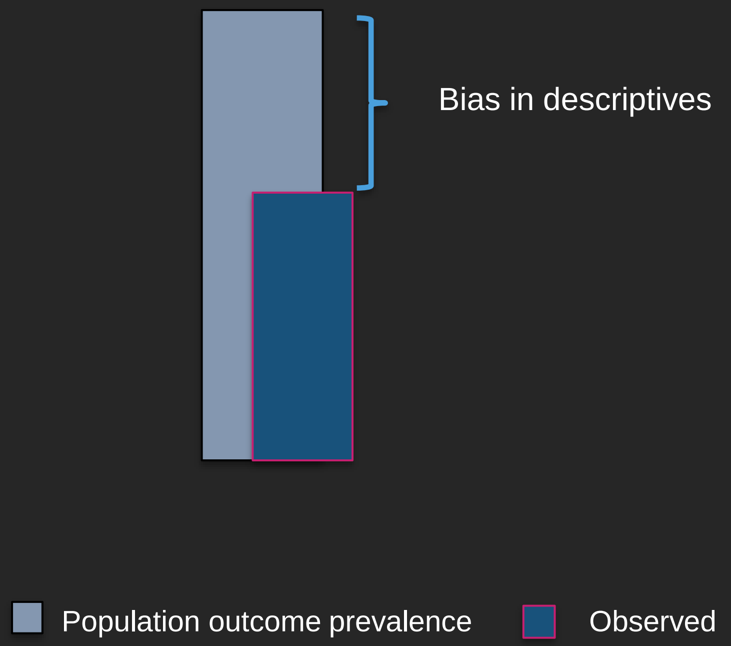
A Capture-Recapture-Based Ascertainment Probability Weighting Method for Effect Estimation with Under-Ascertained Outcomes

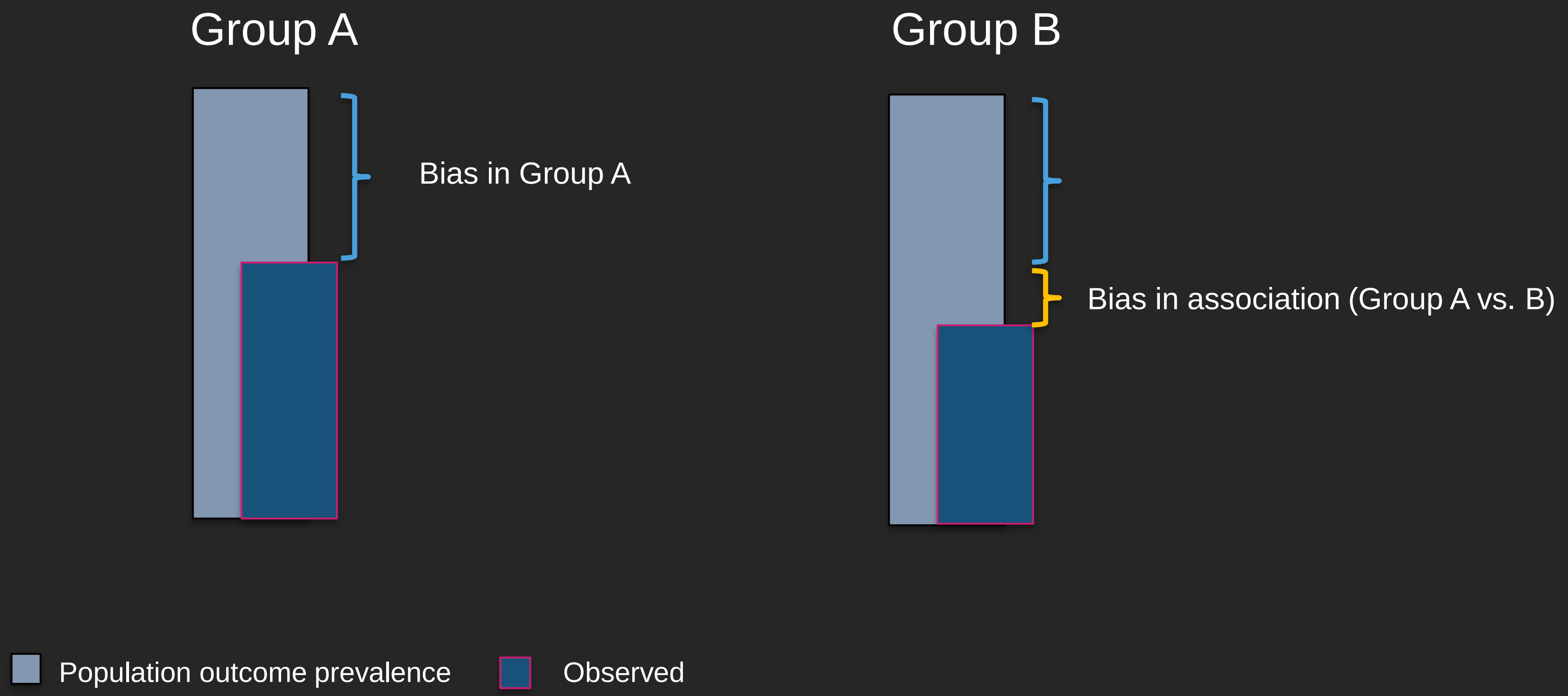
Carl Bonander, Anton Nilsson, Huiqi Li, Shambhavi Sharma, Chioma Nwaru, Magnus Gisslén, Magnus Lindh, Niklas Hammar, Jonas Björk, Fredrik Nyberg

