02 —— THE SHAPE OF A HIDDEN ERROR

When the sample *lies*.

A FIELD GUIDE TO SELECTION BIAS, COLLIDER DISTORTION, AND THE STUDIES THAT MISLEAD US

In 1946, a Mayo Clinic statistician named Joseph Berkson noticed something disturbing in hospital records: diseases that were uncorrelated in the general population appeared linked among admitted patients. The data hadn't lied — the hospital had simply listened to the wrong people. Eighty years later, this single insight underpins a class of error that quietly distorts modern epidemiology, from the obesity paradox to early COVID-19 risk estimates. The mechanism is structural. The fix is unforgiving.

§ 01 A *structural* definition

Selection bias is not a problem of careless sampling. It is a geometric property of the causal graph.

For most of the twentieth century, "selection bias" meant a checklist of failure modes — volunteer bias, response bias, healthy-worker bias, Berkson's bias — each treated as a separate ailment with its own remedy. In 2004, Miguel Hernán, Sonia Hernández-Díaz, and James Robins argued that all of these were one disease wearing different costumes.[1]

> *Bias from selection results from conditioning on a common effect of two variables, one of which is either exposure or a cause of exposure and the other is either the outcome or a cause of the outcome.*
>
> HERNÁN, HERNÁNDEZ-DÍAZ & ROBINS · EPIDEMIOLOGY · 2004

Their structural definition: bias arises from *conditioning on a common effect* of two variables — one a cause of exposure, the other a cause of outcome. Whenever a study restricts itself to a subset of the population (those who were hospitalised, who returned the survey, who survived to follow-up), and that subset is jointly determined by both exposure and outcome, the act of restriction *itself* manufactures association where none existed.

The variable being conditioned on is called a **collider** — two causal arrows converge on it. Conditioning on a collider opens a non-causal path between its parents. The sample, in effect, lies about the population.

§ 02 Anatomy of a *collider*

A causal diagram, before and after the act of selection.

In the **unconditioned** population at left, exposure A and outcome Y are causally independent. They share no common cause; no arrow runs between them. Both, however, influence whether an individual ends up *selected* into the study — admitted to the hospital, returning the survey, surviving to follow-up.

When the analyst restricts the dataset to **S = 1** (the dashed box at right), a non-causal path opens between A and Y. The sample now exhibits an association the population never had. This is collider bias in its purest form — and Berkson's paradox is its oldest example.[3,4]

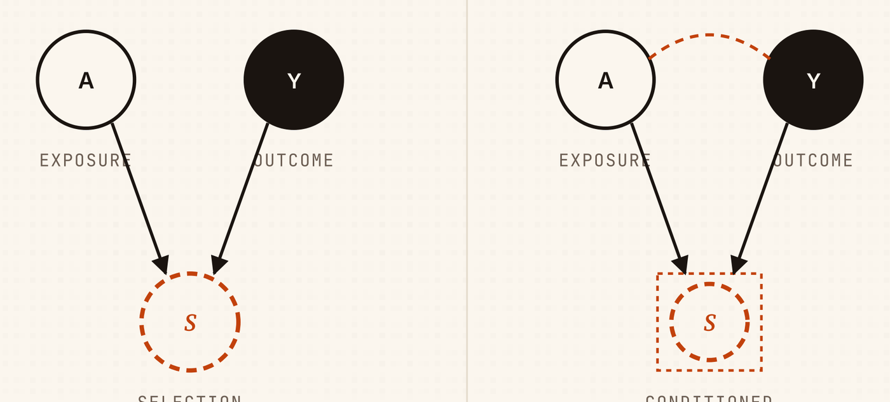


FIG. A · CONDITIONING ON A COLLIDER OPENS A NON-CAUSAL PATH

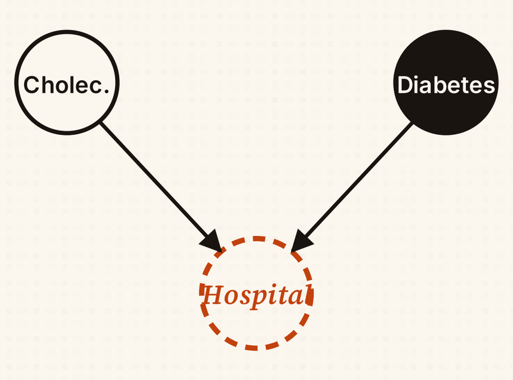
§ 03 Three *cases* from the literature

Different fields, different decades, identical structure: an exposure, an outcome, and a selection node that listens to both.

