Estimating the force of infection and incidence using routine data affected by outcome dependent sampling

Sereina A. Herzog [1], Steven Abrams [2,3], Niel Hens [1,2,3]

1. Centre for Health Economics Research and Modelling Infectious Diseases (CHERMID), Vaccine and Infectious Disease Institute, University of Antwerp, Belgium; 2. Global Health Institute, Department of Epidemiology and Social Medicine, University of Antwerp, Belgium; 3. Center for Statistics, I-BioStat, Hasselt University, Belgium

contact: sereina.herzog@uantwerpen.be

Introduction

In routine data, we are often faced with outcome dependent sampling (ODS), i.e. the probability to have a further observation from the same person depends on the outcome of the current one.

For example, Austrian prenatal care includes screening for the early detection and treatment of toxoplasmosis infection:

- first test negative → up to two further tests
- first test positive → no further tests

The ODS in this example leads to an overrepresentation of negative test results which needs to be accounted for when estimating the force of infection (FOI) from (serial) cross-sectional data.

Objective

To determine an estimator for the force of infection (FOI) - for a Susceptible-Infected (SI) course of infection - using routine data and thereby dealing with outcome dependent sampling (ODS).

Methods

- We derived our proposed estimator theoretically while relying on conditional independence assumptions, i.e. the probability to have a new observation depends only on the outcome (negative or positive test) of the previous observation.
- We compare four FOI (λ) estimators (**Table 1**) differing in considering data structure and ODS. With the estimator '*our*', we estimate also the probabilities to have a 2^{nd} observation given the outcome of the first observation is negative or positive ($\nu_{1|0}$, $\nu_{1|1}$). In contrast, the estimator '*ws*' assumes that the probability to have a 2^{nd} observation is independent of the current outcome ($\nu_{1|\cdot}$).

Table 1: Investigated estimators.

Abbr	Description	Estimators
our	Considers ODS & data structure	$\hat{\lambda}^{our}$, $\hat{\nu}^{our}_{1\|0}$, $\hat{\nu}^{our}_{1\|1}$
ws	Considers data structure but no ODS	$\hat{\lambda}^{ws}$, $\hat{\nu}^{ws}_{1\|\cdot}$
ns	Treats all observations as being independent (no-structure)	$\hat{\lambda}^{ns}$
y_1	Uses only the first observation from each person	$\hat{\lambda}^{y_1}$

Abbr=abbreviation; λ=FOI; $\nu_{1|0}$ = Pr[observe 2^{nd} observation|1^{st} observation negative]; $\nu_{1|1}$ = Pr[observe 2^{nd} observation|1^{st} observation positive]; $\nu_{1|\cdot}$ = Pr[observe 2^{nd} observation]; Pr=probability

- We simulated 5000 datasets for each of the 144 simulation settings (**Table 2**) assuming that all individuals have their 1^{st} and potential 2^{nd} observation at the same time points, i.e. t_1, t_2.

Table 2: Simulation settings.

Parameter	Description	Values
N	Number of individuals	1000
λ	FOI per time unit translates in Pr[seroconversion]	0.01, 0.051, 0.223, 0.511 1%, 5%, 20%, 40%
$\nu_{1\|0}$	Pr[observe 2^{nd} obs.$\|1^{st}$ obs. negative]	0.01, 0.02, 0.04, 0.06, 0.08, 0.09
$\nu_{1\|1}$	Pr[observe 2^{nd} obs.$\|1^{st}$ obs. positive]	0.01, 0.02, 0.04, 0.06, 0.08, 0.09

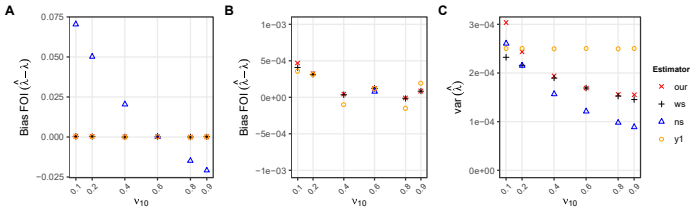
Pr=probability; obs.=observation; seroconversion=1^{st} obs. negative and 2^{nd} obs. positive

- We also analysed a serial cross-sectional serological dataset on toxoplasmosis infection in pregnant women. This dataset contains toxoplasmosis screening results for 21,500 women during their pregnancy (Styria, Austria).[1]
- Maximum likelihood estimates were derived and estimators were compared by the mean squared error (MSE) decomposition.

Results

- **Figure 1A** illustrates that the '*ns*' estimator over- or underestimates the FOI unless there is no ODS (i.e. $\nu_{10} = \nu_{11}$).
- The other estimators ('*our*', '*ws*', 'y_1') have a similar magnitude of bias (**Figure 1B**).
- In general, 'y_1' has the highest variance (**Figure 1C**) as a consequence of information contained in the observations at t_2.

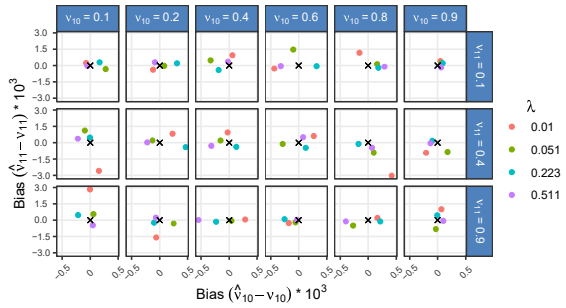
Figure 1: Bias and variance for FOI estimators ($\lambda = 0.223$, $\nu_{11} = 0.6$).



- Usually the MSE(λ) is a bit higher in '*our*' than in '*ws*' (not shown) due to the slightly higher variance in '*our*' (**Figure 1C**). However, our proposed estimator provides additional estimates for the ODS probabilities (**Figure 2**).

Figure 2: Bias for ν_{10} and ν_{11} with estimator 'our'.

The '×' marks where bias for ν_{10} and ν_{11} would be zero.



- We estimated for **toxoplasmosis** $\hat{\lambda}^{our} = 0.01154$ $[0.000157^2]$, this implies an estimated prevalence of 20.6% in 20-year-old women. We observed similar estimates for $\hat{\lambda}^{ws}$ and $\hat{\lambda}^{y1}$ but $\hat{\lambda}^{ns} = 0.00938$ was lower. We estimated $\hat{\nu}^{our}_{10} = 0.765$ $[0.00287^2]$ and $\hat{\nu}^{our}_{11} = 0.378$ $[0.00473^2]$, despite the fact that guidelines imply that those values should have been equal to 1 and 0, respectively.

Conclusion

- Our proposed estimator for the FOI performs well in terms of bias and MSE.
- Our approach yields additional estimates with regard to the ODS probabilities
- Not considering the structure in routine data leads to bias in both directions, either under- or overestimating the FOI.

Acknowledgments

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement 682540 – TransMID).

References

[1] Berghold C, Herzog SA, Jakse H, Berghold A. Euro Surveill. 2016; 21(33):pii=30317.

Interuniversity Institute for Biostatistics and statistical Bioinformatics

Health Economics & Modelling Infectious Diseases
Vaccine & Infectious Disease Institute
University of Antwerp