Causality in the Time of Cholera: John Snow as a Prototype for Causal Inference

Working Paper Version 1.2

Thomas S. Coleman*

March 13, 2019

The story of John Snow's 1855 treatise *On the mode of communication of cholera* is a rollicking good tale – full of heroism, death, and statistics. But more fundamentally Snow's work is a sustained effort to convince skeptics, through argument and a wide variety of evidence, of the waterborne theory of cholera articulated in the 1849 essay of the same name. Snow's data and analysis provide a template for how to convincingly demonstrate a causal effect, a template as applicable today as in 1855. I consider two of strands of Snow's evidence – the Broad Street outbreak and the south London "Grand Experiment" – as pedagogical examples of using non-experimental data to support a causal effect. In doing so I discuss extensions to Snow's analysis using modern techniques and tools: most importantly difference-in-differences regression and count (Poisson) regression for error analysis in quasi-randomized control experiments. These provide clear and compelling examples of the modern techniques and tools, while confirming and strengthening Snow's original conclusion on the causal effect of water supply on cholera mortality.

Keywords: John Snow, cholera, causal inference, epidemiology, statistical methodology, history of science

JEL Classification: C18, N33, N93, B40, C52

Contents

*Harris School of Public Policy, University of Chicago, tscoleman@uchicago.edu

1 Introduction

John Snow's analysis of the cholera epidemic in London in 1854 is justly famous, covered in both popular books and specialist texts. (Johnson [2007], Tufte [1997b], Freedman [1999, 1991], Hempel [2007], Vinten-Johansen et al. [2003], Rothman [2002], McLeod [2000] is a very partial list.) This essay focuses on the evidence and analysis Snow (and others) assembled, particularly in Snow [1855] and Westminster and London School of Hygiene and Tropical Medicine [1855], that ultimately led to the acceptance of polluted water and sewage as the causal agent for cholera infection. The prevailing theory in 1854 was that cholera was caused by miasma or bad air. John Snow (and others, the Reverend Henry Whitehead notably) assembled, analyzed, and presented a variety of data that, taken together, provide a compelling demonstration of the causal effect of drinking water in the cholera outbreaks.

This essay reviews two key components of Snow's evidence and analysis, the Broad Street analysis and the "Grand Experiment" of the multiple suppliers of water to south London. Snow's investigation of cholera in mid-nineteenth century London exhibits multiple aspects of the process by which empirical evidence becomes demonstration of causal effect:

Near-Definitive Demonstrations: Snow's 1855 book contains not one but multiple near-definitive demonstrations that water supply is the causal agent in cholera outbreaks. I highlight two: the 1854 Broad Street outbreak and the South London "Grand Experiment".

Multiple Strands of Evidence: Snow presented evidence and analysis from different events. The 1854 Broad Street analysis was a detailed case study of a specific outbreak. The south London "Grand Experiment" (comparing customers of the Southwark and Vauxhall Waterworks Company versus the Lambeth Water Company) was a large-scale comparison of almost 500,000 individuals, using (in today's language) both differences-in-differences regression and a quasi-randomized natural experiment.

Innovative Data Analysis: The data are simple – much of it easy to work with by pen-and-paper or spreadsheet. Nonetheless the analysis demonstrates multiple data analytic tools and techniques that we use today. The mapping of the Broad Street outbreak has rightly become part of the folklore of epidemiology and data science (see Tufte [1997b], Frerichs [a], Rogers [2013], McLeod [2000], Mackenzie [2010] for a very small sample of the works discussing Snow's spatial analysis). It also provides valuable examples of modern statistical tools such as statistical clustering in spatial analysis and contingency table analysis. Snow's analysis of the South London data was equally innovative. The lack of modern statistical tools precluded Snow from a proper analysis and has resulted in less long-term influence. The comparison of the 1849 versus 1854 outbreaks is credited by Angrist and Pischke [2008] (p. ??) as the first application of the difference-in-difference technique. The comparison of Southwark & Vauxhall versus Lambeth customers is a prototype for randomized control trials. We can extend and complete Snow's analysis using modern statistical methods, and doing so provides both useful pedagogical examples and valuable insight into data and error analysis, particularly when applied to count data.

Failure to Convince: In spite of what has since been recognized as a classic exercise in data, analysis, and argument, Snow failed to convince the medical profession, the policy-making establishment, or the public.

2 Nineteenth-Century London and Cholera

Cholera is a horrible and often deadly disease of the small intestine that causes diarrhea, vomiting, and rapid dehydration. Without treatment roughly 50 percent of people die. Victims suffer dehydration and electrolyte imbalance with symptoms appearing after as little as two hours or as long as five days. Dehydration can be so severe that sufferers' skin turns blue. Nowadays rehydration therapy (developed in the 1960s) can reduce mortality to less than one percent.

Cholera is caused by the bacterium *Vibrio cholerae*, originally identified by Filippo Pacini in 1854 but not widely recognized until re-discovered by Robert Koch in 1883. Cholera is transmitted through water contaminated by untreated sewage. The bacterium reproduces in the human small intestine and escapes into the world through diarrhea and vomiting. The bacterium can travel to new victims when infected sewage is mixed into other people's drinking water. Cholera will generally not reach severe epidemic proportions in a rural and low-density population because the chain of transmission – from the gut of an infected person out into the wider world and then into water that is drunk by others – is not strong enough. When people live close together, however, in crowded conditions with poor sanitation, the bacterium has the opportunity to flourish.

The first modern pandemic (worldwide outbreak) started in 1817 in India, but it was not until 1831-32 that cholera hit London, with 6,536 dying and a mortality of roughly 38 per 10,000 persons.[1] Cholera returned in 1848-49 with 14,137 dying and a mortality of roughly 62 per 10,000 persons, and again in 1853-54 with 10,738 dying and a mortality of roughly 47 per 10,000. (See Westminster [2018])

Today we know that cholera is caused by a waterborne bacterium but in mid-19th century London the prevailing wisdom was that cholera resulted from bad air – vapors or miasma. London in the early and mid 1800s was in many ways a vile and dirty place with raw sewage collected in cesspools under houses or emptied into open sewers and ditches. The theory that disease was transmitted through vile smells and bad odors was wrong but easy to believe, and it was widely accepted as fact.

In the first half of the 19th century urban London provided an ideal environment for cholera to flourish. The population of London was growing quickly (1.4% per year from 1831 to 1851) and living conditions were crowded. Infrastructure to remove sewage and supply fresh water was rudimentary. And perversely, public policy aided the transmission of cholera. Public health officials (led by Edwin Chadwick) strove to improve sanitation in London by connecting house drains and cesspools to city-wide sewers, thus removing localized sources of sewage.[2] Unfortunately this sewage was dumped into the Thames, which also provided drinking water for many in London. The unintended consequence of concentrating the sewage in the Thames was to provide a quick and direct path for the cholera bacterium from the gut of a single person to the mouths of thousands of potential victims.

John Snow's monograph (Snow [1855]) together with the Vestry report (Westminster and London School of Hygiene and Tropical Medicine [1855], to which John Snow (and the Reverend Henry Whitehead) contributed substantially) are justly recognized as providing compelling evidence that cholera was waterborne. They also

[1] London population was 1,729,949 in 1831 and 2,286,609 in 1851. Deaths from the website "Cholera and the Thames" (Westminster [2018]). Population data from the Wikipedia entry "Demography of London", citing "A vision of Britain through time," Great Britain Historical GIS (http://www.visionofbritain.org.uk/data_cube_page.jsp?data_theme=T_POP&data_cube=N_TOT_POP&u_id=10097836&c_id=10001043&

[2] The Public Health Act of 1848 required not only new construction but existing buildings to connect to sewers. See Johnson [2007] p. 118 ff.

provided the solution to London's recurrent cholera outbreaks: move sewage out of the Thames and move water supplies upstream of London. The evidence was largely ignored.

Ultimately the sewage problem was addressed, but not as a result of Snow's work. In the summer of 1858 a heat-wave struck London and the river Thames, full of sewage, stank – an event that earned the moniker "The Great Stink". This finally compelled public officials to commission the civil engineer Joseph Bazalgette to plan and build the Northern and Southern Outfall Sewers, a system of interconnecting sewers that sloped towards outfalls (discharge points) lower on the Thames, below London. This sewer system essentially solved the problem of mixing London sewage with London water and thus put an end to the recurrent outbreaks of cholera.

It was not until the cholera outbreak of 1866, in the east of London, that Snow's ideas were gradually and grudgingly accepted by public officials and the scientific establishment. The 1866 outbreak was confined to the east of London (Bromley by Bow), the last area which was not yet connected to the Bazalgette sewers. Other parts of London saw no sustained outbreak. This evidence – a large-scale outbreak limited to an area with contaminated water – was finally enough to sway skeptics and allow for a more balanced assessment of Snow's data and analysis.

John Snow

John Snow (15 March 1813 – 16 June 1858) was an English physician and a leader in the adoption of anesthesia and medical hygiene. He is considered one of the fathers of modern epidemiology because of his work on cholera, particularly the Broad Street outbreak and the south London "Grand Experiment" discussed in this essay.

Snow devoted much time and effort to the study of cholera, with a fully developed (and correct) theory of the disease presented in Snow [1849]. It is truly startling how well Snow understood the disease, without the benefit of a germ theory of disease or the identification of the bacterium *Vibrio cholerae*:

> The excretions of the sick at once suggest themselves as containing some material which, being accidentally swallowed, might attach itself to the mucous membrane of the small intestines, and there multiply itself by the appropriation of surrounding matter, in virtue of molecular changes going on within it, or capable of going on, as soon as it is placed in congenial circumstances. Such a mode of communication of disease is not without precedent. The ova of the intestinal worms are undoubtedly introduced in this way. The affections [sic] they induce are amongst the most chronic, whilst cholera is one of the most acute; but duration does not of itself destroy all analogy amongst organic processes. The writer, however, does not wish to be misunderstood as making this comparison so closely as to imply that cholera depends on veritable animals, or even animalcules, but rather to appeal to that general tendency to the continuity of molecular changes, by which combustion, putrefaction, fermentation, and the various processes in organized beings, are kept up. (Snow [1849] pp. 8-9)

The outbreak of cholera in 1848-49 was pivotal in Snow's analysis, leading to the publication of his 1849 essay "On the Mode of Communication of Cholera" (Snow [1849]). In this essay he clearly stated both his theory for the waterborne nature of cholera and evidence to support his theory. The evidence was focused on specific incidents, for example the outbreak at Albion Terrace (later renamed Milton Terrace, it was along the Wandsworth Road in London south of the Thames). Albion Terrace was a row of 17 houses that suffered "an extraordinary mortality from Cholera in 1849, which was the more striking as there were no other cases at the time in the immediate neighbourhood; the houses opposite to, behind, and in the same line, at each end of those in which the disease prevailed, having been free from it." Snow described in detail the outbreak (reporting findings from Mr. Grant, the Assistant-Surveyor for the Commission of Sewers), and the piping for water supply and sewage disposal from the Albion Terrace houses and the circumstances that led to contamination of their water supply but not others – how "the water got contaminated by the contents of the house-drains and cesspools; the cholera extended to nearly all the houses in which the water was thus tainted, and to no others." (Snow [1849] p 15 ff, also Snow [1855] p 25 ff)

To modern eyes the evidence in the 1849 essay is convincing, fully supporting Snow's theory of the waterborne nature of cholera transmission. But it was not so for contemporaries, convincing neither the scientific establishment, public health officials, politicians, nor the wider public. The skeptical nature of Snow's audience is important for a full appreciation of Snow's 1855 book Snow [1855]. Snow was compelled to assemble multiple threads of evidence, with every thread carefully supported by evidence and argument. Paradoxically it was the bull-headed nature of his opponents that forced Snow to strengthen his data and analysis to such a degree that it still serves as a prime example of using data and argument to demonstrate a causal relationship.

It is also important to recognize that Snow's analysis, both of the Broad Street outbreak and the south London "Grand Experiment," were confirmatory in nature and not exploratory. Snow was not mapping the Broad Street outbreak to explore the mechanism of cholera transmission – he knew the mechanism already (although he was exploring the *source* of the infection). Rather he was assembling evidence that would further refute the explanations of those who advocated alternative explanations.

Snow never lived to see his ideas widely accepted. He died of a stroke in 1858, just as that summer's "Great Stink" developed. The final London outbreak of cholera in 1866, falling only on areas still afflicted by London sewage, served as the tipping point which eventually led to reconsideration of the cholera transmission mechanism and the recognition of Snow's contributions.

3 Overview of The 1854 Broad Street Outbreak and the South London "Grand Experiment"

Before discussing these two strands of evidence, there are a few points to make about Snow's cholera investigations. In 1849 (with publication of his essay Snow [1849]) Snow had settled on his theory of cholera as an infectious waterborne disease.[3] At the very beginning of that essay he mentions but quickly dismisses doubts about cholera being infectious. He then follows with an extended discussion of the etiology of cholera, argu-

[3] See https://johnsnow.matrix.msu.edu/chronlondon.php (by authors of Vinten-Johansen et al. [2003]) about Snow's investigations in 1848-49 that convinced him of the waterborne nature of cholera. Horsleydown and Albion Terrace seem to be crucial.

ing that since vomiting and diarrhea are the first symptoms, infection is probably through the alimentary canal.[4]

The 1849 essay provides evidence on the waterborne nature of cholera infection, but the arguments and evidence did not convince skeptics; the miasma theory of cholera remained dominant. Snow's 1855 book (Snow [1855]) repeats the theory that cholera is a waterborne infection, and supplies a wealth of additional evidence that (to a modern mind) overwhelmingly demonstrates that cholera is waterborne. The evidence ranges from detailed analysis of single outbreaks (for example Broad Street in 1854) to a large-scale natural quasi-randomized experiment (the south London "Grand Experiment").

3.1 1854 Broad Street Outbreak

On August 31st, 1854, an outbreak commenced in Soho centered in Broad Street (now Broadwick Street) just north of Golden Square. On Friday September 1st 70 died, and on Saturday a further 127. In John Snow's words: "The most terrible outbreak of cholera which ever occurred in this kingdom, is probably that which took place in Broad Street, Golden Square, and the adjoining streets, a few weeks ago. Within two hundred and fifty yards of the spot where Cambridge Street [now Lexington St.] joins Broad Street [now Broadwick], there were upwards of five hundred fatal attacks of cholera in ten days." (Snow [1855] p. 38). This has become justly famous, and is discussed in various publications with (Johnson [2007], Rogers [2013], Tufte [1997a,b], Westminster [2018], Frerichs [a]) being a very short list.

Based on the concentration of cases, Snow quickly identified the Broad Street pump, at the intersection of Cambridge and Broad Streets, as the source of the infection. The evening of September 7th he argued before the vestrymen of the Board of Guardians of St James's Parish that the pump was at fault and, although they did not believe him, they ordered the pump handle removed. The outbreak was already subsiding, as we will see below, but there is a reasonable chance that removing the handle forestalled a reoccurrence.

Over the next months Snow (together with The Reverend Henry Whitehead) built the case against the Broad Street pump. This case was multi-faceted and is discussed in more detail below, with his map being the most famous and lasting component.

Mapping and Snow's Broad Street Data – A Very Incomplete Bibliography

The Broad Street mapping is justly-famous and has been discussed by many authors. This is a (very incomplete) listing of some of the data sets and software available

Li An R package with data (based on Dodson and Tobler's 1992 digitization of Snow's map, not georeferenced) and functions. Computes and visualizes "pump neighborhoods" based on Voronoi tessellation, Euclidean distance, and walking distance. Ability to overlay graphical elements and features like kernel density, Voronoi diagrams, Snow's Broad Street neighborhood, and notable landmarks (John Snow's residence, the Lion Brewery, etc.) via add*() functions. I have used this package for the Voronoi and walking neighborhoods. https://github.com/lindbrook/cholera

Wilson Snow's data in various formats: Cholera Death locations (Vector) with attribute data giving the number of deaths at each point; Pump locations (Vector); John Snow's original map georeferenced to

[4] Snow [1855] also opens with a discussion of the theory of the disease, but I find the discussion in the 1849 essay more concise and more clear than in the 1855 book.