CASE 01

Berkson's *paradox*

HOSPITAL ADMISSION · 1946



In Berkson's hypothetical population, cholecystitis and diabetes were entirely independent. Among the hospitalised — selected because either condition raised admission probability — a spurious +3.83% positive correlation appeared.[3]

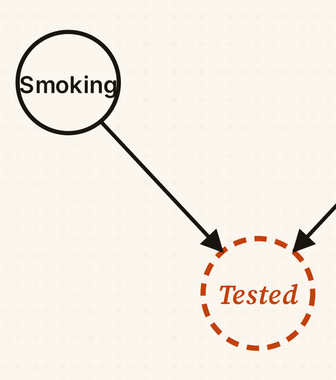
+3.83%

SPURIOUS CORRELATION MANUFACTURED BY SELECTION ALONE

CASE 02

The *smoker's* COVID paradox

GRIFFITH ET AL. · NAT COMMUN · 2020



Early reports showed smokers strangely under-represented in COVID-19 hospitalisations: 5% vs. 27% in the general population. Griffith et al. showed this "protective effect" was an artefact — smokers were selected differently into testing samples.[2]

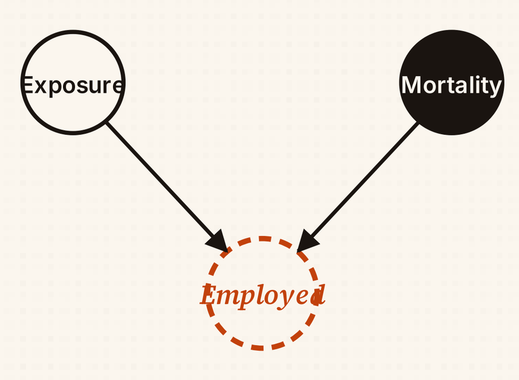
5% / 27%

SMOKING PREVALENCE: HOSPITALISED VS. GENERAL POPULATION

CASE 03

The *healthy* worker

MCMICHAEL · J OCCUP MED · 1976



Workers in occupational cohorts are healthier than the general population — a fact, not an artifact. But comparing their mortality to the population's is a comparison made *after* selection, masking exposure harms by 20–30%.[6,7]

SMR 84

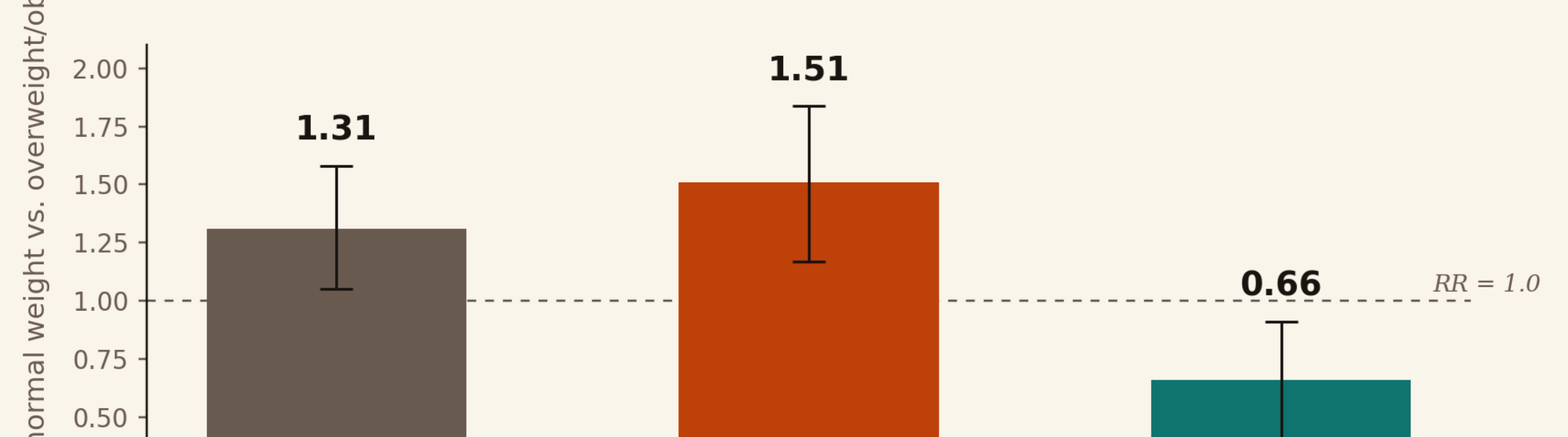
DOLL 1965: UNEXPOSED GAS WORKERS VS. SMR 100 FOR GENERAL POPULATION

§ 04 The *numerical* footprint

When the literature reports a paradox, the structural diagnosis often shows the data simply hadn't been listened to from the right place.

FIG. 1 THE OBESITY PARADOX EVAPORATES

Stratifying by smoking status reverses the apparent mortality benefit of obesity.

