American Journal of Epidemiology
Copyright © 1992 by The Johns Hopkins University School of Hygiene and Public Health
All rights reserved

Vol. 135, No 9
Printed in U S A.

The Value of Capture-Recapture Methods Even for Apparent Exhaustive Surveys

The Need for Adjustment for Source of Ascertainment Intersection in Attempted Complete Prevalence Studies

Ernest B. Hook[1] and Ronald R. Regal[2]

Almost all reported prevalence studies of which we are aware make exhaustive attempts to find diagnosed individuals and report all affected individuals, but make no attempt to estimate or adjust for missing cases. Yet very simple methods introduced in the planning stage of a prevalence study may enable investigators, or at least those subsequently reading their reports, to derive such adjusted estimates. If investigators keep track of the nature of the ascertainment of cases by source and collect and report data that allow calculation of the number of cases by source intersection, then they, or at least others, may derive estimates of missing cases and of the total population affected, by using readily available analogues of capture-recapture methods developed for wildlife populations censuses. Unfortunately, such methods are often inappropriately disparaged or ignored by epidemiologists. The derived estimates are sensitive to assumptions about dependence or independence ("interaction") of various sources, assumptions that sometimes are unprovable, and these estimates have some uncertainty because of statistical fluctuation. Moreover, most investigators who attempt exhaustive prevalence studies apparently believe that they have ascertained all cases and that there is no need to attempt to adjust for, let alone provide data pertinent to, the number of missing cases or to use a statistical method that will at best imply a certain imprecision to their result. Yet a survey that reports prevalence data without adjustment for, or data on, source intersection in essence makes an estimate of missing cases—zero—while providing no quantitative grounds for that claim. The results of all such surveys should be regarded with skepticism because, at best (if the case reports are accurate), they provide only a lower boundary of prevalence. We illustrate the grounds for these views by analyzing data from an apparently exhaustive prevalence study that used at least 14 distinct sources for ascertainment, including advertising, to find cases. Available limited data on source intersection provided in the report enable the plausible inference that the study missed about 25–40% of cases. We urge that no attempted complete prevalence studies be presented without data on ascertainment by source intersection. *Am J Epidemiol* 1992;135:1060–7.

capture-recapture; census; incidence; prevalence

Almost all recent published surveys of the prevalence of disease of which we are aware make extensive attempts to find every individual diagnosed with the condition, without attempting to make adjustment for, or to estimate, the completeness of ascertainment of cases. While investigators' descriptions of the sources and efforts used in prevalence studies may appear to be exhaustive, they provide usually no formal quantitative method of justifying their inference that there are no missing cases. Yet methods pertinent to this issue are readily available,

Received for publication June 7, 1991, and in final form December 13, 1991.

Abbreviations: MLE, maximum likelihood estimate, NUE, nearly unbiased estimator.

[1] School of Public Health, University of California, Berkeley, CA.
[2] Department of Mathematics and Statistics, University of Minnesota-Duluth, Duluth, MN.

if their utility is recognized at the planning stage.

Some years ago Wittes and coworkers (1–4) in a series of papers observed that analogues of methods used by ecologists to estimate wildlife populations may be of value within epidemiology to estimate prevalence. Subsequently, log-linear methods were also developed for this epidemiologic application (5).

Despite the existence of these powerful methods, often termed "capture-recapture" in recognition of their ecologic antecedents, they are relatively little used within epidemiology. They have had perhaps more application in demography. Indeed, the earliest cited use is that of LaPlace (6) who invented the method to estimate the population of France in 1783, and their most prominent application has been in adjustments of the 1990 US census. Most of the few applications within epidemiology are instances in which investigators have available only a few sources known to be incomplete, such as vital record reports of congenital malformations, or small samples (e.g., refs. 1–3 and 7–10). The derived estimates are subject to some sampling error that may be considerable, although it is of interest that rates of Down's syndrome in livebirths derived from birth certificates using very simple capture-recapture methods have provided estimates that in some instances appear more accurate than studies that attempted complete enumeration. (See ref. 11 for references and discussion.)

We suspect that most epidemiologists undertaking prevalence studies regard such capture-recapture methods, if they are aware of them, as useful only when a few, incomplete sources are available. Most investigators who undertake exhaustive attempts at ascertainment, or what we term an "attempted complete prevalence study," do not appear to recognize the utility of these methods. Yet capture-recapture methods applied to data from such surveys can provide some objective evidence to validate their assumptions of complete ascertainment or indicate that the survey has, unexpectedly, fallen short of its goal.