CARL BONANDER, ASSOCIATE PROFESSOR OF EPIDEMIOLOGY

Bias due to under-ascertainment



Bias due to under-ascertainment



Outline of a situation where under-ascertainment could occur

Outcome data linked from an incomplete register

PIN

Study cohort

ID	Y
1	1
3	1

ID	Y
1	1
2	
3	1
4	
5	

Analytical data

ID	Y
1	1
2	0
3	1
4	0
5	0

COVID-19 testing

- Which groups were more likely to get tested for COVID-19?
- Implications for
  - Detection and control strategies
  - Confirmed case data
- No complete register of tests in Sweden
  - SmiNet as good as complete but only for positive tests

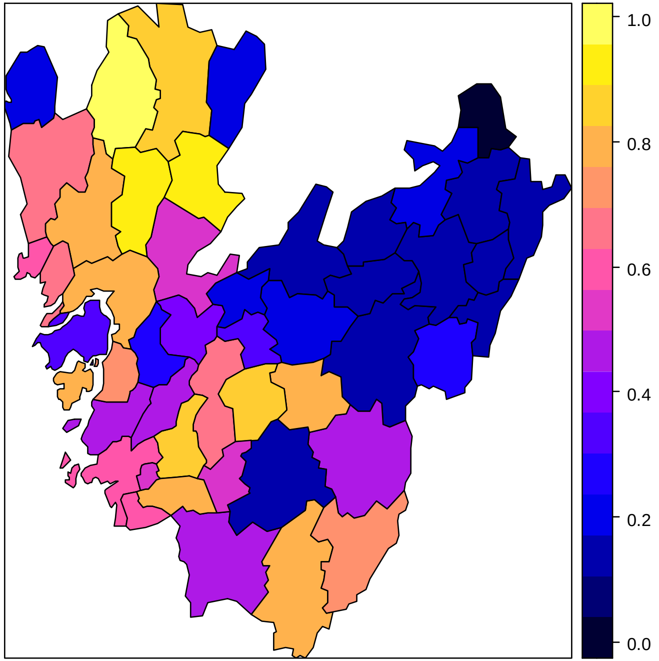
Motivating example

Swedish Covid-19 Investigation for Future Insights – a Population Epidemiology Approach using Register Linkage (SCIFI-PEARL)

