Journal of Physiotherapy 60 (2014) 241–244

Journal of PHYSIOTHERAPY

journal homepage: www.elsevier.com/locate/jphys

Appraisal

Research Note

Cohort studies of aetiology and prognosis: they're different

Introduction

Cohort designs are widely used in epidemiological research. The key features of cohort studies are that: (a) a sample of participants at risk of a particular outcome (the 'cohort') is identified; (b) data on individual cohort members' exposures to certain risk factors and their subsequent outcomes are obtained; and (c) associations between exposures and outcomes are quantified.

Here are two examples of cohort studies.

Paul and colleagues[1] conducted a range of physical and cognitive tests, including balance tests and the Mini-Mental State Examination, on 205 community-dwelling people who had Parkinson's disease. Subsequently, the participants used falls diaries to document any falls that occurred over the next 6 months. The analysis examined associations between physical and cognitive risk factors and the incidence rate of falls.

Moseley and colleagues[2] studied 1549 patients presenting to hospital with an acute wrist fracture. A number of clinical variables such as pain and swelling were assessed in the first week after fracture. Four months later, each participant was contacted by telephone and, where necessary, undertook a clinical examination to determine whether he or she had developed complex regional pain syndrome (CRPS). The analysis examined associations between the clinical variables and the development of CRPS.

Both of these studies were cohort studies, but they had quite different aims. The Paul study was designed to examine factors that *cause* a health outcome (in this case, falls). That is, the aim of the first study was to understand the mechanisms or aetiology of falls in people with Parkinson's disease. In contrast, the Moseley study was designed to identify factors that *predict* a health outcome (the development of CRPS after wrist fracture). Its aim was to generate prognoses for people who have had a wrist fracture.

In general, two types of cohort studies can be distinguished: aetiologic studies, which are concerned with understanding the mechanisms that cause health outcomes, and prognostic studies, which are concerned with prediction of health outcomes.

Even though the same research design—a cohort study—can be used to answer questions about both aetiology and prognosis, the way in which cohort studies of aetiology and prognosis are designed and analysed should be very different. Surprisingly, there have been few explicit discussions of this distinction in the epidemiology literature. Perhaps that is because, historically, epidemiology has overwhelmingly been concerned with questions about aetiology. The purpose of this Research Note is, therefore, to provide an introduction to differences in the design and analysis of cohort studies of aetiology and prognosis.

For more extensive discussions of design and analysis of studies of aetiology and prognosis, the interested reader is referred to the excellent introductory textbook by Grobbee and Hoes.[3] In addition, Holland[4] provides a rigorous grounding in the concept of causation, and Jewell[5] provides a clear, entry-level presentation of design of epidemiological studies aimed at understanding causation. Hernan and Robbins[6,7] provide a lucid presentation of contemporary statistical methods for studying causation, including the use of causal modelling in longitudinal studies. A group of leading methodologists recently published a series of articles that provide an overview of the design and analysis of prognostic studies.[8–11] Steyerberg[11] has written a very accessible textbook on statistical methods for prognostic research.

Differences between cohort studies of aetiology and prognosis

Objectives

A pre-requisite for investigating either causation or prediction is to identify associations between exposures and outcomes. However, the researcher who is concerned with aetiology is ultimately interested in identifying associations that are causal, whereas the researcher who is concerned with prediction does not need to differentiate between causal and non-causal associations; in the latter case, any association, causal or non-causal, can fulfill the role of predicting outcomes.

The study by Paul and colleagues, described above, sought to determine factors that caused people with Parkinson's disease to fall. The task for the investigators was to establish whether putative aetiologic factors (impaired physical capacity or impaired cognition) really did cause falls. The measure of the success of that study is the extent to which it was able to distinguish between putative aetiologic factors that truly do and do not cause falls. In contrast, the study by Moseley and colleagues was explicitly not concerned with determining causes of CRPS; it was designed, instead, to identify prognostic markers. Whether those markers are causal or not was of no relevance, because the goal was to predict, not to understand aetiology. The primary measure of the success of the Moseley study is the extent to which it was able to accurately predict who would go on to develop CRPS.

