

Which Data Reached the Inference: Selection, Missingness, Survivorship and Collider Bias

Juan Segurola

3Dresyns by Resyner Technologies S.L.

© 2026 Juan Segurola. All Rights Reserved.

Correspondence: info@3Dresyns.com · www.3Dresyns.com

INF-003 · Scientific Inference and Experimental Reasoning · Version v0.2

Abstract

Inference is made from the observations that reach the analysed dataset, not from the observations that might have existed under an ideal sampling and measurement process. Entry into the dataset may depend on participation, survival, detection, follow-up, specimen integrity, quality-control rules, complete-case requirements or analyst decisions. When selection depends on variables related to exposure, outcome or their causes, the analysed set can differ systematically from the target population. Conditioning on a collider is one structural mechanism that can generate such distortion, but selection bias is broader than collider terminology alone. Missing data add a related problem: whether complete-case analysis, weighting, imputation or sensitivity analysis is valid depends on the missingness process and the target estimand. This article separates selection, missingness, survivorship and collider-induced selection, and provides a disciplined route from the target population to the analysed sample. It does not treat non-detection and censoring as interchangeable with missingness; those diagnostic measurement problems remain with HEB-007.

Keywords: selection bias; missing data; survivorship bias; collider bias; complete-case analysis; participation; follow-up; target population

1. The analysed dataset is a selected state

Every dataset represents a sequence of inclusion events. Some units are never sampled; some decline participation; some are lost to follow-up; some specimens fail quality control; some outcomes are unmeasured; some observations are removed because required covariates are missing. The scientific question is therefore not only what the observed values show, but why these particular observations became available for analysis.

2. Selection is a causal process

A structural approach treats selection as part of the causal system rather than as a generic property of sample size. Hernán, Hernández-Díaz and Robins showed that selection bias can be represented by conditioning on common effects or other structures that create non-causal associations [1]. Cole and colleagues illustrated how conditioning on a collider can induce an association between otherwise independent causes [2]. The practical consequence is that selection cannot be diagnosed solely from the marginal similarity or dissimilarity of the retained sample.

3. Missingness is not one mechanism

Missing data can arise through different mechanisms and require different assumptions for valid inference. Little and Rubin distinguish settings in which missingness can be ignored under the analysis model from settings in which the missingness process must be modelled or explored through sensitivity analysis [3]. A complete-case analysis can be unbiased in some structures and severely biased in others. Imputation can recover information under appropriate models but cannot manufacture information about an unobserved mechanism without assumptions.

4. Survivorship and conditioning on post-exposure states

Survivorship bias occurs when inclusion requires remaining observable after a process that is itself related to exposure, outcome or common causes. The term is often used loosely; INF uses it only when survival or persistence is a selection event relevant to the target inference. Conditioning on a post-exposure state can also induce selection bias even if no literal biological survival is involved. Examples include analysing only devices that passed a screening step, only patients who remained in follow-up, or only experiments that produced a measurable signal.

5. Selection on a collider

INF-002 defines a collider structurally as a common effect on a specified path. INF-003 owns the inferential consequence when dataset inclusion conditions on such a variable or its descendant. The resulting association is not evidence of a causal relation merely because it appears after stratification or selection. Collider status is graph-relative, so the same variable name cannot be labelled “a collider” independently of the causal structure [5].

6. Distinguish missingness from non-detection and censoring

A missing value may indicate that a measurement was not obtained, whereas a non-detection can be an observed result from a method whose response fell below a capability threshold, and censoring can encode partial information about a value. These states carry different information. HEB-007 owns the diagnostic treatment of non-detection and censoring,

including how to tell them apart from a value that was not measured. INF-003 uses those distinctions only to prevent missing-data procedures from being applied to the wrong observation state.

7. Selection diagrams in words

- Define the target population or complete set of eligible units.
- List each event required for a unit to enter the analysed dataset.
- For every selection event, identify variables that influence it.
- Ask whether conditioning on that event opens a non-causal path or changes the target population.
- Separate pre-exposure sampling from post-exposure selection.
- State which variables are missing and what assumptions make the chosen handling method valid.
- Use sensitivity analysis when plausible alternative selection or missingness mechanisms cannot be ruled out.

8. What weighting and imputation can and cannot do

Inverse-probability weighting, likelihood methods and multiple imputation can reduce bias or recover efficiency under specified models and positivity/support conditions. They do not make selection ignorable by declaration. The analysis remains conditional on the adequacy of the model for participation, missingness or outcome [4]. When unobserved determinants of selection materially affect the target relationship, sensitivity analysis may be more scientifically honest than a single corrected estimate.

9. Reporting the selected state

A defensible report should state the target population, eligible set, recruited or generated units, exclusions, missingness by variable, reasons for exclusion where known, timing of selection relative to exposure and outcome, and the assumptions used by weighting or imputation. This information is part of the evidence chain because the target of inference can change as units are lost.

References

- [1] Hernán, M.A.; Hernández-Díaz, S.; Robins, J.M. A Structural Approach to Selection Bias. *Epidemiology* 15(5), 615-625 (2004). DOI: 10.1097/01.ede.0000135174.63482.43.
- [2] Cole, S.R.; Platt, R.W.; Schisterman, E.F.; Chu, H.; Westreich, D.; Richardson, D.; Poole, C. Illustrating Bias Due to Conditioning on a Collider. *International Journal of Epidemiology* 39(2), 417-420 (2010). DOI: 10.1093/ije/dyp334.
- [3] Little, R.J.A.; Rubin, D.B. *Statistical Analysis with Missing Data*. 3rd ed. Wiley, 2019. DOI: 10.1002/9781119482260.
- [4] Hernán, M.A.; Robins, J.M. *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC, 2020. DOI: 10.1201/9781315374932.
- [5] Pearl, J. *Causality: Models, Reasoning, and Inference*. 2nd ed. Cambridge University Press, 2009. DOI: 10.1017/CBO9780511803161.

Data and materials statement

No new experimental dataset was generated for this article. The article is a methodological analysis based on the cited literature and the frozen ownership architecture of the INF series.