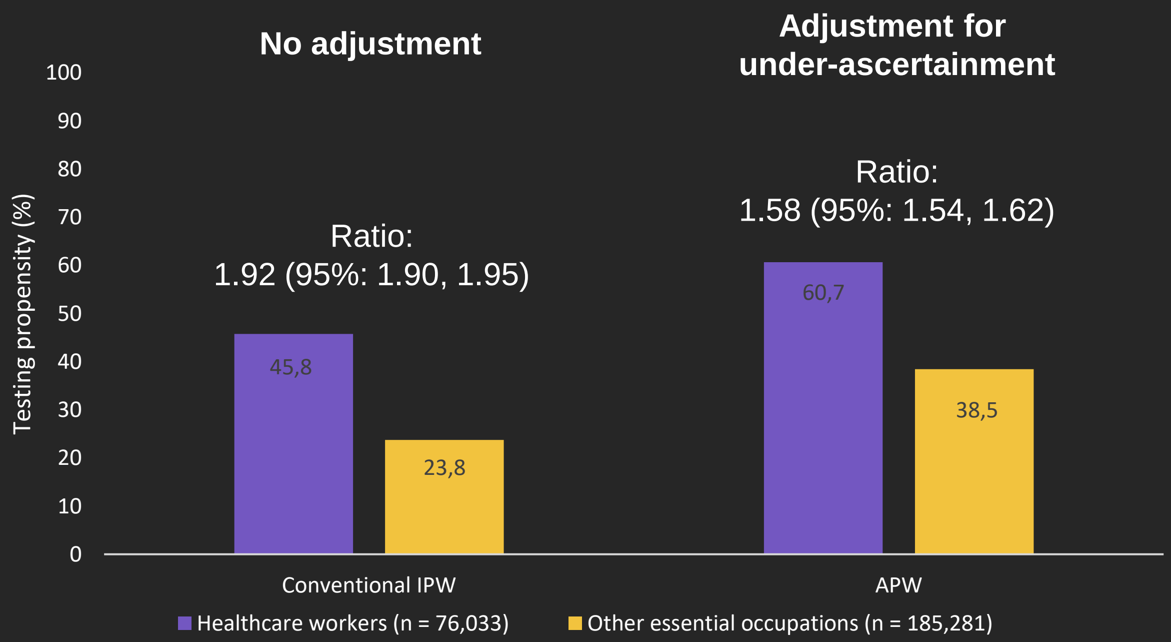
Forskningsprojekt

- **SCIFI-PEARL:** Register-based study on COVID-19
  - Contains data from a vast number of registers
  - Total population data
- Linked to PCR tests in Västra Götaland from database at Sahlgrenska University Hospital
  - Hospital tests from Sahlgrenska + two adjacent hospitals
  - Most (all?) tests performed at public outpatient providers
  - Missing many private healthcare providers + a few hospitals
  - Hence, data is incomplete from a population perspective

