
0895-4356

Effect modification should also be considered in evaluating the effectiveness of medical products by independent authorities such as the Food and Drugs Administration (FDA). This is a relevant aspect in a paper by Ladanie and his team, who focused on novel cancer therapies that are often approved with evidence from a single pivotal trial alone, which raises concerns about the credibility of this evidence. They carried out a metaepidemiologic evaluation of single pivotal trials supporting FDA approval of novel drugs and therapeutic biologicals for cancers. For each trial, the authors determined the presence of five characteristics of pivotal trial evidence described by the FDA, that may indicate higher validity and justify the reliance on a single trial alone. These were operationalized as (1) large and multicenter trial; consistent treatment benefits across (2) multiple patient subgroups, (3) multiple endpoints, and (4) multiple treatment comparisons; and (5) ‘‘statistically very persuasive’’ results. It was found that single pivotal trials typically have some of the above five ‘‘corroborating’’ characteristics, but often only one or two. The authors also conclude that these characteristics need to be better operationalized, defined, and reported. With regard to the evaluation of effect modification, for example, the authors emphasize that the characteristic of consistent effects across subgroups (i.e., no indication for effect modification) is problematic given problems of low statistical power, while on the other hand, reported subgroup differences are often not credible. The authors conclude that it is not yet clear whether single trials, even those fulfilling the above criteria, can provide as strong evidence about benefits and harms of novel treatments as multiple trials would do.

If precise diagnostic decision making is the aim, a strong and consistent predictor of the diagnosis in fact ‘causes’ this to be achieved. Clinical spectrum can then be seen as a ‘diagnostic effect modifier’, as test characteristics such as sensitivity or specificity can vary across strata of development of a clinical presentation or disorder [8]. As spectrum differences can be a result of selection, for instance by referral, these can also be associated with prevalence [9]. Given the need for more empirical data on such diagnostic effect modification, the work of Holtmann et al. to develop practical recommendations for diagnostic accuracy studies in low-prevalence situations, is of special relevance. As they make clear, low disease prevalence poses specific challenges for diagnostic accuracy studies given the large sample sizes that are needed. The authors evaluated design options for diagnostic accuracy studies in low-prevalence situations, by conducting a literature search and discussing reported designs. They identified six designs for diagnostic accuracy studies that they deemed suitable in low-prevalence situations because these reduced the required total number of study subjects and of those undergoing the most burdensome index test or reference standard - procedure. The pros and cons of these designs were described, and the related risk of bias, study efficiency, and alignment with the clinical pathway in routine care were evaluated. It was concluded that in preparing a diagnostic accuracy study in low-prevalence situations, choosing the study design should depend on whether the aim is to limit the number of patients undergoing the index test or reference standard, and the risk of bias associated with particular design types.

To conclude, while we must keep doing our best to eliminate confounding - or when this is not possible, to adjust for it -, we need to identify, measure and manage effect modification as part of reality. We should do more to develop appropriate methods to maximally achieve this, as an important step towards more tailored, individualized care.

J. André Knottnerus
Peter Tugwell
E-mail address: anneke.germeraad@maastrichtuniversity.nl
(J.A. Knottnerus)

References

[1] Miettinen OS. Theoretical epidemiology: Principles of occurrence research in medicine. New York: John Wiley & Sons; 1985.

[2] Guyatt G, Sackett D, Taylor DW, Chong J, Roberts R, Pugsley S. Determining optimal therapy−randomized trials in individual patients. N Engl J Med 1986;314:889−92.

[3] Vohra S. N-of-1 trials to enhance patient outcomes: identifying effective therapies and reducing harms, one patient at a time. J Clin Epidemiol 2016;76:6−8.

[4] Knottnerus JA. Role of the electronic patient record in the development of general practice in The Netherlands. Methods Inf Med 1999;38(4−5):350−4.

[5] Govers TM, Rovers MM, Brands MT, Dronkers EAC, Baatenburg de Jong RJ, Merkx MAW, et al. Integrated prediction and decision models are valuable in informing personalized decision making. J Clin Epidemiol 2018;104:73−83.

[6] Ioannidis JPA, Khoury MJ. Evidence-based medicine and big genomic data. Hum Mol Genet 2018;27(R1):R2−7.

[7] Miettinen OS, Steurer J, Hofman A. Clinical research transformed. 1st ed. Springer Nature Switzerland AG; 2019.

[8] Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. N Engl J Med 1978;299:926−30.

[9] Leeflang MM, Bossuyt PM, Irwig L. Diagnostic test accuracy may vary with prevalence: implications for evidence-based diagnosis. J Clin Epidemiol 2009;62:5−12.