Indeed, we contend that the reports of prevalence studies in the literature must be regarded with skepticism unless such capture-recapture methods are applied. We illustrate here the rationale for this belief by considering the results of application of such methods to the reported data of an attempted complete prevalence study that appeared initially to be exhaustive. This example illustrates, we believe, not only grounds for the need for these methods but also the richness of inference possible from the application of even the simplest capture-recapture methods.

BACKGROUND

Capture-recapture methods derive their name from censuses of wildlife populations in which a prespecified number of animals are captured, marked, released, and subject to recapture (for review, see refs. 12 and 13). Within the epidemiologic literature, the primary focus has been on estimates derived from analysis of overlapping incomplete lists of cases from different sources. In addition to the Bernoulli census (3) and log-linear methods (5), other alternatives, such as application of the truncated binomial distributions, have been proposed (14). An analysis comparing the validity of these methods has appeared (15).

The data used in analyses of incomplete lists are the tabulations of the number of cases found in each possible combination of ascertainment sources. If two sources are used, designated as A and B, there are three possible ways in which a case can be ascertained: only in A, only in B, or in both A and B (which we denote as AB). If there are three sources, A, B, and C, then there are seven nonoverlapping possible modes of ascertainment: A only, B only, C only, AB not C, AC not B, BC not A, and ABC. In general, if there are n possible sources, the number of different possible combinations of nonoverlapping sources of ascertainment is $2^n - 1$. Occasionally, there is an advantage to pooling various types of sources and treating their union as a single source as suggested

by Wittes (3). Thus, in a study of a disorder associated with mental retardation, one might pool all state institutions for the retarded and treat this in the analysis as one single source.

A major issue in the analysis of data on multiple ascertainment is the question of possible dependencies of sources. If data from several different sources are available, then with, for example, log-linear methods (5) one may estimate and adjust for pairwise and higher order dependencies, up to a maximum of $n - 1$, where n is the total number of sources. Yet even if one has data on only two ascertainment types, one may still derive useful information pertinent to population prevalence estimates. If, for instance, two sources (X and Y) are independent and if the structure of the population is as given in table 1, with a cases reported in both X and Y, b cases only in source X, and c cases only in source Y, then the maximum likelihood estimate (MLE) of cases in the population but *not* in X or in Y is

$$d_{\mathrm{MLE}} = \frac{bc}{a}. \qquad (1)$$

The corresponding estimate of the total population (p) is

$$p_{\mathrm{MLE}} = a + b + c + \frac{bc}{a}, \qquad (2)$$

or equivalently,

$$p_{\mathrm{MLE}} = \frac{(a + b)(a + c)}{a}.$$

While these are MLEs, they are biased for small samples, and a preferable, nearly unbiased estimator (NUE) for the unascertained cases is

$$d_{\mathrm{NUE}} = \frac{bc}{a + 1}, \qquad (3)$$

which implies a value for the total population of

$$p_{\mathrm{NUE}} = a + b + c + \left(\frac{bc}{(a + 1)}\right), \qquad (4)$$

or equivalently,

$$p_{\mathrm{NUE}} = \left[\frac{(a + b + 1)(a + c + 1)}{(a + 1)}\right] - 1,$$

a formula suggested initially by Chapman (16) and shown by Wittes (4) to be indeed virtually unbiased for wide ranges of parameter values.

If, however, sources X and Y are *positively* dependent (e.g., a case in Y is more likely to be ascertained in X than a case not in Y), then the values of d and p derived from the above equations will be *underestimates* of the true population value (8). If sources X and Y are *negatively* dependent, then the values derived from these equations will be *overestimates*.

As an anonymous reviewer of this paper has suggested to us, one may demonstrate heuristically the relation between the bias of the maximum likelihood estimate d_{MLE} and source dependencies by considering the relative odds ratio (r) that, if a case is reported in one source, it is reported in the other.

TABLE 1. Data structure by source of ascertainment: two-source case

		Case reported in source Y: Yes	Case reported in source Y: No	
Case reported in source X	Yes	a	b	$a + b$
	No	c	$d = ?$	
		$a + c$		$p = a + b + c + d$

Because

$$r = \frac{ad}{bc}, \tag{5}$$

$$d = (r)\left(\frac{bc}{a}\right) = (r)d_{\text{MLE}} \tag{6}$$

and

$$d_{\text{MLE}} = \frac{d}{r}. \tag{7}$$

Thus, if the sources are positively dependent, then $r > 1$ and d_{MLE} is an underestimate of d, the true value in the population. Negative dependence implies $r < 1$ and, thus, that d_{MLE} is an overestimate.

