Commentary: Estimation of the incidence of stroke using a capture-recapture model including covariates

David Barer

Despite the often dramatic onset and devastating consequences of stroke disease, incidence rates are notoriously difficult to estimate, and for a long time epidemiologists had to rely mainly on mortality data. The first serious attempt to compare stroke incidence in different populations, using standardized clinical definitions, was carried out in the 1970s by the World Health Organization.1 These studies indicated a seven-fold variation in incidence and a two-fold variation in case fatality, but left considerable uncertainty about how much of this variation was due to genetic or lifestyle differences and how much due to problems in case ascertainment. Among those communities where ascertainment appeared to be most complete, the age-adjusted incidence rates were all fairly similar, between 1.5 and 2.0 per 1000 per year.

In fact up to the late 1980s, only a handful of studies could be said to provide reliable estimates of stroke incidence rates in a defined population,2 particularly in higher age groups. Although more rigorous methods have generally been used in more recent studies, cases may still be missed, especially amongst those who die quickly, those who are not admitted to hospital and frail elderly people in whom stroke may be just one of several conditions contributing to disability.

The best studies have all used multiple sources of notification (primary care teams, nursing home and ambulance records, community nursing, therapy and social services, scanning departments, death certificates, etc.), but even then it is impossible to be sure that ascertainment is complete. ‘Capture-recapture’ techniques have long been advocated to estimate the number of cases that might have been missed by comparing the numbers picked up by overlapping notification sources. The simplest models are based on the unlikely assumption that these sources are independent, so more sophisticated approaches have been developed which take account of some of the factors influencing the relative ‘catchability’ of individual cases by different registers.

In this issue Tilling et al. apply such a method to data from the South London Stroke Register, an ongoing study in a multi-ethnic area of inner London which uses 14 different sources of notification.3

Demographic factors and various markers of stroke severity, which might have affected the chances of appearing on the different registers, were fed into a multinomial logistic model, able to take account of continuous covariates as well as interactions between categorical factors. The bottom line estimate was that overall case ascertainment from all sources combined was only 88% complete. Adjustment for missing cases increased the age-standardized incidence rate for stroke from 1.31 to 1.49: outside the 95% confidence interval for the unadjusted estimate (1.21–1.42).

Although the basic idea is simple, the practicalities of multivariate modelling soon introduce complexity. For instance the details of how variables are chosen for inclusion in or elimination from the model can make a considerable difference to the results. Altogether 200 fatal strokes were registered. A 3-source model based on cases picked up from hospital records, death certificates and the research team gave an adjusted estimate of 283 fatal cases. A model based only on information from the last two sources, but including all covariates, estimated 198 cases, whereas a 2-source model using selected variables gave a final estimate of 222 fatal cases.

The authors justify the choices made but an arbitrary element remains. Unstable estimates are produced when there is little overlap between sources, which was unfortunately the case with notifications from hospital-based and primary care teams. Hence the sources of notification chosen as the basis for the final models did not include General Practitioners (though GP notification was one of the cofactors adjusted for). Thus there is still a worry that a disproportionate number of cases might have been missed in the very old and possibly amongst some ethnic or cultural groups less likely to seek admission to hospital.

Stroke severity is likely to influence both the chances of being admitted and of being notified to the research team, so incomplete case ascertainment may introduce substantial bias in estimates of fatality rates (especially within certain subgroups). One wonders how much of the 50% variation in case fatality seen in the recent EROS study,\(^4\) which compared data from community based stroke registers in Britain, France and Germany (including the South London Stroke Register), might be accounted for by this effect.

Similar biases might affect comparative estimates of the incidence and case fatality of different subtypes of stroke.

The methods used by Tilling et al. may not eliminate ascertainment bias but they can at least estimate and probably reduce it, as well as improving overall estimates of stroke incidence. Investigators on other community stroke studies should therefore be urged to re-examine their data using a similar approach, wherever possible.

References