Figure: Estimated data coverage, fall 2020

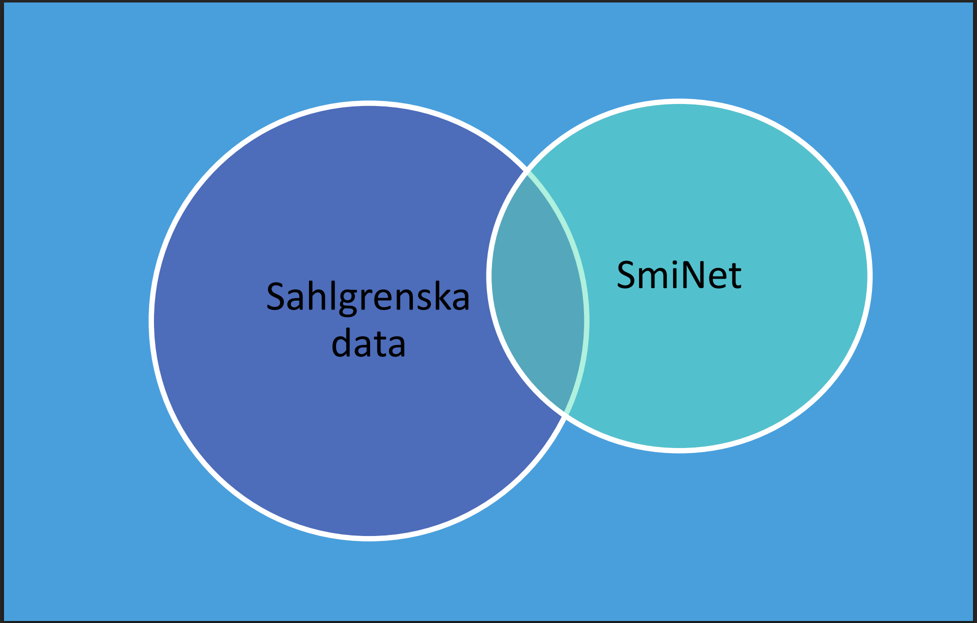


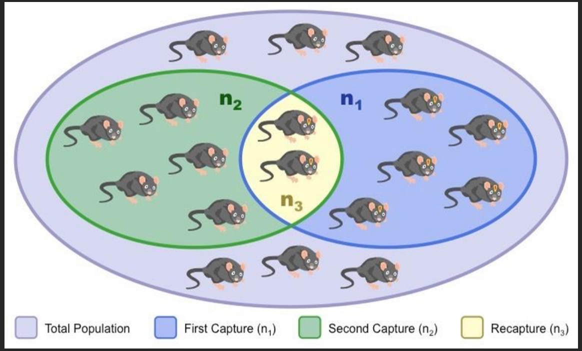
Preview of the empirical example



How to correct for the bias?

All PCR tests





Capture-recapture setup

- Two imperfect data sources
- Partial overlap

Related work

- **Outcome misclassification estimators** (e.g., Gravel & Platt, 2018)
  - Validation data (i.e., reference data where outcome is perfectly observed).
  - No obvious validation data in our setting.

- **Capture-recapture in epidemiology** (e.g., Chao et al., 2001)
  - Estimate the prevalence or incidence of “hidden” public health issues.
  - Not developed for effect estimation.

Gravel CA, Platt RW. Weighted estimation for confounded binary outcomes subject to misclassification. *Stat Med*. 2018;37(3):425–36.

Chao A, Tsay PK, Lin SH, Shau WY, Chao DY. The applications of capture-recapture models to epidemiological data. *Stat Med*. 2001;20(20):3123–57.

Road Traffic Deaths and Injuries Are Under-Reported in Ethiopia: A Capture-Recapture Method

Teferi Abegaz[1]*, Yemane Berhane[2], Alemayehu Worku[3], Abebe Assrat[4], Abebayehu Assefa[5]

1 School of Public and Environmental Health, College of Medicine and Health Sciences, Hawassa University, Hawassa, Ethiopia, **2** Addis Continental Institute of Public Health, Addis Ababa, Ethiopia, **3** School of Public Health, College of Health Sciences, Addis Ababa University, Addis Ababa, Ethiopia, **4** Ministry of Transport, Addis Ababa, Ethiopia, **5** World Health Organization (WHO) Ethiopia, Addis Ababa, Ethiopia

Abstract

In low and middle income countries road traffic injuries are commonly under-reported. This problem is significantly higher among those less severely injured road users. The objective of this study was to determine the incidence and the level of ascertainment of road traffic injuries and deaths by traffic police and hospital registry. In this study two-sample capture-recapture method was applied using data from traffic police and hospital injury surveillance, through June 2012 to May

AIDS Behav (2018) 22:2248–2257
https://doi.org/10.1007/s10461-017-1883-6

CrossMark

ORIGINAL PAPER

Evaluating the Completeness of HIV Surveillance Using Capture–Recapture Models, Alameda County, California

Paul Wesson[1,3] · Richard Lechtenberg[2] · Arthur Reingold[1] · Willi McFarland[3,4] · Neena Murgai[2]

‘Are you on the market?’: a capture–recapture enumeration of men who sell sex to men in and around Mombasa, Kenya