Even if the sources are not independent, the derived values of d and p may be very useful in evaluating estimates of prevalence rates for the population. One may have, for instance, some independent knowledge of the likely direction of the dependence of any of two sources that have been used. Thus, if one knows that two particular sources are likely to be positively dependent, then the estimates derived from the equations above and the boundaries of their confidence intervals may be regarded as plausible lower limits of the true values. If there is likely to be a negative dependence, then the estimates and the boundaries of their confidence intervals become plausible upper limits of the true values.

We apply these simple concepts below to a published attempted complete prevalence study that enabled us to make multiple separate estimates of the population size using the above equations and, from some independent judgments about the nature of the sources of ascertainment, to make estimates of the total number of cases unascertained by the study. While the study attempted an exhaustive prevalence survey, it also reported some limited data on ascertainment. Although there was no adjustment for source of ascertainment, the data provide evidence that the true prevalence is about 30 percent greater than that reported.

THE STUDY

An extensive investigation of Huntington's disease, a genetic neurodegenerative disorder, included a study of the prevalence of cases in Maryland on April 1, 1980 (17, 18). A search for all possible cases was undertaken, using 14 different ascertainment sources, including genealogical records from pedigrees. All cases identified as affected or possibly affected in each source were examined by the principal investigator, a neurologist, to confirm the diagnosis. The authors reported a total of 217 cases alive and resident in Maryland on the prevalence day with the characteristic motor signs for the disorder. There were 212 definite and five probable cases. For the probable cases, no affected relative could be documented. The authors did not present data on the numbers of cases by all possible intersections of sources of ascertainment. They did, however, provide sufficient data to allow calculation of the intersection of each *particular* source will all other sources pooled. This allowed us to estimate the number of missing cases.

In table 2, we present data by ascertainment on particular sources given (18). Within each group, the three entries in each row define values of a, b, and c. Here X, with b unique cases, may be regarded as the source specific to the row, and Y, with c unique cases, refers to cases listed in *all other sources* pooled.

THE ANALYSIS: METHODS AND RESULTS

From data on each ascertainment source, we derived nearly unbiased estimates, d_{NUE} and p_{NUE}, from equations 3 and 4 above, treating all other sources pooled as another separate source. We derived confidence intervals about these estimates using a goodness-of-fit-based method (19, 20). (This results in an asymmetric, more accurate interval than one derived from multiples of the standard error of the estimate.) The nearly unbiased estimates are presented in table 3. We reiterate that any estimate is probably too low if the sources are *positively* dependent and probably too high if the sources are *negatively* dependent.

The derived estimates of missing cases

TABLE 2. Cases by sources*

Source X	*b* Unique to source X	*a* In source X and some other source	*c* Not in source X, in some other source	*b* + *a* Total in X
Genealogical investigation	45	53	119	98
Voluntary health organizations	8	38	171	46
General hospital discharge diagnosis	6	36	175	42
Johns Hopkins genetics clinic	0	36	181	36
Johns Hopkins discharge diagnosis	1	34	182	35
Radio and newspaper spots	9	19	189	28
Urban medical specialists	8	14	195	22
Veterans Administration hospitals	9	12	196	21
Nursing homes	2	15	200	17
Department of Social Services	1	15	201	16
Rural physicians	3	10	204	13
State psychiatric hospitals	1	11	205	12
County health departments	4	2	211	6
National Institutes of Health	1	4	212	5

* Calculated from table 6.5 of ref. 18 (Folstein SE. Huntington's disease: a disorder of families. Baltimore: The Johns Hopkins University Press, 1989:103). Sources were listed here in the order of the total number of cases reported in each source.

range from zero to 281. (Recall that the total number of cases ascertained was 217.) The highest estimate is derived from a source with a small total number of cases (county health department, six cases). To diminish sampling fluctuation, we confine further consideration to estimates derived from the eight sources with 20 or more total case reports ($b + a$) as noted in the last column of table 2. The estimates of missing cases derived from these, as noted in table 3, run from zero (derived from the Johns Hopkins genetics clinic data) to 136 (derived from Veterans Administration hospitals' data).

