Chronic disease prevalence estimates from electronic medical records: Applying capture-recapture to clinical data

Carol Conell, PhD
Division of Research, Kaiser Permanente Northern California

Background

Allocating scarce healthcare resources requires reliable estimates of relative disease prevalence. Routine clinical data is a good source of information, but obtaining estimates from clinical data can be challenging:

- **Disease registers** need to be supplemented with estimates of completeness for different diseases and populations.
- **Intensive screening** can estimate the fraction of unidentified cases, but is expensive and only provides a point estimate.
- **Capture-recapture** is an alternative**.** The number of cases identified by at least one of several distinct indicators and the associations between indicators are used to estimate the number of unidentified cases.

Objective

- Demonstrate how capture-recapture models can provide prevalence estimates on a regular basis
  - For underreported chronic diseases
  - Without increasing healthcare costs
  - Based on explicit models of how routine medical data are generated
- Illustrate for alcohol and substance use disorders

Capture-recapture applied to routine clinical data

- Modeling how information gets into the EMR yields valid and robust estimates using data from routine clinical care.
- Independently recorded diagnoses and treatments can be aggregated biannually, producing 4 distinct disease indicators each year: I_{D1} (I_{D2}) indicate disorders diagnosed and I_{T1} (I_{T2}) disorders treated during the first (second) half year.
- Classifying cases by pattern of identification (I_{D1}, I_{T1} , I_{D2}, I_{T2}) forms a 16 cell identification table.
- One cell (0,0,0,0) contains the totally unidentified cases. N_0 is the number of such cases.
- Hypotheses about the relationships among the indicators are tested to select the best model for disease identification given the social and organizational context.
- N_0 is estimated using the parameters in the best model.
- Total disease prevalence (***N***) is estimated by the sum of identified and unidentified cases.
- Bootstrap methods create .95 confidence intervals.

Application to Alcohol and Substance Use Disorders

Disease: Alcohol and drug use disorders are frequently unidentified and untreated for long periods of time. Like other chronic diseases that do not respond reliably to available medical treatments, use disorders are underestimated in registers. Intermittent efforts at case finding can easily distort the picture of relative prevalence between contexts.

Data Collection: Diagnoses and treatments were aggregated for 2 successive 6 month periods for 31,861 service-using survey respondents in a membership health plan, creating 4 independent indicators of use disorders. Survey information on self-report was used to validate the final model.

Analysis Details:

- Four hypotheses about relationships among the indicators were derived from an explicit data model and tested.
- All were accepted, resulting in a 5 parameter model including a within period interaction for the diagnosis and treatment (D_T) and a cross-period interaction (BothPd) as well as main effects for diagnosis (D) and treatment (T)
- This (Best) model was used to estimate point prevalence and bootstrap-based .95 c.i.
- Two less restricted models produce very similar point estimates but wider confidence intervals.

Findings:

- 1/11 had use disorders.
- 4/5 of the estimated cases were unidentified.

The table displays estimates for unidentified N_0 and total cases N as well as the number of cases identified by diagnosis and/or treatment.

Identified and Estimated Substance Use Disorders During One Year (n = 31,861)

	Not Treated	Treated (T)	Total
Not Diagnosed	$E(N_0)$ *2448 (1359,4505)*	133	
Diagnosed (D)	218	157	
Identified (D and/or T)			508
Model Estimate			$E(N) = 508 + E(N_0)$ *2956 (1867,5012)*

