

Estimating Disease Prevalence from Clinical Data Using Capture-Recapture

Carol Conell, Division of Research, Kaiser Permanente Northern California

ABSTRACT

Capture-recapture analysis provides a natural way to model disease prevalence using the longitudinal diagnostic and treatment data from Electronic Medical Records. This technique corrects for ascertainment bias due to not yet diagnosed incident cases as well as cases not being actively treated. This paper shows how SAS® can be used to transform routinely generated data into a series of repeated measures of disease identification and to apply capture-recapture analysis to the resulting series. A simple example combining survey and clinical data on alcohol and substance use disorders is presented.

INTRODUCTION

Health care planning requires accurate estimates of disease prevalence. Until recently, surveys were the primary source for prevalence data. However, as Electronic Medical Records (EMR) proliferate, longitudinal diagnostic and treatment data have become increasingly available. To date, clinical data have been used mainly to create disease registries and prevalence estimates based on known cases. However an unknown, variable, and for many diseases fairly lengthy period elapses between disease onset and diagnosis. True prevalence is typically underestimated due to cases that are not being treated actively as well as those that have not yet been diagnosed. This confuses the epidemiological picture, leading to lower ascertainment for diseases that are harder to diagnose. Moreover, the likelihood of diagnosis and treatment are often influenced by socioeconomic status (SES), healthcare insurance status, and demographics. As a result, the relative disease burden among subpopulations is misrepresented.

Capture-recapture analysis provides a natural way to model disease prevalence using clinical data (Hook & Regal 1995; 1999). However, this technique continues to be underused, despite two decades of evidence that it can improve prevalence estimates even for diseases like diabetes that are both common and relatively well identified (Bruno *et al* 1994).

Use in public health and healthcare research has probably been discouraged by the heavy reliance in the literature on specialized programs. In addition most healthcare applications have emphasized sporadic indicators rather than routine clinical data. This paper shows how SAS can be used to transform diagnostic and treatment data into a series of repeated measures of disease identification and to estimate disease prevalence by applying capture-recapture analysis to the resulting series. A simple example using information from a clinical database on alcohol and substance use disorders (UD) is presented.

CAPTURE-RECAPTURE TECHNIQUES: FIVE STEPS.

The basic problem in applying capture-recapture analysis to disease prevalence involves estimating the number of active cases of the disease that are not currently registered, which we will refer to as \check{N}_0 . Equally important, given the uncertainty attached to the estimate of the unknown cases, one needs to estimate lower and upper bounds: $L_{\check{N}_0}$ and $U_{\check{N}_0}$. By convention, we will use these terms to refer to the upper and lower confidence bounds associated with a 95% confidence interval for the number of unknown cases. The “ascertainment corrected” estimate is $\check{N} = n + \check{N}_0$, i.e., the sum of the number of known cases (n) and the estimated number of unknown cases (\check{N}_0). In most applications, as a first order approximation we can ignore the uncertainty attached to the number of known cases, since the uncertainty associated with \check{N}_0 dominates that associated with n . This

enables us to calculate upper and lower bounds $L_{\check{N}_0}$ and $U_{\check{N}_0}$ for \check{N}_0 and add these to the observed cases n to estimate upper and lower bounds for the ascertainment corrected prevalence: $L_{\check{N}} = n + L_{\check{N}_0}$, $U_{\check{N}} = n + U_{\check{N}_0}$. Once we have the ascertainment corrected prevalence, we can calculate the sensitivity of various approaches to diagnosis as well as the sensitivity (completeness of registration) of the diagnostic process. And we can check whether sensitivity varies among subgroups.

We can divide the process of obtaining a useable prevalence estimate into five steps. For each step, we will discuss what needs to be and how this is accomplished within SAS: create a capture-recapture record, create an analysis database, select a model, estimate the number of unknown cases and calculate confidence bounds for the number of cases, explore the plausibility of the resulting estimates. For ease of exposition, we will discuss each step in the context of a specific problem: identifying the number of individuals in a membership based health plan who have UD. We will point out features of this case that could affect the modeling. Hopefully, this will make it easy to adapt to other cases.