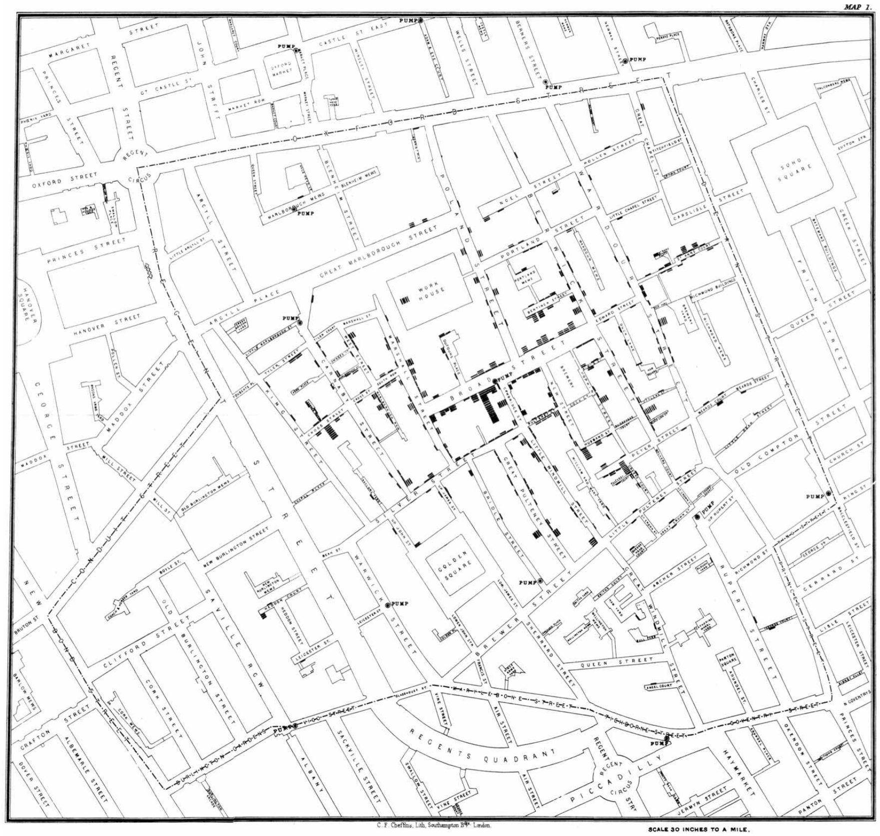

Figure 1: John Snow's Map, Version 1 (Snow [1855])

the Ordnance Survey National Grid (Raster); Current Ordnance Survey maps of the area (from those released under OS OpenData; Contains Ordnance Survey data © Crown copyright and database right 2013; Raster) http://blog.rtwilson.com/john-snows-cholera-data-in-more-formats/

Tufte [1997b] The classic chapter on Snow's mapping and data analysis.

Rogers [2013] A very good piece by Simon Rogers in the Guardian about Snow's map, with extensive links to data and maps.

Mackenzie [2010] Analyzes Snow's maps with ArcGIS and provides data for using an arbitrary (not geo-referenced) scan of Snow's map. See https://www1.udel.edu/johnmack/frec682/cholera/ and https://www1.udel.edu/johnmack/frec682/cholera/cholera2.html

Dimaggio on-line slides for a course of spatial analysis in R. Slide 20 is "Point Process Data: Broad Street Pump Cholera Outbreak" http://www.columbia.edu/~cjd11/charles_dimaggio/DIRE/styled-4/code-16/#(20)

Shiode [2012] and Shiode et al. [2015] are academic papers using Snow's data and extending the analysis with additional mapping techniques and additional data.

McLeod [2000] examines Snow's maps and some of the myths that have developed over the role of the maps in understanding the cholera outbreak.

3.2 South London "Grand Experiment"

London in the early and mid nineteenth century saw the growth of water companies that supplied piped water to residents. In one part of London south of the Thames two companies, the Southwark and Vauxhall Waterworks Company and the Lambeth Water Company, competed and supplied customers in nearby and overlapping districts. During the 1849 epidemic both companies drew their water from the lower Thames, water that was subject to pollution from the drains and cesspools of greater London. In 1852 the Lambeth Water Company moved its source upstream, above the main source of London sewage.[5]

Snow recognized that these circumstances provided a valuable natural experiment, what Snow called his "Grand Experiment". Lambeth's change between the 1849 and 1854 epidemics provides a comparison, a treatment effect, that Snow highlighted. We will be able to re-cast and somewhat strengthen this comparison using the more modern statistical technique of difference-in-differences regression.[6]

Snow also recognized and highlighted the effectively random distribution of customers supplied by one company versus the other in the overlapping districts, and the exogenous nature of the change in 1852:

> In many cases a single house has a supply different from that on either side. Each company supplies both rich and poor, both large houses and small; there is no difference in the condition or occupation of the persons receiving the water of the different companies ... The experiment too, was on the grandest scale. No fewer than three hundred thousand people of both sexes, of every age and occupation, and of every rank and station, from gentlefolks down to the very poor, were divided into two groups without their choice, and, in most cases, without their knowledge; one group being supplied water containing the sewage of London, and amongst it, whatever might have come from the cholera patients, the other group having water quite free from such impurity. Snow [1855] p 75.

Snow himself went house-to-house in the summer of 1854 to identify the supplier (Southwark & Vauxhall versus Lambeth) for every cholera death. This allowed Snow to compare mortality rates between the two suppliers, finding dramatically lower mortality for customers of the Lambeth Company. We can re-cast Snow's analysis to conform more closely to modern standards of a randomized trial. In doing so we simply strengthen Snow's conclusion that there is an overwhelming effect on mortality that is difficult to ascribe to any mechanism other than the source of the Lambeth Company's water.

4 Causation versus Correlation

Causality and the measurement of causal effects is one of the most important goals of social science research. Correlation and prediction have their place but in policy-oriented research causality is paramount: "The reason we seek causal explanations is in order to *intervene*, to govern the cause so as to govern the effect: 'Policy-thinking is and must be causality-thinking.'" (Tufte [1997a,b] p. 6, quoting Dahl [1965])

[5] The move was in response to legislation that required water sources be moved above the London sewage sources. The deadline for moving was 1855. Lambeth moved early, Southwark and Vauxhall delayed. See Johnson [2007] p. 105.

[6] Angrist and Pischke [2008] p 227 credit Snow with the (probable) first use of the difference-in-differences idea.

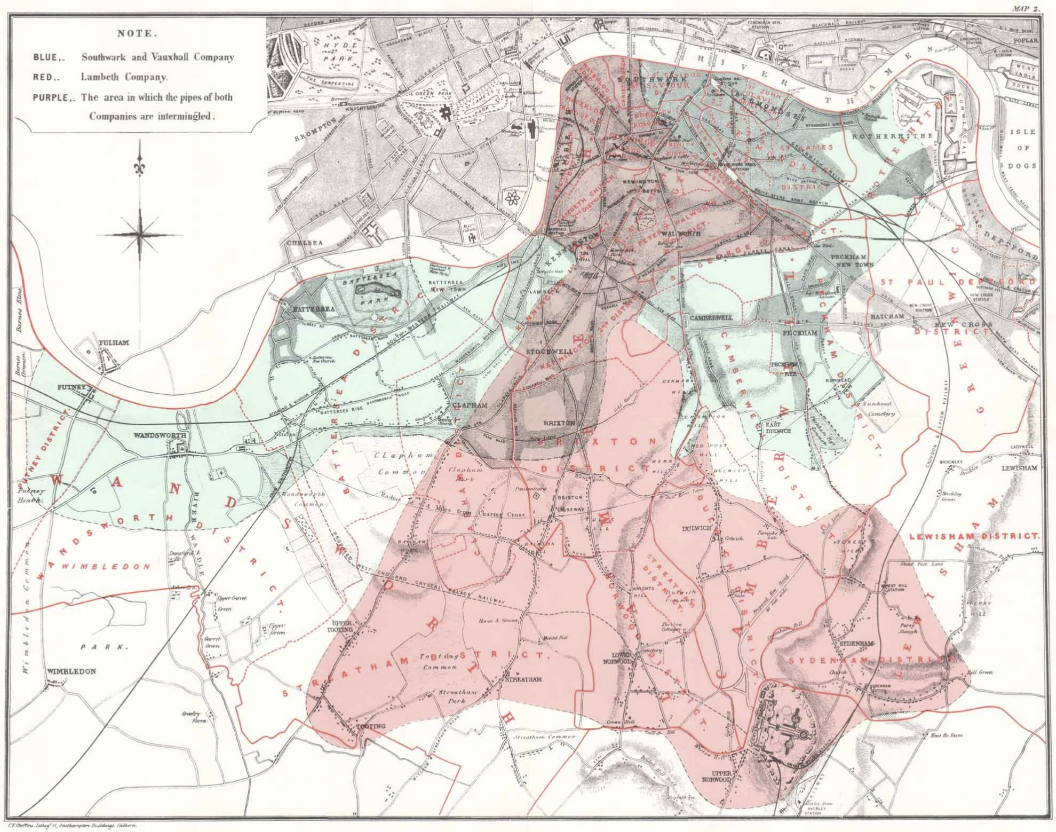

Figure 2: Regions of South London Served by the Southwark & Vauxhall and the Lambeth Companies (Snow [1855])

Demonstrating a causal relationship is notoriously difficult. Rarely if ever is there a "definitive experiment" (Johnson [2007] follows Snow in calling it an *experimentum crucis*) that proves a causal effect, that demonstrates without question the mechanism we are investigating.[7] Any decision on a question of public policy, and indeed scientific knowledge itself, is instead based on the accumulation of multiple strands of evidence each pointing in the same direction. We all know that correlation does not imply causation, and empirical evidence is really nothing more than correlations. But as the author of the XKCD comic says: "correlation doesn't imply causation, but it does waggle its eyebrows suggestively and gesture furtively while mouthing 'look over there'." Casual effects are demonstrated by the accumulation of evidence, when so much evidence tells us to "look over there" that we eventually concede that there is no other place to look but "over there".

The simple and well-worn example of cats, rats, and plague demonstrates why causation is important and why evidence of correlation alone is not sufficient to warrant policy interventions. In 1665 plague erupted in London, the last major outbreak of a series of pandemics that first struck Britain in 1348. Today we know the cause of plague: the bacterium *Yersinia pestis* that is resident in rodent hosts (such as rats) and is transmitted by the bite of fleas. We know the mechanism. The fleas themselves are not harmed by the bacteria but the flea's foregut is blocked by a biofilm and this causes infected fleas to regurgitate bacteria

[7] The 1887 Michelson-Morley experiment is sometimes proposed as the definitive experiment that disproved the existence of a stationary aether and paved the way for Einstein's theory of special relativity. Lakatos [1980] (p 73 ff) shows that the experiment was nothing of the sort, and it was a number of years before it was recognized as providing evidence against the aether theory.

and infect any host they bite. Plague was transmitted throughout the medieval world by rats (and oriental rat fleas) stowing on ships. When rats died the fleas would seek out other hosts and thus infect human populations.

Domestic cats are hunters, with rats one of their prey. An increase in rat population would naturally be associated with both an increase in the number of cats and with plague, leading to an observed correlation between cats and plague. The story goes that authorities noted the correlation, took it for causal, and ordered the destruction of cats. The natural result? With an important predator gone the rats and their fleas grew unchecked, and the plague outbreak worsened. A perfect story of the evil consequences of mistaking correlation for causation. [8]

Causality is a difficult animal, difficult to define and difficult to find, a deep issue in philosophy. In the back of my mind I have a physics-oriented approach where we have alternative models, each of which connects to observables in the real world, and we use experiments and data to rule out some of those alternative models. This approach is clearly indebted to Karl Popper's ideas of empirical falsification (and Imre Lakatos's on the methodology of scientific research programmes), the central idea being that a causal effect can never be proven, although it can be disproven or falsified.

The issue of falsification and counterexamples is not simple, of course. Lakatos (in Lakatos [1980], particularly section 1.3 p 47 ff) lays out the idea of a scientific research programme consisting of a "hard core" together with "protective belt" of auxiliary hypotheses built around the central core. Lakatos argues that anomalies or counterexamples can be accommodated by adjusting the protective belt rather than rejecting the hard core. In fact the death of Susannah Eley in the Broad Street outbreak (a widow from Hampstead discussed below) provides a near-perfect example. Her case is a strong counterexample to airborne transmission (miasma). Nonetheless the official Cholera Commission's report dismisses the anomaly by invoking a strained hypothesis about airborne influences poisoning the water – an auxiliary hypothesis that we now recognize as outlandish.

As a matter of logic there is no evidence, no one incident or set of data, that can prove the waterborne theory.[9] For Snow's skeptical audience this was certainly the case. The only option open to Snow, the only option open to any researcher aiming to demonstrate a causal effect, is to amass so much evidence from multiple sources, approach the problem from so many alternative directions, that skeptics are driven to relinquish one alternative after another and eventually forced to admit that the proposed causal mechanism is the only possibility. For this approach to "proving" causal effects we might adopt the legal language of proof "beyond reasonable doubt" where a case is proven when there is no plausible reason to believe otherwise.

It seems likely that John Snow had a similar view of the world. He had a well-articulated theory of cholera and its transmission, developed by 1849. Snow [1849] laid out this theory explicitly and clearly. Snow apparently developed his theory and turned away from the predominant "effluvial theory" or miasma view during the period 1848-49. It appears that his change in view resulted from examining various sources of evidence (with the Albion Terrace and Horsleydown outbreaks playing an important role), informed by his

[8] This is a wonderful story and should be true, but most likely is not. Although a correlation between cats and plague could plausibly have been observed, I have no found evidence that it was. Defoe (?) states that cats and dogs were killed "according to the advice of physicians" but the reason seems to have been a belief that they were carriers of infection for other reasons, "carrying the infectious streams even in their furs and hair" – i.e. based on logical reasoning rather than empirical observation To make the story even more complicated, recent research suggests that human fleas and lice rather than rats and their fleas caused epidemic outbreaks. (??) Which just goes to show that even a well-proven causal mechanism may not in fact be true.

[9] For a modern reader, knowing already the mechanism for cholera transmission, each of the pieces of evidence on their own is convincing. But we need to put ourselves back in the 1850s when bacteria were unknown, the scientific community was largely committed to the miasma theory and airborne transmission, and the question was truly unsettled.

knowledge and research on gases and anesthesia. Snow [1849, 1855] also also touched on some of the alternate theories (for example the miasma theory that cholera was transmitted through the air).

I want to distinguish between the development of a theory and the testing of a theory, as does Popper ("I shall distinguish sharply between the process of conceiving a new idea, and the methods and results of examining it logically." Popper [1985] p 134, originally from *The Logic of Scientific Discovery*) In this essay I focus on the testing of Snow's waterborne theory and Snow [1855] as an example of the accumulation and presentation of multiple strands of evidence that disprove alternative theories and in the end, after ruling out every other possible cause, leaves the reader no choice but accepting the waterborne theory. I will not delve into the development of his theory (as interesting as that might be).

Role of Statistics and Hypothesis Testing

When we view Snow's 1855 book as a collection of multiple forms and types of evidence disproving alternatives and supporting the waterborne theory, hypothesis tests and statistical tests of significance play a role somewhat different from that discussed in standard texts. Rather than "proving" or "rejecting" a hypothesis, statistics play more the role of calibrating how likely or unlikely are the observed data. Such statistics and associated probability levels serve as tools for judging the quality of our evidence rather than as definitive tests.

Of course one major contrast between Snow and modern empirical work is precisely in the use of statistics. We have the benefit of the development of 150 years' worth of tools and in discussing Snow's analysis we can examine how we might apply some of these tools. One example, showing both how we add statistics to Snow's analysis and how these statistics strengthen one strand of evidence rather than test for causal effects directly, would be the clustering of deaths around the Broad Street pump in the 1854 Soho outbreak.

