

## The use of capture–recapture methods in public health

Eugene M. Laska<sup>1</sup>

To make policy on disease control and to put it into practice, public health officials need to know the size and composition of target populations. Incidence and prevalence estimates provide the foundation for the design and evaluation of essential health programmes. Compulsory reporting to registers is the broadest approach to obtaining these estimates, but it is expensive and the goal of complete enumeration is rarely achieved. Births and deaths, in industrialized countries at least, are well documented, but rates of HIV infection much less so. Health surveys, on the other hand — whether periodic or ad hoc, cross-sectional or longitudinal — identify cases from a random or nationally representative sample. Inferences about the whole population are then drawn by extrapolation. Both approaches have their limits in practice.

Fortunately, a third alternative — a capture–recapture census — may provide more reliable estimates. The technique was first used in 1662 to estimate the population of London, but it was not until 150 years later that Laplace laid out its mathematical foundations. In 1896, Petersen used the approach to estimate the harvestable stocks of Danish fish populations, and in 1930 Lincoln used it to estimate waterfowl abundance in US flyways. In wildlife applications, the method is often called the Lincoln–Petersen estimator; in demography, it is known as dual-system estimation. Its continued refinement over the last 75 years has been stimulated mostly, but not exclusively, by the needs of wildlife research and management (1).

In the original form of this method, a sample of individuals from a target population is captured, marked and released, and a second sample captured at some later time. The number of individuals observed in each sample, as well as the number observed in both, is noted. There are four critical assumptions implicit in the statistical analysis: the population is closed, individuals captured on both occasions can be matched, capture in the second sample is independent of capture in the first, and the probabilities of capture are homogeneous across individuals. Under

these conditions, the data make it possible to estimate the probability of being captured and the size of the population. Log-linear models provide a convenient representation of this basic census and its extension to multiple capture–recapture occasions.

By analogy, the *k*th list in a set of *K* independent lists of members of a target population may be thought of as identifying individuals “captured” on the *k*th occasion. If individuals can be matched, a count of the number that appear on only one list, on two lists and so on up to the number on all *K* lists can be compiled. These are the data needed for a capture–recapture analysis. In health applications, lists can be obtained with little added cost from many sources such as hospitals, doctors, laboratories, insurers, social service agencies, religious institutions, and schools.

The earliest modern application of this method in demography appears to have been to estimate vital rates in India in 1949 (2); its earliest use in epidemiology was to estimate the number of hospital patients using methicillin in 1966 (3). Capture–recapture methods have been used to estimate population size in a wide variety of health applications including, for example, infants born with birth defects, women with preclinical cancer, persons with severe and persistent mental illness, drug abusers, and persons with sexually transmitted infections. Such methods have also been used to evaluate the degree of ascertainment of various disease monitoring systems. Yip et al. document epidemiological applications and include an extensive bibliography (4).

The study on the annual incidence of acute flaccid paralysis reported in this issue of the *Bulletin* by Whitfield & Kelly (pp. 846–851) puts the method to good use. Their estimate is based on a statistical analysis of data from two lists: routine surveillance, and hospital records. In addition, the study documents the ascertainment shortcomings of the surveillance system.

In recent years, related new methods have been developed, stimulated in part by problems involving human rather than wildlife populations. For example, “single

sample methods” in the spirit of the capture–recapture approach have appeared. The plant-capture method was used to estimate the number of street-dwelling homeless people in southern Manhattan as part of the 1990 US decennial census. In lieu of a first capture, study participants whose behaviour and appearance are indistinguishable from the target population are distributed or “planted” throughout sites occupied by the target population (5). Using data on time since last service, a one-week sample survey was used to estimate the number of individuals served annually by a large mental health services system (6). A recently developed method relaxes the requirement that individuals can be exactly matched from capture to capture. Instead, the likelihood of a match based on demographic variables is used.

Capture–recapture techniques are often much less expensive and may be more informative than classic approaches to case-finding. Public health officials face limited budgets in both the developing and the industrialized world. Those interested in the size of difficult-to-identify populations will undoubtedly find estimation procedures based on these methods appealing. ■

1. Pollock KH. Modeling capture, recapture and removal statistics for estimation of demographic parameters for fish and wildlife populations: past, present and future. *Journal of the American Statistical Association* 1991;86:225-38.
2. Sekar C, Deming EW. On a method of estimating birth and death rates and extent of registration. *Journal of the American Statistical Association* 1949;44:101-15.
3. Wittes J, Sidel VW. A generalization of the simple capture–recapture model with applications to epidemiological research. *Journal of Chronic Diseases* 1968;21:287-301.
4. Yip PSF, Bruno G, Tajlman, Seber GAF, Buckland ST, Cormack RM, et al. Capture–recapture and multiple-record systems estimation 1. History and theoretical development, 2. Applications in Human diseases. *American Journal of Epidemiology* 1995;142:1047-68.
5. Laska EM and Meisner M. A plant-capture method for estimating the size of a population from a single sample. *Biometrics* 1993;49:209-20.
6. Laska EM, Meisner M, Siegel C. Estimating the size of a population from a single sample. *Biometrics* 1988;44:461-72. Correction: *Biometrics*, 1989; 45:1347.

<sup>1</sup> Director, Statistical Sciences & Epidemiology Division, Nathan S. Kline Institute for Psychiatric Research, 140 Old Orangeburg Road, Orangeburg, New York 10962, USA (email laska@nki.rfmh.org).