STEP ONE: CREATE IDENTIFICATION HISTORIES

The basic data used in capture-recapture estimation consists of capture or identification histories for a set of individuals who are identified at least once. Each record in this database consists of an identifier for a specific individual, say individual x and a string of j identification or indicator variables $I_1 \dots I_j$, where $I_i(x) = 1$ if individual x was identified on the i^{th} occasion and 0 otherwise. Obtaining a useful identification history (which we also refer to as a capture-recapture record) requires understanding how the data are generated in order to ensure that the assumptions justifying the technique are reasonable. The technique requires identifying a population (here the individuals who have UD and belong to the health plan) that could potentially be identified by each of several indicators. This requires a theory about the social organizational process producing each of the potential indicators or bases for diagnosis. Capture-recapture is explicitly designed to produce reliable estimates when missed cases (false negatives) are frequent, i.e., when the sensitivity of any specific diagnostic occasion is fairly low. Handling false positives is far more problematic. Closed capture-recapture also requires that all individuals are at risk of identification by each indicator. It is possible to model situations in which the population is not closed but doing so is much more complicated and beyond the scope of this paper.

Fortunately, when we are trying to estimate the number of cases of chronic disease within an insured population, the basic requirements for applying capture-recapture are approximated. Any identification associated with insured services will be tagged with a standard identifier for the patient, so one can easily determine whether the same individual is identified on multiple occasions. And each individual can receive service for the same condition and/or be given the same diagnosis as often as they use services and can use as many services as they require, so we can assume sampling with replacement. If an individual receives services that are not covered the sensitivity of identification will decline, but this is a problem that the method handles well.

The first preliminary task involves identifying a universe of subjects S within which to identify prevalent cases. For a chronic condition it makes sense to limit consideration to individuals enrolled for the entire period observed. We refer to this as the universe or super-population, because the modeling will focus on the “diseased” or “affected” population N , i.e., the subset of members of S who have the condition. (For convenience, we will sometimes refer to the population with the condition, whether identified or not, as cases.) In our example, the subject universe consisted of all individuals 18 or older who were enrolled in the health plan for the full year they were observed, used services at least once during that period, and responded to a survey about their health ($N=31,861$).

The universe never appears explicitly in the capture-recapture analysis, but the epidemiological and health services applications all require a clear identification of the universe of subjects as well as of the population of cases to be estimated. Although health service planners may actually want to estimate the number of unknown cases, most of the time interest focuses on relative prevalence, e.g., the per cent of the insured population with UD. In the example considered here coverage was similar for all individuals, but if not one would need to stratify based on different plans.

The second preliminary task involves identifying several distinct occasions on which each patient can be identified as having the condition. Although the concept of capture-recapture applies to situations in which there are only two distinct indicators, in practice useful applications to epidemiology require at least three indicators. As will be clear when we introduce the modeling, the technique can only be applied with two indicators if there is no dependence between the indicators: a condition that is almost never met. This means that we will want at least 2 periods when we have information on two distinct types of identification, resulting in a 4 indicator identification history.

There is considerable flexibility possible in defining an occasion, i.e., the period of time for which one will record identification as a single occurrence. For healthcare utilization and clinical data, the choice can range from a single encounter with the healthcare system to a period as long as a year. The decision about whether to aggregate across encounters and how long a total period should be considered in modeling prevalence should be made empirically depending on the disease's relative prevalence (number of cases per individual in the universe) and chronicity. As the period lengthens, the problem of individuals moving into the diseased population during the period becomes a more important source of potential problems.

For use disorders, chronicity is high. Put differently, during a one year period incident cases constitute only a small the fraction of all prevalent (potentially identifiable) cases. Within the insured population, UD is fairly rare and even if a subject has a use disorder, it is unlikely to be diagnosed by the physician during a substantial fraction of encounters. Consequently, we decided to model prevalence for a one year observation period divided into two half-year identification periods: January 1-June 30 and July 1-December 31. Once the indicators are defined a basic identification history record is created for each subject.

Health plans that use electronic medical records (EMR) and insurance companies routinely produce at least two types of information indicating the existence of a specific chronic illness: diagnoses and treatments. We assume that everyone who had the disorder could potentially be identified both based on diagnosis and on treatment during each of the 2 periods and created 4 indicators: Assuming a hypothetical semi-raw database (RAWDATA) with information about treatments and diagnoses, one can produce an analysis database suitable for estimating the number of individuals with the disease in 3 steps.