Snow's map provides visually compelling evidence for clustering but we can supplement this with a chi-squared statistic comparing the observed versus expected number of deaths around different pumps. This statistic formalizes and quantifies what is obvious from the map. And the statistic confirms that the clustering is far from random (simple random clustering has a probability far less than 10^{-16}). This shows that the clustering is non-random but does not demonstrate (much less prove) that cholera is waterborne. Maybe there was a miasma emanating from the Broad Street pump but not others; this would equally well explain the clustering. Additional evidence that differentiates between the two theories is necessary.[10]

Freedman [1999, 1991] discusses Snow's work with particular focus on the methodology, arguing that "Snow's work is ... a success story for scientific reasoning based on nonexperimental data." (Freedman [1991] p 291) Freedman also argues that "statistical technique can seldom be an adequate substitute for good design, relevant data, and testing predictions against reality in a variety of settings," (Freedman [1991] p 291) and that Snow's work provides a wonderful example of good design, relevant data, etc..

I have written this essay precisely to support and extend Freedman's assertion. Nonetheless I differ from Freedman on his somewhat negative view of the value of regression and other statistical tools ("I do not think that regression can carry much of the burden in a causal argument," and "Arguments based on statistical significance of coefficients seem generally suspect." (Freedman [1991] p 292) I argue below that a careful

[10] As discussed below (and laid out in detail in Johnson [2007]), Snow and The Reverend Henry Whitehead provide such evidence: the low number of deaths at the St. James workhouse at 50 Poland Street and the Lion Brewery at 50 Broad Street, and differential death rates among those who drank versus did not drink from the pump.

statistical analysis of Snow's south London data, applying exactly the regression tools and significance tests for coefficients that Freedman derogates, substantially strengthens Snow's conclusions by providing the context we need for judging our confidence in the observed measurements.

Unobservables

For statistical analysis of data, maybe the biggest roadblock to uncovering causal effects is the problem of unobservables. We observe some association (correlation) between variables of interest: source of drinking water and cholera infection in Snow's case. We can often control for observables such as age, gender, location, income (although these were largely not observed in any detail by Snow and contemporaries) but there always remains possible variables we have not measured and cannot observe. Borrowing Lakatos's language, there will always be an auxiliary hypothesis that can account for the observed correlation *without* the necessity for our proposed causal effect.

Unobservables and alternative (auxiliary) hypotheses are the fundamental and insurmountable barrier to measuring and proving causal effects – there is simply no mechanical or statistical procedure that will absolutely remove the risk of unobservables contaminating our measurement. Randomization, instrumental variables, difference-in-differences regression, selection corrections, regression discontinuity designs, these are all tools we use in our effort to remove or adjust for unobservable variables. But none of these are foolproof – randomization may not be perfect, instruments may be correlated with the unobservables after all, etc.

Two Sets of Data and Three Techniques in Snow's *On the Mode of Communication of Cholera*

I focus on two sets of evidence and three techniques that Snow presents in his 1855 book. The first set of data and associated technique is the justly-famous mapping of the Broad Street outbreak. This is a case study with attention to every detail of the outbreak. Over 600 people living within a few blocks of the pump died over a period of roughly two weeks. Snow (and the Reverend Whitehead) walked the streets of the neighborhood, matched almost all deaths to addresses, tracked many of those who didn't die, tracked down the reason for multiple anomalous observations, and eventually identified both the index case and the mechanism for the original contamination of the Broad Street well. Snow's map is famous but what distinguishes the evidence and makes it truly compelling is the detailed attention to anomalies and apparent contradictions.

The second set of data, the south London "Grand Experiment", are the deaths observed in the 1849 and 1854 epidemics in the south London regions served by the Southwark and Vauxhall Waterworks Company and the Lambeth Water Company – a large number of deaths for a large population (almost 500,000 persons). There are two aspects that make this set of data valuable. First, the two companies competed in nearby and overlapping regions with (according to Snow) no observable differences between customers. Snow recognized the value of this effectively randomized population. Second, in 1852 the Lambeth Company switched from a dirty water source to clean. This data design allowed Snow to perform two separate analyses, what we would now term a difference-in-differences regression and a randomized control trial. Re-stating Snow's analysis in these terms allows us to somewhat improve on his analysis.

For the difference-in-differences we can compare before-and-after treatment (1849 versus 1854) and control versus treatment (Southwark-only region versus Lambeth-only region). For the quasi-randomized trial we consider only 1854 and, within the jointly-served region, compare Southwark customers (untreated or control group, dirty water) with Lambeth customers (treated group, clean water).[11]

The contrast in the data and the approach is striking. For the Broad Street outbreak, Snow focuses on every death. The details, of location, family circumstance, daily routine, all matter. For the south London "Grand Experiment" the aggregates are the focus. It is not that individual circumstances are unimportant, but Snow recognizes (and verifies through observation) that the mixing over customers removes the effect of individual circumstances.

The Power of Snow's Analysis

Snow's evidence and analysis is impressive. He provides multiple perspectives on the problem and considers data from the small-scale and individualistic (Albion Terrace, the Broad Street outbreak) to the large-scale and aggregate (the "Grand Experiment" with close to 500,000 subjects). Part of the power of Snow's evidence is the grounding in a theory that provides the necessary framework for both data collection and data analysis:

> The strength of his model derived from its ability to use observed phenomena on one scale to make predictions about behavior on other scales up and down the chain. ... If cholera were waterborne then the patterns of infection must correlate with the patterns of water distribution in London's neighborhoods. Snow's theory was like a ladder; each individual rung was impressive enough, but the power of it lay in ascending from bottom to top, from the membrane of the small intestine all the way up to the city itself. Johnson [2007] p. 148

The power of Snow's work also derives from the accumulation of multiple strands of evidence. Any single piece might be gainsayed, but taken together they provide a consistent and compelling story with a conclusion (cholera is waterborne) that cannot be denied.

One final point is the importance of good presentation and narrative. Snow's map is still used in textbooks and popular narratives. With the advent of computer technology and geographic information systems (GIS) Snow's map and data have been translated into various formats (see Wilson, Li). The map has resonated across the centuries:

> The map may not have had the impact on its immediate audience that Snow would have liked, but something about it reverberated in the culture. Like the cholera itself, it had a certain quality that made people inclined to reproduce it, and through that reproduction, the map spread the waterborne theory more broadly. In the long run, the map was a triumph of marketing as much as empirical science. It helped a good idea find a wide audience. (Johnson [2007] p 199)

[11] Snow did not perform exactly these analyses, but it is only a small step from Snow's published tables to our more modern presentation.

5 Broad Street August 31 – September 18

The story of the Broad Street cholera outbreak and Snow's mapping have been told many times – in bestsellers (Johnson [2007]), careful discussions of the mapping and visual display (Tufte [1997a,b]), epidemiological textbooks (e.g. Rothman [2002]), and discussions of statistical methodology (Freedman [1999, 1991]). My goal is to place the evidence Snow presents in the context of his effort to demonstrate a causal effect – to show how the case study of the Broad Street outbreak provides powerful evidence that is complementary to the large-scale south London "Grand Experiment".

From the start of the outbreak Snow suspected the Broad Street pump: "As soon as I became acquainted with the situation and extent of this irruption of cholera, I suspected some contamination of the water of the much-frequented street-pump in Broad Street." (Snow [1855] pp 38-39). Snow obtained a list of deaths from the General Register Office (left panel of Figure 3) and then Snow used the list to track the locations of the deceased: "On proceeding to the spot, I found that nearly all of the deaths had taken place within a short distance of the pump" (Snow [1855] p 39)

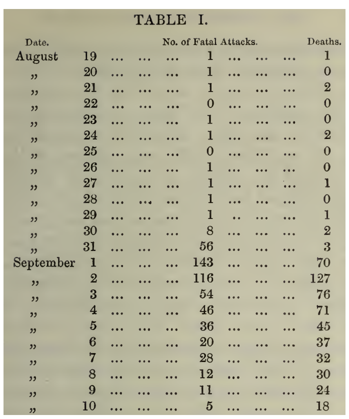

TABLE I.

Date.		No. of Fatal Attacks.	Deaths.
August	19	1	1
,,	20	1	0
,,	21	1	2
,,	22	0	0
,,	23	1	0
,,	24	1	2
,,	25	0	0
,,	26	1	0
,,	27	1	1
,,	28	1	0
,,	29	1	1
,,	30	8	2
,,	31	56	3
September	1	143	70
,,	2	116	127
,,	3	54	76
,,	4	46	71
,,	5	36	45
,,	6	20	37
,,	7	28	32
,,	8	12	30
,,	9	11	24
,,	10	5	18

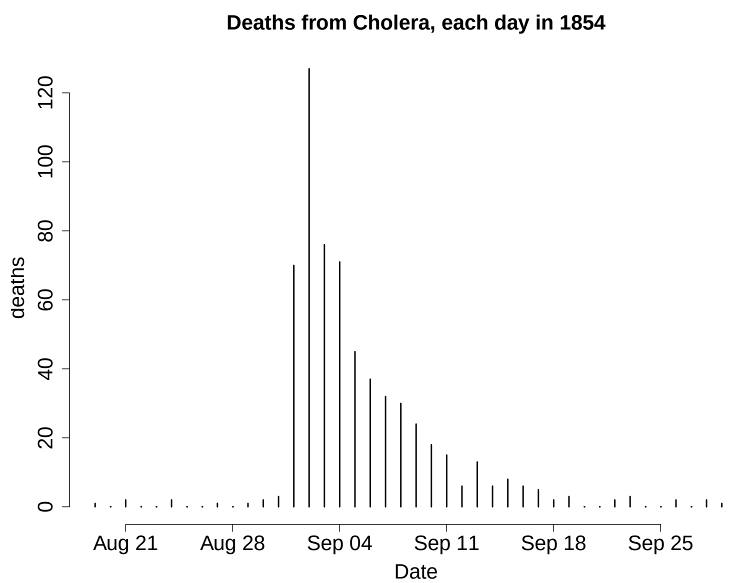

Figure 3: List (top part) of deaths in Soho August and September 1854 (Snow [1855] p 49)
Graphic produced using software from Li

It is important to recognize that most of Snow's work on the Broad Street outbreaks (together with The Reverend Henry Whitehead) revolved not around *identifying* the source of the outbreak (which Snow recognized early) but in building the case that it was water from the pump and not any other mechanism that could explain the characteristics of the outbreak. The first part of Snow's case and the one most remembered today is his famous map, published in December (Snow [1855]), shown above and here as Figure 4 (see Snow [1855], Frerichs [b] for high resolution versions of the map).

Snow's map was not the first to be published. Edmund Cooper, an engineer for the Metropolitan Commission of Sewers, published a map in September 1854 (shown as figure 12.4 in Vinten-Johansen et al. [2003]) but Cooper's map had too much detail – addresses, sewers, gully-holes, and deaths marked in light grey – to provide very useful information. Further, the goal for Cooper and the Metropolitan Commission was not to

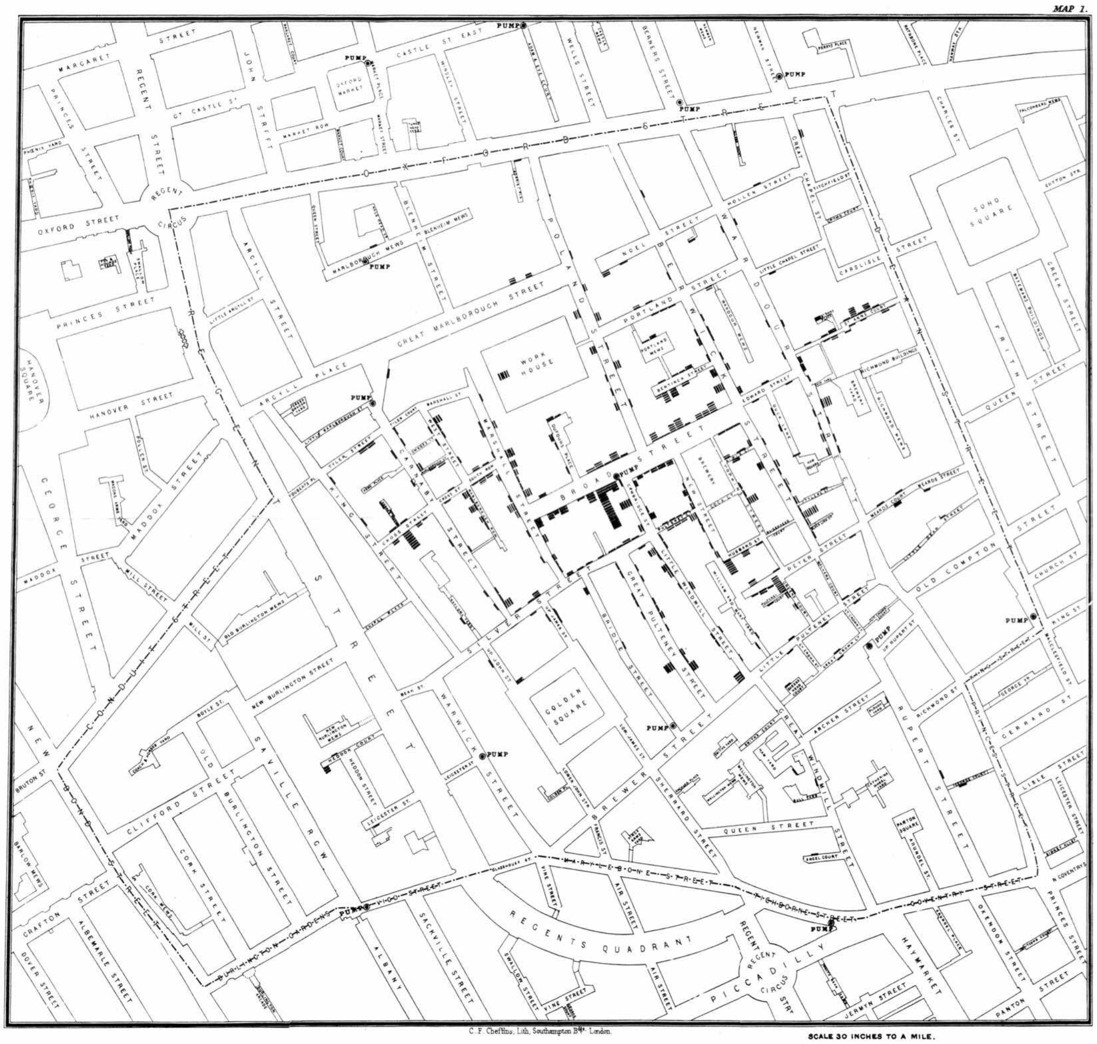

Figure 4: John Snow's Map, Version 1 (Snow [1855])

highlight the pump as the source but to rule out their sewers and grates as sources.

Snow's map was masterful in stripping away extraneous detail and focusing on the important information: location of deaths and the pumps. The dark bars of death against the otherwise spare and simple map is visually arresting and clearly demonstrates the clustering around the Broad Street pump. It has stood the test of time both as a visually compelling piece of design and an important component of Snow's argument for causality.

The concentration around the pump was in itself evidence against a number of then-current ideas or hypotheses. Snow was able to ascertain that mortality was associated with proximity to the pump but did not seem to be associated with elevation (lower floors versus upper floors), crowding, or social standing of the residents. None of these hypothesized effects survived confrontation with Snow's map.

Nonetheless, Snow knew that the map, the concentration around the pump, and his clear understanding of the source and mechanism was not sufficient:

> [Snow] could see at a glance that he'd be able to demonstrate that the outbreak was clustered

around the pump, yet he knew from experience that that kind of evidence, on its own, would not satisfy a miasmatist. The cluster could just as easily reflect some pocket of poisoned air that had settled over that part of Soho, something emanating from the gulley holes or cesspools – or perhaps even from the pump itself. Snow knew that the case would be made in the exceptions from the norm. Pockets of life where you could expect death, pockets of death where you would expect life. Johnson [2007] p. 140

5.1 Confronting the Waterborne Theory with Evidence

Snow (and Whitehead) assembled a mass of detail that corroborated the waterborne theory and ruled out alternatives. Snow and Whitehead assembled and marshaled this evidence piecemeal, often in response to anomalies or exceptions, although we can put these in four broad categories:

1. Those who should have died but escaped – Those close to pump who did not die
2. Those who should have escaped but died – Those far from the pump who nonetheless died
3. Details on the mechanism for contamination of the pump-well – The index case and decaying brick-work
4. Comparing number and rates of infection for those who drank versus did not drink from the Broad Street pump

The first three can be viewed as confronting the predictions of the waterborne theory and the miasma theory (the prime alternative) against evidence. In all cases the waterborne theory predictions survived the comparison, the miasma theory predictions did not.