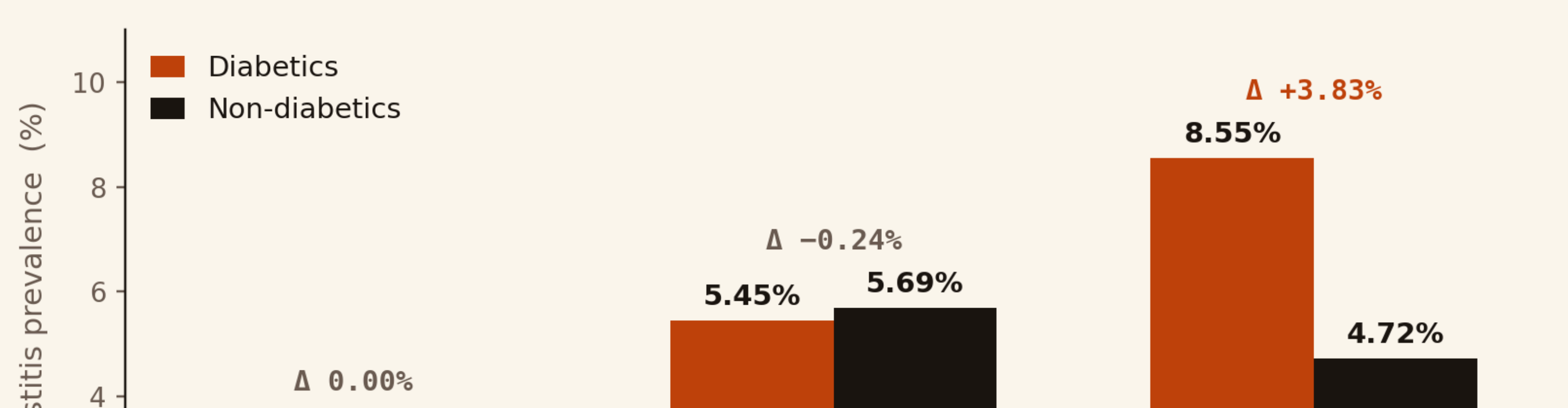


Source: Stokes & Preston, Epidemiology 2014;25(3):454-461. NHANES III + 1999-2004; n=4,642 dysglycemic.

Stokes and Preston re-examined the so-called obesity paradox — the puzzling finding that overweight and obese patients with diabetes appear to outlive their normal-weight peers. Once the analysis stratified by smoking status, the paradox disappeared. Among *never-smokers*, normal-weight patients had **34% lower** mortality risk, exactly the direction biology predicts. The paradox was a collider artefact: smokers are over-represented among lean, ill patients because smoking causes both lower body weight *and* the disease that selects them into the study. [10]

FIG. 2 BERKSON'S NUMERICAL THOUGHT EXPERIMENT

Under zero true correlation, hospital selection alone manufactures a +3.83% spurious association.

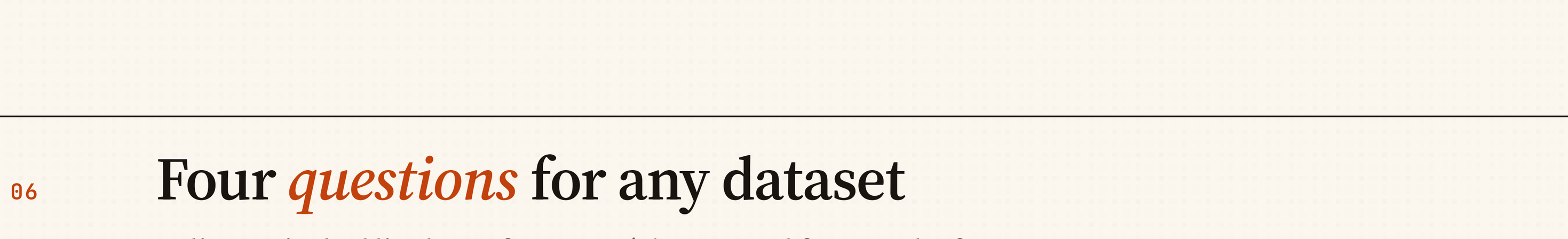


Source: Berkson J. Biometrics Bulletin 1946;2(3):47-53. Hypothetical N = 10,000,000 with independent disease probabilities.

> *Calculating over this parameter space, 40% of all possible combinations lead to an artefactual two-fold protective or risk association operating through this simple model of bias alone.*
>
> — GRIFFITH ET AL. · NATURE COMMUNICATIONS · 2020

§ 05 A *taxonomy* of selection

Sackett (1979) catalogued thirty-five biases. Five of his sampling-stage subtypes recur in the modern literature with surprising regularity.



Codes follow Sackett's appendix taxonomy (Stage 2, sampling).[5]

§ 06 Four *questions* for any dataset

A diagnostic checklist drawn from Hernán's structural framework. If two or more answers are yes, treat the analysis as compromised until proven otherwise.