We assume that RAWDATA contains records with the following 4 variables: Id, Date, Tx, and Dx, where Dx is positive if a relevant diagnosis was recorded on a given Date and Tx is positive any day that the subject received relevant treatment: visited a clinic specializing in the disease, received a prescription for some pharmaceutical used to treat the disease, received a procedure specific to treating the disease. In practice, of course, the raw data will have a variety of diagnostic and treatment codes, but a preliminary step in creating a database to use with Proc GENMOD for a capture-recapture analysis is producing a stylized RAWDATA file.

For each type of indication and each period, we want a period specific indicator. There are a variety

of ways to create an identification history for each subject with the two types of period-specific indicators. In the SAS code examples, we will assume that subjects are uniquely identified by their subject identifier (Id). The code below illustrates one approach, which requires one sort for each indicator and one merge. It produces a database with one record for each subject who was identified by at least one of the indicators and a 4 indicator identification history. E.g., if subject 1 was only identified by treatment in period 2 the first record would be as follows: (Id,D1,T1,D2,T2)=(1,0,0,0,1). And if subject 2 was never identified there would be no record with Id=2.

```
PROC SORT data=RAWDATA (WHERE =(Begdate_1 <=Date <=Enddate_1) and Dx=1)
out=D1;
BY ID;

DATA IdHx_bysubj;
MERGE D1(in=D1) T1(in=T1) D2(in=D2) T2(in=T2);
by Id; D1=max(0,D1); T1=max(0,T1);D2=max(0,D2); T2=max(0,T2);
```

STEP TWO: CREATE AN ANALYSIS DATABASE

We will use the GENMOD procedure to fit a series of models to the pattern of identification, obtain the information we need to select the best model, estimate the number of missed cases, i.e., in our application, the number of cases for which (D1,T1,D2,T2)=(0,0,0,0), estimate upper and lower bounds for the number of missed cases, and estimate the total number of cases.

This requires building an analysis database (ANALDATA) with one cell for each potentially observable history, including those that do not actually occur in the observed data, and one additional cell for the totally and inevitably unidentified case. An important distinction needs to be preserved, between the sampling and the structural zeros.

First, create a table (ObsIdHx) with all possible identification histories and the number of subjects (N) with each history.

```
PROC SUMMARY data=IdHx_bysubj nway; CLASS D1 T1 D2 T2; OUTPUT out=ObsIdHx
(rename=_Freq_=N);
```

The resulting table will have up to 15 observations, one for each identification history that was observed.

Second, create an identification frame database with 1 observation for each identification history that could occur with 4 indicators. Given the four indicators in this case, this is a 16 cell table (2 x 2 x 2 x 2) with one observation for each possible identification history (D1,T1,D2,T2). Order the histories from (1,1,1,1) to (0,0,0,0) and identify the final (0,0,0,0) cell as a structural zero. This database will have at least one more observation than ObsIdHx, since it includes a cell for the structural 0 associated with no identification by any indicator: for the 2 period and 2 types of identification (4 indicator) case (D1,T1,D2,T2,N)=(0,0,0,0,). The basic data in each observation consists of the 4 identification vectors and a count variable (N) indicating how many times this identification history was observed.

Third and last, merge the FRAME table with the ObsIdHx table to get the analysis database: ANALDATA. Set N to 0 for any cell, other than the structural zero, where N is missing, i.e., for the sampling zeros if any in the table.

PROC GENMOD will fit each model to the complete cases in ANALDATA. This means that if there are sampling zeros (i.e., potentially observable identification histories that did not occur in the

observed data), the program will treat these just like the other observed histories. As a result, the analyst DOES NOT need to worry about adjusting for empty cells by adding small amounts.

In practice, there will be other variables included in this database, but they are all functions of the identification history variables. For example, I like to include a variable $STRUCZERO = 1$ for the structural 0 for ease of working with the data, but this is not strictly necessary as this is easily calculated for any observation as $1 - \max(D1, T1, D2, T2)$. I include it to facilitate checking the data structure and estimation routines especially when more complicated models are introduced with covariates to control for subject heterogeneity. The other variables that we need to create to fit this application are discussed in the context of model selection.

