

Statistical Literacy: Simpson's Paradox and Covid Deaths

Milo Schield, University of New Mexico

Abstract: Those vaccinated were more likely to die than those unvaccinated among UK Covid Delta cases. Fortunately the data was broken out by age so that one could easily see that the reverse was true for both age groups. This is a classic case of Simpson's paradox. But simply showing the existence of Simpson's paradox does not explain why it occurs. And it doesn't provide any way to solve or resolve the problem. Finally, students need to be able to describe all this using ordinary English. This paper addresses each of these issues.

INTRODUCTION

Confounding is arguably the #1 problem in dealing with observational data. Yet, it is "the elephant in the room": it is all but absent from our intro statistics courses.

Simpson's paradox is the most extreme form of confounding: the association of rates in all the subgroups have the opposite direction or sign as the same association involving the entire group.

Consider the following data on UK Covid Delta cases between February and early August, 2021. The following excludes the Unlinked Cases. See Appendix.

Figure 1: Covid Delta Cases, Deaths and Death Rates for Vaccinated and Unvaccinated

Population Group	---- Covid Delta Cases ----	
	Vaccinated	Unvaccinated
Cases	117,114	151,054
Deaths	481	253
Mortality Rate	0.41%	0.17%
Risk Ratio (Vac/UnV)	245.2%	

Looking at this table, students can see that the mortality rate for the vaccinated is higher than that for the unvaccinated. This is certainly unexpected.

The risk ratio of 245.2% is confusing. It isn't a percent compare (0.41% is 245.2% more than 0.17%)? It isn't a part-whole compare; those are never more than 100%. It is a simple ratio expressed as a percentage: 0.41% is 245.2% of 0.17%. Percentage are typically used for simple ratios that seldom exceed 100%. So, convert the 245.2% to 2.45.

Statistically-literate students should be able to describe this two group arithmetic association.

In 2021, among those in the UK who were Covid Delta cases, the vaccinated were 2.45 times as likely to die as [were those who were] the unvaccinated.

In 2021, among those in the UK who were Covid Delta cases, dying was 2.45 times as likely among those who were vaccinated as among those who were unvaccinated.

What could explain this 0.24 point difference? This large difference is unlikely to exist if due to randomness.¹ It could be due to confounding. Is this association a mixed-fruit comparison: an apples and oranges comparison? If so, it could be true, but still be very misleading.

¹ 95% margin of error: $2 \cdot \sqrt{p/n} = 2 \cdot \sqrt{0.41\%/117,000} = 0.037\%$. 0.24 pt. difference is statistically significant.

What might confuse this comparison? The simplest confounder is age. Fortunately the data was broken out by age into two age groups as shown in Figure 2.

Figure 2: Covid Delta Cases, Deaths and Death Rates for Vaccinated and Unvaccinated by Age Group

Population	---- Delta Cases <50 ----		---- Delta Cases >=50 ----	
Group	Vaccinated	Unvaccinated	Vaccinated	Unvaccinated
Cases	89,807	147,612	27,307	3,440
Deaths	21	48	460	205
Mortality Rate	0.02%	0.03%	1.68%	5.96%
Risk Ratio (Vac/UnV)	71.9%		28.3%	

These risk ratios are less than 100%. Using just the direction, statistically literate students should be able to form an ordered (ordinal) two-group comparison for both age groups:

For those under 50, dying was more likely among the unvaccinated (0.03%) than among the vaccinated (0.02%). For those under 50, those who were unvaccinated were more likely to die than those who were vaccinated.

For those who were at least 50, dying was more likely among the unvaccinated (5.96%) than among the vaccinated (1.68%). For those who were at least 50, those who were unvaccinated were more likely to die than those who were vaccinated.

Students who had never heard of Simpson's paradox might be very confused. How can the vaccinated have a lower death rate than the unvaccinated for both age groups but have a higher death rate overall?

It's as though a football team won the first half and the second half of a game, but lost the game. That is impossible with counts. But it is quite possible with rates or ratios.

So, students may see Simpson's paradox, but they don't understand how or why it occurs, and they don't understand how to untangle this very confusing situation.