5.1.1 Those who should have died but escaped

Snow's map shows the general concentration around the pump, but that general pattern could be predicted by both waterborne and miasma. The pattern is not uniform, however. I focus on two locations having anomalously low deaths. These stood out for Snow in his walking the neighborhood and are apparent in the map once pointed out: The "Work House" just north of Broad Street and the "Brewery" east of the pump on Broad Street itself. The St. James workhouse on Poland Street had 535 inmates with only five dying. As Snow points out (Snow [1855] p 42) if the death rate had been as high for the workhouse as the surrounding houses more than 100 would have died. The explanation was simple: the workhouse had its own well and was also supplied by piped water (from the Grand Junction Water Works); residents did not visit the Broad Street pump.

For the brewery (the Lion Brewery) the same problem and explanation hold: seventy workmen but no cases of cholera, but the workmen were "allowed a certain quantity of malt liquor," had access to an in-house well, and never drank from the pump.

Both of these locations were close to the Broad Street pump. Under either the theory of waterborne transmission (with the auxiliary hypothesis that residents drink from the closest pump) or the miasma / airborne theory, both locations should have had high death rates. Close examination of the circumstances ruled out the auxiliary hypothesis that residents of these building drank from the Broad Street pump, meaning these low death rates did not falsify the waterborne theory (only the auxiliary hypothesis). These observations do, however, help distinguish between the waterborne and miasma theory since there is no easy explanation for why these two locations and these alone would be spared airborne infection. These observations do not definitively rule out miasma, but make it difficult to sustain.

5.1.2 Those who should have escaped but died

Most of the deaths cluster around the Broad Street pump, but as Snow recognized there are deaths not particularly close to the pump. A fair number of deaths cluster near the Little Marlborough Street pump, nearer Marlborough than Broad Street. This should not be the case if water were the cause (and with the auxiliary hypothesis that residents drink from the closest pump). But Snow states:

> It requires to be stated that the water of the pump in Marlborough Street, at the end of Carnaby Street, was so impure that many people avoided using it. And I found that the persons who died near this pump in the beginning of September, had water from the Broad Street pump. (Snow [1855] p 46)

Snow and Whitehead tirelessly tracked the anomalous cases. There is a cluster of eight deaths at 10 Cross Street, closer to Marlborough than Broad Street. Their story is told in the Vestry report: a tailor aged 50 and his 12 year-old son died September 1st, and within three days four more of his children, all "great drinkers of pump water" who often drank from the Broad Street pump.[12]

Two little girls (one from Ham Yard the other from Angel Court, both off Great Windmill Street far to the south of Broad Street) went to school in Durfours Place (off Broad Street) and drank from the Broad Street pump on the way to or from school. (Westminster and London School of Hygiene and Tropical Medicine [1855] pp 112-113)

One of the most famous cases concerned Susannah Eley, a widow in Hampstead and her niece in Islington who died in early September when there were no other cholera deaths in those areas.[13] Snow discovered from the widow's sons, who owned a factory at 37 Broad Street near the pump, that she had lived in Soho, thought the Broad Street pump water delicious, and regularly had water from the pump brought to her in Hampstead. Both she and her niece drank Broad Street water the day before falling ill (Snow [1855] pp 44-45, also discussed in Tufte [1997b], Johnson [2007], Hempel [2007] p 217 ff as well as others).

These cases provide further evidence that fail to contradict or falsify the waterborne theory but do contradict predictions from miasma or other theories. Although it would be hard to connect the Hampstead death with Broad Street via miasma there were attempts. Johnson [2007] p 186 and Hempel [2007] p 242 (at greater length) quote the Cholera Commission's report as acknowledging that water was the *vehicle* of contamination, but not the ultimate cause:

> The water was undeniably impure with organic contamination; and ... if, at the times of epidemic invasion there was operating in the air some influence which converts putrefiable impurities into

[12]"This family were great drinkers of pump water, and used to send for it every day, but more especially to drink during the night, as they were thirsty in the warm weather, owing to the great number sleeping in one room. The children fetched the water from various pumps, but frequently from Broad Street." Westminster and London School of Hygiene and Tropical Medicine [1855] p 112

[13]Snow credits a Dr. David Fraser for alerting him to these anomalous deaths.

> a specific poison, the water of the locality ... would probably be liable to similar poisonous conversion. Thus, if the Broad Street pump did actually become a source of disease to persons dwelling at a distance ... this ... may have arisen, not in its containing choleraic excrements, but simply in the fact of its impure waters having participated in the atmospheric infection of the district.

This is a nice example (following Lakatos [1980], particularly section 1.3 p 47 ff) of a core theory (miasma) being protected by invoking auxiliary hypotheses (airborne influences contaminating water). By such means it is possible to dismiss what we now recognize as decisive evidence that contradicts the miasma theory (and that supports the waterborne nature of cholera).[14]

5.1.3 Index case and mechanism for pump-well contamination

One critique of Snow's theory for the Broad Street pump as the source of infection was that the well had been used for years, and there was no evidence of earlier contamination. Without a reasonable mechanism for contamination in late-August 1854 there was still room for doubt.

The work of the Reverend Whitehead eventually led to important evidence that provided the likely explanation. Whitehead identified a baby girl Frances Lewis at 40 Broad Street, the building next to the pump, who had fallen sick a day prior to the outbreak (and died September 2nd). Sarah Lewis, the mother, had rinsed diapers and poured the water into a cesspool at the front of the house. The existence of the cesspool was unexpected (drains were supposed to be connected to sewer lines) and further inspection showed that the cesspool was only inches from the well, there was decaying brickwork, and the ground was saturated with water from the cesspool. (See Frerichs [a], http://www.ph.ucla.edu/epi/snow/indexcase.html and Westminster and London School of Hygiene and Tropical Medicine [1855] p 159 ff and p 170 ff).

5.1.4 Comparing infection for pump drinkers versus non-drinkers (survivorship bias)

Comparing those who drank from the pump versus those who did not is crucial and provides a sharp test between the waterborne and miasma theories. Snow's map showed clustering around the pump but such clustering would be predicted by both theories. Where the two theories differ is in predictions about mortality among drinkers versus non-drinkers in the vicinity of the pump: the waterborne theory predicting higher mortality among drinkers, miasma similar between the two groups. Observing higher mortality among drinkers would contradict the miasma prediction, and similar mortality would contradict the waterborne prediction.

Strictly speaking this evidence, focused specifically on Broad Street residents, was not Snow's, being collected and reported by the Reverend Whitehead in the Vestry report (Westminster and London School of Hygiene and Tropical Medicine [1855] p 128 ff). Whitehead "inquired ... concerning 497 of 896 persons resident in Broad Street at the time of the pestilence." Apparently Whitehead was initially skeptical of Snow's theory but discovered that illness among drinkers was substantially higher than non-drinkers:

[14] As Hempel [2007] pp 217-218 states: "When Snow published his first cholera paper in 1849, the *London Medical Gazette* had said that the key test of his theory would be to show that water 'conveyed to a distant and unaffected locality would produce the disease in all who drank it'. Here was the proof." Except that the Cholera Commission did not accept the proof.

> Whitehead began his assault on the pump-contamination theory by examining a crucial absence in Snow's original survey of the neighborhood. Snow had focused almost exclusively on the Soho residents who had perished in the outbreak, detecting that an overwhelming majority of them had consumed Broad Street water before falling ill. But Snow had not investigated the drinking patterns of the neighborhood residents who had *survived* the epidemic. If that group turned out to have drunk from the Broad Street pump at the same rate, then the whole basis for Snow's theory would dissolve. (Johnson [2007] p. 173)

Essentially, Snow was at risk of being subject to "survivorship bias" (or in this case non-survivor bias).[15] Snow focused on those who died and found a high rate of drinking from the pump. But the death rates might have been high for some reason other than water – a confounding or omitted variable problem. If death rate were high among non-drinkers this would indeed indicate confounding variables. To avoid survivorship or selection bias we need to look at the whole population – those who died and those who did not – to measure whether they differed in the key characteristic of drinking from the Broad Street pump.

> In the end, he [Whitehead] tracked down information on 497 residents of Broad Street, more than half the population that had lived there in the weeks before the outbreak. ... Among the pump-water-drinking population, the rates of infection were along the lines that Snow had outlined in his original survey: for every two Broad Street drinkers who were not affected, there were three who fell ill. That ratio seemed even more striking when you compared it to the infection rates among those who had not drunk from the well: only one in ten of that group had been seized with the cholera. (Johnson [2007] pp. 173 and 175)[16]

Table 1 shows the count of the 497 Broad Street residents Whitehead tracked down, categorized by their drinking status and whether they fell ill.[17] Focusing only on those cases where the drinking status is clearly

[15] Abraham Wald's analysis of airplanes returning from bombing runs in World War II is a classic analysis that discusses and controls for survivorship bias. See Casselman [2016], Mangel and Samaniego [1984], Press [2016], Wald [1980].

[16] Johnson's quotes are slightly confusing. In the first paragraph he discusses comparing the drinking status of those who were not ill versus those ill (infected). His critical point is that we have to consider the full population, not just the deaths (non-survivors). Regarding Tables 1 and 2 he is essentially comparing within columns. (Only 17% of those not infected drank from the pump, while 81.5% of those infected drank.) In the second paragraph he provides statistics comparing the illness status of those who did drink versus did not drink: comparing within rows. (Only 6.7% of those who did not drink were infected, while 60.7% of those who did drink were infected.) The following tables, derived from Table 2, show the within column and within row percentages. These do correspond to the first two of the concepts for causation in disease postulated by Evans [1978] p. 254 (in reverse order): 2) exposure to the putative cause should be present more commonly in those with the disease than in those without the disease; 1) the prevalence of the disease should be higher in those exposed than in those not exposed. Note, however, that the contingency table analysis discussed in the text is a more comprehensive statistical testing framework.

within column	Not ill	Yes Ill
No drink	83.0%	18.5%
Yes drink	17.0%	81.5%
Total	100.0%	100.0%

within row	Not ill	Yes Ill	Total
No drink	93.3%	6.7%	100.0%
Yes drink	39.3%	60.7%	100.0%

[17] Whitehead's reporting is not as well laid-out and clear as a reader might wish. The following table shows the source of the numbers in Table 1. Note also that I believe Whitehead's statement (p 132) that "the ratio of those attacked to those who escaped is at least 80 to 57" should be "88 to 57". In reporting 80 Whitehead is including the 35 "recoveries" reported in the list on p 129 but not the additional 8 noted in the text at the top of p 130. I believe those 8 should be included as drinking from the pump, falling ill, then recovering. (See also the verbal summary on p. 79, which I also believer misses the 8 noted on p 130.)

	Not ill	Ill, recovered	Ill, died
Not drink	Table p 130	List p 129	Table p 128
Drank	Table p 131	List p 129 & text p 130	Table p 128
Probably drank	–	List p 129	Table p 128
Uncertain	List, p 131	List p 129	Table p 128

established, the counts form a two-by-two contingency table as shown in Table 2. We can perform a standard Pearson chi-squared test or a Fisher exact test for independence of drinking and illness status. The tests strongly reject that drinking from the pump and illness are not associated. The right or "Expected Count" panel of Table 2 shows why independence is so strongly rejected: under independence we should expect far more illness among those who did not drink, and dramatically fewer illnesses among those who did drink.

Table 1: Count of Residents of Broad Street Categorized by Drinking and Illness

	Not ill	Ill, recovered	Ill, died	**TOTAL**
Did not drink from pump	279	7	13	**299**
Drank from pump	57	43	45	**145**
Probably drank from pump	–	2	10	**12**
Uncertain or Unknown	13	6	22	**41**
TOTAL	**349**	**58**	**90**	**497**

Counts of Broad Street residents collected by the Reverend Whitehead and reported in Westminster and London School of Hygiene and Tropical Medicine [1855] p 128 ff. See text and footnotes for details on source for individual cells

Table 2: Contingency Table Analysis for Drinking versus Illness

Actual Counts	Not ill	Yes ill	**TOTAL**
No drink	279	20	**299**
Yes drink	57	88	**145**
TOTAL	**336**	**108**	**444**

Expected Counts	Not ill	Yes ill	**TOTAL**
No drink	226.3	72.7	**299**
Yes drink	109.7	35.3	**145**
TOTAL	**336**	**108**	**444**

Using cases for which drinking status (drinking from the pump versus not) could be determined. "Expected Counts" are expected if drinking and illness were independent (conditional on row and column sums). The Pearson chi-squared statistic is 154.7. Both the Pearson chi-squared and the Fisher exact test strongly reject the hypothesis that drinking and illness are independent (p-value far less than 0.0001). The Phi coefficient (a measure of association, the same as Cramér's V in this case) is +0.59, showing strong positive association between illness and drinking from the pump.

The conclusion from this analysis is that the water-borne hypothesis (water causes cholera) survives while alternative hypotheses do not. Tables 1 and 2 show that there is a strong positive association between drinking and illness: the Phi coefficient (Cramér's V) is +0.59. In other words the data are consistent with water causing cholera. In contrast hypotheses that other factors (miasma or airborne causes, social class, house crowding, etc.) cause illness do not survive. Among those who did not drink very few fell ill[18] so non-water hypotheses can survive only with an ancillary hypothesis that the non-water factor is *very* strongly associated with drinking from the pump – an ancillary hypothesis that does not have empirical or logical support.

5.1.5 Digression on Statistical Analysis of Spatial Clustering

Nowhere in his work did Snow dwell upon statistics or hypothesis testing, and in some sense this is refreshing – Snow focuses on the multiple forms of evidence that support his theory and contradict alternatives. But statistics has an important role to play, forcing us to always ask "how likely is this set of data, how much confidence should I have in the observation?" In Tufte's words, we always need to make *quantitative comparisons*, always ask *compared with what*?

[18] Conditional on drinking from the pump the fraction falling ill was only 6.7% – see footnote above.

We can ask this question of Snow's mapped data, comparing the actual data with an alternative. To compare we need both an alternative and a measurement, a quantification of the map data. For the alternative we can take the most basic, that deaths are randomly distributed across the map. There are multiple pumps on the map (13 in total), and for the numerical measurement we can ask how many deaths cluster around each pump. For this clustering we have to assign each death to one pump (versus all others). The simplest criterion we could use is distance: assign a death to the closest pump. This is reasonable if we assume that people go to the closest pump for drinking water – not a certainty but a good auxiliary hypothesis.

The criterion of shortest distance defines a region, a neighborhood, around each pump, the neighborhood that is closest to that pump. Mathematicians call this partitioning of the map into neighborhoods a Voronoi diagram or a Voronoi partitioning. Figure 5 shows such a Voronoi partitioning (this partitioning has been done by many authors, starting with Snow (as discussed below); this figure is produced using the R software of Li).