The structural zero cell will only be used in estimation. Including it in the analysis database will facilitate obtaining an estimate of the number of missing cases, and the confidence limits on that number.

The UD analysis began by collecting raw data on treatment and diagnosis for 31,861 subjects in the universe within which we wanted to identify use disorders. Of these, 508 individuals were identified (registered) as having use disorders by at least one of the 4 indicators for a relative detected prevalence of 1.6%. Two critical assumptions justifying capture-recapture analysis are that some unknown number of the 31,353 subjects who were not identified as having UD actually had potentially identifiable use disorders and that we can estimate the volume of unidentified cases by examining the pattern of interaction among the indicators within the identified cases. Although no sampling zeros occurred in this case, there was one identification history cell that contained only 1 case and four of the 15 histories contained less than 4 cases, suggesting the importance of a set up that allows for possible sampling zeros, as well as the desirability of identifying a structural model with a moderate number of parameters.

STEP THREE: SELECT A MODEL

Since we assume that identification on each occasion is possible for any individual, whether or not they have been previously identified, we can assume an underlying multinomial distribution and use a loglinear or logistic regression model (Alho 1990, Tilling and Sterne 1999). For the four indicators, if we let i index the first indicator (D1), j the second (T1), k the third (D2) and l the fourth (T2), we fit a model $\mathbf{M}(i,j,k,l)$ as the sum of a main term: μ_0 , a first-order effect for identification by each of the four indicators, possible 2-way interactions for each pair of indicators, possible 3-way interactions for each triple of indicators and so on. Since the objective is to identify a structural model, i.e., one that is not simply fitting the random associations but reflecting the structure of identification, the best model is usually selected by identifying the model that minimizes Akaike's Information Criterion, corrected for the limited number of observable identification histories and the relatively large number of model parameters (AICc) (Burnham and Anderson 1995; 2002). For each model \mathbf{M} , $AICc(\mathbf{M}) = G^2_M + 2(rk_M)/(r-k-1)$, where G^2_M is the likelihood ratio chi-sq statistic associated with the model, k_M is the number of model parameters and r is the number of observable identification histories. This criterion makes it possible to compare models that are not hierarchical by offsetting improvements in fit by a penalty for model complexity. Using AICc to select a model favors complexity slightly more than selection based on log likelihood (using a significance level of .05 as is usual).

Obtaining reliable estimates for the number of missed cases requires that the model fits the observed counts well, i.e., that the deviance is low. This means that the model accounts for the distribution of cases among the identification histories almost as well as the saturated model. Consequently, we initially fit a saturated model and compare the Log Likelihood of each potential

structural model to the saturated model (maximal) log likelihood, which has 0 deviance. The easiest way to specify a saturated model is to treat the variables as classificatory, since GENMOD will create all the necessary interaction terms automatically. On the other hand, the easiest way to constrain the impact of parameters and explore heterogeneity involves treating identification variables as continuous. The examples give sample codes for both approaches.

The reason that we can switch back and forth from classificatory to continuous predictors is that identification is intrinsically binary. Specifying the reference group parameterization with 0 (no identification) as the reference group on the class statement, as in the example below, results in the same parameter estimates whether we use classificatory variables or continuous variables. In practice, the parameters for most of the models are not really of much interest, although it is helpful to be able to compare the parameters for the identification variables in the two periods to verify that imposing equality constraints on the corresponding parameters in the two periods makes sense.

Since GENMOD is a generalized linear modeling procedure it has a large number of parameters that need to be specified. All the models we fit take the form below with the only changes involving the predictors included in the MODEL statement and whether a CLASS statement is included. Omitting the CLASS statement means the indicators are treated as continuous. Specifying **DIST=POISSON** and **LINK=LOG** fits the loglinear model. Specifying **OBSTATS** creates a database with the predicted value for every identification history in **ANALDATA**, including the all identification missing cell. This database can be used to check residuals. In addition, since the structural zero observation is included, GENMOD calculates the expected value for the missed cases as well as approximate ninety-five percent confidence limits. Specifying **LRCI** calculates upper and lower likelihood limits for all model parameters. Although the parameterization does not affect the results, by including the formatting code and using the **PARAM=REF** and **REF='No'** options as shown, one will produce the same parameters as when we treat the identification variables as continuous.

```
PROC FORMAT;VALUE NOY1F 0='No' 1= 'Yes';
ODS OBSTATS=OBSTATS;
PROC GENMOD DATA=ANALDATA ORDER=FORMATTED;
ODS EXCLUDE OBSTATS;
CLASS D1(REF='NO') T1 D2 T2 TXS0 /PARAM=REF ;
FORMAT dxs0 dxs1 txs0 txs1 n0y1f. ;
MODEL N=D1|T1|D2|T2 / DIST=POISSON LINK=LOG TYPE3 OBSTATS LRCI;
```

