This was published in *Chance*, 2018, 31(4), 16-23.

A large biased data system may contain within it
a smaller, much less biased natural experiment.
Isolation finds or constructs the natural experiment.

A simple example of isolation in building a natural experiment

José R. Zubizarreta, Dylan S. Small, and Paul R. Rosenbaum[1]

What is isolation?

Large data systems increasingly record outcomes and exposures to treatments, but often the occurrence of an exposure/outcome event, the timing of the event, the physical location of the event, and the attending circumstances are biased in ways that cannot be controlled using data recorded by that system. In our example of isolation — an especially simple example — the data are from the US Fatality Analysis Reporting System (FARS) which records information about motor vehicle accidents with at least one fatality. The system records interesting information about such accidents, information about safety equipment, safety belt use, airbag

[1]José R. Zubizarreta is assistant professor of statistics in the Department of Health Care Policy of the Harvard Medical School. Dylan S. Small and Paul R. Rosenbaum are professors of statistics at the Wharton School of the University of Pennsylvania. 2 May 2018.

deployment and much more. Arriving after the fact, the system does not record reliable information about speeds, the initial distance between vehicles before one vehicle started to brake, efforts the drivers made or failed to make to limit the extent of a collision, the promptness and effectiveness of braking, these being important determinants of the forces involved in the crash. In a crash, as in any physical event, forces matter. If Volvo station wagons are driven more safely than Dodge Challengers, at slower speeds, with fewer high-speed lane changes, and at a greater distance from the car ahead, then none of this is recorded in FARS, so superior safety equipment in the Volvo may be credited with effects actually produced by safer driving. If convertible coups are more often driven in weather with excellent road traction, while four-wheel drive SUVs are driven with less attention to weather, then convertibles may appear safer than they would be if driven with poor traction and hence worse braking. Importantly, FARS records nothing about the crash that did not occur because of anti-lock brakes, nor the crash in which all fatalities were prevented by a treatment under study, say safety belts and airbags. Every crash in FARS in which a lone driver hit a tree is a lethal crash, not because trees are so dangerous, nor because safety equipment is helpless against trees, but because, in these circumstances, only a lethal crash would be recorded in FARS. In FARS, there are many obstacles to judging the crashworthiness of a Volvo station wagon, or a Ford F450 pickup truck, because it is difficult to separate the vehicles from the way people drive them; however, these are not insurmountable obstacles.

Within a large but biased data system, isolation means constructing or finding a smaller set of comparisons that are likely to be much less biased. Inside a biased

data system, there may be a fairly large and not very biased natural experiment. One isolated comparison found, not constructed, in FARS compares a driver and a right-front passenger in the same car, one wearing a safety belt, the other unbelted, thereby assessing the effects of safety belts while controlling unmeasured speeds, braking, road traction, and some consequent forces; see Evans (1986) and Rosenbaum (2015). Another isolated comparison assesses the effects of motorcycle helmets by comparing two riders of the same motorcycle, one with a helmet, the other without; see Norvell and Cummings (2002).

> Isolation means that the occurrence of a treatment/outcome event, its timing, its location, its circumstances may be severely biased, but conditionally given that one treatment is given in lieu of another at that time, in that place, and under those circumstances, the conditional distribution of exposure to these two treatments is either not biased or much less biased. Zubizarreta et al. (2014, §2.3) show this formally: certain types of unmeasured biases in the hazard of treatment drop out when conditioning on the receipt of one treatment in lieu of another.

Found comparisons are immediately easy to appreciate, but the possibility of optimal construction substantially enlarges the scope of isolation. An example follows. Usually, the effect of an additional child on a woman's career is difficult to assess, because many women are active in controlling their fertility to fit their career plans. That control may be lost when a mother finds herself pregnant with twins, and a study by Joshua Angrist and William Evans used this fact to obtain estimates

of the effects of an additional child on a woman's career. Building upon their idea, Zubizarreta, Small and Rosenbaum (2014) matched mothers with the same past history of fertility and education at the moment that one mother had twins and her five matched controls had single births.