Scott Geibel[a], Elisabeth M. van der Elst[b], Nzioki King’ola[c], Stanley Luchters[c], Alun Davies[b], Esther M. Getambu[d], Norbert Peshu[b], Susan M. Graham[e], R. Scott McClelland[e] and Eduard J. Sanders[b,f]

AJPH OPEN-THEMED RESEARCH

Estimated Prevalence of Opioid Use Disorder in Massachusetts, 2011–2015: A Capture–Recapture Analysis

Joshua A. Barocas, MD, Laura F. White, PhD, Jianing Wang, MSc, Alexander Y. Walley, MD, MSc, Marc R. LaRochelle, MD, Dana Bernson, MPH, Thomas Land, PhD, Jake R. Morgan, PhD, Jeffrey H. Samet, MD, MA, MPH, and Benjamin P. Linas, MD, MPH

Ascertainment Probability Weighting (APW)

- We propose a method that combines capture-recapture with propensity score weighting
  - Simultaneous adjustment for observed confounding and under-ascertainment bias
  - Uses two imperfect conditionally independent sources of the outcome variable to account for under-ascertainment

Framework: key ingredients

Potential outcomes

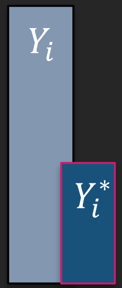
$Y_i(X = 1)$ $Y_i(X = 0)$

$Y_i(x)$, if exposure/treatment is set to *x*.

Assume standard conditions

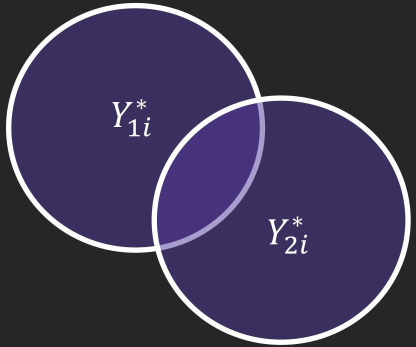
- conditional exchangeability
- exposure positivity
- counterfactual consistency

Under-ascertainment



- $Y_i^* = 1$ implies $Y_i = 1$
- $Y_i^* = 0$ does not imply $Y_i = 0$
- No false positives permitted

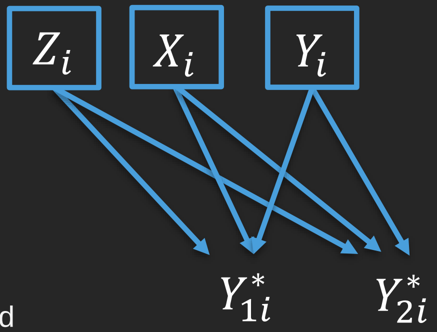
Data setup



- $Y_{ji}^* = 1$; ascertainment in source *j*
- $Y_i^* = 1$ if $Y_{1i}^* = 1$ OR $Y_{2i}^* = 1$
- Must be overlap!

- X is exposure/treatment of interest
- Z are other variables.

Capture-recapture: key assumption



- **Conditional source independence**
  - No correlation between ascertainment in source 1 or 2, conditional on Y=1 and observed characteristics (X and Z).

Example:

- *SmiNet:* Factors that influence probability that a test is positive. Could be things that act as ”barriers” to get tested even with symtoms (e.g., occupation access, education, health behaviors).
- *Sahlgrenska:* Where you live (proximity to included test facilities), choice of healthcare provider (public/private), etc. Tied to socioeconomic factors?
- We need to account for selection mechanisms that are common to both. Likely: occupation, sociodemographics, geography.

Estimation

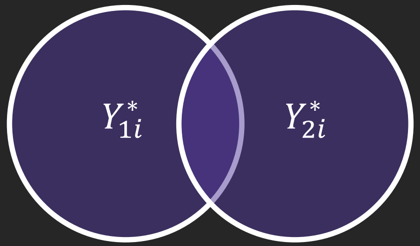
1. Estimate propensity score for exposure level *x*, controlling for Z.
2. Estimate ascertainment probabilities using three models fitted among Y*=1, controlling for X and Z (Eq. 2).
3. Use APW estimator to estimate potential outcome probability for Y(x), e.g., X=1 (Eq. 1).
4. Repeat for group x', e.g., X=0.
5. Compute desired effect estimate (e.g., risk diff., ratio).
6. Bootstrap entire procedure for CIs.

