Protocolo reproducible de gráficos acíclicos dirigidos por capas para estudios con datos secundarios

A reproducible layered directed acyclic graph protocol for secondary-data studies

DOI: https://dx.doi.org/10.17981/ingecuc.22.1.2026.02
Scientific Research Article. Date Received: 11/02/2026, Date Accepted: 01/03/2026

Diego Rivera-Porras https://orcid.org/0000-0003-2169-3208
Universidad de la Costa. Barranquilla, (Colombia)
drivera23@cuc.edu.co

Yulineth Gómez-Charris https://orcid.org/0000-0003-3630-3276
Universidad de la Costa. Barranquilla, (Colombia)
Universitat Politècnica de València
ygomez6@cuc.edu.co; yugocha@upv.edu.es

Valmore Bermúdez https://orcid.org/0000-0003-1880-8887
Universidad Simón Bolívar. Barranquilla, (Colombia)
valmore.bermudez@unisimon.edu.co

To cite this paper

D. Rivera-Porras, Y. Gómez-Charris & V. Bermúdez "A reproducible layered directed acyclic graph protocol for secondary-data studies," INGE CUC, vol. 22, no. 1, 2026. DOI: *https://dx.doi.org/10.17981/ingecuc.22.1.2026.02*

Resumen

Introducción/Contexto: Los estudios con datos secundarios (registros clínicos, administrativos o de sensores) incorporan procesos de selección, medición y datos faltantes que a menudo se tratan como "preprocesamiento", dejando supuestos implícitos y sesgos difíciles de auditar.
Objetivo: Proponer un protocolo reproducible para explicitar, documentar y auditar supuestos causales en estudios con datos secundarios mediante gráficos acíclicos dirigidos por capas.
Metodología: El protocolo separa un sistema causal material (exposición→resultado) y tres capas adicionales: (i) selección/observabilidad, (ii) medición u operacionalización (constructos vs. proxies registrados) y (iii) mecanismos de datos faltantes. Incluye alineación temporal del estimando, reglas de extracción y construcción de cohortes, y control de calidad basado en código fuente del gráfico y trazabilidad de figuras.
Resultados: Se entregan una lista de verificación de reporte (Tabla 1), un mapeo decisión→amenaza→firma en el gráfico con mitigaciones condicionales (Tabla 2), una síntesis de trabajo relacionado y brecha (Tabla 3), ejemplos de gráficos por capas (Figuras 1–9) y evaluaciones reproducibles que cuantifican sesgos típicos y sensibilidad a faltantes (Tablas 4–7).
Conclusiones: El enfoque por capas hace explícitos los supuestos de selección, medición y faltantes; aumenta la replicabilidad al exigir trazabilidad entre código, figura y texto; y operacionaliza la alineación entre decisiones de diseño/análisis y el estimando, limitando inferencias que exceden la evidencia. Un caso aplicado mínimo ilustra que decisiones sobre observación/faltantes pueden cambiar de forma material la magnitud del estimando, por lo que debe reportarse sensibilidad (Δestimación) y denominadores.

Palabras clave

Análisis causal; Análisis de datos; Procesamiento de datos; Recopilación de datos; Estadística; Visualización de datos.

Abstract

Introduction: Secondary-data studies (clinical, administrative or sensor records) embed selection, measurement and missing-data processes that are often treated as "preprocessing", leaving key causal assumptions implicit and hard to audit.
Objective: To propose a reproducible protocol that makes causal assumptions explicit in secondary-data studies using layered directed acyclic graphs.
Methodology: The protocol separates a material causal system (exposure→outcome) and three additional layers: (i) selection/observability, (ii) operationalisation and measurement (constructs vs. recorded proxies) and (iii) missing-data mechanisms. It includes estimand and time alignment, extraction rules and cohort construction, plus quality control based on graph source code and figure traceability.
Results: Outputs include a reporting checklist (Table 1), a decision→threat→graph-signature mapping with conditional mitigations (Table 2), a related-work gap synthesis (Table 3), layered DAG examples (Figs. 1–9) and reproducible evaluations quantifying typical biases and sensitivity to missingness (Tables 4–7).
Conclusions: The layered approach makes selection, measurement and missingness assumptions explicit; improves reproducibility through code–figure–text traceability; and operationalises alignment between design/analysis choices and the estimand, limiting claims that exceed the available evidence. A minimal applied case shows that observation/missingness decisions can materially shift the estimand, motivating explicit sensitivity reporting (Δ) and denominators.

Keywords

Causal analysis; Data analysis; Data processing; Data collection; Statistics; Data visualization.

I. INTRODUCTION

Directed acyclic graphs have become established as a formal language for expressing causal assumptions and deriving identification conditions in observational studies [1], [2]. They facilitate reasoning about confounding, mediation, and colliders and, consequently, guide adjustment and design decisions [3], [4].
In studies with secondary data (clinical, administrative, transactional, or sensor records), the analytical sample and observed variables result from processes additional to the material causal system: selection/observability (here: inclusion/recording in the data source, not "observability" in control theory; e.g., entering the system, being measured, being linked), operationalization of constructs in recorded proxies, and missing data mechanisms. Treating these layers as 'cleaning' or 'pre-processing' is equivalent to imposing undeclared assumptions and can induce selection biases, temporal misalignment (e.g., immortal time), biases associated with proxies dependent on the recording process, as well as truncation due to death in survival studies [5], [6], [7], [8], [9], [10].
Although there are general standards for reporting observational studies, if the causal graph is limited to the material system (exposure→outcome) and does not separately represent (i) the selection/observability process, (ii) the generation of measured proxies, and (iii) the mechanisms of missing data, the relevant assumptions remain implicit and the diagram loses value as an auditable artefact. Furthermore, when the diagram is treated as an image (without versioned specification), small structural changes (e.g., adding/removing an arrow) may go unnoticed and alter open/closed paths and adjustment sets.
Research question: How can the assumptions of selection, measurement, and missing data that condition inference in studies with secondary data be made explicit and audited in a reproducible manner?

Related work and differential contribution

The literature on directed acyclic graphs (DAGs) often emphasizes the material causal system and the derivation of adjustment sets to block non-causal paths [1], [2], [11]. In secondary data, there are specific developments for (i) selection bias when conditioning on inclusion/observability mechanisms (S) [5], [8], (ii) missing data mechanisms and recoverability using R indicators [6], [9], and (iii) temporal alignment to avoid self-induced errors (e.g., immortal time) using 'objective trial' logic [7]. These pieces are complementary but are typically reported in a fragmented manner: selection (S), operationalization/proxies, and missing data (R) are treated as 'pre-processing' rather than as auditable structural assumptions. The differential contribution of this article is to integrate them as a layered DAG with a reproducible contract (code → figure + quality control), quantitative reporting rules (denominators, windows, and Δestimation between specifications), and artefacts ready for peer review.
This article presents a reproducible protocol for layered graphics for studies with secondary data in engineering domains (Internet of Things (IoT), manufacturing, embedded systems, digital platforms) and health. The protocol requires (i) defining the estimator and its temporal alignment (objective trial logic when applicable) [7], [12], [13]; (ii) separating the material causal system from the selection, measurement, and missing mechanisms; and (iii) materialising the diagram as a verifiable artefact (source code + quality control (QC) manifesto + automatic validation). A reporting checklist (Table 1), a decision–threat–mitigation mapping with conditional rules (Table 2), a synthesis of related work

and gaps (Table 3), nine example diagrams (Figures 1–9), three illustrative simulations (Tables 4–6), and a minimal applied case in plant data (Table 7) are reported.

II. METHODOLOGY

The protocol is geared towards studies with secondary data where (a) observability depends on contact, measurement or linkage processes, (b) analytical variables are recorded proxies of underlying constructions, and/or (c) there is relevant missing data. It does not propose a new estimator: its objective is to make assumptions explicit and improve the auditability of the design and analysis.
Minimum outputs of the protocol: (i) a material graph for the effect of interest, (ii) additional layers for selection/observability, measurement, and missing data, (iii) a set of reportable decisions (Table 1), (iv) predefined alternative specifications with comparison of estimates, and (v) traceability between graph code, figures, and text.

