



FEATURE

CHRISTMAS 2016: FOOD FOR THOUGHT

Is caviar a risk factor for being a millionaire?

Anders Huitfeldt argues that the answer depends on your definition of “risk factor” and calls for greater clarity in research

Anders Huitfeldt *postdoctoral scholar*

Meta-Research Innovation Center at Stanford, Stanford University School of Medicine, Palo Alto, CA, USA

The risk factor approach to epidemiology was introduced by the Framingham Heart Study investigators,^{1 2} who first alluded to the idea in 1951.³ The first use of the term “factor of risk” appeared in 1961,⁴ but it was not precisely defined. The resulting semantic confusion has hindered precise communication about study design and data analysis. To illustrate the problem, let us suppose that you want to study the causes and distribution of personal wealth. You have a secretive friend, and, among other questions, you are interested in knowing whether he is a millionaire. You are aware that there are some attributes, or risk factors, that are thought to be linked to being a millionaire. You decide to investigate.

What is a risk factor?

The first step is to choose your definition of risk factor. Clinical research can generally be divided into four broad objectives based on the intended use of the information obtained by the study: diagnosis, prognosis, treatment effects, and aetiology. Each of these research objectives is associated with a different definition. Table 1¹ gives examples of how these four definitions of risk factor are used in the scientific literature and shows how each definition describes a different relation between the dependent variable and the independent variable.

A variable may qualify as a risk factor under more than one definition of the term. For example, cholesterol is believed to be a risk factor for heart disease under each of the four definitions. However, it is generally not plausible to assume that a variable that is a risk factor according to one definition will always be a risk factor under the other definitions. Commonly used statistical techniques do not automatically differentiate between the types of relation described in table 1¹,¹⁵ and a model adapted to studying one type may not be appropriate for another.¹⁶ Therefore, when conducting observational studies, data analysis needs to be designed to match the particular definition that is being considered.

No study design can resolve a semantics discussion

Returning to the secretive friend, you decide to find out whether he is a millionaire (diagnosis) by exploring the association between caviar and wealth. You conduct a large scale prediction study to answer this question and confirm that people who eat caviar are more likely to be millionaires. Therefore, when you observe that your dining companion orders a fine beluga, you heighten your suspicion that he is wealthy.

Next, hoping to piggyback off the success of another promising young acquaintance, you turn your interest to predicting whether a person will become a millionaire in the future (prognosis). However, after a follow-up study lasting several years, you conclude that there is little to be gleaned from knowing a person’s baseline caviar consumption. You suspect that this is because there are several mechanisms that are active simultaneously. For example, some people who have acquired a taste for caviar are fortunate to have rich parents and may expect a princely inheritance in the near future. Others who enjoy caviar do not have rich parents and are at risk of going broke. As it turns out from your study, these two phenomena cancel each other out on average. Caviar consumption in the present is therefore not a reliable prognostic factor for future wealth.

After falling out with your young friend, you conclude that you will have to make your first million on your own. In looking for insight on what actions you can take to bring about this objective (treatment effects), you conduct several randomised trials to estimate the causal effect of eating caviar. In these trials you recruit a large number of non-millionaires and randomly assign half of them to spend most of their income on caviar whereas the other half avoids it altogether. Regrettably, you find that not only does caviar not make you a millionaire, it has an unfortunate tendency to bankrupt a person.

After spending a few years resolving some problems with your bank and the institutional review board, you turn your attention

to understanding the role of caviar in the mechanism of wealth creation (aetiology). You suspect that outside randomised trials, people who are susceptible to the detrimental effects of caviar always become addicted to the delicious black stuff and therefore never become millionaires. It is therefore possible that among those who become millionaires, nobody can attribute their success to having avoided caviar—that is, the absence of caviar had no role in the aetiology. However, after consulting numerous textbooks on causal inference, you discover that this hypothesis is not testable with currently available statistical theory without strong and often unrealistic assumptions. Therefore, you fail to reach a conclusion.

Your extensive studies have shown that caviar is useful for predicting if someone is a millionaire (diagnosis/detection) but not for predicting if they will become a millionaire in the future (prognosis). Furthermore, you conclude that excessive consumption may reduce your probability of becoming a millionaire (treatment effect) but you are unable to answer questions about caviar's role in the mechanism of wealth creation (aetiology) without relying on questionable assumptions.

This brings us to the first crucial point: If scientist A asserts that caviar is a risk factor, it is unclear which type of relation he is referring to. Therefore, if scientist B disagrees but uses a different definition of risk factor, they both may be right. There is no study design that can resolve this disagreement: the scientists are not arguing about the underlying reality but about who gets to define the term risk factor.

Epidemiological studies with ambiguous research objectives

The BMJ declines to publish the findings from your four studies, and your painstaking work is lost to the dustbin of history.

Several years later you are asked to serve as a peer reviewer on a large observational study that aims to determine whether caviar is a risk factor for being a millionaire.

After you make some initial requests for clarification, the editor tells you that risk factor is a standard term in epidemiology, that, of course, you understand what it means, and why do you have to get so pedantic all of a sudden? Your task is simply to determine whether the study supports its conclusions or, in other words, whether the study design and data analysis can be expected to find “the truth.”