Only two derived estimates imply complete or nearly complete ascertainment. One is that just cited, derived from the Johns Hopkins clinic data, and the other is from the Johns Hopkins discharge diagnoses ($d_{NUE} = 5$). If each of these sources is independent of all others pooled, then the total survey result would probably be close to complete. This would also imply that each of the *other* sources had negative dependencies with all the others pooled, since the estimates of missing cases derived from each of the other sources are much higher. However, if anything, it appears far more plausible that ascertainment of a case at the Johns Hopkins genetics clinic or through discharge diagnoses is *positively* associated with the ascertainment mode in all the other

TABLE 3. Numbers of cases reported in each source (n), nearly unbiased estimates of unascertained cases (d_{NUE}), and total number of cases (p_{NUE}), on assumption of the independence of each source X with all other sources pooled

Source X	n	d_{NUE}	90% confidence interval	p_{NUE}	90% confidence interval
Genealogical investigation	98	99	67–149	316	284–366
Voluntary health organizations	46	35	17–68	252	234–285
General hospital discharge diagnosis	42	28	12–59	245	229–276
Johns Hopkins genetics clinic	36	0	0–7	217	217–224
Johns Hopkins discharge diagnosis	35	5	1–21	222	218–238
Radio and newspaper spots	28	85	43–174	302	260–391
Urban medical specialists	22	104	50–231	321	267–448
Veterans Administration hospitals	21	136	68–307	353	285–524
Nursing homes	17	25	6–80	242	223–297
Department of Social Services	16	13	1–54	230	218–271
Rural physicians	13	56	17–169	273	234–386
State psychiatric hospitals	12	17	2–78	234	219–295
County health departments	6	281	106–2,113	498	323–2,330
National Institutes of Health	5	42	5–270	259	222–487

sources pooled, since patients seen elsewhere subsequently would probably continue to carry the diagnosis made there, and cases already diagnosed elsewhere would probably be referred to that center. Moreover, ascertainment at the genetics clinic would also be likely to have a positive association with ascertainment through genealogical investigation of reported cases, the source with the largest number of reported cases. Thus, we regard estimates derived from these sources as likely to be underestimates because of plausible positive dependencies. The range of estimates of the missing cases generated from considering each of the other six sources with 20 or more cases is 28 (general hospital discharge diagnoses) to 136 (Veterans Administration hospitals). It would appear plausible that cases seen in Veterans Administration hospitals would be less likely than the average case to be seen in all other sources pooled because of the relatively closed nature of this organization, consistent with a negative source dependence. Thus, the 136 cases missed would appear, if anything, to be an overestimate. "Urban specialists," the source providing the next highest estimate, 104 cases missed, might, on similar grounds, provide a negative dependence and overestimate.

The next highest estimate of missing cases, 99, is derived from genealogical investigation of reported cases. We believe that this source is, if anything, not negatively dependent with other sources but, rather, is either independent of them or, more plau-

sibly, positively dependent, for the presence of a case reported in genealogical investigation already implies that at least one relative was first ascertained in some *other* source. The derived value of $d_{NUE} = 99$, therefore, is either a nearly unbiased estimate of the true number or an underestimate. The confidence interval 67–149 derived from this source thus may be regarded as, if anything, biased downward.

The next highest estimate, 85 missing cases, is derived from responses to radio and newspaper spot announcements. Presumably, individuals responding to media solicitations had some reason to believe they had or were concerned about Huntington's disease. This implies that they had already learned of the diagnosis through some medical source or perhaps were concerned because of some affected relative. All cases responding were investigated thoroughly by the authors, and only those confirmed were included in the survey. Yet, nine of 28 (38 percent) cases in this source were ascertained *only* here, not through medical sources or family genealogical investigation. We see no obvious reason why this source should be either positively or negatively dependent of all other sources. The values derived from this source provide a nearly unbiased estimate of the missing ($n = 85$) and total ($n = 302$) cases, consistent with the 28 percent missed. Of the number of missing cases, the 90 percent confidence interval 43–174 is wider than and includes the 90 percent confidence interval 67–149 derived from genealogical records. The latter was, if anything, biased *downward.* Thus, a plausible 90 percent confidence interval about missing cases is bounded by 67–149, implying a 90 percent interval about the prevalence bounded by 284–366. This suggests that, with 90 percent confidence, somewhere between 24 percent and 41 percent of cases were missed.

DISCUSSION

Our goal has been to illustrate the additional inferences possible from data on the intersection of sources of ascertainment for an attempted complete prevalence survey, not to define a precise prevalence estimate for the disease considered, nor to draw criticism to the study itself. Its extensive search for cases, including advertising, and the other aspects of its investigation are superior to those of most reported prevalence surveys of genetic or congenital disorders and probably of other disorders as well. Indeed, the very inclusion of limited data on ascertainment that enabled this reanalysis indicates a recognition of the possible utility of such data, a recognition absent from most other attempted complete surveys of which we are aware.

