Bias caused by missing values in primary predictor variables. (Does smoking kill the drinking classes or *vice versa*?)

Roger Newson (King's College, London, UK)

`roger.newson@kcl.ac.uk`

28 January 2002

- What drew my attention to this subject.
- A fantasy cohort study on a fantasy population.
- Levels of bias caused by missing primary predictor values.
- Implications for real-world studies.

What drew my attention to this subject

- The ALSPAC study involves 14060 children, of whom a minority of 2973 (selected in a way unknown) have cord trace element assays. (The money ran out before all the cords were assayed.)

- We want to assess the association of cord trace elements (eg selenium) with atopic disease outcomes, controlling for a long list of established confounders.

- Initially, we used the whole cohort in our analyses, having a huge category with unknown values for trace elements.

- This was done in order to use data on the majority, without cord element assays, to estimate confounder effects.

- It was objected that this might in some way cause bias. (With a bit of thought, I decided that this objection was probably valid.)

A fantasy study on a fantasy population

- Imagine a hypothetical population where 50% are cigarette smokers and 50% are beer drinkers, and the two habits may be positively associated.

- Imagine also that, in that population, cigarette smoking doubles the death rate, whereas beer drinking has no effect on the death rate.

- Imagine, in addition, that the epidemiologists in that country view beer drinking as an established risk factor, but have not yet investigated smoking.

- A group of epidemiologists design a cohort study to investigate smoking and drinking as predictors of dying, to investigate a recently-proposed hypothesis that smoking is killing the drinking classes.

- The data are analysed using Poisson regression to establish death rates per person-year at risk. Smoking is included as the predictor of primary interest, and drinking is included as a necessary counfounder. However, a lot of subjects have unknown smoking status, possibly because the money ran out before all the cotinine assays were done.

Simulations carried out

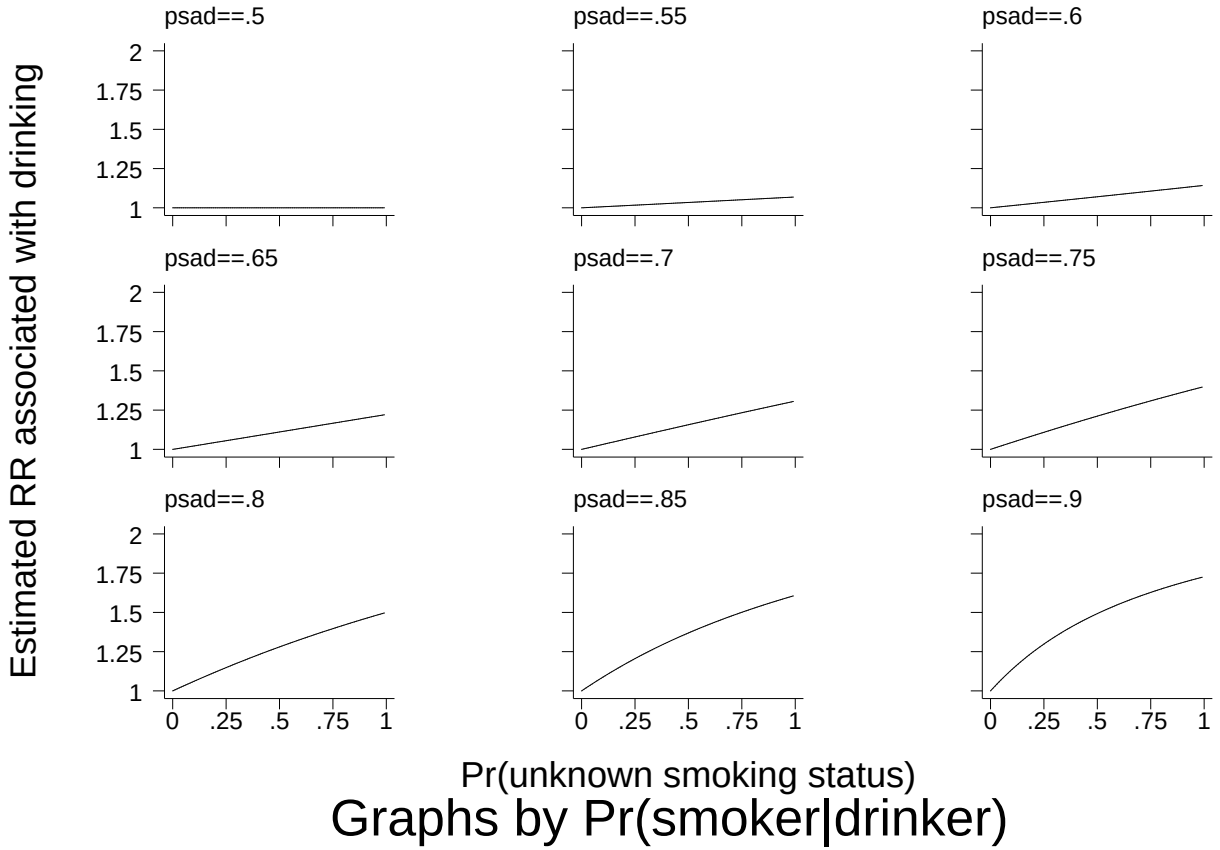
In all simulations, I assumed that, in fact, 50% of the cohort smoke, and 50% drink, and that the two habits are non-negatively associated. However, for some reason, drinking habits are recorded for the whole cohort, whereas a *random* subset of the cohort have unknown smoking habits. Two parameters were varied independently between simulations:

- The association between smoking and drinking. The percent of drinkers who also smoked (and therefore the percent of non-drinkers who were also non-smokers) was varied from 50% to 90%, in increments of 5%.