Stated more precisely, isolation refers to a *differential comparison* in a *risk-set match*; see Zubizarreta et al. (2014). A *differential comparison* compares one treatment given in lieu of another treatment. In an isolated comparison, a risk-set match compares people who received one treatment in lieu of another at one of those rare and brief moments when the biases that promote one treatment and not the other are absent. Perhaps unworn safety belts are more common in recklessly driven cars, but that matters much less, if at all, given two people in the front seat of the same car in the same crash, one belted, the other unbelted. Perhaps wearing a motorcycle helmet and riding cautiously typically go together, but that matters much less, if at all, given two people on the same motorcycle, one with a helmet, the other without. Perhaps a woman's plans for career and family strongly bias the occurrence and timing of births, but that matters much less, if at all, when comparing two pregnant women with the same history of fertility and education, one currently having twins, the other currently having a single child.

Outside of randomized experiments, the effects caused by treatments are difficult to see, shrouded in fog. For brief moments, on rare occasions, the fog lifts; an equitable comparison emerges from the mist. Isolation refers to noticing and collecting the brief and rare occasions, separating them from everything else.

A footnote concerning which cars are safe in crashes

The Insurance Institute for Highway Safety (IIHS) conducts experimental crash tests in which cars strike fixed objects, reporting results on its webpage. We applaud their valuable efforts, but we consider an aspect that they mention but do not emphasize. They wrote: (http://www.iihs.org/iihs/ratings/TSP-List, visited 5 August 2017)

> IIHS conducts vehicle tests to determine crashworthiness — how well a vehicle protects its occupants in a crash. ... Models that earn Top Safety Pick+ or Top Safety Pick are the best vehicle choices for safety within size categories. Size and weight influence occupant protection in serious crashes. Larger, heavier vehicles generally afford more protection than smaller, lighter ones. Thus, a small car that's a Top Safety Pick+ or Top Safety Pick doesn't necessarily afford more protection than a bigger car that doesn't earn the award.

The role of vehicle mass is often noted. The IIHS page then gives "Top Safety Pick" awards to two "minicars," a 2017 Mini Cooper and a 2017 Toyota Yaris, and to one "large pickup," the 2017 Honda Ridgeline. Other types of cars, minivans and SUV's also won awards. The Ford F450 pickup won no award for safety, but has a shipping weight that can be more than three times the weight of a Mini Cooper. How much does that matter?

Is safety within automotive category an important indicator of safety when vehicles in different categories collide? Which vehicles are safer in accidents involving two vehicles? There are many interesting questions about automotive safety, and FARS

can speak to some of them. This particular question is not about what causes vehicles to collide, nor about what causes them to collide with particular forces; rather, it asks: What happens conditionally given that two specific vehicles from different categories did collide with the forces that were actually present? What is the differential effect of being in one type of vehicle in lieu of another type given an accident involving these two vehicles in particular, largely shared, circumstances? Many things that might bias a comparison between different accidents are made similar by comparing two vehicles in the same accident.

What is the differential effect of being in one type of vehicle in lieu of another type given an accident involving these two vehicles in particular, largely shared, circumstances?

Perhaps convertible coups are more often driven on sunny days with excellent road traction, while four-wheel drive SUV's are often driven in snow, so a collision between a convertible and an SUV is likely to occur with better braking than many collisions between two SUV's. However, conditionally given that a convertible and an SUV have collided, with the forces that did exist in that crash, what is the differential effect of being in the convertible rather than the SUV? The variation in forces between different collisions at different times and places may be biased by weather conditions and driving behaviors that increase or reduce the hazard of certain types of accidents. However, in any collision, the opposing forces are dissipated, perhaps in a complex way, in bringing the two colliding vehicles to a halt. Think of adding opposed force vectors in high school physics. Perhaps Volvo station wagons are typically driven

more safely than pickups, at slower speeds with greater distance from the car ahead, so perhaps the typical accident involving a Volvo wagon is a lower speed, lower force event than the typical accident involving a pickup. However, this matters much less, if at all, conditionally given that the Volvo and the pickup have collided in a particular way. What is the differential effect of being in the Volvo wagon rather than the pickup in this accident as it actually happened?