Step-by-step protocol

Step 1: Define the estimator and time alignment. Specify the target population, exposure contrast, outcome, baseline, and follow-up period. State whether the estimator is intervention-analogous or associational. Document retrospective and evaluation windows (in days/months) and any use of post-baseline information.
Step 2: Document data source and cohort construction. Describe source and coverage, extraction rules, and inclusion/exclusion criteria with their timing. Report the flow from raw records to the analytical sample (counts by stage).
Step 3: Construct the material graph (Layer 1). Define constructs and causal order for the effect of interest (exposure→outcome) and locate candidates for confounding, mediation, and effect modification. Justify assumptions with domain knowledge and literature.
Step 4: Incorporate selection/observability (Layer 2). Add selection nodes for mechanisms such as 'being in care,' 'being measured,' 'being linked,' or 'being monitored.' Specify causes of selection and explicitly state whether the analysis conditions selection (observable sample) and what target population it involves.
Step 5: Separate constructs and proxies (Layer 3). When exposure/outcome/covariates are recorded proxies, distinguish between construct (underlying) and observed variable, and include determinants of the recording process (e.g., utilisation, coding, intensity of observation).
Step 6: Model missing data (Layer 4). Introduce observation indicators (R) for key incomplete variables and specify their causes. State the assumed mechanism in causal terms and report denominators (total n vs. observed n) per analysis [6], [9].
Step 7: Derive identification and modelling strategy. Based on the complete graph, justify the adjustment set and/or strategy (stratification, weighting, g-methods, imputation, observation process models). Avoid adjusting for colliders or mediators unless the estimator requires it. Explicitly state assumptions of positivity and common support when using weighting [14], and justify the treatment of baseline covariates (adjustment for baseline vs. change) when applicable [15].
Step 8: Predefine alternative specifications and quality control. Define alternative analyses (windows, definitions, exclusion rules) and report changes in the estimate compared to the base analysis. Provide the graph code and a quality control (QC) manifest (counts + hashes) and perform automatic verification (integrity, counts, and acyclicity) to enable reproducible auditing.

Formal conventions and artefact agreement (layered DAG)

To ensure that the DAG is not just an image but an auditable artefact, a minimum contract is adopted: (1) each figure has a stable ID (Figure x) and a source file with the same name (Figx.dot); (2) the graph must be directed and acyclic; (3) semantics are reinforced with naming conventions (e.g., _true/_obs (or */_obs) for construct vs. observed, R_ for observation indicators, and S for selection/observability); (4) layer↔colour and layer↔shape mapping is kept fixed and declared in the manuscript (for greyscale interpretation); (5) node/arc counts are reported in the manuscript and SHA-256 hashes of .dot and .jpg are recorded in the quality control (QC) manifest of the supplementary material; (6) a script automatically verifies integrity, counts, and acyclicity against the quality control (QC) manifest. This contract allows traceability without requiring access to microdata. Quality control (QC) verifies traceability and integrity of the artefact (code↔figure), but does not itself validate causal validity, identification of the estimator, or absence of unmeasured confounding.

Reproducibility and supplementary files

For each figure, the source code for the graph (DOT/Graphviz format) and raster and vector exports (JPG 300 dpi, PDF, and SVG) are included as supplementary material. A quality control manifest is provided with node and arc counts and SHA-256 hashes to unambiguously link code and figure. The supplementary material also includes: (i) verification scripts (verify_qc_manifest.py and verify_manuscript_figures.py) to check integrity/structure and manuscript↔figure matching against the QC manifest; (ii) reproducible simulation code and tabular outputs associated with Tables 4–6; (iii) a minimal reproducible applied case (Table 7) based on the 'stackloss' dataset (Brownlee [16]) loaded from statsmodels, with an induced missing mechanism to illustrate Layer 4; and (iv) a one-command execution script (reproduce.sh) that regenerates figures, verifies QC, and reproduces Tables 4–7.

III. RESULTS AND DISCUSSION

The results are presented as methodological artefacts: a checklist (Table 1), a mapping of frequent decisions in secondary data with conditional mitigations (Table 2), a synthesis of related work and gaps (Table 3), three illustrative simulation evaluations (Tables 4–6) and a minimal applied case (Table 7), in addition to nine example diagrams (Figures 1–9). Each example illustrates a typical inferential threat, its signature in the graph, and mitigations consistent with the estimator. The threats and mitigations synthesise literature on selection bias, missing mechanism, target trial emulation, and support/positivity assumptions [5], [6], [7], [8], [9], [10], [13], [14].

Table 1. Minimum checklist for reporting and auditing studies with secondary data using layered graphics.

Domain	*Items to report*
Estimand definition and temporal alignment	Specify the target population, exposure contrast, outcome, time zero, and follow-up; state whether the estimand is intervention-analogous or associational; define baseline and retrospective/assessment windows in days or months; document any post-baseline information used to define exposure or covariates and justify temporal alignment (target trial logic where applicable).
Data provenance and construction of the analytic cohort	Describe the data source, coverage, and extraction rules; list inclusion and exclusion criteria and their temporal placement; represent eligibility/observability as a selection node when appropriate; report the flow from raw records to the analytic sample (counts at each stage).
Layer 1: Material causal system	Define constructs and causal ordering; propose the material graph for exposure → outcome including candidate confounders, mediators, and effect modifiers; justify assumptions (expert knowledge, literature, domain expertise).
Layer 2: Selection/inclusion system	Encode selection mechanisms (e.g., "measured", "in care", "linked", "followed") and their causes; state whether the analysis conditions on selection (observable sample) and which target population this implies; indicate limits to generalisability and transportability.
Layer 3: Measurement/operationalisation system	Distinguish underlying constructs from recorded proxies when proxies depend on the recording process; list determinants of observation/coding; indicate when exposure or outcome is defined or conditioned using proxies.
Layer 4: Missing data system	Introduce missingness indicators (R) for key variables with incomplete data; state assumptions about the missing data mechanism in causal terms; specify denominators (total n vs. observed n) for each analysis.
Identification, adjustment, and modelling	List covariates actually adjusted for and map each to its role in the graph (confounder/mediator/collider/selection criterion); justify the adjustment set and strategy (stratification, weighting, g-methods, imputation, models of the observation process) in relation to the estimand.
Alternative specifications and diagnostics	Pre-specify alternative windows and definitions; for each, report changes in the estimate relative to the primary analysis; include sensitivity diagnostics (placebo/falsification tests where appropriate) and discuss results that appear robust versus fragile.
Reproducibility and reporting	Provide the graph code and scripts used to generate each figure; export figures in the required format; maintain traceability between code, figure, and manuscript text (node/edge counts, hashes, manifest).

Note: "Selection" refers to eligibility/observability; *R* denotes the observation indicator (*R* = 1 observed, *R* = 0 missing).
Source: Authors.

Table 2. Typical decisions in secondary data: inferential threat, signature in the graph, and conditional mitigations to the estimator.

Decision	*Threat*	*Graph signature*	*Mitigation (examples)*
Restricting to "measured" laboratory data	Selection / collider bias	Conditioning on S_meas influenced by utilisation and morbidity	If the analysis conditions on S_meas = 1 and S_meas is a collider (causes of E and Y), then: (a) redefine the estimand to the observable population; (b) weight by P(S_meas = 1 | C) when identifiable; (c) report alternative specifications varying selection assumptions (Δestimate).
Exposure defined using a post-baseline window	Immortal time bias	Use of future information to define exposure; misalignment of time zero	If E_obs uses information after t_0, then: (a) redefine time zero and the assignment window; (b) emulate a "target trial" (eligibility, assignment, and follow-up aligned); (c) treat E as time-dependent if the estimand requires it.
Proxies affected by observation intensity	Measurement / proxy bias	Proxy with parents in the construct and in determinants of recording	If the proxy depends on recording determinants that also affect Y, then: (a) refine or validate the proxy; (b) predefine alternative proxies and report Δestimate; (c) adjust for recording determinants only if the DAG indicates confounding (not if they are mediators/colliders).
Excluding cases with missing variables	Selection due to missingness	Conditioning on R = 1 when R depends on causes of E/Y	If R depends on causes of E/Y (or on X itself), then: (a) multiple imputation under an explicit mechanism (MAR/MNAR) or modelling of the observation process; (b) sensitivity analyses when MNAR is plausible; (c) report denominators and, where appropriate, redefine the estimand to R = 1.
Standard adjustment for time-dependent confounders	Feedback bias	Post-exposure confounder affecting future exposure and outcome	If time-dependent confounding affected by prior exposure is present, then: (a) marginal structural models with weighting, the g-formula, or g-estimation; (b) explicit temporal indexing and update schedule; (c) avoid standard adjustment for post-exposure covariates.
Eligibility defined by contact with the system	Contact-based selection	S depends on utilisation and morbidity; changes the target population	If contact/observability S depends on factors related to E and Y, then: (a) define the estimand in the population with contact (S = 1); (b) weight by P(S = 1 | C) if plausible/identifiable; (c) discuss transportability and perform sensitivity analyses.
Analysing only devices/users with complete telemetry or logs (S_tx = 1)	Selection / observability bias (collider) and possible positivity violation	E → S_tx ← system state/utilisation → Y; conditioning on S_tx alters denominators and opens spurious paths	If S_tx is a collider and/or induces positivity violations, then: (a) declare the target population as "observable units"; (b) model S_tx with infrastructure/operational covariates; (c) apply weighting with diagnostics (tails/extremes) and trimming where appropriate; (d) sensitivity to unobserved failures; (e) report telemetry loss rates (denominators).
Record linkage across sources (S_link = 1) to construct exposure/proxies	Selection bias (linkage) + conditional measurement	E_obs exists only if S_link = 1; S_link depends on identifier quality (Q_id) and factors related to Y; conditioning on S_link may open spurious paths and change denominators	If E_obs requires S_link = 1, then: (a) report linkage rates and compare covariates before/after linkage; (b) model or weight by P(S_link = 1 | C) when plausible; (c) sensitivity to MNAR linkage failure; (d) declare the change in target population if correction is not feasible.
Sensor/telemetry failure dependent on load or battery (R_X)	Missing data bias (MNAR plausible) + local positivity degradation	R_X depends on Load/Bat and causes of E/Y; complete-case analysis conditions on R_X = 1 and induces informative selection	If R_X is informative, then: (a) model the observation process (e.g., weighting by P(R_X = 1 | C)); (b) imputation under an explicit mechanism and sensitivity analyses when MNAR is plausible; (c) report denominators by Load/Bat strata; (d) report Δestimate and changes in precision.

