Healthy Epidemiology

Christian Torp-Pedersen

2025-01-18

Rules are made to be disobeyed, so the following advice should not necessarily be followed. Nevertheless, they are basic reflexes to editors and reviewers with insight, and therefore it is wise to have a good explanation when the advice is ignored.

Healthy epidemiology may result in healthy or unhealthy programming. On www.heart.dk You can find "Rules of engagement", "Good programming practice" and "R_on_DST" with useful advice.

Analysis plan and intermediate results

It is frustrating to be told to start all over with a different method and it is frustrating for advisors to bring such input. The solution is to start by producing an analysis plan. This does not need to include more than a single page for most problems, but discussion of this in the potential author group can avoid many problems. The critical sections of this analysis plan are the selection process for a population, definitions of exposure, outcomes and covariates, as well as the proposed analyses, including the time frame.

On a similar note, the sharing of intermediate results with the author group can equally protect a study from deviating in a direction that does not make sense.

A relevant population

The population for a study should be selected such that the reader can recognize a relevant population. The population should also attempt to represent those who are at risk of having the events of interest. Two extreme scenarios can help understand the potential issues. 1) In a population study where blood pressure versus cardiovascular disease is the target, the population included to estimate the predictive value of blood pressure should most likely include people both with and without established cardiovascular disease simply because those with established cardiovascular disease contribute with the majority of events. 2) In contrast, a study of cardiovascular consequences of hormonal contraception should exclude all women with almost any prior disease simply because young women generally are healthy such that the few with cardiovascular disease will only distort the result.

For our register populations we are very often selecting a specific exposure in order to examine that exposure in relation to an outcome. We therefore want to select such a population that we can interpret the relation between exposure and outcome in a meaningful way.

Try always to imagine a physician sitting under a palm tree on a tropical island who reads your paper. This person should be able to recognize his own patients if he is to take your paper seriously. My editorial experience is that uncertainty or obvious bias in the population selection is a very common reason for plain rejection. # Less is more with large data When working with the Danish national registers it is easy to be misled by the vast numbers available, but the real strength of these large data is to be able to select a highly recognizable population. Even formally conducted population surveys are subject to participation bias and the large registers can therefore achieve wide trust if a subpopulation that appears credible is selected. Potassium among hypertensives treated with a diuretic is trustworthy, but any potassium for any reason is not.

Select variables with care

A directed acyclic graph (DAG) should always be made. It makes it easier to ensure correct selection of variables. The DAG is particularly useful to avoid adjustment (condition on) of a collider which may distort the relations to be studied.

The DAG also makes it easier to explicitly show the assumed temporal order and relation between the variables included in the analysis. Any potential confounder that occurs between the time of the exposure and the outcome is a mediator. Mediators can be examined with mediation analysis which is complex and leads to results that are very difficult to understand. Therefore a solution should be sought where all confounders are defined prior to the exposure definition, often referred to as baseline/time zero. As an example bystander resuscitation starts AFTER a cardiac arrest and is a mediator. But if the arrival of an ambulance is selected as time zero, then the bystander resuscitation can be included as a confounder.

Understand adjustment

Adjustment for multiple variables is often conceived as a magical statistical powder that makes the consequences of these variables disappear when applied. The reality is easily understood when the actual formulas are examined.

In general we perform regressions on a logarithmic scale

$$log(X) = \beta_0 + \beta_1 * age + \beta_2 * sex + \beta_3 * exposure.......$$

Where X is the odds ratio, hazard ratio or rate ratio depending on whether you are using logistic regression, Cox regression or Poisson regression.

You will in each case be interested in exp(3) to represent odds ratio, hazard ratio or rate ratio.

The consequence is that we assume age and sex to shift log(X) up or down, but the term 3 is independent of age and sex. Therefore the odds ratio, hazard ratio or rate ratio we calculate is assumed to be equal for people with identical age and sex. Two people can be assumed to have the same odds ratio, hazard ratio or rate ratio providing they otherwise have the same covariate pattern. An obvious consequence is that we would expect to find the same regression parameter if we conditioned on a specific value of one of the covariates. It is important to realize that this is a strong assumption we cannot easily expect to hold. We can compensate for this assumption by examining interaction terms between the exposure and relevant confounders, but we can rarely truly adjust for all possible interactions. It is possible to bypass these problems by using prediction models that are not dependent on the linear combination of covariates and regression coefficients such as random forest models, but in most cases we try to stick with standard models.

The solution is to select a relevant population where the covariates hopefully do not distort our results and carefully examine multiple subgroups and/or interactions, preferably defined in the analysis plan to not let the data guide us.

Choose the right design

Most studies can be categorized as a cohort study or a case-control study. The definitions are simple: A cohort study has time starting at exposure and a case-control study starts at outcome.

A cohort study can produce predictions whereas a case-control study can only produce odds ratios or hazard ratios.

Time-dependent variables in a cohort study can be tricky to handle, but in a nested case-control study the same variables are easily captured.

If the objective is risk estimation, then a cohort study is absolutely necessary. If the objective is just finding a direction of importance of a variable, then the nested case-control study is very easy to conduct.

If risk is the goal then calculate risk

Most attempts to calculate risk use odds ratio or hazard ratio that are only with difficulty interpreted as risk. It is well known that the odds ratio is close to relative risk when the odds ratio is close to one / when the outcome is rare, but the hazard ratio is more difficult. It is in reality the average relative rate over the time period studied and it completely ignores a competing risk - most often of death from other causes. First, it is easy to calculate average risk at selected time horizons from simple fractions without censoring or cumulative incidence with the Aalen-Johansen estimator.

If there is none or little censoring multivariable prediction can easily be determined with a predict function from a logistic regression - also in the presence of competing risk. If there is censoring it is necessary to use two Cox models.

The advantage of this approach is that the average risk at a defined time horizon is shown using G-estimation, a causal inference approach of a hypothetical randomised experiment where the whole population first has one value of exposure and then the other. The approach is very useful when in theory a randomised experiment could have been done, but often useful even when this is not possible.

Do not condition on the future

It is a common mistake to define a population not only based on variables available at baseline, but also variables that are defined later. An example could be the exclusion of people who immigrate without outcome during the time of a study. This is an example of conditioning on the future and the most common problem is "immortal lifetime bias". You are excluding observation time for people who do not have an outcome prior to immigration and thereby introducing a biased estimate.

Approach the result gradually

It is tempting with a large trove of data to shove everything in a multivariable model and thereby go directly from millions of data points to a single hazard ratio or similar. Such a result has a high chance of being misleading. A strategy where the path from raw data to a final result should be gradual and ensure that the final models chosen are relevant. Descriptive statistics and intermediate results are necessary. # Descriptive statistics Any analysis should start with very simple descriptive statistics of variables of interest. Visual examination of distribution plots and scatter plots is the first step. Barplots of for example calendar time for time zero by the exposure groups can also reveal potential problems.

If time to outcome is of interest, the simple cumulative incidence is important. If most events occur very early a different analysis strategy is necessary compared to a situation where most events occur gradually over time or predominantly late.

An examination of outcomes in subgroups is important since it shows whether the outcome of interest is mostly isolated to a subpopulation. This examination can change the course of a study.