Some readers might think that the distinction between cohort studies of aetiology and prognosis is simply that studies of aetiology monitor healthy participants in order to determine who develops the disease of interest, whereas studies of prognosis monitor people who already have a health condition in order to determine who develops particular disease-related outcomes.[13] However, as these examples show, that way of distinguishing studies of aetiology and prognosis is problematic: cohort studies of initially healthy participants could either investigate factors that cause disease to develop, or factors that predict who will develop disease; and cohort studies of people who already have a disease could either seek to identify exposures that cause sequelae of the disease, or exposures that predict who will develop sequelae of the disease.

Many epidemiologists are reluctant to use the word 'cause' when they write reports of their cohort studies. Instead, they claim that they are seeking to identify 'associations'. The justification for this practice appears to be that epidemiological studies use observational designs, which provide a less rigorous foundation for

http://dx.doi.org/10.1016/j.jphys.2014.07.005

inferences about causation than experimental (ie, randomised) studies, so observational studies should not be used to support claims about causation. However, demonstration of the existence of an association between an exposure and a health outcome is of little or no intrinsic interest. Ultimately, demonstration of an association is only useful if the association can be shown to be either causal or predictive.[14] It may be difficult to establish causation, but establishing causation is, nonetheless, often the ultimate objective.

The reluctance of some epidemiologists to make claims about causation means that it is often not clear whether particular cohort studies are designed to investigate aetiology or prognosis. The lack of clarity is not helped by the use of vague terminology: reports of epidemiological studies refer to the exposures of interest as ‘risk factors’ or ‘predictors’ without ever making it clear if the interest is in *causal factors* (for the purposes of understanding disease aetiology) or *predictive factors* (for the purposes of making prognosis). Epidemiologists should not be shy about the objectives of their cohort studies; they should be explicit about whether the aim is to investigate aetiology or prognosis.

Choice of exposure variables

In a classic paper,[4] Holland described how he and Donald Rubin, pioneers in the development of research methods for establishing causation, proposed the slogan ‘No causation without manipulation’. They meant that it is not possible to talk meaningfully about the causal effects of a variable unless it is clear how that variable could be manipulated. For example, Holland contends that it is usually meaningless to talk about the causal effect of race because it is not possible to conceive of how a person’s race could be changed. The causal effect of an exposure variable must be understood as the difference in an individual’s outcomes with and without exposure to that variable,[4] yet it is hard to conceive what a person would be if he or she was other than his or her own race. In contrast, we can imagine that a person might or might not be exposed to different levels of physical activity, so we can validly ask questions about the causal effects of, say, training schedules on risk of developing lower limb overuse injuries in distance runners. Rubin has a slightly more nuanced interpretation; in his view, it is only possible to define the causal effects of a variable when it is clear what it means for a particular individual to be both exposed and not exposed to that variable.[15] He would contend that epidemiologists should not seek to determine the causal effect of exposures that do not satisfy this criterion. The same constraint does not apply to studies of prognosis. Variables that cannot be manipulated, such as race and gender, are potentially strong predictors in many contexts.

Confounding

Many associations are not causal. One of the reasons that associations between exposures and outcomes occur, even when the exposure and outcome are not causally related, is confounding. (Another reason is the closely related concept of selection.[6]) Confounding occurs when an exposure and its outcome share a common cause.[6] Consider, for example, the weak association between cognitive impairment and the subsequent risk of falling. That association may be causal: cognitive impairment could cause falls. But it is also possible that the association between cognitive impairment and risk of falling is a non-causal association caused by confounding. Confounding might arise if, for example, severe Parkinson’s disease causes both cognitive impairment and falls. In that case, disease severity would be a confounder of the relationship between cognitive impairment and falls. In the presence of confounding, the strength of the association between an exposure and an outcome does not provide a measure of the strength of the causal effect of the exposure on the outcome.