> At any moment while driving, each of us has a momentary risk or hazard of being involved in a lethal crash, a risk that varies with time, weather, driving behavior and many other factors. As a safe driver, your hazard of being in a lethal crash may be lower than mine, and you may be more likely than I to be in a slower speed, lower force crash, but all that matters less, if at all, conditionally given that we have collided in a single crash. Before the fact, our hazards of being in a crash of this kind may have been very different, but all this matters less conditionally given that this crash has occurred.

Is there evidence of bias in which crashes occur under what circumstances? Is there evidence that it is unwise to look at vehicles one at a time? Indeed there is. The biases we can see are less worrisome than the biases we cannot see, because we can often correct for biases we can see, but the biases we can see suggest such biases are commonplace. Later, we examine 2923 crashes involving two moving vehicles, one a car, the other a light truck. Of the 2923 crashes, 64 crashes involved a convertible car, but only 4 crashes involved a convertible car in the winter in the

northern US — specifically, in November, December, January and February, at or above latitude 40 degrees or north of Philadelphia. In these same 2923 crashes, pickups were 75% more likely than SUVs to be in a fatal crash on a country road (an odds ratio of 1.75); however, presumably this is because of where pickups tend to be driven, not because of their unusually poor performance on country roads. In contrast, within a single crash involving two vehicles, many things, measured or unmeasured, are typically the same or similar for the two vehicles: ambient light, weather, location, road traction, inter-vehicle distance prior to braking, and much more.

Because different types of vehicles, say a Volvo station wagon and a Ford F450 pickup, are driven by different people, in different ways, under different circumstances, the type of vehicle may predict both the absolute risk of a crash and unmeasured aspects of a crash, such as the distance between vehicles prior to the start of any braking and the reduction in relative speed due to braking. However, given that a Volvo station wagon and a Ford F450 pickup have collided, the type of vehicle predicts fewer features of the crash; for instance, the distance between the two vehicles before any braking is the same.

Isolated comparisons of crash safety

Description of the comparison

Using data from the 2015 FARS, we examine lethal accidents involving exactly one car and one so-called "light truck". A light truck is an SUV, minivan or pickup. Each individual accident, each such isolated comparison, produces its own 2×2 table of counts, car or light truck, survived or died. There are 2,923 such 2×2 tables, and Table 1 displays five of the 2×2 tables.

For both light trucks and cars in this comparison, the minimum and median number of occupants was 1, the upper quartile was 2, but the maximum for light trucks was 12 and for cars was 7.

A few details of the comparison follow. We consider only accidents in which there are exactly two vehicles (VE_TOTAL==2), both of which are in transit — e.g., not parked — (VE_FORMS==2), and the only people involved in the crash were in one of these two vehicles — no pedestrians or bicyclists — (PERNOT-MVIT==0). Additionally, there had to be data in FARS about both vehicles. A few two-vehicle accidents were not collisions — perhaps a vehicle went off the road to avoid a collision — but we did not remove these from the analysis. The vehicles had to be either passenger cars or trucks (VEHTYPE_T=="Passenger Car" or VEHTYPE_T=="Truck"), thereby, for instance, excluding motor cycles. Finally, attention was restricted to cars and light trucks by requiring BODYSTYL_T to be one of the following categories, "Convertible", "Coupe", "Hatchback", "Pickup", "Sedan", "Sport Utility Vehicle", "Van Cargo", "Van Passenger", "Wagon". Then

light trucks were defined to have BODYSTYL_T of "Pickup","Sport Utility Vehicle", "Van Cargo", or "Van Passenger", the rest being cars. By definition, every 2×2 table contains at least one person in a light truck, at least one person in a car, and at least one death, but there were no survivors in 68 of the 2,923 accidents.

Our comparisons illustrate the statistical concept of isolation, but do not endorse or critique particular vehicles or brands, for which numerous considerations are relevant.

Does it matter that FARS only records lethal crashes?