- The percent of the cohort who had missing smoking status. This was varied from 0% to 99% in increments of 1%.

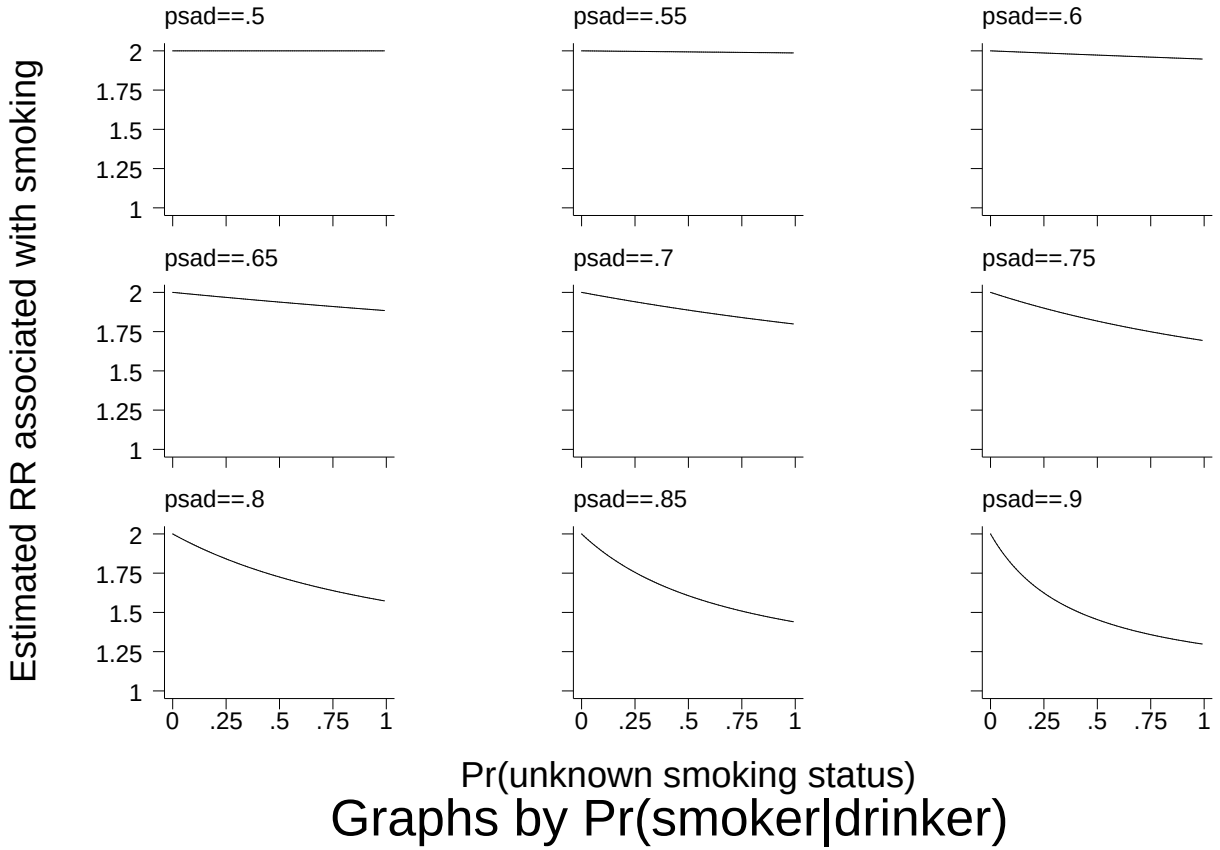
There were therefore $9 \times 100 = 900$ simulations. In each simulation, the mutually-adjusted smoking-related and drinking-related relative risks were measured, assuming no interaction. Smoking was a 3-level factor (yes, no or missing), whereas drinking was a two-level factor (yes or no).

Adjusted relative risks from beer drinking estimated from the cohort study



If smoking and drinking are positively associated, then random missing smoking values cause a spurious positive association between drinking and dying, even after adjustment.

Adjusted relative risks from cigarette smoking estimated from the cohort study



If smoking and drinking are positively associated, then random missing smoking values cause smoking to appear less hazardous than it is, even after adjustment.

What does this imply for real-world studies?

In our fantasy study:

- Smoking played a role similar to selenium in the ALSPAC study, namely the novel risk factor being investigated.

- Drinking played a role similar to the established confounders in the ALSPAC study, namely the established risk factor which will be invoked by a skeptical audience if not controlled for.

- If a lot of individuals have missing smoking (or selenium) values, then this may cause some of the effect of smoking (or low selenium) to be spuriously attributed to drinking (or established confounders such as mother's education and prematurity).

- Therefore, if the novel predictor is the true cause of effects traditionally attributed to the confounders, then a large category with the novel predictor unknown will cause the novel predictor effect to be underestimated.

- Note that this is true even if the missing category is chosen at random, in a way unrelated to what the values would have been, had they not been missing.