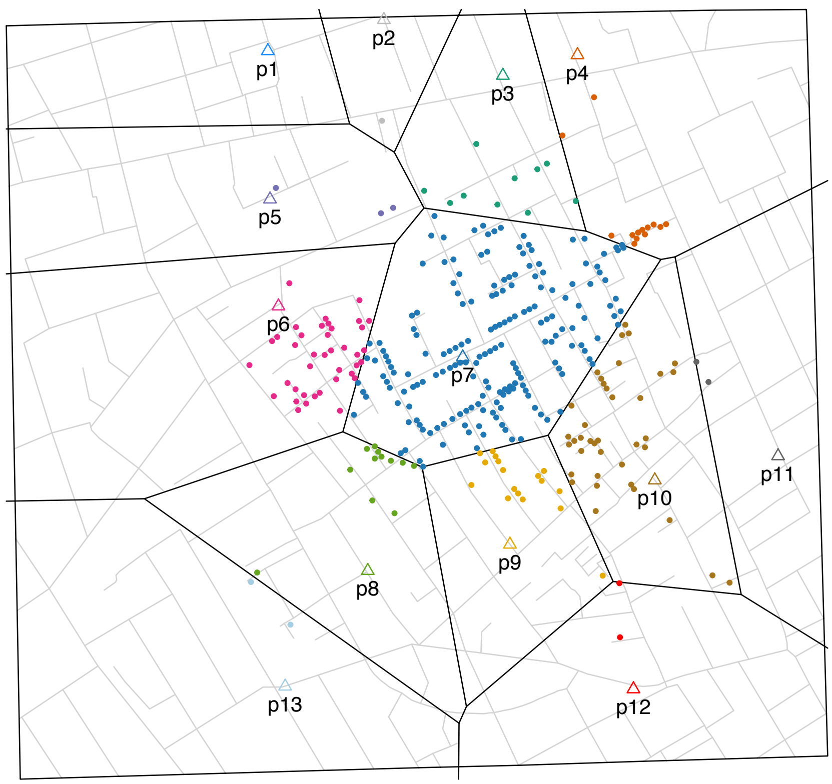

Figure 5: Voronoi Diagram (Euclidian Neighborhoods)

We can now count how many actual deaths occur in each pump's neighborhood, and how many would occur if deaths were randomly or evenly distributed across the map. Before doing so, however, we want to address an important issue regarding the Voronoi diagram. In Figure 5 we are ignoring roads, buildings, walls and assuming that distance is measured strictly by Euclidian or map distance. Much better would be to account for the walking route that a person would need to follow. John Snow recognized this and in the second version of his map, published in the Vestry report (Westminster and London School of Hygiene and Tropical Medicine [1855]) Snow drew in a "walking neighborhood" that he measured by walking the streets from the pump.

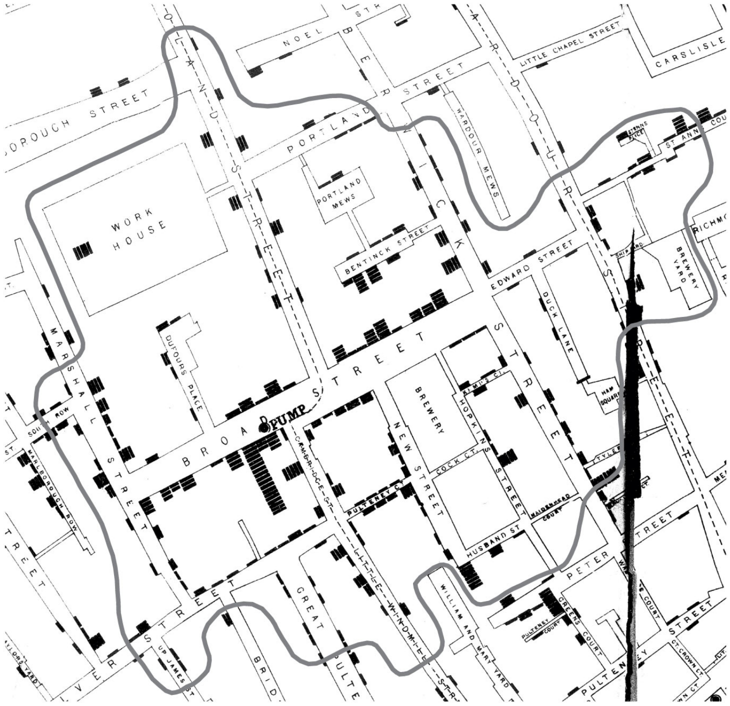

Figure 6: Enhanced Detail from Snow's Second Map (from Vinten-Johansen et al. [2003] figure 12.6)

We can do this in a more automated manner, by converting the street network into a network map and "walking" the network to find the closest pump – this is done in the package by Li. When we start with a large number of "simulated deaths" that are randomly distributed across the map then this walking algorithm defines walking neighborhoods around each pump, as shown in Figure 7. This formalizes or automates Snow's manual walking shown in Figure 6. We can also count the fraction of these randomly-distributed deaths that fall in each walking neighborhood, thus producing a quantitative measure of how many deaths we expect in each neighborhood (if deaths were random and people walked to their nearest pump). When we use the actual deaths and walk to the nearest pump then we assign deaths to pumps – we assign deaths to a walking neighborhood.

Table 3 shows the results for walking neighborhoods. "Actual" shows the assignment of actual deaths (those shown on Snow's map) to neighborhoods. "Expected" shows how many we would expect if there were 321 total deaths, deaths were randomly distributed across the map, and everyone walked to their nearest pump. We now have a quantitative measure of the actual map versus what we would expect on a simple "random" alternative. This is in the form of an observed and expected distribution and asymptotically the sum $\sum \frac{(act-exp)^2}{exp}$ should be chi-squared distributed (with 12 degrees of freedom). The observed statistic is over 1000, while the 1% level for a chi-squared variable is 26, meaning that there is a very low probability we would observe this actual distribution if deaths were random.

The statistical analysis reinforces what we already know – the clustering around the Broad Street pump is highly unusual.[19] The value of the exercise is that it formalizes the conclusion, providing a quantitative and reproducible way to measure how likely or unlikely is the observed clustering. The statistics help us

[19]Note that this is by no means the only statistical tool that can be applied to these data. For example Kernel Density Estimation can be used to show the concentration of cases around the Broad Street pump – see for example Li, Shiode et al. [2015], Shiode [2012].

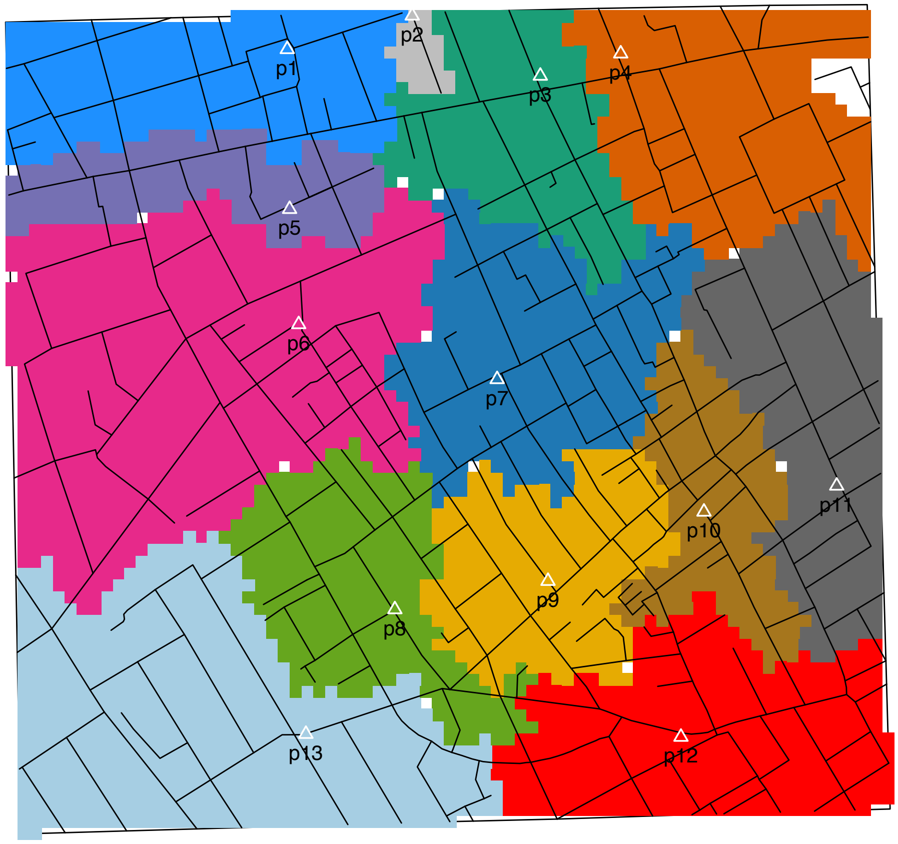

Figure 7: Quasi-Voronoi Diagram (Walking Neighborhoods, produced using Li)

determine how convincing is the evidence provided by the map (in this case very convincing), information that we can then bring to bear on the question of the source of the cholera epidemic.

The clustering analysis is also valuable in highlighting anomalous evidence. Snow points out the large number of cases clustering around the Little Marlborough Street pump,[20] and explains that the pump was not popular ("the water of the pump in Marlborough Street, at the end of Carnaby Street, was so impure that many people avoided using it." Snow [1855] p. 46). Table 3 makes clear that the clustering near the Bridle Street and Rupert Street pumps is equally, possibly more, anomalous. Further analysis of the outbreak might try to find an explanation for the high number of cases near the Bridle Street and Rupert Street pumps.[21]

5.2 Weight of Evidence

The evidence concerning the source and mechanism for the Broad Street outbreak is, to a modern reader, overwhelming. This is partly because we know the causal agent for cholera (*Vibrio cholerae*) and we know

[20]A large number but fewer than expected. In this case the "expected" count is probably too high because of vagaries of the map. Consider pumps 5 and 6 and the Voronoi partitioning shown in Figure 5. Regions for those two pumps are open to the west – Snow had no need to show pumps to the west. The size of those regions and thus the number of expected cases would be reduced by either including pumps to the west or reducing the extent of the map by moving the western boundary inwards. Regions for pumps 7, 8, 9, and 10, in contrast, are bounded by further-outlying pumps.

[21]Note from Figure 5, however, that virtually all the cases for the Bridle Street and Rupert Street pumps (9 and 10 in the figure) are in the part of their respective regions closest to the Broad Street pump.

Table 3: Actual versus Expected Deaths by Pump Neighborhood – Walking Distance

	pump id & name	Actual	Expected	chi-sq
1	Market Place	0	23.0	23.0
2	Adam and Eve Court	0	1.7	1.7
3	Berners Street	12	19.3	2.8
4	Newman Street	6	26.6	16.0
5	Marlborough Mews	1	13.8	11.9
6	Little Marlborough Street	44	55.8	2.5
7	Broad Street	189	27.6	942.4
8	Warwick Street	14	21.4	2.5
9	Bridle Street	32	19.9	7.4
10	Rupert Street	20	15.0	1.7
11	Dean Street	2	25.0	21.2
12	Tichborne Street	1	28.6	26.6
13	Vigo Street	0	43.2	43.2
	Sum	321	321	1102.8

it is waterborne, so our standards for evidence are lower. But even so, the evidence is substantial – it is hard to come up with a convincing alternative hypothesis, a hypothesis that is not contradicted by multiple strands of evidence.

In concluding we should recognize that the map is the center of the Broad Street narrative, but only one piece of an overall body of evidence. Nonetheless it is the map that is most remembered:

> The map may not have had the impact on its immediate audience that Snow would have liked, but something about it reverberated in the culture. Like the cholera itself, it had a certain quality that made people inclined to reproduce it, and through that reproduction, the map spread the waterborne theory more broadly. In the long run, the map was a triumph of marketing as much as empirical science. It helped a good idea find a wide audience. (Johnson [2007] p. 199)

6 South London "Grand Experiment" – Snow [1855]

With the south London evidence, what Snow called his "Grand Experiment," the data, scale, and methodology shift dramatically from the case studies of Broad Street or Albion Terrace. Instead of detailed examination of every case, here Snow relies on averaging over or randomizing over large numbers of Londoners with varying characteristics. Snow recognized that circumstances of the water supply companies and their response to mandated changes in water supply provided an ideal experiment – in fact, as we will see, two related forms of experiment.

The Southwark and Vauxhall Waterworks Company and the Lambeth Water Company competed over some regions south of the Thames (shown in Figure 2). In 12 sub-districts Southwark and Vauxhall alone supplied customers. In 16 sub-districts, with a population of roughly 300,000, the two companies competed directly, supplying customers side-by-side:

> In many cases a single house has a supply different from that on either side. Each company supplies both rich and poor, both large houses and small; there is no difference in the condition or occupation of the persons receiving the water of the different companies. Snow [1855] p 75

During the 1849 cholera epidemic both companies drew water from lower in the Thames, water that was polluted with the sewage of London and thus cholera bacteria from other infected Londoners. In 1852 the Lambeth Company moved their water supply.[22] Snow recognized that the close proximity and mixing of customers, together with the change in Lambeth's source, fortuitously provided a natural experiment.

The fundamental problem Snow was trying to address, the fundamental problem that arises in any investigation of causal effects, is the influence of unobserved or confounding variables. There will be variables that we either cannot measure or have not considered, variables that could be the underlying cause and that are associated with our proposed causal agent in such a way as to generate an observed correlation that is no more than a spurious (rather than causal) association. In Snow's time there were various proposed causal agents for cholera: miasma or bad air (effectively airborne transmission); elevation above sea level and the River Thames; rank and station of the individual. None of these were prima facie silly: London at the time had open sewers and cesspools and thus plenty of bad air and stink; higher-elevation areas of London such as Hampstead had lower cholera rates than lower-lying areas around the Thames (because areas around the Thames tended to take drinking water from the Thames); poor and crowded neighborhoods of London had higher infection rates than wealthy areas (because of poor sanitation and poor-quality water supplies). For some of these variables Snow could control, but for many he could not.

Snow spent considerable time looking for cases that would naturally control for unobservables, cases where the only reasonable or possible difference between infection versus not was the water supply. Albion Terrace is one of the early examples: contiguous houses with one set infected and another not, arguable the same in all respects except water supply. The circumstances of the Southwark & Vauxhall versus Lambeth companies in south London provided a large-scale case where Snow could reasonably argue that individuals were mixed across any and all unobserved characteristics, where "people of both sexes, of every age and occupation, and

[22] As mentioned above, the move was in response to legislation that required water sources be moved above the London sewage sources. The deadline for moving was 1855. Lambeth moved early ("The Lambeth Company removed their water works, in 1852, from opposite Hungerford Market to Thames Ditton; thus obtaining a supply of water quite free from the sewage of London." Snow [1855]p 68), Southwark and Vauxhall delayed. See Johnson [2007] p. 105.

of every rank and station, from gentlefolks down to the very poor were divided into two groups without their choice, and, in most cases, without their knowledge." (Snow [1855] p 75)

We can follow Snow in using this data, but we can sharpen his analysis by considering two related forms of analysis. The first is difference-in-differences regression where we compare aggregate groups – a control group (the 12 sub-districts supplied only by Southwark where people received contaminated water in both 1849 and 1854) versus a treatment group (the 16 sub-districts supplied jointly by Southwark and Lambeth and thus containing a proportion of people who received contaminated water in 1849 and clean water in 1854). In this case we use regression techniques to control at the aggregate level for possible unobserved differences between the control and treatment groups and for differences across time.[23] Reference for difference-in-differences

The second form of analysis focuses specifically on the randomization and the jointly supplied sub-districts. Given the circumstances of the supply companies and the change in Lambeth's water source, the 16 jointly-supplied sub-districts provide a good approximation to a randomized control trial.

The difference-in-differences or aggregate analysis corresponds to what epidemiologists might call a "retrospective study" – looking backwards at non-experimental or observed data and using survey or (as in this case) mortality register data to compare groups or disease exposure. The quasi-randomized analysis corresponds more closely to what we might call a "prospective" study. The data are not truly experimental – the assignment of customers to Southwark and Vauxhall versus Lambeth was not randomized by a researcher, nor was the "treatment" of Lambeth moving to a clean water supply imposed or designed as part of a clinical experiment. But overall they might just as well have been because water customers generally did not know about the change nor did they choose their supplier (in the 1840s) based on the treatment that was instituted in 1852.