That depends on what the authors are trying to find out: a method that will simultaneously answer questions about diagnosis, prognosis, treatment effects, and aetiology cannot exist. How could such a method exist, when these questions can have different answers? In other words, the relevant methodological questions the reader engages with to determine if the conclusions are supported depend on what the authors are trying to achieve—that is, in which definition of risk factor they are interested.

This brings us to the second crucial point. Unless the research objective is clearly defined in terms of an explicitly stated definition of risk factor, it is not possible to evaluate whether the study design and data analysis are appropriate to answer the research questions, and therefore not possible to evaluate the credibility of the study or its conclusions.

Semantic ambiguity leads to methodological confusion

The authors of the observational study describe their methods. They have a large dataset with many relevant variables, and use a prediction model with covariates selected by an automated algorithm. Unsurprisingly, they conclude that caviar is a risk factor for wealth, but take great care to avoid using the word “cause” other than in the context of a cliché about the logical implications of correlations.

However, you notice that the paper contains some discussion of seemingly causal questions, such as mechanism of action and policy implications. You also notice that they control extensively for confounding. This is puzzling. Confounding is a phenomenon that complicates our attempts to estimate causal effects, but is not relevant if the goal of the research is to reduce diagnostic or prognostic uncertainty.¹⁵ Some obvious questions are therefore immediately raised: Why do the authors control for confounding when their goal is not to estimate a causal effect? Was this not a prediction model?

At this stage, you begin to wonder whether it was such a great idea to unify four types of relation into a single term. Are the authors mixing up distinct concepts, each having different implications for how the study should be designed and analysed? Does the paper answer neither question because it mixes methods that were intended to answer distinct questions?

Implications for observational research

As the example shows, the appropriate analytical approach depends on which definition of risk factor the investigators had in mind. If this is left unspecified, the paper does not contain enough information to evaluate whether the conclusions are credible, which brings the purpose of peer review into question.

Some have advocated reducing ambiguity by settling on a single definition of risk factor. For example, Miquel Porta's *Dictionary of Epidemiology*¹⁷ defines a risk factor as “a factor that is causally related to the change in the risk of a relevant health process, outcome, or condition.” However, this approach can only solve the problem if all researchers agree to use the term only in this sense. Moreover, this definition implicitly assumes that epidemiologists are only interested in causality, to the exclusion of other worthy research objectives such as reducing diagnostic¹⁸ or prognostic¹⁹ uncertainty.

Instead, I suggest that journal editors should enforce a taboo²⁰ on the term “risk factor,” thereby forcing investigators to spell out exactly what they mean by the term. For example, authors could be required to specify whether they are interested in a diagnostic factor, a prognostic factor, an aetiological factor, or a treatment effect. Only then will it be possible for readers to understand exactly what the investigators intended to learn, and to engage in productive scientific conversation about whether they succeeded in accounting for the biases associated with that particular research objective.

I thank Sonja Swanson, Galit Shmueli, Andrew Watt, Steve Goodman, Sander Greenland, Ada Cohen, and Etsuji Suzuki for helpful comments.

Competing interests: I have read and understood BMJ policy on declaration of interests and have no relevant interests to declare.

Provenance and peer review: Not commissioned; externally peer reviewed.

1 Dawber TR, Moore FE, Mann GV. II. Coronary heart disease in the Framingham Study. *Int J Epidemiol* 2015;44:1767-80. doi:10.1093/ije/dyv346 pmid:26705414.

Key messages

The definition of "risk factor" will vary depending on whether a research question is exploring diagnosis, prognosis, treatment effects, or aetiology

Unless a definition is specified, it is not possible for readers of research papers to understand what the investigators attempted to learn or evaluate whether they succeeded in their objectives

Journal editors should require authors to specify the intended use of the research findings and ensure that the methods were appropriate