RECASTING THE DATA

To help students really understand what is happening, we need to do seven things:

1. Form risk ratios that are greater than unity.
2. Eliminate the deaths. Keeping the cases and death rates is sufficient.
3. Create total data: cases, deaths and death rates.
4. Write a two-group arithmetic comparison for each of the three tables.
5. Calculate and compare the prevalence of vaccinated in each age group
6. Standardize on group prevalence of vaccinated.
7. Write a two-group arithmetic comparison for the standardized association

The first three steps generate the table in Figure 3

Figure 3: Covid Delta Cases and Death Rates for Vaccinated and Unvaccinated by Age Group

Population	----- Covid Delta Cases -----			
Group	Vaccinated	Unvaccinated	Total	
Cases	117,114	151,054	268,168	
Mortality Rate	0.41%	0.17%	0.27%	
Risk Ratio (Vac/UnV)	2.45			
Population	----- Delta Cases <50 -----		----- Delta Cases >=50 -----	
Group	Vaccinated	Unvaccinated	Vaccinated	Unvaccinated
Cases	89,807	147,612	27,307	3,440
Mortality Rate	0.02%	0.03%	1.68%	5.96%
Risk Ratio (UnVac/V)	1.39		3.54	

Step 4: Write a two group arithmetic comparison for each of the three tables.

Among all Delta cases, vaccinated are 2.5 times as likely to die as unvaccinated.

Among Delta cases under 50 years old, unvaccinated are 1.4 times as likely to die as are vaccinated.

Among Delta cases age 50 and up, unvaccinated are 3.5 times as likely to die as are vaccinated.

These relative risks are all less than four, so they are somewhat vulnerable to the influence of confounders.² But our minds want to know what explains this Simpson's paradox reversal.

Step 5. Calculate the prevalence of vaccinated in each age group: the right column in Figure 4.

Figure 4: Covid Delta Death Rates by Vaccinated and Age Group along with the Prevalence of Seniors

	Crude			Number of Cases			----Weights----	
Death rates	<50	50+	All	<50	50+	All	<50	50+
Un-vac	0.03%	5.96%	0.17%	147,612	3,440	151,054	0.977	0.023
Vaccinated	0.02%	1.68%	0.41%	89,807	27,307	117,115	0.767	0.233
			2.47	237,419	30,747	268,169	0.885	0.115

Describe the two-group association involving seniors arithmetically:

“Seniors (age 50 and up) are 10 times as prevalent among the vaccinated as among the unvaccinated.”

This factor of 10 disparity is what makes the original association a crude association: a mixed fruit comparison or an apples and oranges comparison.

Step 6. Standardize on group prevalence of vaccinated.

Our minds look for solutions to problems. Can we take into account (control for) this confounding disparity -- without a computer? Yes! Use standardizing: a weighted average.

What would the death rates be if both groups had the same mixture of seniors? The only thing different is the far right column. In this case, both groups used the combined mixture. The

² Schield (2018). Confounding and Cornfield; Back to the Future. www.StatLit.org/pdf/2018-Schild-ICOTS.pdf

calculation of these standardized averages is shown in the lower-right corner of the figure. Now the ratio of the standardized death rates is 3.4 for unvaccinated versus vaccinated.

Figure 5: Covid Delta Crude and Standardized Death Rates by Vaccinated and Unvaccinated

Death rates	Crude			Number of Cases			----Weights ----			Standard
	<50	50+	All	<50	50+	All	<50	50+	All	
Un-vac	0.03%	5.96%	0.17%	147,612	3,440	151,054	0.977	0.023	0.71%	
Vaccinated	0.02%	1.68%	0.41%	89,807	27,307	117,115	0.767	0.233	0.21%	
			2.47	237,419	30,747	268,169	0.885	0.115	3.38	
Crude Comparison: mixed-fruit comparison						Standardized: Both groups have same mix				
0.17% = 0.977*0.03% + 0.023*5.96%						0.71% = 0.885 *0.03% + 0.115 *5.96%				
0.41% = 0.767*0.02% + 0.233*1.68%						0.21% = 0.885 *0.02% + 0.115 *1.68%				
50+ are 10 times as prevalent among the vaccinated (23%) as among the unvaccinated (2.3%).										
https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1009243/Technical_Briefing_20.pdf										

Step 7: Write a two-group arithmetic comparison of the standardized association:

Among UK Covid Delta cases, unvaccinated are 3.4 times as likely to die as are vaccinated *after controlling for age*.

STUDENT LEARNING OUTCOMES

Statistically-literate students should be able to:

- recognize Simpson’s paradox
- describe it using ordinary English
- recognize that it may be a crude comparison
- calculate the appropriate weights for a measured confounder
- calculate adjusted weighted-averages
- describe the results in ordinary English
- understand “control for” or “take into account”

In this case study, students can see how a crude comparison (a mixed fruit comparison) can be converted into an adjusted comparison (an apples and apples comparison). Now, teachers can put these problems on a test and expect students to work out the standardized results.


A CONFOUNDER-BASED STATISTICAL LITERACY COURSE

Statistical educators should offer a new confounder-based statistical literacy course that:

- asserts that *Association is Not Causation*; that *Disparity is Not Discrimination*
- focuses on the *Story Behind the Statistics*
- shows how a *crude association* (mixed fruit comparison) may conceal the real story
- shows students how to *control for* confounders
- shows students these things *without computers*

The University of New Mexico is now offering such a course: over 100 students now enrolled.