An accident is in FARS only if someone died. The FARS system attempts to describe every crash in which someone died, and suppose for this discussion that it was successful in that attempt. How does our $2 \times 2 \times 2923$ table from FARS relate to the much larger $2 \times 2 \times K$ table that records all accidents involving one car and one light truck in 2015? The larger table has $K - 2923$ subtables in which no one died. That is, the 2×2 tables not in FARS differ from those in Table 1 in that each column total in the second or "Died" column is zero.

The best test of no association between vehicle type and mortality within crashes — more precisely, the uniformly most powerful unbiased test against a constant odds ratio not equal to 1 in a $2 \times 2 \times K$ table — is a test that conditions on the four row and column totals in each 2×2 subtable, the so-called marginal totals, and the familiar Mantel-Haenszel test is one of several large sample approximations to this best, exact test; see Birch (1964). In both the exact test and its large sample approximation, a table with a zero in the row or column totals does not contribute to the test. The

reason such a table does not contribute is that, if there is a zero in the margin, then there is only one 2×2 table with the same margins as this table, so conditioning on the margins has fixed the entire table, so it is constant and not random. Intuitively, a 2×2 table in which no one died may be a crash in which slow speeds and successful braking limited the forces involved in the crash, and such a crash provides little or no information about crashworthiness and mortality. In parallel, tables with a zero marginal total do not affect the conditional maximum likelihood estimate of a constant odds ratio or of coefficients describing the dependence of odds ratios on covariates. The conditioning in the $2 \times 2 \times K$ tables is the simplest case of the more general type of conditioning in Zubizarreta et al. (2014, §2.3).

In brief, FARS provides all of the 2923 tables with at least one death, omits the $K - 2923$ tables without a death, but the omitted tables would not have affected the conditional analyses, even had they been present. The FARS system is not a census of US vehicles in 2015 — presumably, vehicles driven by cautious drivers are underrepresented — nor is it a census of vehicle crashes in 2015 — presumably, less forceful crashes are underrepresented; however, the $2 \times 2 \times 2923$ table is essentially a census of lethal crashes involving exactly one car and one light truck, so it is a census of the informative part of the $2 \times 2 \times K$ table recording all such crashes.

Cars versus light trucks

When a car and a light truck collide, who dies? We express the answer in terms of odds ratios. If you flip a fair coin and roll a fair die, then the odds of a "head" on the coin are 1-to-1 or 1/1, and the odds of "6" on the die are 1-to-5 or 1/5, so the

odds ratio is $(1/1)/(1/5) = 5$, so the "head" is 5-times more likely than the "6"; its odds are 5-times greater. (We calculate odds ratios within crashes using the `clogit` function in the `survival` package in `R`, and we report 95% confidence intervals, CI, for odds ratios within parentheses.)

In the $2 \times 2 \times 2923$ table of 2923 lethal crashes of a car and a light truck, the estimated odds ratio linking occupancy in the car with death is 4.76 (with 95% CI [4.34, 5.23]). In brief, the odds of death are almost five times higher in the car than in the light truck.

Safety belts matter also. We distinguish four types of restraints: lap/shoulder belts (71.7%), child restraints (3.0%), no belt restraint (21.5%) and other (3.9%). Being unrestrained rather than wearing a lap/shoulder belt was associated with a 6.19-fold higher odds of death. Taking this into account — taking into account both restraint use and vehicle type — being in a car rather than a light truck was even more strongly associated with death: the odds ratio is 5.35, rather than 4.76, (with 95% CI [4.79, 5.97]).

Of the 2923 lethal crashes involving a car and a light truck, there were 14 crashes in which the car was one of the two brands of minicars mentioned by the IIHS, the Mini Cooper or the Yaris, though the 14 cars were not the specific years and models IIHS recommended. There were 48 people involved in these 14 crashes, of whom 14 died. All 14 deaths occurred in the minicars, although 8 occupants of minicars survived. Using the exact test of Birch (1964), the two-sided P-value testing no association within crashes is 0.0000643.