APW estimator (1)

$$\widehat{\Pr}(Y_i(x) = 1) = \frac{1}{N}\sum_{i=1}^{N}\frac{I(Y_i^* = 1, X_i = x)}{\hat{a}(x_i, z_i)\,\widehat{\Pr}(x|z_i)}$$

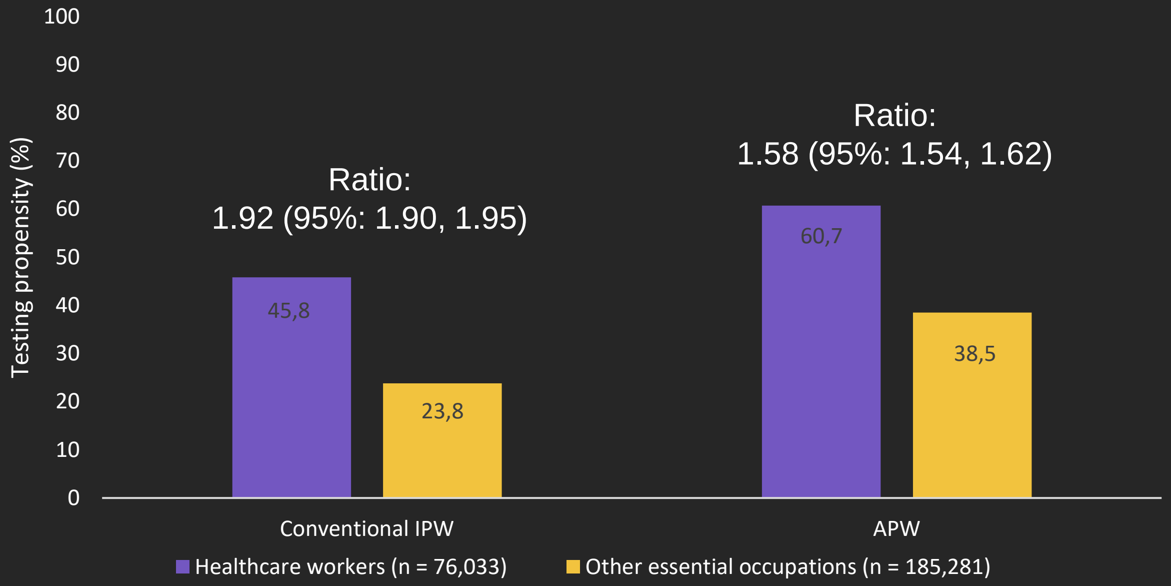
Ascertainment/capture-recapture (2)

$$\hat{a}(x_i, z_i) = \frac{\widehat{\Pr}(Y_{1i}^* = 1, Y_{2i}^* = 1 | Y_i^* = 1, x_i, z_i)}{\widehat{\Pr}(Y_{1i}^* = 1 \middle| Y_i^* = 1, x_i, z_i)\,\widehat{\Pr}(Y_{2i}^* = 1 \middle| Y_i^* = 1, x_i, z_i)}$$



Healthcare workers vs. other essential occupations in VGR

2020-07-01 to 2020-12-26: *PCR-tested at least once during period*



Adjustment for:

- Age
- Sex
- Occupation
- Income
- Country of birth groups (World Bank classification)
- Education
- Marital status
- Children living at home
- Household type
- Charlson comorbidity index (pre-pandemic)
- Municipality fixed effects (49)

Other occupations: teaching, social care, service sector, postal/delivery, transport services, police/security, cleaning staff

Concluding remarks

- **Key benefits:**
  - Extends outcome misclassification adjustment to settings without access to proper validation data
  - Could be useful tool for register research on outcomes that are poorly ascertained

- **Limitations:**
  - Assumes no false positives
  - Requires at least two data sources
  - May be difficult to be fully convinced that source independence holds