1 Does selection depend on the *exposure*?

If exposed individuals are more (or less) likely to enter the dataset, the sample's exposure distribution is no longer the population's. By itself this is benign — but it is the first ingredient.

`DAG test: A → S`

2 Does selection depend on the *outcome*?

If sicker (or healthier) participants are over-represented, S becomes a function of Y. Combined with question 1, this is the structural fingerprint of a collider.

`DAG test: Y → S`

3 Are we conditioning on a *collider*?

Restriction, stratification, matching, and adjustment are all forms of conditioning. If a covariate, filter, or strata is a common effect of A and Y, the analysis has opened a backdoor path.

`DAG test: A → S ← Y`

4 Is loss-to-follow-up *differential*?

In cohort studies, dropout is a real-time selection process. If sicker exposed participants leave the study at higher rates, the surviving sample is a collider-conditioned subset. Censoring and selection share the same structure.

`DAG test: A → C ← Y`

§ 07 Three *remedies*

No statistical adjustment can fully retrieve a population from a biased sample. But three approaches, in order of preference, can recover most of what was lost.

REMEDY A · DESIGN

Restriction & sampling

The cleanest fix is to never create the collider in the first place. Population-based cohorts, target trial emulation, and pre-disease enrolment windows avoid the joint dependence of selection on exposure and outcome.

"The optimal way to mitigate the problem is to use appropriate sampling strategies at the study design stage." — Griffith et al., 2020[2]

REMEDY B · STATISTICS

Inverse probability weighting

If selection occurs and its drivers are measurable, weight each participant by the inverse of their selection probability. The weighted pseudo-population approximates the unselected target — provided every common cause of S, A, and Y is observed.

$$W^S = 1 / P(S = 1 \mid A, C)$$

REMEDY C · BOUNDS

Sensitivity analysis

When neither design nor IPW is possible, bound the bias. Smith and VanderWeele's E-value for selection asks: how strong would an unmeasured selection-driver have to be to fully explain away the result? Implausibly strong values lend confidence; modest ones do not.

$$\text{E-value} = RR^S + \sqrt{RR^S(RR^S - 1)}$$

REFERENCES

1 **Hernán MA, Hernández-Díaz S, Robins JM.** A structural approach to selection bias. *Epidemiology.* 2004;15(5):615–625. doi:10.1097/01.ede.0000135174.63482.43

2 **Griffith GJ, Morris TT, Tudball MJ, et al.** Collider bias undermines our understanding of COVID-19 disease risk and severity. *Nat Commun.* 2020;11:5749. doi:10.1038/s41467-020-19478-2

3 **Berkson J.** Limitations of the application of fourfold table analysis to hospital data. *Biometrics Bulletin.* 1946;2(3):47–53. Reprinted: *Int J Epidemiol.* 2014;43(2):511–515. doi:10.1093/ije/dyu022

4 **Westreich D.** Berkson's bias, selection bias, and missing data. *Epidemiology.* 2012;23(1):159–164. doi:10.1097/EDE.0b013e31823b6296

5 **Sackett DL.** Bias in analytic research. *J Chronic Dis.* 1979;32(1-2):51–63. doi:10.1016/0021-9681(79)90012-2

6 **McMichael AJ.** Standardized mortality ratios and the "healthy worker effect": Scratching beneath the surface. *J Occup Med.* 1976;18(3):165–168. doi:10.1097/00043764-197603000-00009

7 **Shah DH.** Healthy worker effect phenomenon. *Indian J Occup Environ Med.* 2009;13(2):77–79. doi:10.4103/0019-5278.55123

8 **Banack HR, Kaufman JS.** The "obesity paradox" explained. *Epidemiology.* 2013;24(3):461–462. doi:10.1097/EDE.0b013e318290abe7

9 **Banack HR, Kaufman JS.** Does selection bias explain the obesity paradox among individuals with cardiovascular disease? *Ann Epidemiol.* 2015;25(5):342–349. doi:10.1016/j.annepidem.2015.02.008

10 **Stokes A, Preston SH.** Obesity paradox: Conditioning on disease enhances biases in estimating the mortality risks of obesity. *Epidemiology.* 2014;25(3):454–461. doi:10.1097/EDE.0000000000000075

11 **Hernán MA, Robins JM.** *Causal Inference: What If.* Boca Raton: Chapman & Hall/CRC; 2020. Chapter 8: Selection bias. hsph.harvard.edu/miguel-hernan/causal-inference-book

12 **Smith LH, VanderWeele TJ.** Bounding bias due to selection. *Epidemiology.* 2019;30(4):509–516. doi:10.1097/EDE.0000000000001032

13 **Levy NS, Keyes KM.** Causal knowledge as a prerequisite for interrogating bias: Reflections on Hernán et al. 20 years later. *Am J Epidemiol.* 2021;192(11):1797. doi:10.1093/aje/kwab274