Note: S_meas denotes "being measured/recorded". S_tx/S_link denote observability via telemetry or linkage. "Utilisation/Load" refers to the intensity of system contact/load generating the record. C denotes baseline covariates used to model P(S = 1 | C) or P(R = 1 | C).
Rule: Mitigations apply only under the role (confounder/mediator/collider) derived from the layered DAG and the declared

estimand.
Source: Authors.

Table 3. Summary of related work and gap: from isolated approaches to an auditable integrative protocol.

Methodological component	*Typical reference*	*What it contributes*	*Gap in secondary data that this work addresses*
Material DAG and adjustment sets	Pearl; d-separation–based guidelines	Material causal structure; adjustment sets; avoidance of colliders	S/R/proxies often remain implicit; limited traceability when the DAG is edited manually.
Selection bias (S-nodes)	Hernán–Robins; Bareinboim–Pearl	Formalises how conditioning on S may open causal paths; transportability/recoverability	Does not standardise reporting of denominators and alternative specifications linked to the DAG.
Missing data (R-nodes)	Pearl; Mohan–Pearl	Recoverability and explicit assumptions regarding observation (R)	Usually discussed separately from selection and proxy measurement; rarely translated into reporting rules (Δestimate).
Temporal alignment / "target trial"	Hernán–Robins (target trial framework)	Defines t_0, exposure windows, and follow-up; prevents immortal time bias	Does not require integration of temporal windows with S/R/proxies within a single auditable artefact.
Graph reproducibility	DAGitty / code-based graph specification	Verifiable derivation of adjustment sets from an explicit graph	Without an artefact contract (hashes/counts/QC), drift may occur between manuscript text and figure.
This work: Layered DAG + auditable contract	—	Integrates layers 1–4; reporting rules (denominators/windows/Δ); code → figure with QC	Transforms secondary data decisions into assumptions that can be inspected and compared during peer review.

Note: Conceptual synthesis intended to situate the distinctive contribution. Validity and applicability depend on the estimand and the declared DAG (layers 1–4).
Source: Authors.

Figures 1–9 use redundant colour and shape coding to facilitate reading and accessibility: material system (blue circle), selection/observability S (yellow diamond), recorded proxies *_obs (green box), missing R_* (red octagon) and determinants/context (grey circle).

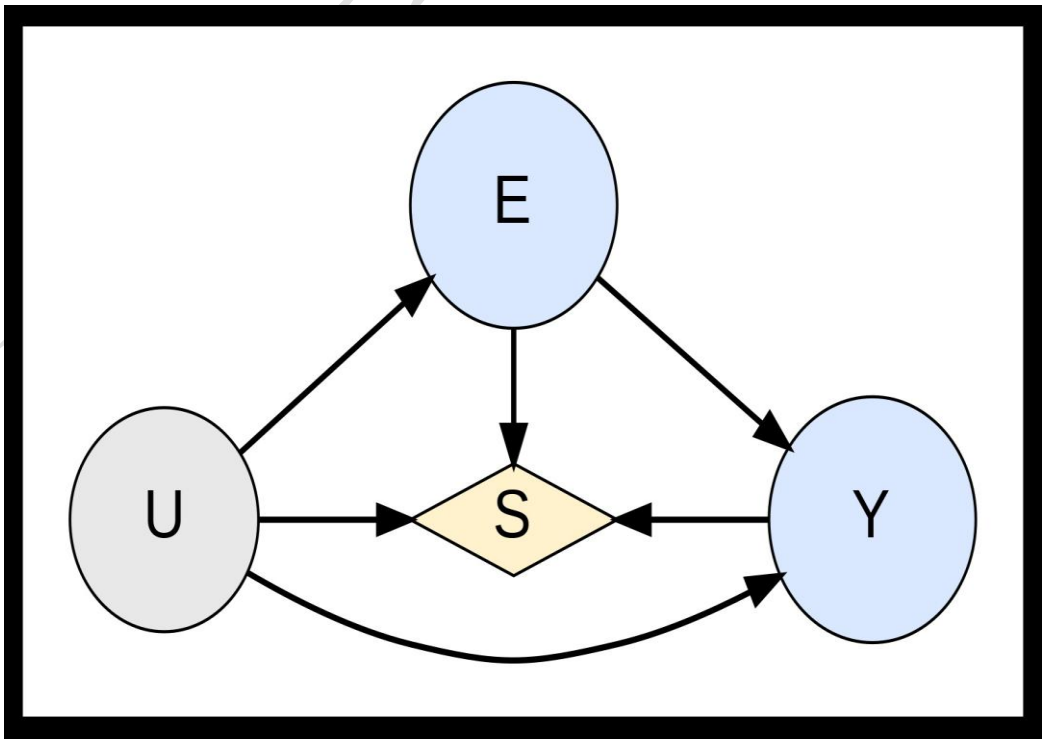

Figure 1. Selection bias when conditioning on observability (Layer 2).
Note: U: unmeasured factor; E: exposure; Y: outcome; S: selection/observability (S=1 indicates that the individual is observable/included). QC count: 4 nodes, 6 arcs.

Source: Authors.

In Figure 1, conditioning the analysis on S=1 induces collider bias by opening the non-causal path E→S←Y and may modify the target population. If U also affects S, the restriction S=1 introduces dependence between E and U and limits transportability. The protocol requires explicitly stating the causes of S and justifying any adjustment/weighting consistent with the estimator (e.g., estimator defined in observable population).

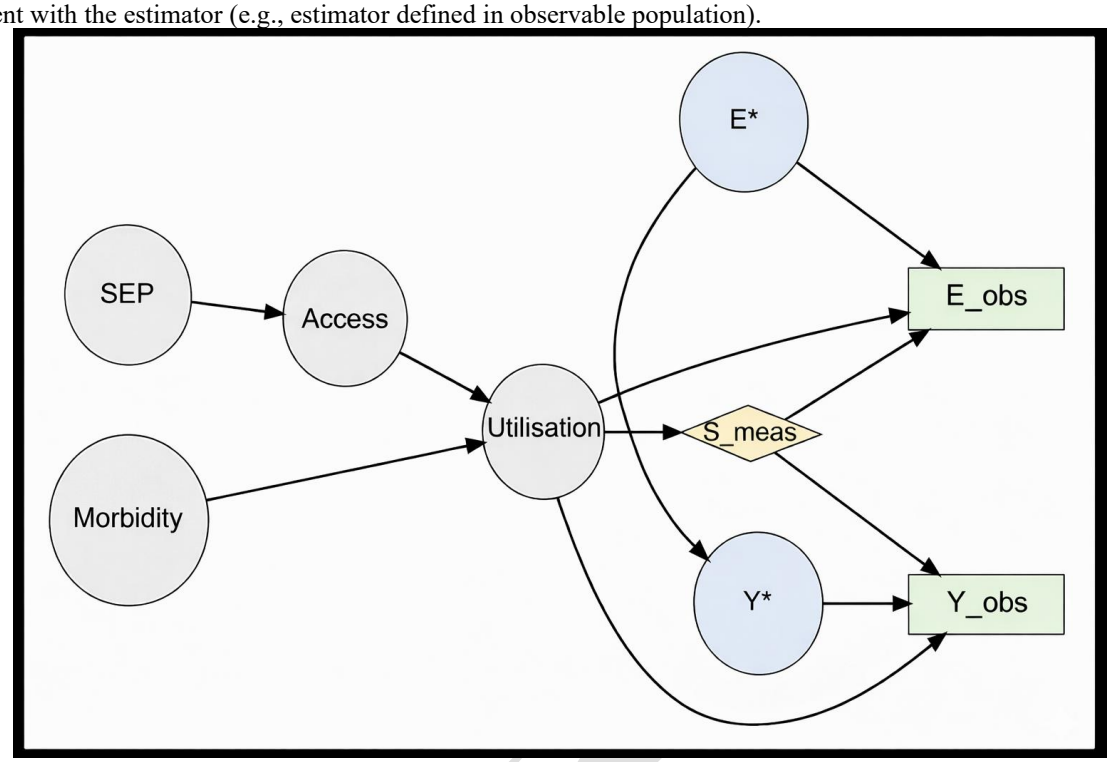

Figure 2. Usage-guided observation: conditional measurement and recorded proxies (Layers 2–3).
Note: SES: socioeconomic status; Access: access to services; Morbidity: disease burden; Util: intensity of use/contact; S_meas: being measured/recorded; E*: exposure construct; Y*: outcome construct; E_obs/Y_obs: recorded proxies. QC count: 9 nodes, 11 arcs.
Source: Authors.