Figure 6: Univ. of New Mexico MATH 1300 Statistical Literacy Catalog Description



Statistical Literacy

NM UNIVERSITY CATALOG

MATH 1300 (3)
Participants will study the social statistics encountered by consumers. Investigate the story behind the statistics. Study the influences on social statistics. Study the techniques used to control these influences. Strong focus on confounding.

Meets New Mexico General Education Curriculum Area 2: Mathematics and Statistics.

RECOMMENDED ARTICLES:

Statistical educators should study confounder-based statistical literacy. Here are some articles that focus on confusing results – such as those due to confounding.

- *The Diabolical Denominator*: www.StatLit.org/pdf/2021-Schield-MathFest.pdf
- *Teaching Confounding*: www.StatLit.org/pdf/2021-Schield-USCOTS.pdf
- *Confounding and Cornfield*: www.StatLit.org/pdf/2018-Schield-ICOTS.pdf
- *U. of New Mexico offers MATH 1300*: www.StatLit.org/pdf/2021-Schield-ASA.pdf

ACKNOWLEDGEMENT:

This paper is based on Schield's presentation at the 2021 ASA JSM Birds of a Feather session. Thanks to Mathew Brenneman (Emory Riddle) and Rebecca Pierce (Ball State) for finding and sharing this data on the ASA Isolated Statisticians list on August 10, 2021.

BIBLIOGRAPHY

UK Data Source:

https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1009243/Technical_Briefing_20.pdf

APPENDIX: UK Data Source

Table 1: The Table 5 Data in UK Technical Briefing 20

Table 5. Attendance to emergency care and deaths of confirmed and provisional Delta cases in England by vaccination status (1 February 2021 to 2 August 2021)

Variant	Age group (years)**	Total	Cases with specimen date in past 28 days	Unlinked	<21 days post dose 1	≥21 days post dose 1	Received 2 doses	Unvaccinated
Delta cases	<50	265,749	84,772	28,330	23,822	40,449	25,536	147,612
	≥50	33,736	13,803	2,989	195	5,640	21,472	3,440
	All cases	300,010	98,722	31,841	24,018	46,089	47,008	151,054
Cases with an emergency care visit§ (exclusion‡)	<50	8,449	N/A	70	756	1,127	694	5,802
	≥50	1,940	N/A	10	15	326	1,098	491
	All cases	10,391	N/A	82	771	1,453	1,792	6,293
Cases with an emergency care visit§ (inclusion#)	<50	10,975	N/A	119	953	1,368	864	7,671
	≥50	3,342	N/A	24	30	486	1,815	987
	All cases	14,319	N/A	145	983	1,854	2,679	8,658
Cases where presentation to emergency care resulted in overnight inpatient admission§ ((exclusion‡)	<50	1,970	N/A	35	136	203	153	1,443
	≥50	1,059	N/A	7	12	125	620	295
	All cases	3,030	N/A	43	148	328	773	1,738
Cases where presentation to emergency care resulted in overnight inpatient admission§ (inclusion#)	<50	3,084	N/A	61	211	298	224	2,290
	≥50	2,074	N/A	20	23	230	1,131	670
	All cases	5,159	N/A	82	234	528	1,355	2,960
Deaths within 28 days of positive specimen date	<50	71	N/A	2	4	4	13	48
	≥50	670	N/A	5	6	65	389	205

18

https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1009243/Technical_Briefing_20.pdf

Table 2: Calculation of Vaccinated (omitting Unlinked)

A	B	C	D	E	F		1	
UK Data		----- Table 5 -----						2
		Total	Unvaccinated	Unlinked	Vaccinated*			3
Delta cases	<50	265,749	147,612	28,330	89,807	=C4-D4-E4		4
	>=50	33,736	3,440	2,989	27,307	=C5-D5-E5		5
	Missing**	525	2	522	1			6
	ALL	300,010	151,054	31,841	117,115			7
								8
Deaths***	<50	71	48	2	21	=C9-D9-E9		9
	>=50	670	205	5	460	=C10-D10-E10		10
								11
Death Rate	<50	0.027%	0.033%	0.007%	0.023%	=F9/F4		12
	>=50	1.986%	5.959%	0.167%	1.685%	=F10/F5		13
* Data used by Mathew Brenneman								
** ALL Cases total shown in Table 5 minus the total of the two age groups (above)								
*** within 28 days of positive specimen date								
Original source of UK Data:								
https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1009243/Technical_Briefing_20.pdf								