Lighter cars versus heavier light-trucks

Some vehicles are labeled light trucks because of their body styles, but are not hefty vehicles. For instance, the Chrysler PT Cruiser is nominally a passenger van, a type of light truck, but at 3165 lbs, it weighs somewhat less than a Honda Accord, and the same is true of some Ford Ranger pickups. Additionally, some of the larger cars from Lincoln and Mercedes-Benz weigh as much as the typical light truck, about 4450 lbs. The comparison in Figure 1 is not quite the intended comparison of hefty SUV's with commuters' cars.

The 2923 crashes in Figure 1 are divided roughly in half, into 1565 crashes on left in Figure 2, and 1358 crashes on the right in Figure 2. Figure 2 labels these as typical or atypical, but it would be more correct to speak of prototypical and ambiguous. The division is based on the dashed lines in Figure 1, at the upper quartile for cars and the lower quartile for light trucks. On the left in Figure 2 are 1565 crashes of a car whose weight is below the upper quartile for cars with a light truck whose weight is above the lower quartile for light trucks. On the left in Figure 2, the median car weights about 3000 lbs., and the median light truck weighs about 4750 lbs, about 50% more. On the right in Figure 2 are the remaining 1358 crashes: a heavier car, a lighter light truck, or both, or a missing vehicle weight. Although the situations on the left and right of Figure 2 are about equally common in FARS, the situation on the left is the intended or prototypical situation, in which a car that is not especially large collides with a light truck that is not especially small. A collision between a Lincoln Towncar and a Chrysler PT Cruiser would be on the right of Figure 2, and it would roughly reverse the median weights on the left, 3000

versus 4500. If you were trying to decide, with safety in mind, between a hefty SUV and a small or mid-sized car, then the left side of Figure 2 is more relevant than Figure 1.

For the comparison on the left in Figure 2, the odds ratio is 8.10 (with 95% CI [6.99, 9.39]), while on the right it is 2.91 (with 95% CI [2.57, 3.30]). Taking account of safety belt use, as above, yields on the left an odds ratio of 9.16 (with 95% CI [7.70, 10.90]), while on the right it yields an odds ratio of 3.24 (with 95% CI [2.80, 3.75]). When a hefty SUV, van or pickup collides with a small or mid-sized car, the odds of death are about 8 or 9 times higher in car.

Could it be that robust, survival prone individuals buy pickups, and frail, mortality prone individuals buy small and mid-sized cars, so that the results above are entirely due to bias in who buys light trucks, not an effect caused by the differences in mass in Figure 2? Actually, the comparison on the left in Figure 2 is insensitive to large unmeasured biases. It is easiest to think about a matched pair, a car and a light truck whose only occupants are their drivers. Of the 2923 crashes, 1129 were matched pairs. Consider the magnitude of unmeasured bias needed to explain away the survival difference between hefty light trucks and conventional cars on the left of Figure 2. In a matched pair, to explain away the survival difference, that bias would increase the odds of buying a light truck 13-fold, and increase the odds of death 15-fold; see Rosenbaum (2017, Table 9.1, with $\Gamma = 7$). To explain the difference in survival in terms of the unmeasured frailty of people who buy cars instead of pickups, there would have to be an enormous difference in frailty closely tracking who buys what.

Some Reminders, Considerations and Limitations

- Sometimes, the timing and occurrence of a treatment comparison — its hazard — is biased by unobserved covariates, but some of these biases may briefly vanish in a differential comparison of two treatments conditionally given that a particular comparison has occurred at a particular time. This is isolation. Cautiously driven cars may typically be involved in fewer, slower, less forceful crashes, but this matters less conditionally given that two particular vehicles have collided. Isolation may occur naturally, scattered throughout an otherwise biased data system, or isolation may be deliberately constructed by the analyst; see the Further Reading for an example of a construction.

- Isolation may remove some important unmeasured biases, but cannot be relied upon to remove all unmeasured biases. Differential biases may remain. Perhaps cautiously driven cars are typically struck, while recklessly driven cars typically strike, so that isolated comparisons may remain somewhat biased. For this reason, analytical adjustments and sensitivity analyses are needed to address any differential biases that may remain; see the Further Reading.