The saturated model automatically omits the four-way interaction: with one missing cell, there are only 15 observed cells; consequently, a model that allows 1 main effect, 4 first order effects, 6 second order effects and 4 third order effects has 15 parameters and fits the observed cells perfectly. For the UD example, the associated log likelihood statistic was 1510.6 with 15 model parameters. More constrained models are compared against this one to see whether a parsimonious model can be identified with sufficiently low deviance. If so, one selects the model with the lowest AICc. At this point, where sporadic indicators are used, the modeling process becomes complicated and ad hoc.

Fortunately, in the case of two distinct indicators (Dx and Tx) based on experience in two non-overlapping periods of the same length, a much simpler modeling process is justified by the logic of the indicators. Since each of the two types of indicators is defined in the same way for each period, we expect the effect of identification by Dx (Tx) and the relationship (if any) between identification by Dx and Tx will be invariant across periods. If this assumption is correct, instead of allowing a marginal effect for each of the four indicators and period specific interactions, we can fit a single first-order effect $\mu_D(\mu_T)$ for identification by Dx(Tx) in each period, respectively and a single

second-order interaction effect μ_{D_T} for the interaction of diagnostic and treatment based identification. In addition, it is reasonable to allow for the fact that either due to individual heterogeneity in the ease of diagnosis or the likelihood of treatment or to ongoing processes leading to association across periods, there is likely to be some association of diagnosis across periods and also of treatment but much less likely to be a direct association between identification based on diagnosis in one period and identification based on treatment in the other period. This results in a fairly straightforward model selection process. The same basic logic extends to any number of periods for data gathered based on EMR or similar health plan or insurance databases.

Following through on the structured nature of the data, we fit four, successively more restricted, models to determine whether these assumptions about the pattern of identification were justified. As we explained above, in order to fit the more constrained models in GENMOD, we first created some additional variables in ANALDATA.

First, since we wanted to treat the indicators as continuous rather than classificatory, we created the continuous interaction measures duplicating those that the program automatically creates when we specify class measures. Specifically, for each pair of identification measures, we created a second-order interaction effect that was 1(positive) only when both types of identification were positive. E.g., the interaction between D1 and T1 was coded

$$D1_T1 = D1 * T1;$$

Second, for each pair of parameters where we wanted to impose an equality constraint, or equivalently substitute a single parameter, we created a single measure by adding together the parallel indicators. E.g., since we wanted to be able to constrain the parameters on D1 and D2 to take a single value, we created an indicator D that summed up identification by D1 and D2. In all, we created four new indicators, each adding up a pair of the original indicators

$$D = D1 + D2; \quad T = T1 + T2; \quad D_T = D1_T1 + D2_T2; \quad I_{\text{bothpd}} = D1_D2 + T1_T2;$$

Using these new variables, we can easily specify four successively more restricted models. Model B1 allows for all second order (2-indicator) interactions but no three-indicator interactions. We fit model B1 two different ways. Initially, we treated the indicators as classificatory and used the following model specification.

$$N = D1 | T1 | D2 | T2 @ 2;$$

Subsequently, we treated the indicators as continuous, and used the following model specification, which requires the interaction measures we coded explicitly..

$$N = D1 \ T1 \ D2 \ T2 \ D1_D2 \ T1_T2 \ D1_T1 \ D2_T2 \ D1_T2 \ D2_T1 \ ;$$

The results were, as expected, identical.