6.1 Comparison of Aggregate Groups Using Difference-in-Differences Regression

Snow's Table XII has the basic count (death) data for 1849 and 1854 by sub-district. These data are reproduced in the first four columns of my Table 32. Snow points out that "The table exhibits an increase of mortality in 1854 as compared with 1849, in the sub-districts supplied by the Southwark and Vauxhall Company only, whilst there is a considerable diminution of mortality in the sub-districts partly supplied by the Lambeth Company." (Snow [1855] p 89) We can emphasize this point by converting death counts to mortality rates.[24]

Table 4 shows the mortality rates summarized by sub-districts ("First 12" supplied by Southwark & Vauxhall only, "Next 16" supplied jointly) and the numbers make Snow's point starkly clear – the 1854 mortality rate for the jointly-supplied "next 16" sub-districts fell by a factor of 1.5 (130/85).[25] But Table 4 also highlights that

[23] For the difference-in-differences analysis I do not use the "last four four sub-districts" supplied only by the Lambeth company in 1854 because three of these subdistricts were not supplied by any water company in 1849: "It is necessary to observe, however, that the supply of the Lambeth Company has been extended to Streatham, Norwood, and Sydenham, since 1849, in which year these places were not supplied by any water company." Snow [1855] p 89

[24] Converting to rates using the population by sub-district for 1851 (the year of the census closest to 1849 and 1854) from Snow's Table VIII, shown in the fifth column of Table 32. For some reason Snow did not convert the counts in Table XII to rates, a conversion which I think would have strengthened his argument.

[25] Angrist and Pischke [2008] p 227 credit Snow with the (probable) first use of the difference-in-differences idea. Interestingly, Snow does not seem to have expressed the deaths as mortality rates (as in Table 4), an expedient that would have made his point clearer by emphasizing the similarity in 1849 and the stark difference in 1854, thus highlighting the "treatment effect" of Lambeth moving to a clean water source in 1852.

we should consider changes over time (mortality for Southwark-only sub-districts rose somewhat from 1849 to 1854) and differences across districts (mortality was somewhat lower for the jointly-supplied subdistricts in 1849). We want to control for or remove these time and region effects, in other words we want to apply a formal difference-in-differences regression analysis. Table 4 lends itself to a particularly clean and simple example of a difference-in-differences analysis that controls for these effects.

Table 4: Mortality Rates from Cholera per 10,000 Persons in 1849 & 1854, Summary from Snow Table XII & Table VIII

Region or Sub-District Subtotals (Supplied by)	1849 Deaths per 10,000	1854 Deaths per 10,000
First 12 (Southwark & Vauxhall Water Company Only)	135	147
Next 16 (Joint Southwark & Vauxhall and Lambeth Companies)	130	85
Last 4 (Lambeth Water Company only)	85	19
Total	130	104

Table 5 shows the intuition, how we perform a difference-in-differences analysis for the rates from Table 4. (We work with natural logs of rates, for reasons detailed in the appendix.) First, to control for changes across time (which we assume are the same for both regions) we difference across columns. Second, to control for inherent differences in regions (which we assume are the same across time) we difference across rows. We are left with a pure treatment (clean water) effect of -0.511, which translates into a reduction in mortality of 1.67 times (exp(+0.511). Note, of course, that we can either difference first across columns (time) and then regions (rows) or first across rows (regions) and then columns (time). The answer is (and has to be) the same.

Table 5: Difference-in-Differences Calculations for Log Rates (rates from Table 4)

Region or Sub-Districts – Supplied by	1849 Death Rate (log)	1854 Death Rate (log)	Diff 1854 less 1849
First 12 – Southwark Only	$\ln(.0135) = -4.306$	$\ln(.0147) = -4.223$	0.084
Next 16 – Joint Southwark and Lambeth	$\ln(.0130) = -4.342$	$\ln(.0085) = -4.769$	-0.427
Diff Joint less Southwark	-0.036	-0.547	**-0.511**

One question Table 5 *cannot* answer is how much confidence we should have in the estimate of -0.511. Table 32 shows that there is a total population of 486,936, with 167,654 and 300,149 in the two sub-regions. These are large numbers and from this we might like to conclude that the mortality rates (or log-rates) are very precisely estimated, but this conclusion is wrong. Table 33 shows that there is substantial variation in the data – mortality varies across both sub-districts and time (1849 versus 1854).[26] We need a statistical framework in which we can embed both the treatment effect measured in Table 4 and the underlying sub-district variation shown in Table 33. This framework is laid out in the following sub-section and the appendix.[27] Ultimately we will see that there is an effect that is both statistically and economically significant, but the evidence is not as overwhelming as might appear from Table 5.

[26]For example, mortality in two of the joint Vauxhall/Lambeth sub-districts (Kennington 1st and Clapham) actually increases, by quite a bit, from 1849 to 1854.

[27]We want to use data internal to the observations to assess variability of the estimates, what Stigler [2016] calls "intercomparison".

6.1.1 Regression Analysis for Difference-in-Differences

The analysis for Tables 4 and 5 is particularly simple because we have only "before-versus-after" (1849 versus 1854) and "control-versus-treatment" (Southwark-only versus joint regions). This makes Snow's data particularly simple and clear as an example of difference-in-differences, fully described by Table 5.

More generally we will have many variables – multiple regions, many time periods, and possibly covariates which we want to control for.[28] To generalize this difference-in-differences analysis we can write the rates as a function of year, region, and treatment fixed effects:

$$\ln(R_{region,yr}) = \mu + \delta_{54} \cdot I_{yr=1854} + \gamma_J \cdot I_{region=joint} + \beta \cdot I_{region=joint} \cdot I_{yr=1854} + \alpha \cdot Covariates + \varepsilon \quad (1)$$

δ_{54} Year effect – increase in rate for 1854 versus 1849

γ_J "Joint Region" effect – increase (or decrease) in rate for jointly-supplied region versus Southwark-only

β Treatment effect – increase (or decrease) in rate for joint (treatment) region when treated versus untreated

$I_{yr=1854}$ *etc.* Indicator or dummy variables that are 1 for year=1854, region=joint, etc.

Table 6 shows the mapping for this simple example between the data (Table 4) and Equation 1. This example is simple and the mapping from coefficients to entries in the table are one-to-one.

Table 6: Difference-in-Differences Structure for Rates in Table 4

Region or Sub-Districts – Supplied by	1849 Death Rate (log)	1854 Death Rate (log)	Diff 1854 less 1849
First 12 – Southwark Only	$\ln\left(R_{S,49}\right) = \mu$	$\ln\left(R_{S,54}\right) = \mu + \delta_{54}$	δ_{54}
Next 16 – Joint Southwark and Lambeth	$\ln\left(R_{J,49}\right) = \mu + \gamma_J$	$\ln\left(R_{J,54}\right) = \mu + \delta_{54} + \gamma_J + \beta$	$\delta_{54} + \beta$
Diff Joint less Southwark	γ_J	$\gamma_J + \beta$	β

To address the question of statistical variability of the estimates in Table 5 we use the difference-in-differences Equation 1 for a regression with the data in Tables 32 and 33. The error term ε allows for variation across sub-districts and time and using a regression and statistical framework allows us ask to questions about the precision of the estimated coefficients.[29]

Analyzing the data in the regression framework of Equation 1 is a powerful idea, providing the statistical framework for asking and answering crucial questions. It is in this context that I disagree with Freedman and his opinion of regression analysis: "regression models are not a particularly good way of doing empirical

[28] For this example we could include, if we had them, measures such as average age, gender, income, social rank or class, housing density; characteristics that we might think would be associated with or we would like to test for association with cholera mortality.

[29] Technically, regression Equation 1 should be a count regression with the error term ε being, for example, Poisson or Negative Binomial instead of normal. This is discussed in the appendix. For the moment we will think of Equation 1 as a standard linear regression (normal errors). The ideas are the same, the results not too different, and I will quote the count regression results as necessary.

work in the social sciences today" (Freedman [1991] p. 304). Further, for Freedman "Snow's work exemplifies one point on a continuum of research styles; the regression examples mark another" (p. 304). My point is that while Snow's 1855 work is a powerful example of good research, it is incomplete without the statistical testing provided by embedding it in a regression framework. Snow's work and regression are not contrasting points on a continuum but rather necessary complements, two components of a complete research program.[30]

Column 1 of Table 7 shows the linear regression of (log) rates in Equation 1 with no covariates. The estimated coefficient is -0.567. This does not match the -0.511 in Table 1 for a good reason: Table 1 is based on counts and effectively aggregates rates (averages) across sub-district based on population, while Equation 1 estimated as a linear regression on log rates does not weight by population. The solution, discussed in detail in the appendix, is to use counts and Poisson and Negative Binomial regression. Column 2 of Table 7 shows the Poisson regression corresponding to the column 1 linear-in-rates regression – the coefficient matches Table 5.

Table 7: Regressions for Sub-District Difference-in-Differences 1849 vs 1854

	Linear regression (log rate), single treatment effect	Count (Poisson), single treatment effect	Count (Neg Binomial), Single treatment effect	Linear regression (log rate), two treatment effects	Count (Neg Binomial), two treatment effects
Treatment – "less Lambeth" (ln)	**-0.567**	**-0.511**	**-0.500**	**-0.355***	**-0.338***
standard error	**0.311**	0.039	**0.246**	**0.325**	**0.248**
z value (coeff/SE)	-1.8	-13.2	-2.0	-1.1	-1.4
p-value	7.5%		4.2%	28.1%	17.3%
treatment (ratio)	1.76	1.67	1.65	1.43	1.40
Treatment – "more Lambeth" (ln)				**-1.203**	**-1.132**
standard error				**0.460**	**0.353**
z value (coeff/SE)				-2.6	-3.2
p-value				1.18%	0.14%
treatment (ratio)				3.33	3.10
Residual deviance (df, p-value)		1542 (52, p<1e-10)	59.8 (52, p=22%)		60.0 (51, p=18%)
theta ("size" from Gamma mixing)			4.96		5.57
Single region / less Lambeth fixed effect	0.078*	-0.036*	-0.032*	0.029*	-0.064*
More Lambeth fixed effect				0.223*	0.059*
Time fixed effect	0.074*	0.084*	0.057*	0.074*	0.057*

Deaths by sub-district from 1849 and 1854 for the 28 sub-districts ("first 12" Southwark-only and "next 16" jointly-supplied) shown in Snow [1855] Table XII and my Table 32, with population from Snow's Table VIII. Total 56 observations. * = ***not*** significant at the 10% level (robust errors for Poisson regression). The Poisson and Negative Binomial regressions are fitted with the R function glm (family=poisson) and the glm.nb function from the MASS package. The parameter "theta" is the size or θ for a "parametrization (1)" Negative Binomial (see appendix). Robust standard errors are calculated with the R "sandwich" package, using the default "HC3" for the adjusted variance-covariance matrix (see Zeileis [2004] and the R "sandwich" manual).

The Poisson regression, however, has a serious issue that is highlighted by the large value for the "Residual deviance". The variation in the observed counts (variation across and within sub-districts) is much more than can be accounted for by assuming that counts are Poisson – what is called in the literature "overdispersion". This is discussed in more detail in Appendix Section 10.1 but the immediate implication is that the usual standard errors are dramatically too small. We could calculate robust standard errors (and we

[30]In fairness to Freedman, he does highlight examples that *are* poor uses of regression.

do so in the appendix), but a better solution is to build in the necessary random variation with Negative Binomial regression, which then provides realistic standard errors. Column 3 of Table 7 shows that the overall treatment effect is large (coefficient -0.505, a factor of 1.66 reduction) but is only estimated with a moderate degree of precision (p-value 4.1%).

The regression approach has an additional benefit, besides quantifying the treatment effect and the uncertainty implied by observed (in-sample) variation – we can easily estimate multiple treatment effects. Snow says that four of the jointly-supplied sub-districts had more Lambeth-supplied customers, and we should therefore expect to observe a larger reduction in mortality in these sub-districts.[31] The regression framework of Equation 1 allows us to decompose the overall treatment effect into a "more-Lambeth" (the four named sub-districts) and a "less-Lambeth" effect (the remaining jointly-supplied sub-districts.

The fourth and fifth columns show the regressions (linear log-rates and Negative Binomial) with two treatment effects. The fifth column (Negative Binomial) shows a very large and precisely-estimated more-Lambeth treatment, reduction by a factor of 3.12 with p-value 0.14%.

Putting the aggregate group comparison of Table 5 into the difference-in-differences a regression framework of Equation 1 (and 5) provides both more generality (allowing for covariates, for example) and a statistical framework for answering questions about the precision of the estimates. This regression framework allows us to use the substantial variability shown in Table 33 to assess our confidence in the observed reduction in mortality. (This is using what Stigler [2016] calls "intercomparison".)

With these statistical tools, tools that Snow did not have, we can say with a high degree of confidence that the introduction of a clean water supply by the Lambeth Water Company led to substantially lower death rates in 1854 compared with both 1849 and the dirty water supply of the Southwark and Vauxhall Company.

6.2 Quasi-Randomized Comparison: 1854 Cholera Mortality Within Joint-Supply Sub-Districts

6.2.1 Snow's Comparison: Table IX

Snow recognized that a fortuitous combination of circumstances, together with his own hard work, provided a near-perfect experiment to compare the effect of dirty versus clean water. As already mentioned, the Lambeth Water Company moved its water source in 1852 to Thames Ditton, above London's major sewage outflows.[32] Furthermore, in the 16 sub-districts supplied jointly by the Southwark & Vauxhall Company and the Lambeth Water Company, customers were mixed in an apparent random manner.[33] The change of Lambeth's water source in 1852 seems to be independent of customers' choice of water company (a choice made years previously) and can plausibly be taken as exogenous to individuals' choice of water company.[34]

[31]"In certain sub-districts, where I know that the supply of the Lambeth Water Company is more general than elsewhere, as Christchurch, London Road, Waterloo Road 1st, and Lambeth Church 1st, the decrease of mortality in 1854 as compared with 1849 is greatest, as might be expected." Snow [1855] p 89

[32]"The Lambeth Company removed their water works, in 1852, from opposite Hungerford Market to Thames Ditton; thus obtaining a supply of water quite free from the sewage of London." Snow [1855]p 68

[33]"In many cases a single house has a supply different from that on either side. Each company supplies both rich and poor, both large houses and small; there is no difference in the condition or occupation of the persons receiving the water of the different companies." Snow [1855] p 75

[34]Customers "were divided into two groups without their choice, and, in most cases, without their knowledge; one group being supplied water containing the sewage of London, and amongst it, whatever might have come from the cholera patients, the other group having water quite free from such impurity." Snow [1855] p 75.

This combination of circumstances meant that the 16 sub-districts jointly supplied by Southwark and Lambeth provided a natural or quasi-randomized control trial for the 1854 cholera outbreak: customers were effectively randomly assigned to the control group (Southwark & Vauxhall, contaminated water supply) or the treatment group (Lambeth, clean water supply) with no opportunity to self-select.

To measure the treatment effect with this trial, however, Snow needed to go beyond the statistics collected by the official agencies (the data shown in Table 32). He needed to ascertain both the water source for each death (to obtain a count of deaths by supplier) and the relevant population served by the two companies (for measuring exposure to suppliers and converting counts to rates).

For measuring exposure or population served by the companies, Snow had data on the number of *houses* supplied by each company: "A return had been made to Parliament of the entire number of houses supplied with water by each of the Water Companies." (Snow [1855] p78) Note that this is *houses* and not number of people, so we have to change our interpretation of the rates somewhat, although I do not think this is a substantial issue.

To ascertain the water source for each death, Snow himself went house-to-house in the jointly-supplied sub-districts to identify the supply for each house in which a death occurred. In many (even most) cases the residents did not know which company supplied their water,[35] but Snow quickly developed a reliable chemical test to distinguish between the water of the two companies, based on salinity.[36]

Snow's exposure measure (houses supplied by the two companies) covered all 32 sub-districts and did not distinguish between the 12 sub-districts supplied only by Southwark versus the 16 jointly supplied. Because some houses in the Southwark-only sub-districts were supplied by well or Thames or even "ditches", Snow also had to ascertain the source for the Southwark-only sub-districts. For this Snow enlisted the assistance of a Mr. John Joseph Whiting, L.A.C.