- Isolation attempts to find a natural experiment inside a population, but it does not attempt to represent or describe the population as it naturally occurs. Experiments that estimate causal effects and surveys that faithfully describe populations are different things, and they can be biased in different ways. Because certain types of cars may often be driven more cautiously than others, equitable comparisons of crash safety may occur only in atypical crashes, that

is, crashes that are unrepresentative of the population. If Volvo station wagons are typically driven more safely than large pickups, then it might be wise to charge less to insure the station wagon, even though you might be safer in the large pickup conditionally given a collision between these two vehicles. If the population does not resemble an equitable experiment, then an equitable natural experiment will not resemble the population.

Conclusion

A large but biased data system may contain within it a smaller, much less biased, natural experiment. Some isolated comparisons occur naturally and simply need to be noticed and separated from other, more numerous, but biased comparisons. Other isolated comparisons combine a random element that needs to be noticed and a construction that pulls together similar situations in which that random element picked one treatment in lieu of another.

Further Reading

Angrist, J. D. and Evans, W. N. (1998). Children and their parents' labor supply. *American Economic Review*, 88, 450-477.

Birch, M. W. (1964). The detection of partial association, I: The 2 x 2 case. *Journal of the Royal Statistical Society*, B, 313-324. Develops the optimal test for a $2 \times 2 \times K$ table.

Evans, L. (1986). The effectiveness of safety belts in preventing fatalities. *Accident*

Analysis and Prevention, 18, 229-241. Compares belted and unbelted individuals in the front seat of the same car in the same crash.

National Highway Traffic Safety Administration (2015) *Fatality Analysis Reporting System*. www.nhtsa.gov/research-data/fatality-analysis-reporting-system-fars

Norvell, D. C. and Cummings, P. (2002). Association of helmet use with death in motorcycle crashes: a matched-pair cohort study. *American Journal of Epidemiology*, 156(5), 483-487. Compares two people riding the same motorcycle, one with a helmet, the other without.

Rosenbaum, P. R. (2006). Differential effects and generic biases in observational studies. *Biometrika*, 93, 573-586. Observes that the differential effect of one treatment in lieu of another is immune to certain biases, called generic biases, that promote both treatments. Includes discussion of sensitivity analysis for differential biases.

Rosenbaum, P. R. (2015). Some counterclaims undermine themselves in observational studies. *Journal of the American Statistical Association*, 110, 1389-1398. Compares belted and unbelted individuals in the front seat of the same car in the same crash.

Rosenbaum, P. R. (2017). *Observation and Experiment: An Introduction to Causal Inference*. Cambridge, MA: Harvard University Press. Isolation is discussed briefly in Chapter 12.

Zubizarreta, J. R., Small, D. S. and Rosenbaum, P. R. (2014). Isolation in the construction of natural experiments. *Annals of Applied Statistics*, 8, 2096-2121. Introduces the concept of isolation and develops its key properties. Illustrates the

active construction of isolated comparisons for the Angrist/Evans study. Includes discussion of sensitivity analysis for differential biases.

Table 1: Five of the 2,923 tables recording mortality in crashes of one car and one light truck. Weight refers to shipping weight, that is, the weight of the vehicle without passengers or cargo. The ID number is STCASE from FARS.

ID	Type	Mortality		Total	Make/Model	Weight (lbs)
		Survived	Died			
10060	Light Truck	2	0	2	Honda Odyssey	4365
	Auto	4	1	5	Ford Escort	2419
	Total	6	1	7		
		Survived	Died			
10114	Light Truck	0	1	1	Ford Explorer	3952
	Auto	1	0	1	Acura TL	3623
	Total	1	1	2		
		Survived	Died			
10129	Light Truck	1	0	1	Nissan Pathfinder	3961
	Auto	1	1	2	Pontiac Gand Am	3102
	Total	2	1	3		
		Survived	Died			
10130	Light Truck	2	0	2	GMC Yukon	4947
	Auto	0	1	1	Pontiac Bonneville	3633
	Total	2	1	3		
		Survived	Died			
10195	Light Truck	1	0	1	GMC New Sierra	5284
	Auto	0	2	2	Ford Fusion	3526
	Total	1	2	3		