Model C1 is the first model to impose constraints on parameters across the two periods. Specifically, we constrained the main effects for diagnosis and treatment as well as the within period interaction of diagnosis and treatment to be invariant by period, but allowed the 4 measures of identification s to have type and period specific cross-period interactions.

$$N = D \ T \ D_T \ D1_D2 \ T1_T2 \ D1_T2 \ T1_D2;$$

Model C2 dropped the cross-type and period interactions.

$N=D \quad T \quad DT \quad D1_D2 \quad T1_T2;$

Model C3 constrained the within type interaction to be identical for the two types of identification.

$N=D \quad T \quad D_T \quad Idbothpd;$

Model C4 omitted the interaction of identification across period.

$N=D \quad T \quad D_T;$

Table I below summarizes the information used to select the best model along with the estimates for the missing cells for each model. It also includes estimates for the number of cases missed and the ascertainment corrected prevalence estimates, discussed in the next section.

When we examined the parameters in model B1, as expected, the parameters associated with the different indicators in the two periods were almost identical. In fact, the MLE (.95) confidence intervals on the parameters for diagnosis in the two periods were identical to 2 significant figures and both the period-specific diagnosis-treatment interaction and the treatment coefficients differed only very slightly more.

<i>Table I: Comparing Alternative Models.</i>				
Model	AICc	Df*	N_0	Est(N)
Saturated	>430**	0	-	-
B1	198.6	4	3043(1539-6020)	3551 (2049-6528)
C1	141.6	7	2476(1410-4346)	2984 (1918-4854)
C2	128.1	9	2393 (2340-4270)	2901 (2848-4778)
C3	122.5	10	2448 (2380-4344)	2956 (2888-4852)
C4	193.9	11	370 (282 - 486)	-878 (790 - 994)

* Df is $15-k_M$, i.e., the difference between number of parameters in the model and the number of cells in the table or the number of parameters in the saturated model.

** The AICc cannot be calculated because $r - k - 1$ is 0 but if we assume that $r - k - 1$ we get a penalty of 430. If we omit the 2 least significant parameter, fitting a nearly saturated model, the deviance is extremely low and there is almost no reduction in G^2 , and the calculated AICc is 460.4.

STEP FOUR: CALCULATE ASCERTAINMENT CORRECTED PREVALENCE ESTIMATES.

The estimated value and confidence intervals for the number of missed cases are obtained from the final observation of OBSTATS. The value of Pred for this cell, the structural zero, is the maximum likelihood point estimate for the estimate of N_0 , and Lower and Upper are the lower and upper endpoints for an approximate 95% confidence intervals. See SAS documentation for the calculation of these. The ascertainment corrected prevalence estimates simply add the estimates for the structural zero to the observed cases. An alternative approach to calculating confidence intervals is to use the SAS supplied BOOTCI macro (<http://support.sas.com/kb/24/982.html>). Miller (2004) provides details on calculating boot-strap confidence intervals in SAS.

Table I demonstrates that the AICc value is minimal for model C3, somewhat higher for models C1 and C2, and much higher for all the other models. Since the estimates for the missing cases from models C1-C3 are all extremely close, we can use the estimates produced by model C3. If a number of models have similar AICc and widely varying values for $E(N_0)$, one should consider averaging the

results, which requires calculating AIC weights reflecting how close other AIC values are to the minimal AICc averaging (Giatting G *et al* 2007).

STEP FIVE: MODEL PLAUSIBILITY AND EXTENSION.

The best model (C3) leads to an estimate of UD that is about 5 times as high as the number of cases registered medically during the year. One way to determine whether this is plausible is to compare it to estimates based on entirely different approaches: in this case the estimate is consistent with Mertens *et al* (2005), which compares identification during routine care with the results of screening specifically for the disorders. Another approach involves fitting the model within a population independently identified as having the disorder. As mentioned earlier, everyone in the universe of members we considered responded to a survey on their health at the end of the year. Three hundred forty reported UD at that time. If we apply the best model structure (N3) separately to those who did and did not report UD, we find the following. Among those who did not report use disorders, 1.3% (350/31,521) had UD identified by the medical indicators and we estimate 7.7% (2417/ 31,521) had them. Among those who did report use disorders, 33.2% (113/340) had UD identified by the medical indicators and we estimate 89.2% (303/340) had them.

