Capture-recapture methods—useful or misleading?

Kate Tilling

Disease registers are used for two main purposes: to measure the incidence or prevalence of a disease, or to study its natural history. For example, the WHO MONICA collaboration was established in the early 1980s to register myocardial infarction and stroke in different populations worldwide, and thus allow comparisons of incidence to be made.1 Similarly, cancer registries are routinely used to provide data for comparisons of incidence of different cancers between areas of the UK. These purposes clearly require a different breadth of data than a register intended to study the natural history of a disease. For instance, a stroke register in South London, UK, was established not only to measure the incidence of stroke in this area, but also to follow stroke patients over time in order to examine factors affecting outcome and risk factor management.2 Quality criteria for stroke incidence registers have been defined, emphasizing the importance of complete, community-based case ascertainment.3

In either case, a register needs to provide an accurate, unbiased estimate of the number of cases missed by the register. In practice however, it is often assumed that if the register has been conducted ‘carefully’ then it will be 100% complete, or at least have missed so few cases that the implications of the study will be unaffected by incompleteness. For example, the MONICA study set out criteria (more than 10% of fatal cases not hospitalized, more than 5% of non-fatal cases not hospitalized, 28-day case fatality less than 40% and a ratio of fatal cases to stroke deaths from routine mortality statistics greater than 1) for assuming equal completeness of case ascertainment across different centres.4 However, an estimate of the completeness of the register was not made, and there was no evidence that the criteria were necessary or sufficient for completeness. Using the number of cases from a register to estimate incidence assumes that no cases are missing.5 In addition, if the cases missed differ from those observed (e.g. less severe cases are less likely to be registered) then using only the observed cases will lead to biased inferences.

Capture-recapture methods have been advocated for use in estimating completeness of a register.6 These methods were originally developed to estimate the size of a closed animal population. The procedure is that at one time as many animals as possible in an area are captured, tagged and released—the ‘capture’ stage. At a later time this is repeated—the ‘recapture’ stage. The number of animals in each sample, and the number common to both, are used to estimate the number in the total population (assuming that capture and recapture are independent) (Table 1).

Table 1 The simple capture-recapture problem

<table>
<thead>
<tr>
<th>‘Capture’</th>
<th>‘Recapture’</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>a</td>
<td>a+b</td>
</tr>
<tr>
<td>No</td>
<td>x</td>
<td>c</td>
</tr>
<tr>
<td>Total</td>
<td>a+c</td>
<td>N=a+b+c+x</td>
</tr>
</tbody>
</table>

If the capture and recapture are independent, then the estimated probability of being captured on both occasions is equal to the product of the probabilities of being captured on each occasion. The number of animals missed (x) can then be estimated, using this independence assumption, as follows:

\[ \frac{a}{N} = \left( \frac{a+c}{N} \right) \times \left( \frac{a+b}{N} \right) \]

\[ a(a+b+c+x) = (a+c)(a+b) \quad \text{(since } N = a+b+c+x \text{)} \]

\[ \frac{x}{a} = \frac{bc}{a^2} \]

This simple technique was first applied to the epidemiological problem of estimating the size of a human population in the 1940s,7 and became more widespread following the work of Wittes in the early 1970s.8 In the epidemiological context the two sources may be two lists, for instance hospital records and death certificates relating to the same disease. The methodology has been extended to include more than two sources.6

Capture-recapture models have been used to estimate incidence of many diseases and health-related problems, including cancer,9,10 stroke,11,12 homelessness,13 mental illness13 and drug use.14 It has been recommended that a simple capture-recapture analysis, with corresponding estimate of completeness, should accompany any results obtained from a register.5 However, two important and related assumptions which are made when using the simple capture-recapture method cast doubt on its use in epidemiology. The first is that when there are two sources, they are assumed to be independent, and more generally, that there is no dependency between all k sources in a k-source model. The second is that all individuals have the same probability of being captured. Neither can be directly tested, and violation of either could lead to over- or under-estimation of the true population size.6
It is unlikely that these assumptions will hold in an epidemiological study. For example, in a cancer registry two sources often used are death certificates and hospital discharge records. Cases captured by one source are more likely to be also captured by the other, leading to dependence between the sources and violating the first assumption. More severe cases are more likely to be admitted to hospital and identified correctly as cancer, which would thus appear on the discharge record. They are also more likely to die, and to have cause of death recorded as cancer. Thus, more severe cases will be more likely to be captured, violating the second assumption. The likelihood that subject characteristics would be associated with probability of capture has been identified by several authors.15,16 These problems have led to distrust of capture-recapture as a method for estimating the completeness of an epidemiological survey or register.