Snow [1855] Table VIII (p 85) provides the count of deaths for seven weeks ending 26th August 1854, categorized by water source (Southwark&Vauxhall, Lambeth, pump-well, Thames water, or unascertained). In his Table IX (my Table 8) Snow summarized and tabulated the data together with the number of houses supplied by the two companies, Southwark & Vauxhall and Lambeth. Snow calculates a mortality rate per 10,000 houses and finds a large difference between those supplied by Southwark & Vauxhall versus Lambeth: 315 for Southwark & Vauxhall versus 38 for Lambeth-supplied customers. This is, obviously, a large difference. (The "Rest of London" is interesting but not of immediate import here.)[37]

There is an important issue, however, with Snow's Table IX (my Table 8). Snow combines deaths for a supplier across all sub-districts. Table 9 summarizes the deaths from Snow's table broken out by sub-districts: the first 12 sub-districts supplied solely by Southwark & Vauxhall, the next 16 supplied jointly by Southwark & Vauxhall and Lambeth, and the final 4 supplied solely by Lambeth.

[35]"Even when the water-rates are paid by the residents, they can seldom remember the name of the Water Company till they have looked for the receipt." Snow [1855] p 77

[36]The Southwark and Vauxhall Company obtained their water from lower down in the Thames estuary, at Battersea Fields about a half mile above Vauxhall bridge. At that point the water had relatively high salt content, "37.9 grains of common salt per gallon" according to Snow. The Lambeth Water Company's water was from Thames Ditton, above London, and thus had low salinity ("0.95 grains of chloride of sodium in the water"). Snow [1855] pp 77-78) It turns out (Hempel [2007] p 173) that Snow was lucky. The test depended on the higher salt content of Southwark and Vauxhall's water, which resulted from a salt wedge intrusion from the North Sea. The salinity varies depending on the weather and other conditions, but "while [Snow] was carrying out his research the weather stayed exceptionally hot and dry, which meant the Southwark and Vauxhall salt content remained consistently high. ... He later admitted that he had been fortunate."

[37]Snow reported counts for the four weeks ending 5th August separately in Table VII (p 84). The conclusion from those data are the same: mortality for Lambeth-supplied households was dramatically lower. See the appendix for detailed discussion of the Table VII data.

Table 8: Houses, Deaths, and Mortality per 10,000 Households, First Seven Weeks of 1854 Cholera Epidemic – Table IX Snow [1855] p 86

Water Supplier	Number of houses	Deaths from Cholera	Deaths in each 10,000 houses
Southwark and Vauxhall	40,046	1,263	315
Lambeth Company	26,107	98	38
Rest of London	256,423	1,422	59

Note that this corrects a rounding error in the "Deaths in each 10,000 houses" for Lambeth in Snow's original table

Table 9: Deaths from Cholera in the First Seven Weeks of 1854 Cholera Epidemic – Supplier along rows, Region (sub-district) along columns

Water Supplier	Southwark & Vauxhall "first 12" sub-districts	Southwark and Lambeth jointly-supplied "next 16"	Lambeth only "final 4"	All sub-districts total
Southwark and Vauxhall	738	525	–	1,263
Lambeth Company	–	94	4	98
Both Suppliers	738	619	4	1,361

Data on deaths by sub-district and by supplier (Southwark & Vauxhall Co versus Lambeth Co) from Snow [1855] Table VIII (and also reported in Snow [1856] Table I and II). Death counts are for the seven weeks ending 26th August.

For a randomized control trial we need to compare randomized control versus treated subjects, focusing only on the 619 deaths in the "next 16" jointly-supplied sub-districts in the second column; Snow's Table 8 does not do this. The 1,263 deaths reported for the Southwark & Vauxhall Company combine customers living in the "first 12" sub-districts served solely by the Southwark and Vauxhall company together with customers living in the "next 16" joint sub-districts. Only those in the "next 16" joint sub-districts are randomly mixed with Lambeth customers; those in the "first 12" Southwark-only sub-districts chose to live there and may differ from those in the "next 16" joint sub-districts in ways we cannot determine.[38] Similarly for 98 deaths reported for the Lambeth company, which combines randomly-mixed customers in the joint sub-districts with customers choosing to live in the Lambeth-only sub-districts.

6.2.2 Extending Snow's Quasi-Randomized Trial

There are two directions we can extend Snows results. First, with some reasonable assumptions we can isolate, at least approximately, the quasi-randomized control-versus-treated comparison within the jointly supplied sub-districts. Second, we can apply some rather basic statistical analysis to ask how confident we should be that the observed differences in mortality rates are due to water supply and not random variation.

To compare Southwark versus Lambeth for only the jointly-supplied sub-districts we require the houses supplied by each in the joint region. We can in fact impute a reasonable estimate. The last column of Table 10 shows the reported number of houses by supplier. The final row of Table 10 shows the 1851 population. We can distribute houses to sub-district according to population (essentially assuming the persons-per-house are the same in all sub-districts), giving the third row of the table. Now, the key trick is that in the "first 12" Southwark & Vauxhall-only sub-districts, all of those 22,777 imputed houses must be supplied by Southwark

[38]The difference-in-differences analysis discussed above is a method to control for this, and the analysis there implies that there are not large observed differences between the Southwark-only versus jointly-supplied districts, apart from water supply. Nonetheless the point remains that a clean randomized control trial compares only the randomly-mixed subjects.

& Vauxhall. Given we know the total houses supplied by Southwark & Vauxhall (40,046) we can infer that the remaining 17,269 houses are in the jointly-supplied sub-districts. Similarly, for the "last 4" Lambeth-only sub-districts, all the 2,599 imputed houses must be supplied by Lambeth, implying that the remaining 40,777 are in the jointly-supplied sub-districts.

Table 10: Actual (bold) and Imputed Houses by Supplier and Sub-District

Water Supplier	Southwark & Vauxhall "first 12" sub-districts	Southwark and Lambeth jointly-supplied "next 16"	Lambeth only "final 4"	All sub-districts total
Southwark & Vauxhall houses	22,777	17,269	–	**40,046**
Lambeth Company houses	–	23,508	2,599	**26,107**
Both Suppliers houses	22,777	40,777	2,599	**66,153**
POPULATION	**167,654**	**300,149**	**19,133**	**486,936**

We can now combine Tables 9 and 10 to get imputed mortality rates by supplier and sub-district region, shown in Table 11. The final column reproduces the mortality rates shown in Snow's Table IX but it is the bolded entries in the second column in which we are interested. These measure mortality for the control versus treatment customers, for only the randomly mixed jointly-supplied sub-districts.[39][40] If we take Snow's assurances about the mixing of customers, then this is close to a randomized control trial. And the results show dramatically lower mortality for customers drinking clean water supplied by the Lambeth Water Company.[41]

The next question is what confidence do we have in the large differences shown in mortality rates in Tables 8 and 11? This is a difficult question but one approach is to assume that in 1854 the probability of death follows the same Negative Binomial distribution as estimated above for the 1849 versus 1854 difference-in-differences analysis. In other words, assume that each individual's mortality risk is Poisson (approximating a Binomial) while across the population the mortality rate is Gamma-distributed with shape $\theta = 5.48$. This implies that the count will be Negative Binomial with mean equal to the observed mortality count and Negative Binomial size (Gamma-mixing shape) $\theta = 5.48$.[42] With this assumption we can calculate 95% confidence bands and get some idea how different the observed mortality rates are, relative to random variation.

[39] It is only the exposure measure – the number of houses in the jointly-supplied sub-districts – that is imputed; counts by sub-district regions are observed. Because of the population distribution (particularly the large population for the jointly-supplied sub-districts relative to the last-4 Lambeth-only sub-districts) the resulting mortality rates are not overly sensitive to the imputation assumption. The imputation assumes that the housing density (number of people per house) is the same across all sub-districts at 7.36 people per house. Houses are then distributed across sub-district regions (Southwark-only, joint, Lambeth-only) according to observed 1851 population. Alternatively, if we assumed the Southwark-Only (first 12 sub-districts) had a density 1.5-times higher and the Lambeth-only (last 4 sub-districts) 1.5-times lower, this would change the mortality rates to 211 and 42, still substantially different from each other.

[40] Snow claims that at the beginning of the epidemic mortality was concentrated particularly among customers of the Southwark and Vauxhall Company: "In the beginning of the late epidemic of cholera in London, the Thames water seems to have been the great means of its diffusion, either through the pipes of the Southwark and Vauxhall Company, or more directly by dipping a pail in the river." Snow reports deaths for the first four weeks (ending August 5th) in his Table VII (p 84) and applying the analysis discussed here supports Snow's assertion. Mortality for the Southwark versus Lambeth companies in the jointly-supplied sub-districts were 66.6 and 5.96 per 10,000 for the first four weeks, different by a factor of 11.2 and larger than for the full period (per Table 11, mortality 304.0 and 39.99 per 10,000 or a factor of 7.6).

[41] Further results in the appendix, using population data by sub-district from Snow [1856], reinforce the conclusion that mortality for Lambeth Company customers, supplied with clean water, was dramatically lower than for Southwark & Vauxhall customers.

[42] This likely over-estimates the dispersion in the actual distribution because the theta parameter $\theta = 5.48$ is probably too low (meaning variance too high). The estimate $\theta = 5.48$ embeds the variation across sub-districts including both the Southwark-only and the jointly-supplied sub-districts, variation across time (1849 versus 1854), and variation in treatment (dirty versus clean water in 1854) that is not captured by the two treatment effects estimated in the model of the fourth column of Table 26.

Table 11: Mortality per 10,000 Households, First Seven Weeks of 1854 Cholera Epidemic – Calculated from Tables 9 and 10

Water Supplier	Southwark & Vauxhall "first 12" sub-districts	Southwark and Lambeth jointly-supplied "next 16"	Lambeth only "final 4"	All sub-districts total
Southwark & Vauxhall	324.0	**304.0**	–	315.4
Lambeth Company	–	**40.0**	15.4	37.5
Both Suppliers	370.6	159.9	69.2	228.9

The right-most column reproduces Snow [1855] Table IX (note that the "37.5" rounds to "38", correcting a minor rounding error in Snow's original table). The first three columns extend Snow's table. Data on deaths by sub-district and by supplier (Southwark & Vauxhall Co versus Lambeth Co) are from Snow [1855] Table VIII and summarized in my Table 9. Houses are imputed and shown in my Table 10.

Table 12 shows these 95% confidence bands, with counts converted to rates by dividing by imputed houses. This implies that although there could be substantial variation in mortality rates arising simply from statistical variation, such random variation is very unlikely to lead to the large observed difference of 304 versus 40 per 10,000 households. This give a strong indication that the calculated mortality rates result from the different water supply.

Table 12: Mortality per 10,000 Households and 95% Confidence Bands, First Seven Weeks of 1854 Cholera Epidemic

Water supplier (in joint sub-districts)	2.5% quantile	Mean	97.5% quantile
Southwark & Vauxhall	104	**304**	608
Lambeth Company	13	**40**	81

6.3 Summary For South London "Grand Experiment" – Snow [1855]

Snow took advantage of a set of fortuitous circumstances that provided two types of experimental design. For the first, what we would now call a difference-in-differences design, Snow had aggregate or group observations on mortality (number of deaths) for both an untreated control group (12 sub-districts supplied only by Southwark and Vauxhall) and a treatment group (16 sub-districts supplied jointly by the Southwark and Vauxhall Company and the Lambeth Company). Snow observed those groups both before and after a change in treatment, a change that was applied differentially to the two groups. The observations on the groups over time and across differential treatment regimes meant that Snow could plausibly argue that the resulting differences in mortality were due solely to differences in treatment – clean versus contaminated water.

The regressions in Table 7 and Figure 17 provides compelling statistical evidence that the change in water supply from 1849 to 1854 produces a very large fall in mortality, even given the substantial variation in the observed 1849 and 1854 mortality (Tables 32 and 33). This difference in mortality shows up strongly even after we control for observed differences between the treatment versus non-treatment sub-districts and observed change over time from 1849 to 1854.[43] It is difficult to envision any alternative explanation (apart

[43]The regressions in Table 7 control for region (untreated (Southwark-only) versus untreated (jointly-supplied)) and time (1849 versus 1854). These effects are shown in Table 7, are very small relative to the water treatment effect, and not statistically significant. This provides evidence that there are not substantial or consistent observed differences (in mortality) between the untreated versus treated regions or between 1849 and 1854, apart from the water source.

from water supply) for the observed fall in mortality. Any confounding effect would need to be imposed in the same manner as Lambeth's change in water source – between 1849 and 1854, only on the jointly-supplied districts, and more on the "more-Lambeth" sub-districts. Snow's arguments that there was no such change (apart from water source) seem reasonable.

For the second experimental design, Snow exploited the plausibly random mixing and assignment of customers between control (contaminated water, Southwark and Vauxhall customers) versus treatment (clean water, Lambeth customers). Snow's "shoe-leather" work to collect data on the supplier for each cholera death, observations that supplemented the official statistics, provide the raw material for the analysis. By somewhat extending Snow's analysis we can perform what is as close to a truly randomized control trial as one is likely to find in a natural setting.

The observed difference in mortality for customers supplied by Southwark & Vauxhall (contaminated water) versus Lambeth (clean water) is large whether measured using Snow's Table IX (Table 8, combining the randomly-mixed population with others) or using imputed houses for exposure (Table 11). The differences are large (a factor of roughly 8) and would be extremely difficult to ascribe to purely random variation.

In Snow's day the accepted theory for the cause for cholera was airborne transmission. Hypothesized contributing or differentiating factors included social status, elevation, weather, housing conditions, innate susceptibility, fear of infection, and others. For the south London data considered here the uniformity of conditions (such as elevation or weather) within the relatively compact south London sub-districts, together with the well-mixed and random-assignment nature of customers' 1854 water source, allowed Snow to control for and thus average out the effect from any and all of these sources. The large observed difference in mortality between control (Southwark customers) versus treatment (Lambeth customers) is very difficult to ascribe to any source apart from the different water source. Although no scientific experiment, whether a controlled or natural experiment, can prove a scientific theory, this natural experiment comes as close as one could hope.

7 Extending South London "Grand Experiment" With Detailed Population Data – Snow [1856]

Snow in 1855 was limited by a lack of data on population by supplier. For the difference-in-differences analysis he could not precisely measure the treatment effect because the proportion of households supplied by Southwark versus Lambeth varies across sub-districts, and he did not have the population proportions. For the quasi-randomized experiment Snow did not have the population by supplier for sub-districts, and so could only compare at the aggregate level of all sub-districts combined (Snow [1855] Table IX, my Table 8).

Snow [1856] made an attempt to rectify this short-coming, using population data published by the Board of Health (Simon [1856]).[44] Snow provided an interesting analysis but one that, by current statistical standards, is insufficient. This section extends Snow's analysis using more modern statistical tools.

It falls into three sections:

1. Review of Snow's analysis
2. Extension of difference-in-differences analysis (1849 vs 1854) using population proportions to estimate more accurate treatment effect
3. Analysis of quasi-randomized trial with first seven weeks (through 26th August 1854) by sub-district and full outbreak (through October 1954) by Registration District

7.1 Review of Snow [1856]

Snow wanted to show that difference in water supplier was predominant, overriding other factors such as crowding or proximity to the Thames. His approach was

- Calculate a mortality rate for the whole region (all sub-districts together) but separately for individuals drinking Southwark-supplied versus Lambeth-supplied water
- For each sub-district predict the count for the two suppliers separately, based on the sub-district population by supplier (and the supplier-specific mortality)
- For each sub-district combine the counts for the two suppliers and then divide by the combined population to get a mortality rate, weighted by the sub-district supplier populations

Snow showed the result of these calculations in his Table VI, reproduced here as Table 13.