In studies of aetiology, the objective is to determine the causal effect, so it is essential to control for confounding. This can be achieved at the design stage by sampling study participants from within each stratum of the confounder, and at the analysis stage by estimating associations within strata or using statistical modelling to ‘adjust for’ confounders.[16] These control strategies are usually imperfect, because the confounder may be measured imperfectly (random measurement error produces what is known as ‘regression dilution bias’[17] in statistical adjustments) and because statistical models used to adjust for confounding may be specified incorrectly (eg, non-linear relationships may be treated as linear relationships). To the extent that there are important confounders, failure to properly control for confounding will produce biased estimates of causal effect in aetiologic cohort studies. A researcher who seeks to study aetiology must therefore identify all potentially important confounders, measure them without error, and properly control or adjust for the confounders. To the extent that the researcher is unable to do that, he or she will obtain potentially biased estimates of causal effects.

In prognostic cohort studies, there is no need to control for confounders. To the extent that confounders are strongly associated with outcomes, it may be useful to obtain data on exposure to confounders and incorporate confounders as predictor variables in a prognostic model. However, in prognostic studies it is not obligatory to disentangle causes and confounders; this is because, if the objective is simply to make an accurate prognosis, it does not matter if the prognosis is based on causal variables or confounders. Even if the association between exposure and outcome is due purely to confounding, exposure to confounders can act as a marker of outcome. At least in theory, when the intention is to study prognosis, known confounders can be intentionally omitted from a statistical model without compromising the validity of the predictions. In practice, it may be wise to include known strong confounders in prognostic models, because doing so is likely to increase the generalisability of the model to other populations.

Analysis

The broad approach to analysing aetiologic studies and prognostic studies is necessarily quite different. One of the most important differences is that the analysis of aetiology should be driven by theory, whereas the analysis of prognosis can, to a greater extent, be driven by the data. Here, ‘theory’ is used to mean existing knowledge about causes of, and associations with, the outcome.

Theory can drive the analysis of aetiologic studies in several ways. For example, theory will suggest plausible confounders, which need to be controlled for in the analysis. It might also suggest whether the relationships between particular exposures and outcomes are likely to be non-linear. In addition, theory can indicate where there is the possibility of strong interactions between exposures that need to be modelled. Finally, theory might indicate whether the exposures could lie on the same causal pathway. (Two exposures lie on the same causal pathway when one exposure causes a second exposure, which itself causes the outcome of interest. In that case, the second exposure is called a mediator.) If the aim is to identify causal effects, great care must be taken in the construction of statistical models that incorporate mediators. For this reason, and because it is essential to identify and control for all potentially important confounders, aetiologic studies often use complex statistical models. These models may include many variables, continuous variables may have non-linear relationships with outcomes, there may be interactions between variables, and the model may allow for mediators lying on the same causal pathway.

In prognostic studies, any statistical model that accurately predicts outcomes can do the job. It matters relatively little if the statistical model is specified incorrectly. That is, it is of little consequence if particular causal variables are omitted from the statistical model, or if a non-linear relationship is modelled as a linear relationship, or if an interaction or a mediator effect is ignored, and so on. (The caveat is that an incorrectly specified statistical model is not likely to predict outcomes as strongly as a correctly specified model, and may not perform well when applied to a population that differs from the study sample.)

There are compelling reasons to ignore complexity when developing prognostic models. Complex prognostic models are difficult to use in clinical practice. They can only be used to predict outcomes of individuals if many measurements have been obtained from those individuals. Also, quite a bit of calculation is required to generate predictions from complex models – usually enough to necessitate the use of a computer, but busy clinicians don't always have the time or inclination to use a computer to generate prognoses. The most useful predictive models require data on only a few easily measured variables and are simple enough to be applied to individual patients without a computer. The study by Moseley and colleagues[2] is useful because it identified a very simple prediction rule: people who experience high levels of pain (scores of 5 or more on a 10-point scale) in the week after wrist fracture are at high risk of developing CRPS.

