Informed presence in electronic health record data: bias and bias reduction approaches in longitudinal analyses

Daniel T. Vader,[1] Di Shu,[1,2] Rebecca A. Hubbard,[1] Craig Boge,[2] Anna Sharova,[2] Kevin Downes,[2,3] Yun Li[1,2,3]

[1]Department of Biostatistics, Epidemiology, & Informatics, University of Pennsylvania; [2]Pediatric IDEAS Research Group of the Center for Pediatric Clinical Effectiveness, Children's Hospital of Philadelphia; [3] Department of Pediatrics; Perelman School of Medicine of the University of Pennsylvania

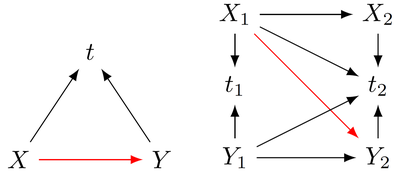
Electronic health record (EHR) data capture is not random

- Patients contribute more data to EHRs when they are sick than when they are healthy.
- These tendencies may cause **informed presence**: systematic differences between captured data and non-captured data.
- Informed presence can affect study results, biasing estimates toward or away from the null.

Objectives

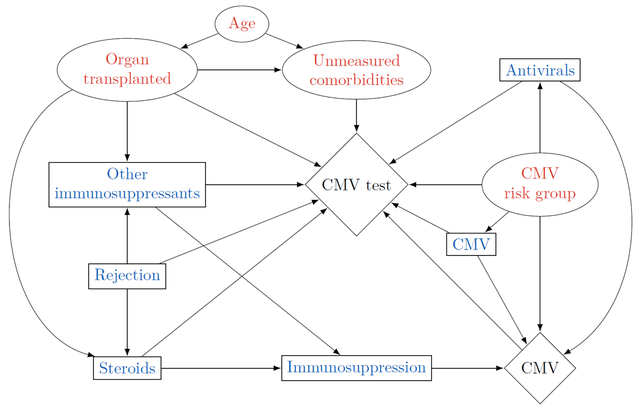
- Focusing on population-level effects in longitudinal analyses, we present:
  - A collider bias framework for understanding informed presence bias
  - Novel graphical approaches for describing irregular measurement
  - Comparison of four alternative approaches to adjusting for informed presence bias
- We illustrate these tools using a recurrent events investigation into the effects of steroid treatment on cytomegalovirus infection.
  - The analysis features real world data from the Children's Hospital of Philadelphia on 271 pediatric solid organ transplant patients.

Informed presence bias as collider bias in longitudinal settings

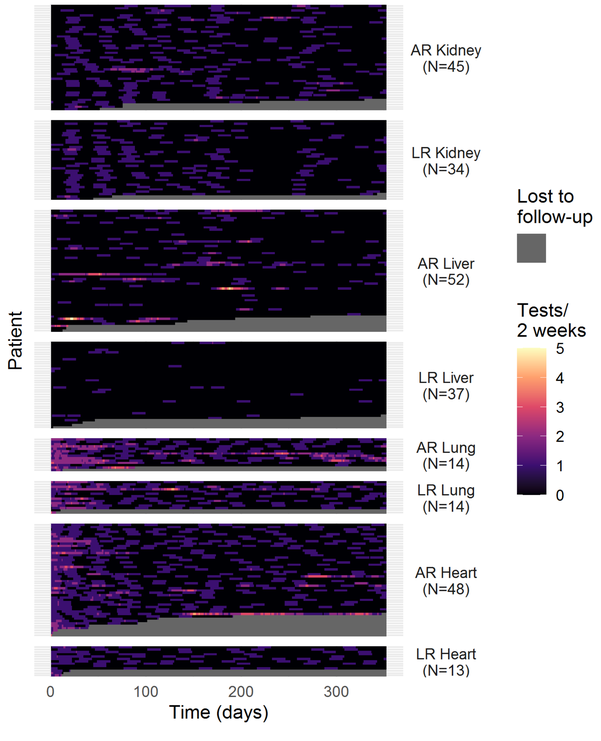


Measurement time (t) is a collider on the path between exposure (X) and outcome (Y) in both cross-sectional (left) and longitudinal (right) settings.

Conceptualizing determinants of cytomegalovirus (CMV) testing

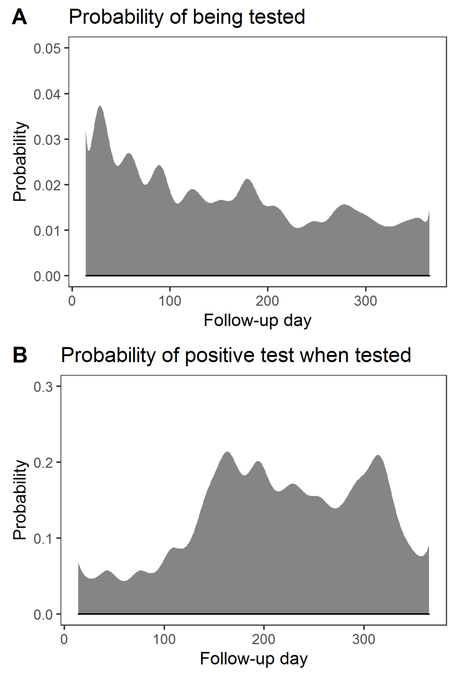


Vizualizing patient-specific testing rates by transplant group



AR, at risk of CMV (organ donor or recipient were CMV positive prior to transplant); LR, low risk of CMV (donor and recipient were CMV negative)

Vizualizing dependency between outcome and testing



Adjusting for informed presence

Model	IRR	(95% CI)
Naïve	1.83	(1.02, 3.28)
TSLO	1.66	(1.03, 2.68)
IIW	1.37	(0.73, 2.57)
BS IIW	1.37	(0.71, 2.27)
MO	1.40	(0.73, 2.68)

IRR, incident rate ratio; CI, confidence interval

Naïve: Generalized estimating equations (GEE) with no adjustment for informed presence.
TSLO: GEE with ad hoc adjustment using time since last observation (TSLO).
IIW: GEE with adjustment using inverse intensity weighting (IIW).
BS IIW: IIW with bootstrapped confidence intervals.
MO: GEE with adjustment using multiple outputation (MO).

Interpretation

- Conceptual diagram that mapped determinants of testing demonstrated that the analysis was at risk of informed presence bias.
- Novel visualization approaches showed dependency between organ type, risk group, outcome status, and testing.
- Adjusting for informed presence attenuated the estimated relationship between steroids and CMV without increasing standard errors.
- Estimate from ad hoc time since last observation (TSLO) approach was inconsistent with other adjusted estimates, suggesting poor bias reduction performance.

Conclusion

- When conducting repeated outcomes analyses with irregularly measured EHR data we recommend:
  1. Using conceptual diagrams of the observation process to assess whether conditioning on observation is likely to induce collider bias
  2. Visualizing dependence between outcome and observation process in the data
  3. If appropriate, accounting for outcome dependence in statistical analysis
- Informed presence is a common but understudied concern in EHR-based analyses. Investigators should consider and address informed presence just as carefully as they consider and address missing data.

Contact Information

Web: CHOP Pediatric IDEAS Research Group
Email: vaderd@pennmedicine.upenn.edu

CCEB | Perelman School of Medicine, University of Pennsylvania

Children's Hospital of Philadelphia