Workshop 5 — Sources of bias in cohort studies

Cohort studies examining recovery, survival, recurrence, etc in a series of patients are called prognostic studies.[1] Cohort studies can also be carried out on subjects without the disease(s) of interest. Subjects in such cohort studies are often recruited on the basis of having and/or not having certain exposure(s) or risk factor(s) and then followed up for a period of time to observe the development of new health outcomes. Examples include many occupational cohorts and subjects with exposures arising from certain environmental incidents, eg atomic bomb survivors. In recent years, more and more cohorts are formed without specifying exposure(s) of interest, but consist of representative samples of defined populations. Cohort studies basically compare the risk of having certain health outcomes among subjects with or without (or with different levels of) specific exposures and hence are subject to all three common sources of bias.

Historical cohort studies

This is usually based on a group of subjects with common exposure(s) or experience(s) for which historical data/records are available, eg employees of a certain company, members of a certain professional organisation, babies born in a certain hospital or having received a certain immunisation, etc. The outcome(s) should have occurred by the time the study is being conducted taking into consideration the likely lag time or latency period; both exposure(s) and outcome(s) are historical events.

Selection bias

As long as the selected exposed group is being defined clearly (inclusion and exclusion criteria) and every eligible subject included, selection bias by the investigator should not be a major concern for internal validity, though the results may not be generalised to other groups/populations. The risk of having a certain health outcome is compared to an unexposed group identified either within the same cohort (internal comparison) or from an outside source, eg the general population (external comparison). The baseline risk, apart from that arising from the exposure, of the selected comparison group may not be comparable to the exposed cohort, and this may introduce bias arising from confounding (see below). Response rate is not a concern, as records of routinely collected data are used, but self-selection bias can be a major concern if the proportion lost to follow-up is substantial (say >20%).

Information bias

Information on exposure(s) is usually not a problem, as this is the basis for identifying subjects into the cohort and such information has been recorded in the history. Of course, misclassification is still possible even if the person or persons involved in this process are blinded to the health outcome(s) due to the usually crude exposure information available in routine records. However, this would be non-differential and the possible resulting bias should be towards the null. Information on outcome(s) is also not a major issue, provided that the methods used to retrieve and ascertain the outcome are uniform in both the exposed cohort, and the comparison group and the person(s) involved are blinded to the exposure status. Some cohort studies try to obtain information on the exposure and potential confounding factors retrospectively through interviews, and such information could be subject to serious recall biases as in the case-referent studies,[2] and as such have very little advantage over the latter.

Confounding

In historical records, the availability of information on potential confounding factors is usually limited, eg absence or incomplete smoking history in an occupational cohort with a certain chemical exposure that might be related to lung cancer. It is unlikely that a historical cohort study can adequately address potential confounding. Furthermore, the choice of an external comparison group that is not comparable to the cohort in terms of baseline risk can also introduce confounding. Thus, depending on the distributions of specific known risk factors (eg age, smoking, socioeconomic status) in the two groups, statistical adjustment may not adequately handle the differences between them.

Prospective cohort studies

The basic approach is to recruit subjects with a common exposure or experience into a cohort and follow them up prospectively to document one or more adverse health outcomes. A comparable cohort with similar baseline risk apart from the exposure can be built in for comparing the risks, or the risk of the exposed cohort can be compared to that of the general population (external comparison). Furthermore, large population-based cohorts examining multiple exposures and multiple outcomes are getting more and more popular in recent years, with both exposed and unexposed groups included in the same cohort.

TABLE. Main sources of bias in cohort studies

Study design	Source of bias*					
	Selection bias		Information bias			Confounding
	Investigator	Self (study subjects)	Exposure(s)	Outcome(s)	Confounding factor(s)	
Historical cohort study						
Internal comparison	+	++	+	+	+	++
External comparison	+	+/-	+/-	+/-	+/-	++
Prospective cohort study						
Internal comparison	+	++	+/-	++	+/-	+
External comparison	+	+/-	+/-	+/-	+/-	++
Population cohort	++	++	+/-	++	+/-	++

* ++ denotes major source, + minor source, and +/- unlikely

As the study is undertaken prospectively, it should be well-planned and the data to be collected, as well as the methods used to do so, should be of better quality than in a historical cohort study. However, bias from different sources can still be present.

Selection bias

Selection bias on the part of investigator(s) is not a major concern for exposure-based cohorts. For population-based cohorts, care should be taken to recruit a representative sample of the target population. As consent is usually required for participating in prospective studies requiring follow-up, non-participation can be a concern. The major challenge is to maintain successful follow-up for most subjects, otherwise, serious self-selection bias can occur.

Information bias

There is little room for information bias on exposure(s), as the relevant data are collected at baseline (without knowledge on outcomes) using standardised methods. As with prognostic studies on patient cohorts,[1] bias arising from ascertainment of outcomes can be a major concern. This is especially problematic when the presence of a certain risk factor may affect how subjects are being followed up and investigated. For example, smokers may have more frequent follow-ups or undergo more investigations when respiratory symptoms occur and this can lead to a higher chance of detecting lung cancer than in non-smokers.

Confounding

One major advantage of prospective cohort studies over historical ones is that in theory, information on all potential confounding factors (known risk factors) can be collected and used for subsequent adjustment. This may not be true for population-based cohorts, as some outcomes of interest may be identified after baseline data collection, in which case not all relevant risk factors for these unplanned outcomes might have been included. If external comparison is used for an exposed cohort, confounding would be a major concern, due to the frequent lack of information on potential confounding factors in the comparison group.

The Table summarises the main sources of bias in cohort studies.

Ignatius TS Yu
Shelly LA Tse
Clinical Epidemiology Group
Hong Kong Medical Journal

References

1. Yu IT, Tse SL. Clinical Epidemiology Workshop 3—Sources of bias in case series, patient cohorts, and randomised controlled trials. Hong Kong Med J 2011;17:478-9.
2. Yu IT, Tse SL. Clinical Epidemiology Workshop 4—Sources of bias in case-referent studies. Hong Kong Med J 2012;18:46-7.