As it is not necessary to properly model the effects of confounders in prognostic studies, prognostic models can be constructed largely without regard to theory. Consequently, the process of developing a predictive model tends to be more data driven than theory driven – the researcher can rely heavily on statistical algorithms to find the best weighting of an optimally selected set of predictors that strongly predicts outcomes. The reliance on data-driven statistical procedures has some important consequences. Most importantly, data-driven processes for building statistical models risk 'over-fitting'. That is, data-driven model building may generate models that generate very accurate predictions in the study sample, but perform relatively poorly when applied to other people from the same population.[18] The simplistic application of stepwise variable selection procedures is known to be particularly problematic.[12] The problem is greatest when many variables are considered as possible predictors and (if the outcome is binary) when relatively few people experience the outcome of interest (ie, when there are few 'cases'), therefore, interpretation is most straightforward when the number of candidate predictors is small and the number of cases is large. A widely used guideline is that there should be no more than one candidate predictor for every 10 cases[19] (also see Vittinghoff and McCulloch [20]). A range of statistical approaches, including bootstrap variable regression and penalised regression, have been developed in an attempt to deal with the problem of over-fitting.[18,21–24]

That is not to say that over-fitting cannot also be a problem in studies of aetiology. An undisciplined approach to the analysis of aetiologic studies can also produce over-fitting. Ultimately, any statistical model that uses data-driven algorithms – even careful, theory-driven analyses of causation (using, for example, the approach described by Hosmer and Lemeshow[25]) – must be replicated in other studies before the findings can be considered credible.

Statistics of interest

There is one more important and under-recognised difference in the way that cohort studies of prognosis and aetiology should be analysed. In studies of aetiology, the aim is to quantify the causal effect – the extent to which an aetiologic factor modifies health outcomes. So studies of aetiology must contrast the outcomes of people who are and are not exposed to the aetiologic factor. When the outcome is measured on a continuous scale, the most common way to quantify the causal effect is in terms of the mean difference in outcomes of exposed and unexposed people. When the outcome is binary, the causal effect is commonly quantified with a ratio statistic, like the relative risk or odds ratio. The uncertainty in (or imprecision of) estimates of the average causal effects are usually quantified with confidence intervals.

In prognostic studies, there is no interest in causal effects, so there is no need to contrast outcomes of exposed and unexposed people. Instead, the primary purpose is to estimate the *expectation* of the outcome. For continuous outcomes, the expectation is the mean outcome, and for binary outcomes, the expectation is the proportion of people experiencing that outcome. Prognostic studies may also generate predictions in specific subpopulations, either by stratifying risk estimates or by using statistical models to generate estimates for any combination of predictive factors.[16] Even then, the ultimate aim should still be to estimate the expected outcome within each stratum, not to contrast the risk in different strata.

It is necessary to quantify the performance of prognostic models. And in prognostic models, performance is determined not by the precision of estimates of effect (as in studies of causation) but by the ability to correctly discriminate individual people's outcomes.[18] When outcomes are measured on a binary scale, discrimination is frequently summarised with the 'area under the curve' statistic. When outcomes are measured on continuous scales, discrimination can be quantified with a correlation coefficient (such as r) or a coefficient of determination (r^2). The amount of uncertainty in an individual's prognosis can be quantified with prediction intervals.

One statistic that is used infrequently in studies of prognosis with binary outcomes, but which would appear to be particularly useful, is the predictive likelihood ratio (see the Moseley study[2] for one example of a study that reported predictive likelihood ratios). Likelihood ratios can be used to predict the outcomes for an individual based both on prior knowledge about the individual's prognosis and the particular values of the predictor variables for that individual. They are widely used in the interpretation of diagnostic tests[26] and arguably should be used much more in studies of prognosis. Likelihood ratios can be estimated directly with conventional logistic regression models.[27]

Interpretation

The primary motivation for studying the aetiology of a health condition is to suggest mechanisms through which health interventions might act and, in that way, to inform development of new interventions. To the extent that an exposure causes a health condition, interventions that reduce that exposure can prevent the health condition from developing. And to the extent that a health intervention has its effects solely by reducing an exposure, the maximal possible effect of the health intervention is equal to the causal effect of the exposure.

Prognostic studies do not provide estimates of causal effects, so they should not be used to guide decisions about selection of health interventions. Instead, the value of prognostic studies is that they can generate prognostic information that can be used to advise patients about the probability of developing particular health outcomes. In addition, data from prognostic studies can be used to identify people with a poor prognosis. Identification of people with a poor prognosis is potentially useful because many health interventions, particularly preventive interventions, are of most benefit for people with the worst prognoses.[28]

Summary

This Research Note has described differences between cohort studies designed to investigate aetiology and prognosis. While it will often be the case that questions about both aetiology and prognosis are best answered with cohort studies, the design and analysis of such studies must be quite different. The differences are summarised in Table 1.