The above results strengthen the plausibility of the estimates. They also suggest that one of the ways in which one might want to extend the model would be to allow for heterogeneity. For the type of application discussed here, demographic characteristics (age, gender, and ethnicity) and comorbidities and related differences in service utilization are the most common, identifiable, sources of heterogeneity. The simplest way to allow for heterogeneity is to stratify within the population and fit models separately to each stratum. Alternatively, one can unfold the identification database to allow separate cells for each of the strata and fit a single model to the counts, allowing parameters to vary by stratum where justified by the AICc. This approach is only slightly more complicated and will have higher power than fitting models separately to each subpopulation. Tilling (1999) discusses incorporating covariates.

CONCLUSIONS

Using SAS 9.2, it is easy to develop plausible estimates for the number of cases of substance and alcohol use disorder found among members. The fact that repeated measures are readily available for each member makes it possible to develop stable and parsimonious models. Consistent with other research on these disorders, we determine that there is substantial under-ascertainment. Since the techniques used are readily applicable to any chronic condition, the ascertainment corrected estimates can also be used to improve planning by clarifying the true relative prevalence of more and less readily ascertained diseases.

REFERENCES

- Alho JM. 1990. "Logistic regression in capture-recapture models." *Biometrics* 46:623-35
- Bruno G, LaPorte RE, Biggeri A, McCarty D, Pagano G. 1994. "National diabetes programs. Application of capture-recapture to count diabetes?" *Diabetes Care* 17:548-56.
- Burnham KP, White GC., Anderson DR. 1995. "Model selection in the analysis of capture-recapture data". *Biometrics* 51, 888-98.
- Burnham KP, Anderson DR. 2002. *Model Selection and Multi Model Inference: A Practical Information-Theoretic Approach*, Springer-Verlag. New York: NY.
- Giatting G *et al* 2007. "Choosing the optimal fit function: Comparison of the Akaike information criterion and the F-test." *Med Phys.* 34 (11): 4285-92.
- Hook EB, Regal RR. 1995. "Capture-recapture methods in epidemiology: Methods and Limitations." *Epidemiol Rev* 17(2): 243-64.
1999. "Recommendations for presentation and evaluation of capture-recapture estimates in

epidemiology.” *J Clin Epidemiol*. 52(10):917-26; discussion 929-33.

International Working Group for Disease Monitoring and Forecasting. 1995a. “Capture-recapture and multiple-record systems estimation I: History and theoretical development.” *Am J Epidemiol* 142:1047–58.

1995b. “Capture-recapture and multiple-record systems estimation II: Applications in human diseases.” *Am J Epidemiol* 142:1059–68.

Lewden C *et al.* 2006. “Number of deaths among HIV-infected adults in France in 2000, three-source capture–recapture estimation” *Epidemiol Infect.* 134(6): 1345–52)

Miller, DP. 2004. “Bootstrap 101: Obtain robust confidence intervals for any statistic.” *SUGI Paper* 193-29.

Mertens JR, Weisner C, Ray GT, Fireman B, Walsh K. 2005. “Hazardous drinkers and drug users in HMO primary care: prevalence, medical conditions, and costs.” *Alcohol Clin Exp Res* 29(6):989-98.

Tilling K, Sterne JAC. 1999. Capture-Recapture Models Including Covariate Effects *American Journal of Epidemiology* 149(4)392-400.

ACKNOWLEDGEMENTS

The research was approved by the KFRI IRB board and supported by NIAAA 1 R21 AA12804 and NIAAA 5 R01 AA014037. Nancy Gordon, Division of Research, Northern California Kaiser supplied the survey data used in the example analysis. Constance Weisner and G. Thomas Ray collaborated on some of the research on use disorders that formed the basis for the example used.

RECOMMENDED READING

A General Methodology for the Analysis of Capture-Recapture Experiments in Open Populations, *Biometrics*, Sep 1996.

CONTACT INFORMATION.

Your comments and questions are valued and encouraged. Contact the author at:

Name: Carol Conell

Enterprise: Division of Research, Kaiser Permanente Northern California

Address: 2000 Broadway

City, State ZIP: Oakland, CA 94612

Work Phone: 510-891-3574

E-mail: Carol.Conell@KP.ORG

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.