A simple means of allowing for inequality of probability of capture is to stratify the analysis by the variable(s) thought to be related to capture (e.g. case severity), and then perform separate capture-recapture analyses for each stratum.7 However, this increases the variability of the estimated population size for each stratum, especially when data in one or more strata are sparse. An alternative is to use more than two sources, and use log-linear methods to model dependence between sources, although this gives no information on the characteristics associated with capture. There are many possible models depending on the interactions included (e.g. eight possible models with three sources), but the best approach is to use the model with all possible interactions.17

An alternative, advocated recently, is the inclusion of covariates in a log-linear or logit model.18,19 In a simulation study, inclusion of capture-related covariates improved accuracy of the estimate of the population size compared to biased estimates from the simple model.19 In addition, this method can identify patient characteristics related to probability of capture by the different sources. Such information could be useful for improvement of an ongoing register by identifying patient subgroups with a high probability of being missed by the register. This approach has been used to identify subsections of the population who were unlikely to be captured by the US Census,20 and characteristics of both individual and company related to probability of reporting industrial accidents.21 The parameters from the model can also be used to estimate the number of cases in different population subgroups. For example, adjusted age- and sex-specific incidence rates could be derived.

Any register, no matter how carefully conducted, is likely to miss some cases. However, complete data are not needed in order to estimate incidence rates, standardized incidence rates or incidence rate ratios. Instead, a register could be designed not to be complete, but to use the combination of sources and covariates that gave the least biased and most accurate estimate of incidence. Rather than aiming to register every single case, care would have to be taken to ensure that the sources were independent, to avoid small numbers of cases being captured by particular combinations of sources. Similarly, covariates would be chosen to explain the maximum amount of the variation in probability of capture between cases. Correct case-definition and matching between lists should be a priority of any register, but especially those designed with capture-recapture analysis in mind.16

The sources and covariates could be chosen a priori, then verified and refined over a period of time by running two registers in parallel—one aiming to recruit all cases, the other aiming to use capture-recapture. The ‘complete’ register could be used to verify sources, covariates and the model used to estimate population size from the capture-recapture register. After the development period, the capture-recapture register alone could be continued. Any loss in precision from an incomplete register could be balanced against the decreased resource needs, particularly for long-term monitoring of trends in disease prevalence or incidence. More practical work on the development of sources and selection of covariates for capture-recapture analysis is needed.

Even where a ‘complete’ register is necessary (e.g. for organizing clinical care), an estimate of completeness should be made. Capture-recapture methods are more likely to produce a biased estimate of the population size if one source (or combination of sources) captures very few cases. In this case, the estimate of the number of cases missed could be close to zero or very large (depending on the model used). One solution might be to perform a capture-recapture analysis using only those cases captured by the two or three main sources of notification. A check on the plausibility of the estimate of number of cases missed could be made by comparing it to the known number of cases missed by these main sources.

Capture-recapture methods offer the potential to reduce the costs of disease registers, and to reduce bias in the estimation of incidence and comparison of population groups. However, the assumptions made when using simple capture-recapture methods are unlikely to be true in epidemiological studies. Modelling of covariate effects may produce better population size estimates and thus overcome some of the current distrust. Practical examples of studies where capture-recapture methods have reduced bias or improved cost-effectiveness are needed. We also need to move away from the idea that all registers need to be as complete as possible. A well-designed, incomplete, register may provide a more accurate, unbiased estimate of incidence than a nearly complete register which fails to identify particular population groups.

References