Table 1
Differences in the design and analysis of cohort studies of aetiology and prognosis.

Aetiology	Prognosis
Objective is to determine the causes of particular health outcomes. Primary interest is in effects.	Objective is to predict health outcomes. Primary interest is in outcomes.
Only exposures that can be manipulated are of interest because only they can have definable causal effects.	Any exposure can be a predictor.
Analysis should be theory driven.	Analysis can be data driven, so care must be taken with 'over-fitting'.
Models must include all non-ignorable determinants of outcome, including all confounders.	Confounding is irrelevant. Elaborate models are often unhelpful. Simple models with few predictors are preferred because they are more useful in clinical practice.
Exposures (putative causes) must be measured with little or no error.	Exposures (predictors) should be easily measured.
Model must be correctly specified or estimates of causal effects may be biased. May have to model non-linear effects, interactions and mediator variables.	Predictions can still be accurate, even if the model is incorrectly specified.
Analysis involves contrasting outcomes with and without exposure.	Analysis involves estimating expected outcomes.
Can be used to develop new health interventions.	Can be used to inform people of their prognoses and to identify those at high risk for whom intervention is most likely to be effective.

Robert D Herbert[a,b]
[a]*Neuroscience Research Australia*
[b]*School of Medical Sciences, University of New South Wales, NSW, Australia*

References

1. Paul SS, et al. *Neurorehabil Neural Repair.* 2014;28:282–290.
2. Moseley GL, et al. *J Pain.* 2014;15:16–23.
3. Grobbee DE, Hoes AW. *Clinical Epidemiology: Principles, Methods, and Applications for Clinical Research.* Sudbury, MA: Jones and Bartlett; 2009:413.
4. Holland PW. *J Am Statist Assoc.* 1986;81:945–960.
5. Jewell NP. *Statistics for Epidemiology.* New York: Chapman & Hall/CRC Press; 2004.
6. Hernan MA, Robins JM. *Causal Inference. I: Causal Inference Without Models* 2013, http://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/.
7. Hernan MA, Robins JM. *Causal Inference. II: Causal Inference With Models* 2013, http://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/.
8. Altman DG, et al. *BMJ.* 2009;338:b605.
9. Moons KG, et al. *BMJ.* 2009;338:b606.
10. Moons KG, et al. *BMJ.* 2009;338:1317–1320.
11. Royston P, et al. *BMJ.* 2009;338:1373–1377.
12. Steyerberg EW. *Clinical Prediction Models: a Practical Approach to Development, Validation, and Updating.* New York: Springer; 2009:497.
13. de Bie RA. *Physiother Theory Pract.* 2001;17:161–171.
14. Kaufman JS, et al. *Soc Sci Med.* 2003;57:2397–2409.
15. Rubin DB. *J Am Statist Assoc.* 1986;81:961–962.
16. Rothman KJ, et al. 3rd ed *Modern Epidemiology.* Philadelphia, PA: Lippincott-Raven; 2008:738.
17. MacMahon S, et al. *Lancet.* 1990;335:765–774.
18. Harrell Jr FE, et al. *Stat Med.* 1996;15:361–387.
19. Peduzzi P, et al. *J Clin Epidemiol.* 1996;49:1373–1379.
20. Vittinghoff E, McCulloch CE. *Am J Epidemiol.* 2007;165:710–718.
21. Harrell FE. *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis.* New York: Springer; 2001:568.
22. Steyerberg EW, et al. *Stat Med.* 2000;19:1059–1079.
23. Steyerberg EW, et al. *Med Decis Making.* 2001;21:45–56.
24. Austin PC. *Stat Med.* 2008;27:3286–3300.
25. Hosmer DW, Lemeshow S. *Applied Logistic Regression.* 2nd edn New York: Wiley; 2000.
26. Grimes DA, Schulz KF. *Lancet.* 2005;365:1500–1505.
27. Knottnerus JA. *Med Decis Making.* 1992;12:93–108.
28. Glasziou PP, Irwig LM. *BMJ.* 1995;311:1356–1359.