In Figure 2, the measurement depends on Util and Morbidity, so restricting to 'measured' individuals (S_meas=1) selects a subpopulation and opens non-causal paths through S_meas. Furthermore, E_obs and Y_obs depend on both the construct and the intensity of observation, implying that proxy-based analyses inherit assumptions about the recording process. Mitigation should be aligned with the estimator (e.g., redefine estimator for observable population, or model/weight by measurement when identifiable).

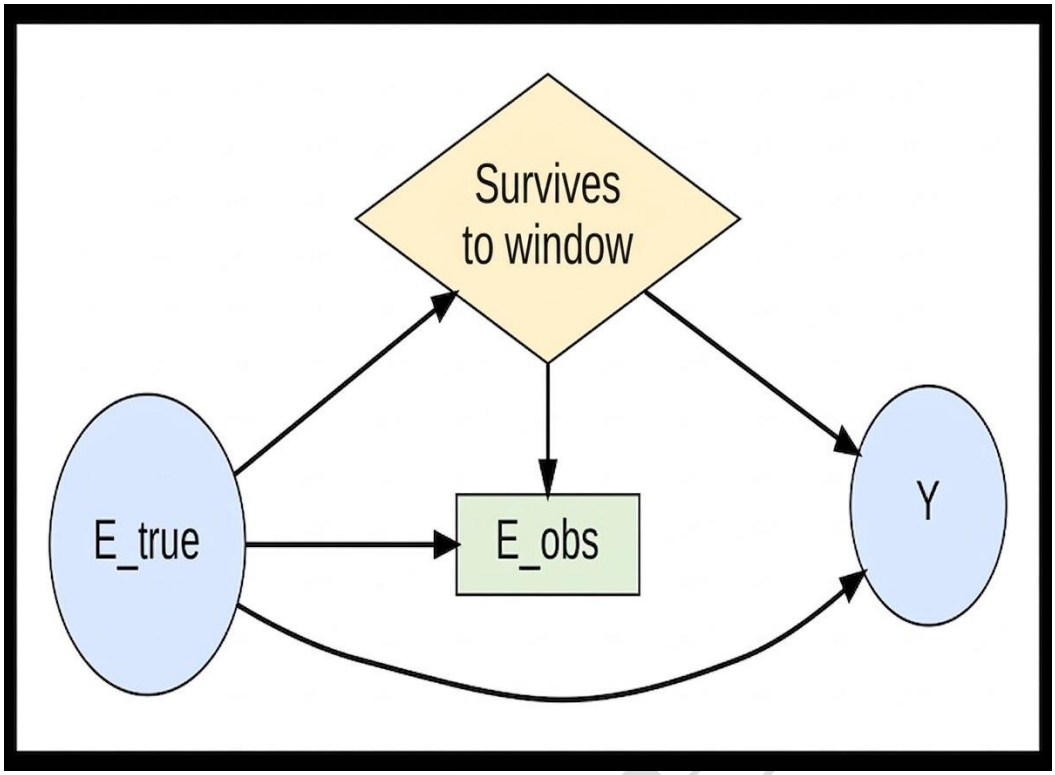

Figure 3. Immortal time and zero-time misalignment when defining exposure with post-baseline window.
Note: E_true: true exposure; E_obs: observed exposure defined using a window; "Survives window": structural requirement of survival/absence of event during the window; Y: outcome. QC count: 4 nodes, 5 arcs.
Source: Authors.

In Figure 3, when E_obs is defined with information after the start of follow-up, a window is created in which the individual must 'survive' to be classified as exposed, structurally excluding early events. This generates zero-time misalignment and can induce bias even without confounding. Correction requires redefining time zero and exposure assignment (objective trial logic) or using time-dependent methods depending on the estimator.

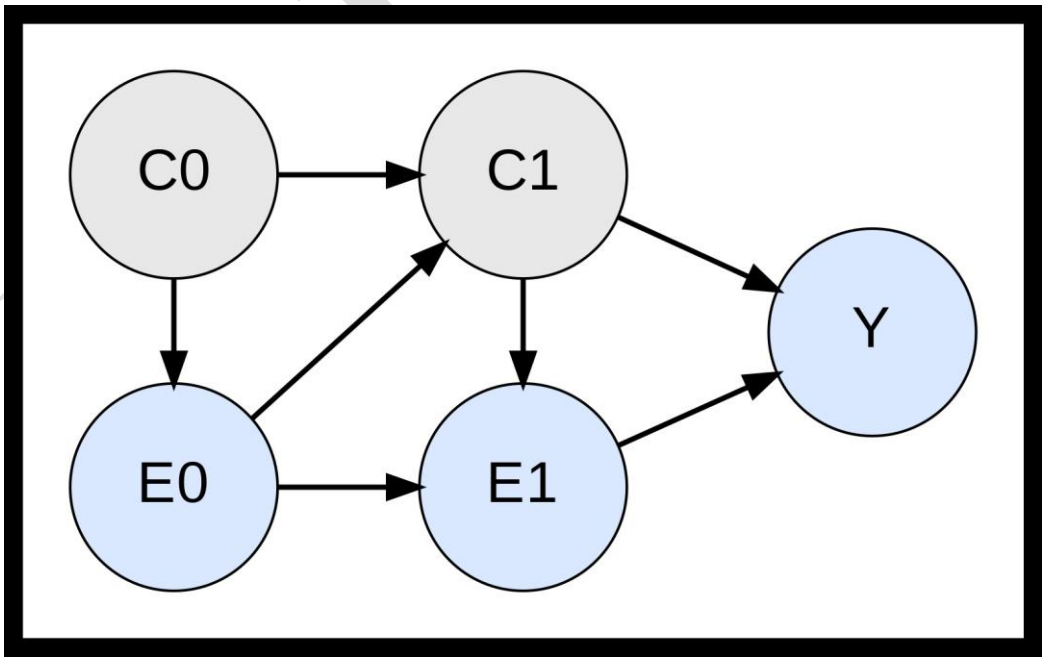

Figure 4. Time-dependent confusion with feedback (g-methods).
Note: C0/C1: confuser at t0/t1; E0/E1: exposure at t0/t1; Y: result. QC count: 5 nodes, 7 arcs.
Source: Authors.

In Figure 4, c1 is a confounder of the relationship E1→Y but is also affected by E0. Standard adjustment for C1 may introduce bias by blocking part of the effect or opening unwanted paths. For intervention-analogue estimators, approaches such as inverse probability weighting or g-methods are required, with explicit time index and covariate update schedule.

