UNCONFOUND LABS · TECHNICAL METHODS PAPER #01 · BIAS SERIES

When the Link Breaks

Quantifying, and controlling, the bias that record-linkage error injects into observational studies.

Methods demonstration on synthetic and public data, not a real-world clinical study. The reproducible implementation is maintained privately; this paper presents the problem, the method, and the results at a level a reviewer can assess and a peer can trust.

ABSTRACT

Modern evidence often depends on joining records that were never designed to be joined, across systems with no shared identifier. The joins are imperfect, and the errors flow silently into the effect estimates that decisions rely on. Working in a synthetic cohort with known ground truth, we show that plausible mother-child linkage error attenuates estimated risk ratios toward the null in a predictable, analytically derivable way: a true relative risk of 2.0 is reported as 1.73 under conventional exact-match linkage. We further show that probabilistic linkage recovers 73% of true relationships at 95% precision where exact matching recovers 36% at 75%, and that exact matching degrades as data volume grows. The contribution is a reusable, auditable method for measuring linkage-induced bias before it corrupts a conclusion, validated against a closed-form prediction and a null control.

1 · THE PROBLEM: ERROR YOU CANNOT SEE

Record linkage underlies pharmacoepidemiology, health-services research, and any analysis that assembles a person from fragments held in different systems. Mother-child linkage is a revealing hard case because a mother and her child share no personal identifier: different birthdates, different names, different identifiers. The connection survives only in inherited and shared attributes, a surname passed down and a common household address, and those are exactly the fields real-world data corrupts through name changes, relocations, transcription error, and administrative substitution.

Linkage fails in two ways with opposite consequences. A **missed link** drops a record; if the failure is unrelated to what is studied, it costs precision but introduces no bias. A **false link** attaches the wrong entity, importing the wrong exposure status; it misclassifies the variable under study and biases the answer. Reporting a single "error rate" conflates the two. Separating them is the first requirement for controlling the bias.

2 · METHOD: KNOW THE TRUTH, THEN BREAK IT ON PURPOSE

The obstacle to studying linkage error in real data is that the correct answer is unknown; if it were known, linkage would be unnecessary. We remove the obstacle by working in a synthetic cohort where the true relationships are known by construction, then degrading the records with a realistic, tunable error model and attempting to recover the truth. Every linkage decision and every downstream estimate is scored against ground truth.

Corpus. A synthetically generated population of 22,802 records yields a linkage benchmark of 9,383 records (3,339 mother and 6,044 child) with 6,044 known mother-child links.

Linkage strategies compared. (1) A probabilistic engine built on the Fellegi-Sunter model (Expectation-Maximization parameter estimation, a multi-tier Jaro-Winkler comparison on the fuzzy name field, exact-match comparison on household attributes, and multi-pass blocking). (2) Deterministic exact-matching on shared quasi-identifiers, the conventional baseline. (3) The oracle, the true links, as an upper bound. Each strategy is resolved to a one-to-one assignment before scoring, so the reported metrics reflect the actual linkage output.

Bias propagation. A known exposure-to-outcome effect is planted in the cohort. The effect is estimated through each linkage strategy using a modified-Poisson model with cluster-robust standard errors, replicated over 200 Monte-Carlo draws per condition. The bias is also derived analytically, and the simulation is checked against the formula.

3 · RESULTS

73%	36%	2.0 → 1.73
PROBABILISTIC RECALL AT 95% PRECISION	EXACT-MATCH RECALL AT 75% PRECISION	TRUE RR ATTENUATED UNDER EXACT MATCHING

3.1 · Recovery

On the one-to-one linkage output, the probabilistic engine recovered **73% of true links at 95% precision** (F1 0.82), tunable to **99.9% precision at a stricter threshold** (at 58% recall). Deterministic exact-matching recovered only **36% of true links at 75% precision**, because requiring exact agreement on a shared key is fatal when roughly a third of those keys have legitimately diverged.

3.2 · Exact matching degrades with scale

As the population grew, deterministic precision fell sharply and its false-link rate rose to **24.9%**: more unrelated individuals coincidentally share the same surname and postal code, so exact matching does not merely miss true pairs, it fabricates false ones. This failure mode is invisible in small pilot datasets and directly parallels dense registries where many members share surnames and location attributes.

3.3 · The bias

Linkage error attenuates the risk ratio toward the null, monotonically in the false-link rate. A true relative risk of 2.0 was estimated at **1.73** under deterministic matching (a ~13% attenuation), driven by its 24.9% false-link rate, while probabilistic linkage at a calibrated threshold remained effectively unbiased. The simulated attenuation **matched the closed-form analytic prediction** (1.732 observed versus 1.716 predicted), and a built-in null control (a true effect of 1.0) stayed at 1.0, confirming that the method manufactures no spurious signal.

4 · WHY THE RESULT IS TRUSTWORTHY

1. **Known ground truth**, so every error and every bias is measured, not inferred.
2. **Closed-form validation**: the simulation reproduces an independently derived analytic formula, so the numbers are not an artifact of one implementation.
3. **A null control**: a true effect of 1.0 must, and does, stay at 1.0.
4. **Independent methods review**: the analysis design was audited by an independent reviewer before results were generated, and two framing errors (an incorrect prevalence axis and a pooled error-rate metric that conflated the two failure modes) were corrected up front.

5 · WHAT THIS GENERALIZES TO

The study is maternal-pediatric drug safety, but the capability is general: wherever decisions depend on linked, imperfect records, undetected linkage error can silently bias the conclusion, and this framework quantifies that bias before it does. Because the method is developed entirely on synthetic data with known ground truth, it can characterize a linkage pipeline's reliability without exposing sensitive records, a property that matters for privacy-constrained health systems and for any domain where the underlying data cannot be handled in the open.

SCOPE AND LIMITATIONS

This is a methods demonstration on synthetic and public data. Causal claims concern linkage-error-induced bias under known data-generating conditions, not real teratogenicity. The core result addresses non-differential linkage error, which attenuates toward the null precisely because the linkage failure is independent of exposure and outcome by construction; health-dependent (differential) linkage error can bias in either direction and is treated as a boundary condition. Residual finite-sample bias in the ratio estimator is bounded at roughly 0.02 risk-ratio units at the reported cohort scale.

HOW TO CITE

Choi H. Quantifying bias from mother-child record linkage error in pharmacovigilance using synthetic and FAERS public data. Unconfound Labs Technical Methods Paper, 2026. https://www.unconfoundlabs.com/notes/when-the-link-breaks