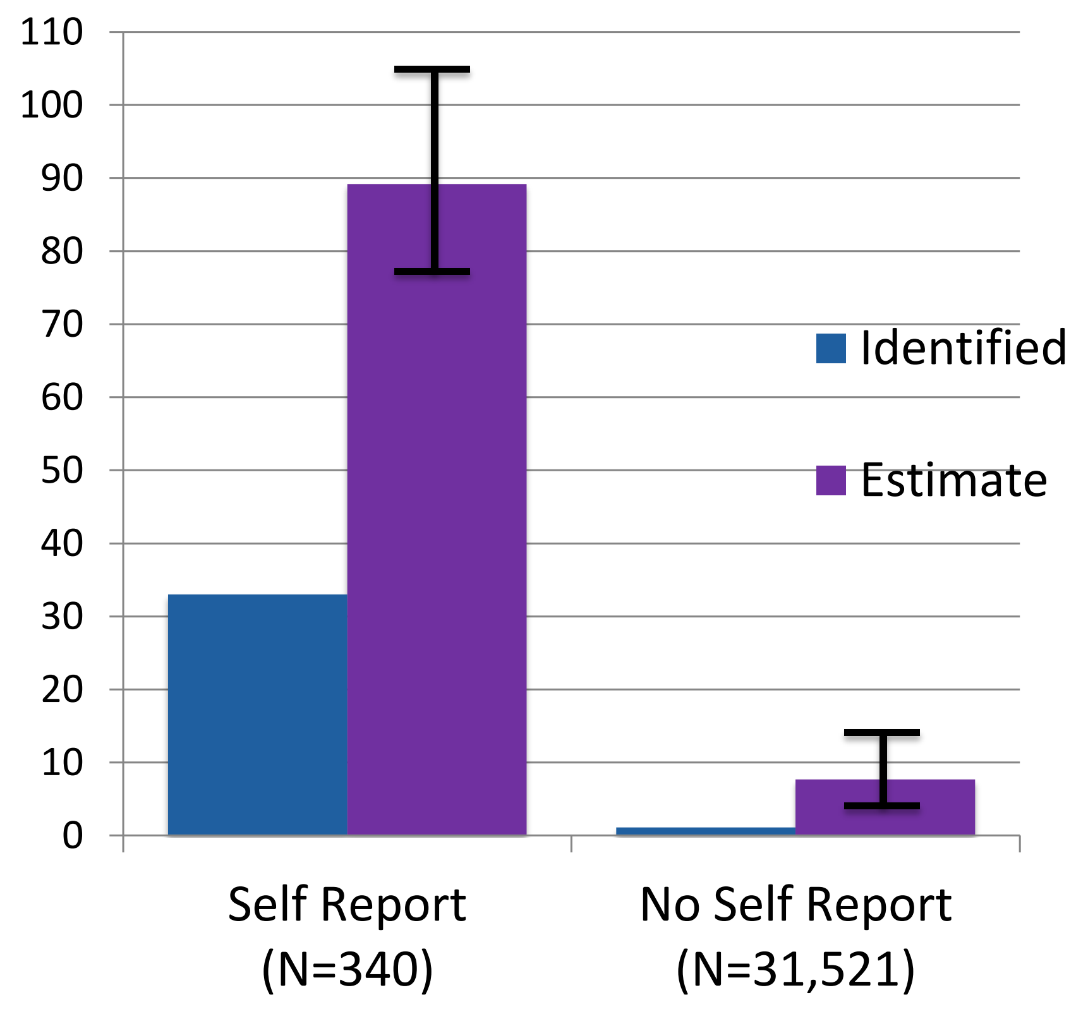
Validation

Estimates agreed with an intensive screening study.
Mertens (2005) found only 1/5 of use disorders identified.

Estimates match the survey self-report subpopulation.
Survey self-reports validated the model.

- Self-reporters were much more likely to be medically identified (33% vs 1.3%)
- Model estimated use disorders for self-reporters
  - Best Estimate: 89%
  - Within the .95 c.i.: 100%
- Even among self-reporters 2/3 were medically unidentified.

Percentage of Substance Use Disorders Clinically Identified and Estimated for Subgroups Defined by Survey Self-Report



The model was fit separately for those who reported disorders on the survey "Self Report" and those who did not "No Self Report." The percent identified is displayed followed by the percent estimated with a .95 c.i.

Conclusions: Utility and Scope of Technique

Why the Specific Model Works Well

- Assessing a chronic condition
- Closed population
- Separate diagnosis and treatment records
- No mutually exclusive identifiers
- Low identification rate

Capture-Recapture Can Estimate Chronic Disease Prevalence from Routine Data

Basic requirements:

- A well-specified data model
- 3+ identifiers (8 cells in identification table)
- The approach can be adapted to violations of the other assumptions given sufficient identifiers.

Limitation

- Does not identify new cases

Conclusions

- Capture-recapture applied to the routine medical data replicated prevalence estimates from intensive screening.
- Survey self-reports validated the capture-recapture prevalence estimates.
- The model of how medical data were produced in this setting was confirmed.
- Validated model estimates can monitor elusive conditions without disturbing patients.
- This is a promising approach for comparing prevalence across settings and diseases.

REFERENCES

Conell, Carol 2011. Estimating Disease Prevalence from Clinical Data Using Capture-Recapture. http://www.wuss.org/proceedings11/Papers_Conell_C_76171.pdf

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.

ACKNOWLEDGMENTS

Approved by the Kaiser Foundation Research Institute's Institutional Review Board. Supported by NIAAA 1 R21 AA12804 and NIAAA 5 R01 AA014037. Nancy Gordon supplied the survey data used in the example analysis. Constance Weisner and G. Thomas Ray collaborated on some of the research on use disorders.