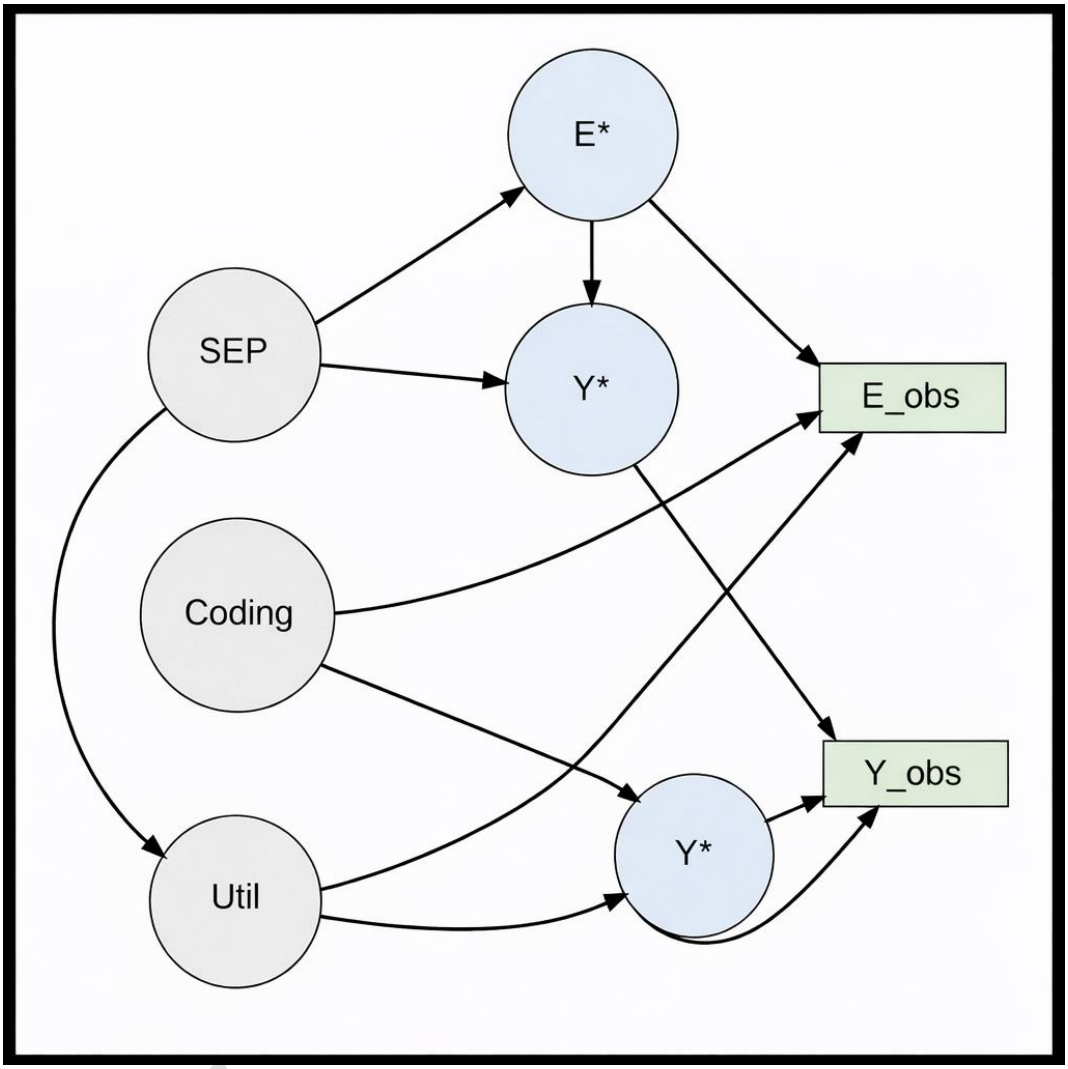

Figure 5. Operationalisation layer: constructs vs. proxies with determinants of registration (Layer 3).
Note: SES: socioeconomic status; Use: use/contact; Coding: recording practice; E*/Y*: constructs; E_obs/Y_obs: recorded proxies. QC count: 7 nodes, 10 arcs.
Source: Authors.

In Figure 5, the proxies (E_obs, Y_obs) depend simultaneously on the construct and on registration determinants (Util, Coding). Therefore, using proxies as if they were the construct can induce spurious association via registration determinants (E_obs←Util→Y_obs) and requires declaring assumptions of validity/reliability. Mitigation includes refining proxy definitions, validating coding rules, and avoiding adjustment for registration determinants unless justified by the graph and the estimator.

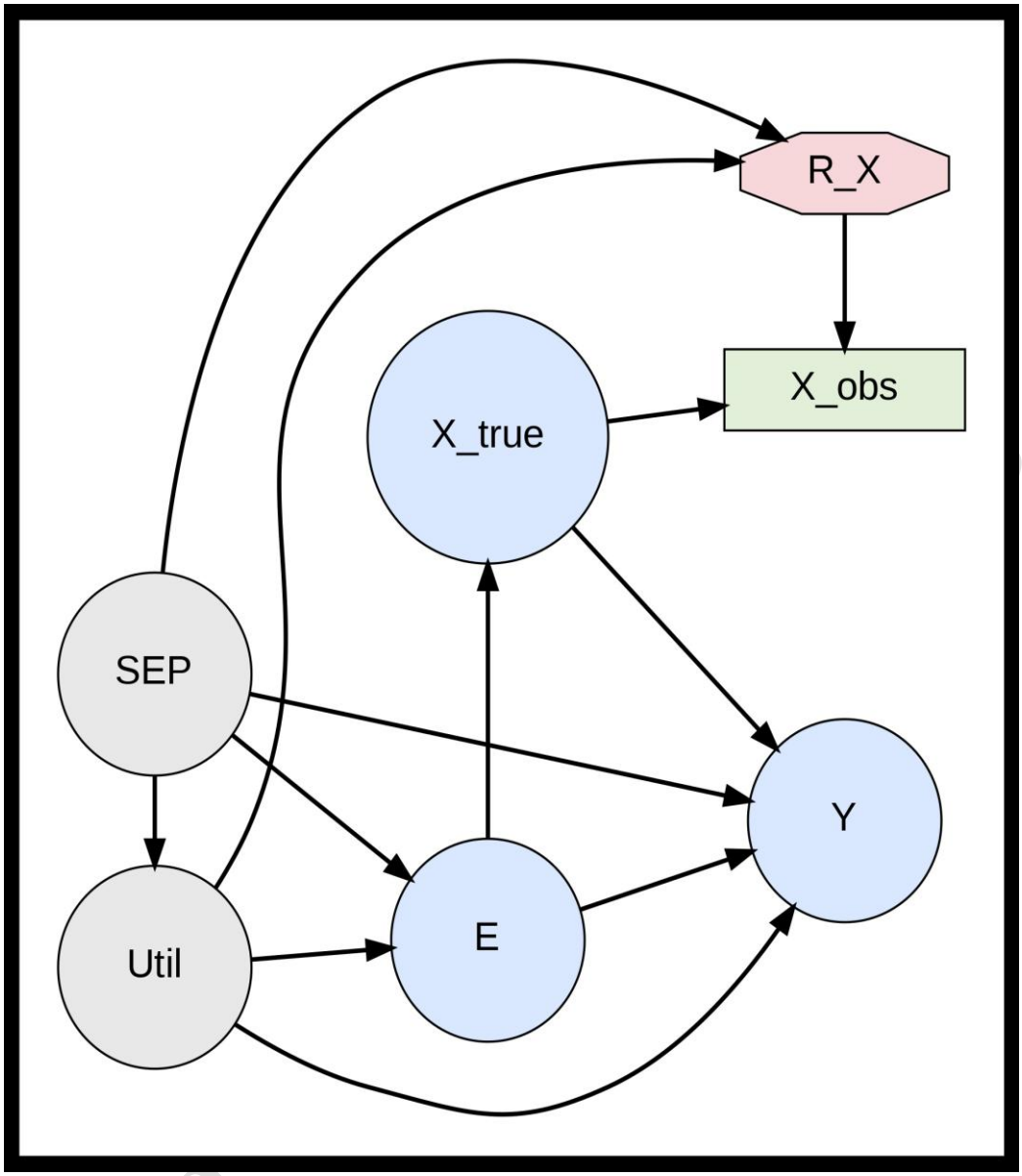

Figure 6. Missing data: observation indicator R and observed variable (Layer 4).
Note: SES: socioeconomic status; Use: use/contact; E: exposure; Y: outcome; X_true: covariate or true mediator; R_X: observation indicator for X; X_obs: observed value of X. QC count: 7 nodes, 12 arcs.
Source: Authors.

In Figure 6, excluding cases with missing X is equivalent to conditioning on R_X=1. If R_X depends on causes of E and/or Y (directly or indirectly via Util/SEP), the analysis of complete cases may bias the estimate. The protocol forces the representation of the missing mechanism and the alignment of the solution (imputation under stated assumptions, observation process models, or redefinition of the estimator for the observable population).

Illustrative simulation assessment: immortal time bias

A reproducible simulation was performed to quantify the bias induced by defining exposure with a post-baseline window (Figure 3). A synthetic cohort of n=200,000 individuals was generated with time to event T ~ Exponential(λ) and potential exposure start time A ~ Exponential(ρ), independent. By construction, exposure does not modify risk (zero causal effect). Observed exposure was defined as 'started before 30 days,' which requires survival until A (immortal time). We compared (i) a naive analysis that classifies from t0 using the window and (ii) a benchmark analysis that starts follow-up on day 30 and classifies according to onset before 30 days. The code and tabular output are included in the supplementary material. The results are summarised in Table 4.

Table 4. Illustrative simulation of immortal time bias when defining exposure with a post-baseline window (window = 30 days).

Analytical design	*Population and follow-up*	*RR (95% CI)*	*Inferential interpretation*
True (reference)	Total population; 365-day horizon	1.000 (by construction)	No causal effect; any deviation reflects bias or sampling variability.
Naïve (classification from t_0)	Total population; window = 30 days	0.965 (0.955–0.974)	Spurious protective effect due to immortal time bias (use of future information).
Landmark (start at day 30)	Restriction to T > 30; follow-up 30–365 days	1.001 (0.990–1.011)	Avoids immortal time bias; estimand interpretable among those surviving the window.