The simple information that we could derive from the data provided in the report on the intersection of each source with all other sources enables the plausible inference that a significant proportion of cases were missed by this attempted complete prevalence survey. Were data available on more complex categories of source intersections, particularly on three-or-more-way intersections of genealogical records, public media responses, and the others, then one could make a more refined estimate and probably narrow the confidence interval.

Of course, any estimate of missing cases may be inaccurate. There may always be some higher order intersection of sources of which one is unaware. However, the assumption of most attempted complete prevalence surveys, that they have found all cases, results also in an "estimate" of missing cases: zero or negligibly small. One may evaluate independently this assumption with quantitative methods if information on ascertainment sources and their intersections is available. If data by ascertainment source intersection are provided in prevalence studies, then one can undertake an independent evaluation of completeness. Without such data, however, no estimate of completion is possible. For this reason, we urge that no attempted complete prevalence studies be published without data on ascertainment intersection and that one should regard with skepticism published results of surveys that do not report such information.

REFERENCES

1. Wittes JT, Sidel VW. A generalization of the simple capture-recapture model with applications to epidemiological research. J Chronic Dis 1968;21: 287–301.
2. Wittes JT, Colton T, Sidel VW. Capture-recapture methods for assessing the completeness of case ascertainment when using multiple information. J Chronic Dis 1974;27:25–36.
3. Wittes JT. Applications of a multinomial capture-recapture model to epidemiological data. J Am Stat Assoc 1974;69:93–7.
4. Wittes JT. On the bias and estimated variance of Chapman's two-sample capture-recapture population estimate. Biometrics 1972;28:592–7.
5. Bishop YMM, Fienberg SE, Holland PW. Discrete multivariate analysis: theory and practice. Cambridge, MA: MIT Press, 1975:229–56.
6. LaPlace PS. Sur les naissances, les mariages, et les morts. (In French). In: Historie de l'Academie Royale des Sciences, Année 1783. Paris: p. 693 (cited by Seber GAF. The estimation of animal abundance and related parameters. London: Charles Griffin, 1982.)
7. Lewis CE, Hassanein KM. The relative effectiveness of different approaches to the surveillance of infections among hospitalized patients. Med Care 1969;7:379–84.
8. Hook EB, Chambers GM. Estimated rates of Down's syndrome in livebirths by one year maternal age intervals for mothers aged 20 to 49 in a New York State study—implications of the "risk" figures for genetic counseling and cost benefit analysis of prenatal diagnosis programs. In: Bergsma D, Lowry RB, Trimble BK, et al., eds. Numerical taxonomy of birth defects and polygenic disorders. Vol 13. New York: Alan R. Liss, Inc, 1977:123–41.
9. Hook EB, Albright SG, Cross PK. Use of Bernoulli census and log-linear methods for estimating the prevalence of spina bifida in livebirths and the completeness of vital record reports in New York State. Am J Epidemiol 1980;112:750–8.
10. Huether CA, Gummere GR, Hook EB, et al. Down syndrome: percentage reporting on birth certificates and single year maternal age risk rates for Ohio 1970–1979; comparison with Upstate New York data. Am J Public Health 1981;71:1367–72.
11. Hook EB. The epidemiology of Down syndrome. In: Pueschel SM, ed. Down syndrome: advances in biomedicine and the behavioral sciences. Cambridge: Ware Press, 1982:11–88.
12. Seber GAF. The estimation of animal abundance and related parameters. London: Charles Griffin, 1982.
13. White GC, Anderson DR, Burnham KP, et al. Capture-recapture and removal methods for sampling closed populations. Los Alamos: Los Alamos National Laboratory, 1982. (LA-8787-NERP).
14. Morton NE, Chung CS, Mi MP. Genetics of interracial crosses in Hawaii. Basel: S. Karger, 1967: 94–103.
15. Hook EB, Regal RR. Validity of Bernoulli census, log-linear, and truncated binomial models for correcting for underestimates in prevalence studies. Am J Epidemiol 1982;116:168–76.
16. Chapman DG. Some properties of the hypergeometric distribution with applications to zoological sample censuses. Univ Calif Public Stat 1951;1: 131–60.
17. Folstein SE, Chase GA, Wahl WE, et al. Huntington disease in Maryland: clinical aspects of racial variation. Am J Hum Genet 1987;41:168–79.
18. Folstein SE. Huntington's disease: a disorder of families. Baltimore: The Johns Hopkins University Press, 1989:103.
19. Regal RR, Hook EB. Goodness-of-fit based confidence intervals for estimates of the size of a closed population. Stat Med 1984;3:287–91.
20. Regal RR, Hook EB. The effects of model selection on confidence intervals for the size of a closed population. Stat Med 1991;10:717–21.