- 2 Aronowitz RA. *Risky medicine: our quest to cure fear and uncertainty*. University of Chicago Press, 2015;doi:10.7208/chicago/9780226049854.001.0001.
- 3 Dawber TR, Meadors GF, Moore FE Jr. Epidemiological approaches to heart disease: the Framingham Study. *Am J Public Health Nations Health* 1951;41:279-81. doi:10.2105/AJPH.41.3.279 PMID:14819398.
- 4 Kannel WB, Dawber TR, Kagan A, Revotskie N, Stokes J 3rd. Factors of risk in the development of coronary heart diseasesix year follow-up experience. The Framingham Study. *Ann Intern Med* 1961;55:33-50. doi:10.7326/0003-4819-55-1-33 PMID:13751193.
- 5 Skinner JS, Smeeth L, Kendall JM, Adams PC, Timmis A. Chest Pain Guideline Development Group. NICE guidance. Chest pain of recent onset: assessment and diagnosis of recent onset chest pain or discomfort of suspected cardiac origin. *Heart* 2010;96:974-8. doi:10.1136/hrt.2009.190066 PMID:20538674.
- 6 Wilson PW, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of coronary heart disease using risk factor categories. *Circulation* 1998;97:1837-47. doi:10.1161/01.CIR.97.18.1837 PMID:9603539.
- 7 Levine GN, Keaney JF Jr, Vita JA. Cholesterol reduction in cardiovascular disease. Clinical benefits and possible mechanisms. *N Engl J Med* 1995;332:512-21. doi:10.1056/NEJM199502233320807 PMID:7830734.
- 8 Hernán MA, Robins JM. *Causal inference*. Chapman and Hall, CRC, 2016.
- 9 Brown MS, Goldstein JL. How LDL receptors influence cholesterol and atherosclerosis. *Sci Am* 1984;251:58-66. doi:10.1038/scientificamerican1184-58 PMID:6390676.
- 10 Smith GD, Ebrahim S. Mendelian randomization: genetic variants as instruments for strengthening causal inference in observational studies. In: Weinstein M, Vaupel J, Wachter K, eds. *Biosocial surveys*. National Academies Press, 2008.
- 11 Gelman A, Imbens G. Why ask why? Forward causal inference and reverse causal questions. NBER working paper No 19614. 2013. <http://www.nber.org/papers/w19614>
- 12 Greenland S, Robins JM. Conceptual problems in the definition and interpretation of attributable fractions. *Am J Epidemiol* 1988;128:1185-97.PMID:3057878.
- 13 Suzuki E, Yamamoto E, Tsuda T. On the relations between excess fraction, attributable fraction, and etiologic fraction. *Am J Epidemiol* 2012;175:567-75. doi:10.1093/aje/kwr333 PMID:22343634.
- 14 Rothman KJ. Causes. *Am J Epidemiol* 1976;104:587-92.PMID:998606.
- 15 Shmueli G. To explain or to predict? *Stat Sci* 2010;25:289-310doi:10.1214/10-STS330.
- 16 Shmueli G, Koppius O. Predictive analytics in information systems research. *Manage Inf Syst Q* 2011;35.
- 17 Porta M. *A dictionary of epidemiology*. Oxford University Press, 2008.
- 18 Hendriksen JMT, Geersing GJ, Moons KGM, de Groot JAH. Diagnostic and prognostic prediction models. *J Thromb Haemost* 2013;11(Suppl 1):129-41. doi:10.1111/jth.12262 PMID:23809117.
- 19 Khullar D, Jena AB. Reducing prognostic errors: a new imperative in quality healthcare. *BMJ* 2016;352:i1417. doi:10.1136/bmj.i1417 PMID:26993146.
- 20 Yudkowsky ES. A human's guide to words. In: *Rationality: from AI to zombies*. Machine Intelligence Research Institute, 2015.

Published by the BMJ Publishing Group Limited. For permission to use (where not already granted under a licence) please go to <http://group.bmj.com/group/rights-licensing/permissions>

Table

Table 1| Objectives of clinical research and associated definitions of risk factor

| Research objective | Definition of risk factor | Suggested term | Example of application | Preferred data analysis or study design | Relevant biases and shortcomings |
|--------------------------|---|---------------------|--|---|---|
| Diagnosis | Any personal attribute that can be used to make a diagnosis more reliable | Diagnostic factor | Serum cholesterol in people presenting with chest pain ⁵ | Prediction model with binary outcome variable (measured at the same time as the diagnostic factor)† | Ascertainment of outcome may have imperfect sensitivity and specificity. Model may be overfit to training dataset |
| Prognosis | Any personal attribute that can be used to make more reliable predictions about future risk of medical conditions | Prognostic factor | Serum cholesterol predicts future cardiovascular disease ⁶ | Prediction model with time-to-event outcome variable | As above |
| Treatment effects | An action that may be taken to increase or decrease the probability of the outcome | Treatment effect | Cardiac risk is reduced by lowering serum cholesterol levels ⁷ | Randomised controlled trials. Observational studies with explicit causal models ⁸ | Confounding, selection bias, etc |
| Aetiology | A phenomenon, action, or substance that has a role in the aetiological mechanism | Aetiological factor | Cholesterol is involved in the mechanism behind atherosclerosis ⁹ | Some aetiological questions can be examined using the same methods as for treatment effects (eg, mendelian randomisation). ¹⁰ For others, there is no consensus on preferred study design. Relevant concepts include reverse causal inference, ¹¹ excess fraction, ¹² aetiological fraction, ¹³ and sufficient component cause models ¹⁴ | Imprecisely stated research questions because of current state of statistical methods |

*Note that not all commonly accepted risk factors for cardiovascular disease meet all four definitions. For example, family history is valid both as a prognostic factor and as a diagnostic factor, but if you attempt to reduce your patient's coronary risk by starting their parents on primary prevention, you are likely to be struck from the register. Some variables even have opposite effects depending on whether we are interested in prediction or causation. For example, if the patient's clinical history shows that he has had a coronary artery bypass graft, your risk estimate increases for the purposes of both diagnosis and prognosis, although the procedure itself almost certainly reduced his risk. †Such models are often termed "detection models" in the data mining literature, where they are used to detect fraud.