Note: Parameters: n = 200,000, λ = 0.002 (per day), ρ = 0.01 (per day), horizon = 365 days, window = 30 days; fixed seed (*seed* = 20260224). RR calculated as cumulative risk at 365 days; 95% CI computed on log(RR).
Source: Authors (simulation).

Naive analysis introduces a protective association with relative risk (RR) <1 despite the causal effect being null; the only source is temporal misalignment that imposes prior survival to be classified as exposed. This pattern corresponds to the immortal time signature that the protocol requires to be explicitly represented (Layer 2 and Figure 3) before selecting the estimator and method.

Illustrative simulation assessment: selection bias/observability

The bias induced by conditioning the analysis on an observable sample (S=1) was simulated, in accordance with the selection bias signature represented in Figure 1. Two scenarios were considered: (A) selection depending on exposure and outcome (S←E, S←Y), and (B) selection depending on exposure and an observed baseline covariate L (S←E, S←L). In both cases, exposure has a true effect β=1 on Y. Estimators in the population were compared with those in the selected sample, with and without adjustment for L when applicable. The code and tabular outputs are included in the supplementary material. The results are summarised in Table 5.

Table 5. Illustrative simulation of selection bias/observability when conditioning on S=1 (Figure 1).

Scenario	*Analysis*	*Selection rate (mean)*	*β̂_E (mean [p2.5–p97.5])*
A: S depends on E and Y	Population (Y ~ E)	59.2%	0.999 [0.946–1.057]
A: S depends on E and Y	Sample S = 1 (naïve, Y ~ E)	59.2%	0.826 [0.753–0.909]
B: S depends on E and L (baseline)	Population (Y ~ E + L)	52.2%	1.000 [0.947–1.068]
B: S depends on E and L (baseline)	Sample S = 1 (naïve, Y ~ E)	52.2%	0.918 [0.775–1.033]
B: S depends on E and L (baseline)	Sample S = 1 (adjusted, Y ~ E + L)	52.2%	1.002 [0.912–1.091]

Note: Simulations with N = 4,000 per replicate, R = 100 replicates, true β = 1. Scenario A: $Y = \beta \cdot E + \varepsilon$, logit P(S = 1) = −0.2 + 0.6·E + 0.7·Y (collider E → S ← Y). Scenario B: $L \sim N(0,1)$, $Y = \beta \cdot E + 1 \cdot L + \varepsilon$, logit P(S = 1) = −0.2 + 0.6·E + 0.7·L. The interval is empirical (2.5–97.5 percentiles) of β̂_E across replicates.
Source: Authors (simulation).

In Table 5, in scenario A, conditioning on S=1 induces a substantial bias (β̂_E≈0.83) even though the true effect is 1. In scenario B, selection causes E and L to correlate in the observable sample, biasing the analysis without adjustment; adjustment for L recovers the effect (β̂_E≈1.00). This operationalises the protocol rule: any restriction to 'observables' requires making S explicit and justifying whether the bias can be blocked with measured information or whether sensitivity is required.

Illustrative simulation assessment: bias due to missing data in a confounder

The impact of excluding records with missing data in a confounder X (conditioning on R_X=1, Figure 6) was simulated. Two mechanisms for missing data were considered: (A) MAR, where the observation of X depends on exposure and outcome (R_X←E, R_X←Y), and (B) MNAR, where it also depends on the true value of X (R_X←X_true). We compared (i) a reference analysis with complete X, (ii) complete cases (only R_X=1), and (iii) multiple imputation under a linear MAR model (m=15). The code and tabular outputs are included in the supplementary material. The results are summarised in Table 6.

Table 6. Illustrative simulation of bias due to missing data in a confounder when conditioning on R_X=1

Missingness mechanism	*Analysis*	*Missingness rate (mean)*	*β̂_E (mean [p2.5–p97.5])*
A (MAR): R_X depends on E and Y	Reference (Y ~ E + X_true)	55.4%	0.998 [0.927–1.064]
A (MAR): R_X depends on E and Y	Complete cases (R_X = 1)	55.4%	0.913 [0.837–1.005]
A (MAR): R_X depends on E and Y	Multiple imputation (MAR; m = 15)	55.4%	1.000 [0.927–1.066]
B (MNAR): R_X depends on E, X and Y	Reference (Y ~ E + X_true)	54.7%	0.998 [0.927–1.064]
B (MNAR): R_X depends on E, X and Y	Complete cases (R_X = 1)	54.7%	0.914 [0.829–1.010]
B (MNAR): R_X depends on E, X and Y	Multiple imputation (assuming MAR)	54.7%	1.086 [0.988–1.174]

Note: Simulations with N = 4,000 per replicate, R = 80 replicates, true β = 1, m = 15 imputations. Data-generating process: $X \sim N(0,1)$, logit P(E = 1) = −0.2 + 0.8·X, $Y = \beta \cdot E + 1 \cdot X + \varepsilon$. Mechanism A (MAR): logit P(R_X = 1) = 0.2 − 0.6·E − 0.4·Y. Mechanism B (MNAR): logit P(R_X = 1) = 0.2 − 0.6·E − 0.6·X − 0.4·Y. The interval is empirical (2.5–97.5 percentiles) of β̂_E across replicates.
Source: Authors (simulation).

In Table 6, under MAR, the analysis of complete cases biases the effect by conditioning on R_X=1, while multiple imputation under the correct model recovers β≈1. Under MNAR, both complete cases and imputation under MAR assumptions can bias (even in opposite directions). This justifies Layer 4 representing R and the report including diagnostics, comparability, and sensitivity analysis when the missing mechanism is not identifiable with observed variables.

Minimum applied case: plant data for auditing measurement layers and missing data

As an applied demonstration in an engineering domain, Brownlee's 'stack loss' dataset [16] was used, with 21 days of measurements from an ammonia to nitric acid oxidation plant. An illustrative estimator (linear model) is defined as the expected change in STACKLOSS per 1 unit of WATERTEMP, adjusting for AIRFLOW (operating rate) and ACIDCONC (acid concentration). The objective here is to show traceability and sensitivity to assumptions (layers), not to establish a generalisable causal conclusion. For reproducibility, the applied case script loads the dataset from statsmodels.datasets.stackloss (source: Brownlee [16]) and does not require distributing a separate data file.
To explicitly activate Layer 4 (missing values), an observation mechanism was introduced where WATERTEMP is more likely to be missing when STACKLOSS and AIRFLOW are high (plausible scenario of event-conditioned monitoring). This generates an indicator R_T and an observed variable T_obs, so that conditioning on R_T=1 may induce collider bias (Y → R_T ← AIRFLOW), as anticipated by the layered DAG.

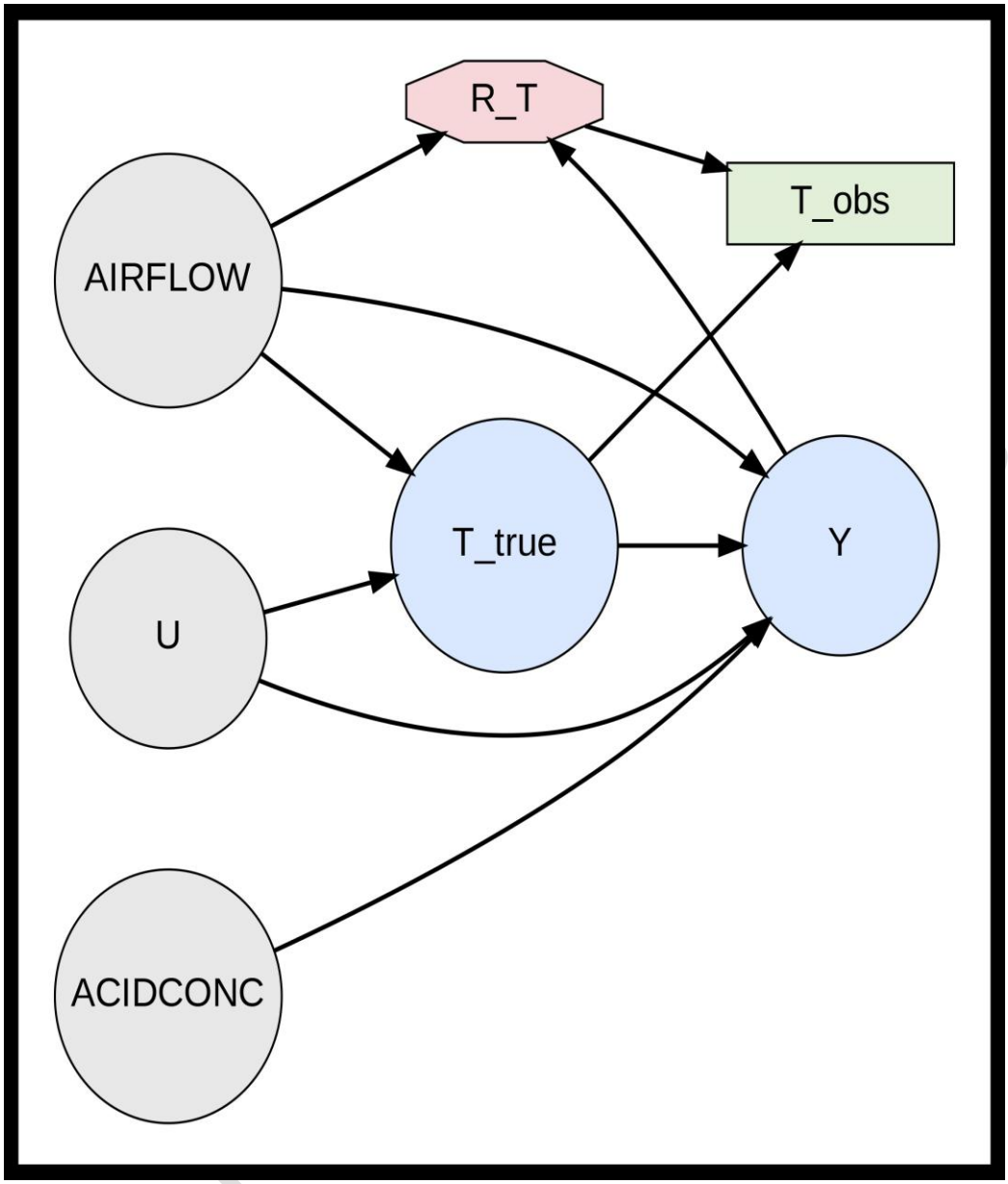

Figure 7. Minimum applied case (stack loss plant): Layered DAG for WATERTEMP and STACKLOSS with process-dependent missing values.
Note: AIRFLOW and ACIDCONC are operational covariates; T_true is the temperature (construct), T_obs is the observed record, and R_T is the observation indicator. QC count: 7 nodes, 10 arcs.
Source: Authors.

In Figure 7, if R_T depends on Y and operational variables, analysing only records with observed T_obs is equivalent to conditioning on R_T=1, opening spurious paths. The protocol requires this mechanism to be declared and alternative analyses consistent with the DAG to be reported.

Table 7. Minimum applied case (Brownlee [16]): sensitivity of the WATERTEMP→STACKLOSS estimator to observation/missing decisions.

Specification	*β_WATERTEMP*	*95% CI*	*Δβ vs baseline*
Baseline (complete data; reference). n = 21; missing = 0.	1.295	[0.574, 2.017]	0.000
Complete cases (conditioning on R_T = 1). n = 15; induced missingness = 6 (28.6%).	0.567	[0.253, 0.882]	−0.728
Multiple imputation (chained equations, MICE; m = 30; MAR conditional on Y and covariates). n = 21; induced missingness = 6 (28.6%).	0.403	[−0.631, 1.437]	−0.892

Note: β_WATERTEMP is estimated using ordinary least squares (OLS), adjusted for AIRFLOW and ACIDCONC. Missingness is induced in 6/21 (28.6%) observations of WATERTEMP to illustrate Layer 4 (fixed seed). Multiple imputation uses **multiple imputation by chained equations (MICE)** (30 imputations; fixed seed), including STACKLOSS and covariates in the imputation model; 95% CIs computed using Rubin's rules.
Source: Authors, based on Brownlee [16].

In Table 7, the estimate of the effect of WATERTEMP on STACKLOSS changes materially depending on how missing values are handled (complete cases vs. imputation), even though the material model is the same. This is the type of inferential drift that the protocol seeks to make visible through (i) layered DAGs (to justify assumptions), (ii) explicit denominators, and (iii) reporting of Δestimation between specifications.
Additional examples to expand coverage of frequent scenarios in engineering and administrative records:

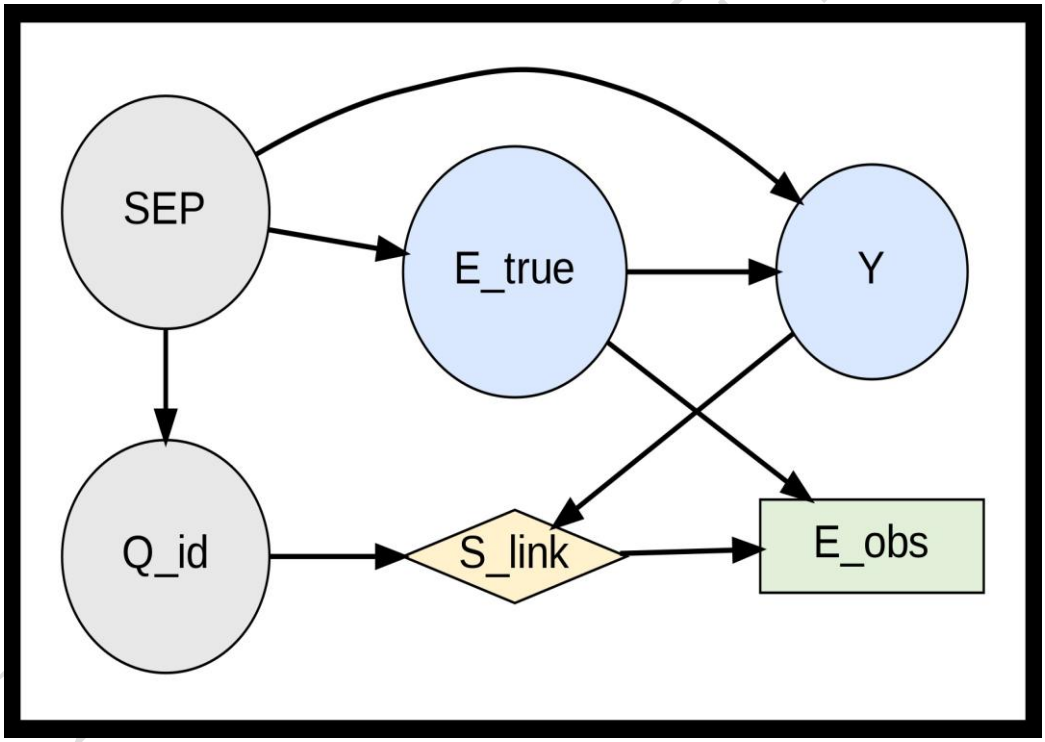

Figure 8. Record linkage: selection by link (S_link) and conditional measurement (E_obs).
Note: SEP: socioeconomic position; Q_id: quality of identifiers; E_true: exposure (construct); E_obs: proxy observed only if S_link=1; S_link: link success; Y: outcome. QC count: 6 nodes, 8 arcs.
Source: Authors.

In Figure 8, if E_obs is constructed only for linked records (S_link=1), the analysis implicitly conditions on S_link. When S_link depends on Q_id and outcome-related factors, the target population changes and selection bias may arise. The protocol requires reporting link rates, denominators, and at least one alternative specification (e.g., weighting by P(S_link=1|C) or MNAR sensitivity) to audit stability (Δestimation).

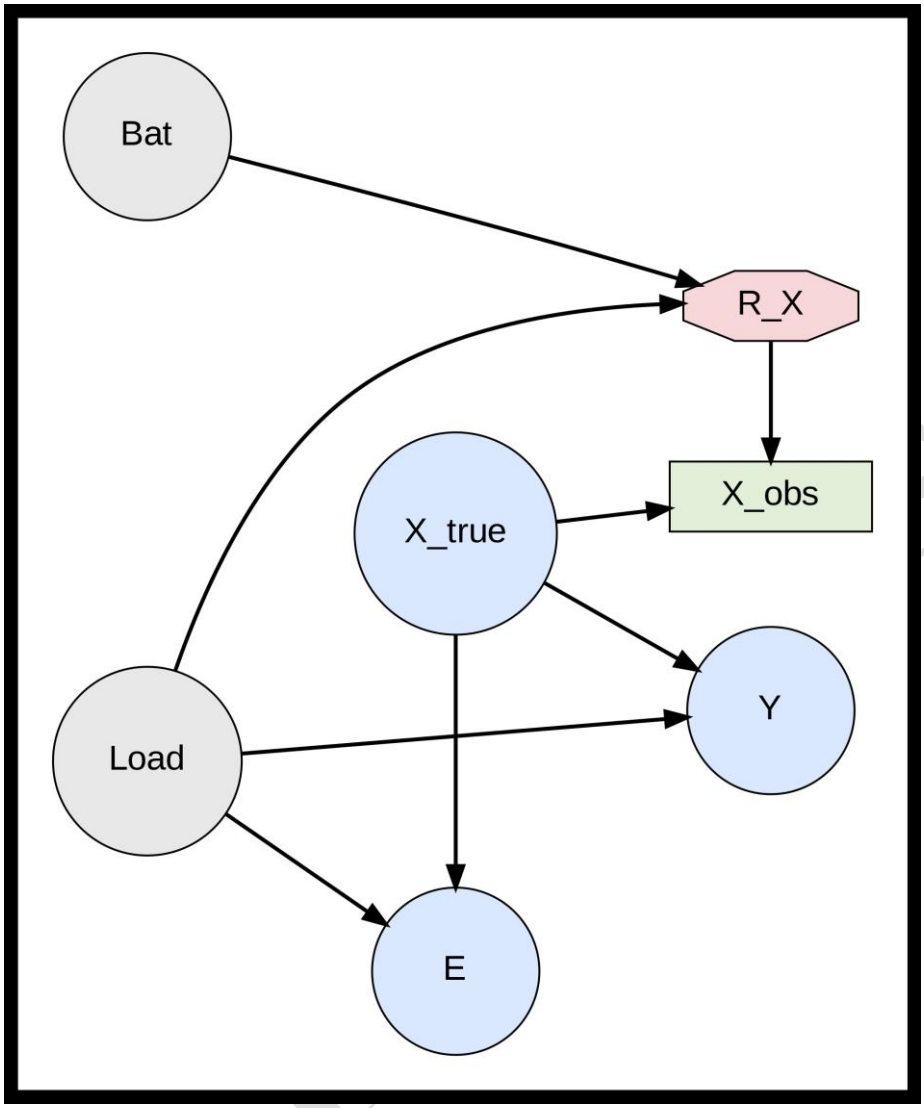

Figure 9. Sensor/telemetry failure: missing information (R_X) dependent on load and battery.
Note: Load: load/operational condition; Bat: battery/energy; X_true: covariate/state not fully observed; R_X: observation indicator (1=observed); X_obs: recorded measurement; E: exposure/control action; Y: outcome. QC count: 7 nodes, 8 arcs.
Source: Authors

In Figure 9, when R_X depends on variables that also affect E or Y (e.g., load or energy), excluding cases with missing X_obs is equivalent to conditioning on R_X=1 and may induce informative selection. Aligned mitigation depends on the estimator: model P(R_X=1|C) and weight, impute under a declared mechanism, and report sensitivity when MNAR is plausible, always with denominators by operational strata.

Taken together, the examples (Figures 1–9) and illustrative simulations (Tables 4–7) show that extraction and definition decisions (windows, eligibility/observability, coding rules, handling of missing data) are not neutral: they change the estimate or introduce non-causal paths if conditioned on selection, proxies, or observation indicators. The layered protocol does not in itself resolve the lack of identification (e.g., unmeasured confounding); its contribution is to force assumptions to be explicit, the analysis to be auditable, and the proposed mitigations to be aligned with the estimator and the target population.

IV. CONCLUSIONS

A reproducible protocol for layer-directed acyclic graphs for studies with secondary data was presented, aimed at separating material causal systems, selection/observability, measurement/operationalisation, and missing data.
The protocol strengthens reporting by requiring temporal alignment of the estimator, traceability between extraction decisions and causal assumptions, and explicit documentation of selection mechanisms, proxies, and missing data.
As a minimum practice, it is recommended to publish the source code of the graph and a quality control manifesto linking each figure to its specification (nodes/arcs and hashes), along with predefined alternative specifications. Three illustrative simulation evaluations—immortal time, selection/observability, and missing data—were included (Tables 4–6), reproducible with the supplementary material, to show the magnitude and potential direction of bias under representative topologies. In addition, a minimal applied case with plant data (Table 7) was included to demonstrate the sensitivity of the estimator to measurement decisions and missing data handling; it is recommended to extend the validation applied in specific domains (health, public administration, sensors/IoT, digital platforms) to quantify the impact on analytical decisions and estimators, and to document assumptions that cannot be identified through sensitivity analysis.

V. CRediT AUTHORSHIP CONTRIBUTION STATEMENT

D. Rivera-Porras: Conceptualization, Research, Data curation, Writing—Original draft, Visualization. **Y. Gómez-Charris**: Conceptualization, Methodology, Writing—Review and editing, Supervision. **V. Bermudez:** Methodology, Writing—Review and editing, Supervision, Project management

VI. FUNDING

The authors declare that they did not receive specific funding for the completion of this work.

VII. CONFLICT OF INTEREST

The authors declare that they have no conflicts of interest.

VIII. ACKNOWLEDGEMENTS

Does not apply.

IX. DATA AND CODE AVAILABILITY

Supplementary material includes: (i) source code for figures (DOT/Graphviz) and exports (JPG 300 dpi + PDF/SVG), (ii) a quality control manifest (counts + SHA-256 hashes) and automatic verification scripts (verify_qc_manifest.py and verify_manuscript_figures.py), (iii) scripts and outputs from reproducible simulations corresponding to Tables 4–6, (iv) a minimal reproducible applied case (Table 7) that loads the "stackloss" dataset from statsmodels (source: Brownlee [16]) and induces a controlled missing mechanism to illustrate Layer 4, and (v) a reproduce.sh script that runs the reproduction pipeline in one command. No sensitive microdata or restricted information is included.

X. REFERENCIAS

[1] J. Pearl, "Causal diagrams for empirical research (with discussion)," Biometrika, vol. 82, no. 4, pp. 669–710, 1995, doi: 10.1093/biomet/82.4.669.
[2] J. Pearl, Causality. Cambridge University Press, 2009, doi: 10.1017/CBO9780511803161.
[3] T. J. VanderWeele and S. Shpitser, "A new criterion for confounder selection," Biometrics, vol. 67, no. 4, pp. 1406–1413, 2011, doi: 10.1111/j.1541-0420.2011.01619.x.
[4] M. A. Hernán, "The C-word: scientific euphemisms do not improve causal inference from observational data," Am. J. Public Health, vol. 108, no. 5, pp. 616–619, 2018, doi: 10.2105/ajph.2018.304337.
[5] M. A. Hernán, S. Hernández-Díaz, and J. M. Robins, "A structural approach to selection bias," Epidemiology, vol. 15, no. 5, pp. 615–625, 2004, doi: 10.1097/01.ede.0000135174.63482.43.
[6] J. Pearl, "Missing-data mechanisms and causal inference," J. Am. Stat. Assoc., vol. 109, no. 507, pp. 987–992, 2014, doi: 10.1080/01621459.2014.951426.
[7] M. A. Hernán and J. M. Robins, "Using big data to emulate a target trial when a randomized trial is not available," Am. J. Epidemiol., vol. 183, no. 8, pp. 758–764, 2016, doi: 10.1093/aje/kwv254.

[8] E. Bareinboim, J. Tian, and J. Pearl, “Recovering from selection bias in causal and statistical inference,” in Proc. 28th AAAI Conf. Artif. Intell. (AAAI-14), 2014, pp. 2410–2416. [Online]. Available: https://ftp.cs.ucla.edu/pub/stat_ser/r425.pdf
[9] K. Mohan, J. Pearl, and J. Tian, “Graphical models for inference with missing data,” Adv. Neural Inf. Process. Syst., vol. 26, pp. 1277–1285, 2013, doi: 10.5555/2999611.2999754.
[10] J. Chubak, S. R. Pocobelli, L. E. Weiss, and D. R. Cook, “Bias due to truncation by death in studies of cancer treatment and survival: A simulation study,” Cancer Epidemiol. Biomarkers Prev., vol. 30, no. 1, pp. 24–33, 2021, doi: 10.1158/1055-9965.epi-20-0234.
[11] J. Textor, B. van der Zander, M. S. Gilthorpe, M. Liśkiewicz, and G. T. Ellison, “Robust causal inference using directed acyclic graphs: the R package ‘dagitty’,” Int. J. Epidemiol., vol. 45, no. 6, pp. 1887–1894, 2016, doi: 10.1093/ije/dyw341.
[12] M. A. Hernán and J. M. Robins, “Per-protocol analyses of pragmatic trials,” N. Engl. J. Med., vol. 377, no. 14, pp. 1391–1398, 2017, doi: 10.1056/nejmra1705389.
[13] G. Danaei, M. García Rodríguez, E. Cantero, and M. Hernán, “Electronic medical records can be used to emulate target trials of sustained treatment strategies,” J. Clin. Epidemiol., vol. 61, no. 6, pp. 512–518, 2008, doi: 10.1016/j.jclinepi.2007.04.014.
[14] D. M. Westreich and S. R. Cole, “Invited commentary: positivity in practice,” Am. J. Epidemiol., vol. 171, no. 6, pp. 674–677, 2010, doi: 10.1093/aje/kwp436.
[15] D. M. Murray, R. C. Pals, J. M. Blitstein, N. L. Alfano, and E. Baker, “A comparison of regression methods for adjusting for baseline in individually randomized and cluster-randomized trials,” Stat. Med., vol. 37, no. 25, pp. 3857–3870, 2018, doi: 10.1002/sim.7842.
[16] K. A. Brownlee, "Statistical Theory and Methodology in Science and Engineering," 2nd ed. New York, NY, USA: Wiley, 1965.