There are a number of issues that make these calculations somewhat difficult, detailed in the appendix. For now we simply follow Snow, who compared the compared the calculated mortality with the actual and argued that they "bear a close relation":

> it will be observed that the calculated mortality [the final two columns of Table 13] bears a very close relation to the real mortality [under "Deaths 1854"] in each subdistrict. This relation exists

[44]Simon [1856], tellingly, did not reference Snow [1855] or credit Snow with collecting deaths by supplier for the first seven weeks or prompting the Registrar-General to collect the last ten weeks' statistics.

with regard both to the gross mortality and to the mortality to each 10,000 living, all through the table, and proves the overwhelming influence which the nature of the water supply exerted over the mortality, overbearing every other circumstance which could be expected to affect the progress of the epidemic. Thus, in the crowded, dirty, and very poor subdistricts of Lambeth Church, first part, and Waterloo, first part, lying by the river side, the mortality was low in consequence of the water supply being chiefly that of the Lambeth Company; whilst in the thinly peopled, and comparatively genteel subdistricts of Clapham and Battersea the mortality was very high, in consequence of the impure water of the Southwark and Vauxhall Company. Taking this inquiry altogether, and considering that the results which were published two years ago, and could only be estimated collectively, are now corroborated in detail through upwards of thirty subdistricts, it probably supplies a greater amount of statistical evidence than was ever brought to bear on a medical subject. (Snow [1856] p. 248)

Table 13: Snow [1856] Table VI: Mortality from Cholera in 1854, in Thirty-one Sub-Districts, as compared with Calculations founded on the Results shown in Table V

	1855	District	Sub-District	Pop	Population Estimates by Supplier			Deaths 1854		Calculated Mortality			
	Seq			1851	Southwark	Lambeth	Both Cos	Count	Rate	Southwark	Lambeth	Both	Rate
1	13	St. Saviour, S.	Christchurch, Southwark	16,022	2,915	13,234	16,149	113	*70.5*	*46.6*	*35.7*	*82.4*	*51.0*
2	1	St. Saviour, S.	St. Saviour, Southwark	19,709	16,337	898	17,235	378	*191.8*	*261.4*	*2.4*	*263.8*	*153.1*
3	2	St. Olave	St. Olave, Southwark	8,015	8,745	0	8,745	161	*200.9*	*139.9*	*0.0*	*139.9*	*160.0*
4	3	St. Olave	St. John, Horsleydown	11,360	9,360	0	9,360	152	*133.8*	*149.8*	*0.0*	*149.8*	*160.0*
5	4	Bermondsey	St. James, Bermondsey	18,899	23,173	693	23,866	362	*191.5*	*370.8*	*1.9*	*372.6*	*156.1*
6	5	Bermondsey	St. Mary Magdalen	13,934	17,258	0	17,258	247	*177.3*	*276.1*	*0.0*	*276.1*	*160.0*
7	6	Bermondsey	Leather Market	15,295	14,003	1,092	15,095	237	*155.0*	*224.0*	*2.9*	*227.0*	*150.4*
8	14	St. George, S.	Kent Road	18,126	12,630	3,997	16,627	177	*97.6*	*202.1*	*10.8*	*212.9*	*128.0*
9	15	St. George, S.	Borough Road	15,862	8,937	6,672	15,609	271	*170.8*	*143.0*	*18.0*	*161.0*	*103.1*
10	16	St. George, S.	London Road	17,836	2,872	11,497	14,369	95	*53.3*	*46.0*	*31.0*	*77.0*	*53.6*
11	17	Newington	Trinity, Newington	20,922	10,132	8,370	18,502	211	*100.9*	*162.1*	*22.6*	*184.7*	*99.8*
12	18	Newington	St. Peter, Walworth	29,861	14,274	10,724	24,998	391	*130.9*	*228.4*	*29.0*	*257.3*	*102.9*
13	19	Newington	St. Mary, Newington	14,033	2,983	5,484	8,467	92	*65.6*	*47.7*	*14.8*	*62.5*	*73.9*
14	20	Lambeth	Waterloo Road (1st)	14,088	3,548	11,939	15,487	59	*41.9*	*56.8*	*32.2*	*89.0*	*57.5*
15	21	Lambeth	Waterloo Road (2nd)	18,348	7,171	12,533	19,704	118	*64.3*	*114.7*	*33.8*	*148.6*	*75.4*
16	22	Lambeth	Lambeth Church (1st)	18,409	3,113	15,878	18,991	49	*26.6*	*49.8*	*42.9*	*92.7*	*48.8*
17	23	Lambeth	Lambeth Church (2nd)	26,784	7,868	16,023	23,891	195	*72.8*	*125.9*	*43.3*	*169.2*	*70.8*
18	24	Lambeth	Kennington (1st)	24,261	15,775	2,708	18,483	305	*125.7*	*252.4*	*7.3*	*259.7*	*140.5*
19	25	Lambeth	Kennington (2nd)	18,848	7,874	5,620	13,494	143	*75.9*	*126.0*	*15.2*	*141.2*	*104.6*
20	26	Lambeth	Brixton	14,610	1,922	9,356	11,278	48	*32.9*	*30.8*	*25.3*	*56.0*	*49.7*
21	29	Lambeth	Norwood	3,977	0	1,066	1,066	10	*25.1*	*0.0*	*2.9*	*2.9*	*27.0*
22	27	Wandsworth	Clapham	16,290	6,747	134	6,881	167	*102.5*	*108.0*	*0.4*	*108.3*	*157.4*
23	8	Wandsworth	Battersea	10,560	6,276	276	6,552	171	*161.9*	*100.4*	*0.7*	*101.2*	*154.4*
24	9	Wandsworth	Wandsworth	9,611	907	94	1,001	59	*61.4*	*14.5*	*0.3*	*14.8*	*147.5*
25	10	Wandsworth	Putney	5,280	74	0	74	9	*17.0*	*1.2*	*0.0*	*1.2*	*160.0*
26	30	Wandsworth	Streatham	9,023	0	3,244	3,244	15	*16.6*	*0.0*	*8.8*	*8.8*	*27.0*
27	31	Camberwell	Dulwich	1,632	0	25	25	0	*0.0*	*0.0*	*0.1*	*0.1*	*27.0*
28	11	Camberwell	Camberwell	17,742	9,139	639	9,778	242	*136.4*	*146.2*	*1.7*	*147.9*	*151.3*
29	12	Camberwell	Peckham	19,444	5,438	392	5,830	175	*90.0*	*87.0*	*1.1*	*88.1*	*151.1*
30	28	Camberwell	St. George, Camberwell	15,849	4,295	5,437	9,732	132	*83.3*	*68.7*	*14.7*	*83.4*	*85.7*
31	7	Rotherhithe	Rotherhithe	17,805	12,218	0	12,218	283	*158.9*	*195.5*	*0.0*	*195.5*	*160.0*
		Houses in streets with no death			28,929	23,338	52,267			*462.9*	*63.0*	*525.9*	*100.6*
		Not identified			2,712	165	2,877			*43.4*	*0.4*	*43.8*	*152.4*
		Totals		482,435	267,625	171,528	439,153			*4,282.0*	*463.1*	*4,745.1*	*108.1*
		Population per Registrar-General			266,516	173,748	440,264			*4,264.3*	*469.1*	*4,733.4*	*107.5*

This reproduces Snow [1856] Table VI with minor modifications. "Pop 1851" is the population estimate from the 1851 Census, reported in various tables in Snow [1855, 1856]. "Population Estimates by Supplier" are from Snow [1856] Table V using population estimates from Simon [1856]. Death counts for 1854 are from the Registrar-General and generally match Snow [1855] Table XII. (I assume changes reflect updates to the Registrar-General's counts subsequent to Snow's 1855 publication.) "Calculated Mortality: Southwark" and "Calculated Mortality: Lambeth" are the count of deaths based on mortality rates of 160 and 27 per 10,000 (Snow's calculation from his Table V) and the Southwark & Vauxhall and Lambeth sub-district populations. The final five columns are calculated based on the earlier columns, rounded to one decimal. Snow reported numbers rounded to zero decimals, and there are a few minor differences due to errors in Snow's rounding. (See http://www.hilerun.org/econ/papers/snow/index.html for postings of Snow's original data.) The only substantive error is for the Christchurch sub-district: Snow reported a mortality rate of 57 per 10,000 where it should be 51. "1855 Seq" is the sequence number from Snow's 1855 tables (where sub-districts are sorted by "first 12" Southwark-only, then "next 16" jointly-supplied).

Snow provided no statistical tests, hardly surprising given that the necessary statistical tools were developed many years after. One reasonable approach is the regression on population proportions detailed below in Section 7.2.1. Another testing procedure, proposed by Koch and Denike [2006] but less than ideal, is to treat the sub-district mortality (actual versus calculated) as paired observations and apply a paired t-test (with 31 observations). This is discussed in the next section.

Before turning to more formal testing, it is valuable to examine Snow's data graphically. Although a paired t-test is not a good testing framework, pursuing the idea reveals and highlights important issues. For a paired t-test we would calculate the difference in mortality rates for each sub-district. The left panel of Figure 8 shows these differences as solid circles. The underlying assumption for the paired t-test is that these differences are drawn from a normal distribution with a constant variance. Figure 8 shows error bars under this normality assumption and using the variance calculated from the observed data.

One important reason the paired t-test is not appropriate is that the assumption of constant-variance normality is not true. Each sub-district mortality rate is itself an estimated mean (the sum of Bernoulli binary events – death or non-death – divided by the population – sample size) and each mean has its own standard error. For a large population the rates will go to normal by the central limit theorem. This fact forms the basis for the standard t-test for comparing rates in clinical trials. We will want to allow different variances if the samples sizes are different. (See the Appendix Section 10.3.1 and BMJ, Jakobsen et al. [2015].) The resulting formula for the variance of the difference in rates is $SE(r1-r2) = \sqrt{r1(1-r1)/n1 + r2(1-r2)/n2}$. When one or both sample sizes are small (as for example Putney with only 74 Southwark-plus-Lambeth customers) the standard error will be large.

The right panel of Figure 8 shows the standard error for the differences based on the counts being generated by a Bernoulli process. (Or, as an approximation, a Poisson process.) Some of the error bars are very wide: Putney (sub-district 10) is particularly wide because there are only 74 people supplied by the Southwark & Vauxhall or the Lambeth Companies. The assumption that all sub-districts have the same variance, embedded in the left panel of Figure and the paired t-test above, is clearly far from true.

But the Poisson error bars are still not the complete story. We want to incorporate both the variation *across* sub-districts (the error bars in the left panel of Figure 8) and *within* sub-districts (the right panel of Figure 8). As discussed in Section 7.2.1 and Appendix Section 10.1 we need to use the statistical framework of counts, Poisson and Negative Binomial regression.[45] Figure 9 previews what we see when we incorporate both the within- and across-sub-district variation: The standard errors are considerably wider than shown in either panel of Figure 8.

There remains, however, a subtle but critically important issue. The left panel of Figure 8 assumes that each sub-district rate (or each difference in rates) is drawn from a random distribution (normal in this case) with constant variance. The right panel of Figure 8 assumes that each sub-district rate is a fixed number, which differs across sub-districts. The difference is essentially assuming random versus fixed effects. Combining the across and within sub-district variation to give a Negative Binomial as mixture of Poissons assumes random effects.

[45]In the framework of Poisson count regressions (which is the appropriate mathematical extension of the Bernoulli assumption) we incorporate random sub-district rates by allowing the underlying sub-district rates to themselves be random variables, something very much like the left panel of Figure 8. We mix the rates that generate the Poisson distributions with an underlying distribution, often using a Gamma which then produces a Negative Binomial count distribution. This is discussed in the Appendix Section 10.1.

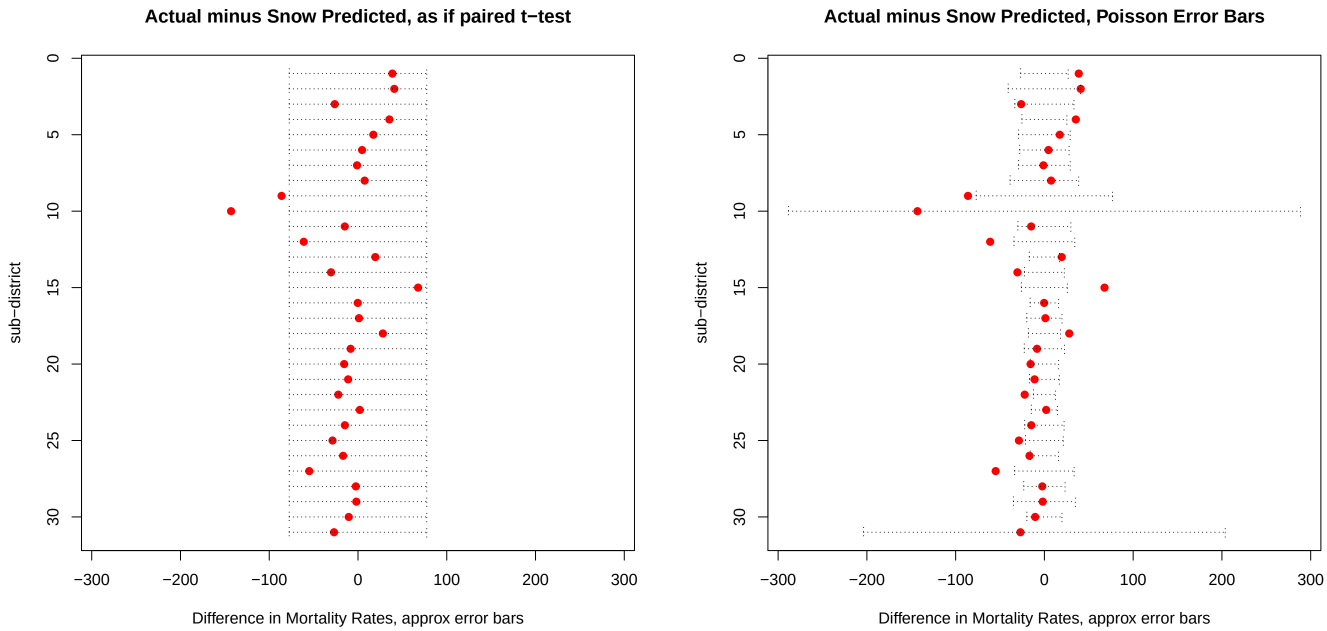

Figure 8: Difference in Mortality Rates (per 10,000) Between Actual and Snow's Predicted for 1854 for the 31 Sub-districts shown in Snow [1856] Table VI (my Table 13)

The red circles are the actual sub-district mortality rate minus Snow's predicted rate based on Southwark and Lambeth populations and Snow's calculated rates (160 per 10,000 for Southwark & Vauxhall and 27 for Lambeth). For consistency with other figures, sub-districts are sorted as in Snow [1855], the "first 12" Southwark-only followed by the "next 16" jointly-supplied sub-districts. The left panel shows approximate 95% error bars *if* the difference in rates were normally distributed random variables with the same variance. The right panel shows approximate 95% error bars assuming the counts are generated by an underlying Bernoulli process with different rates and variances for each sub-district and for the calculated versus actual mortality rates: $SE(r1-r2) = \sqrt{r1(1-r1)/n1 + r2(1-r2)/n2}$.

There are no data in Snow's Table VI (my Table 13) which could distinguish between the two hypotheses of fixed versus random effects. For the difference-in-differences analysis we do have some data (within sub-district comparisons of 1849 versus 1854) but the issue is still delicate. We know that sub-districts differ and so we should expect them to have somewhat different intrinsic rates, but there is also some evidence that rates within sub-district vary. The choice between random versus fixed effects does not have a major impact on estimates of parameters, but it has a dramatic impact on the standard errors and thus the confidence we assign to our estimates.

7.1.1 Koch & Denike

Koch and Denike [2006] undertake a valuable exercise to examine Snow [1856] and Snow's analysis of the South London data. They state as their aim the correction of Snow's analysis: "This paper describes a previously unacknowledged methodological and conceptual problem in Snow's 1856 argument. We review the context of the South London study, identify the problem and then correct it with an empirical Bayes estimation (EBE) approach."

Unfortunately they fail in their goal. They fail for a variety of reasons. First, the statistical test they examine (paired t-test shown in Tables 14 and 15) is both inappropriate and not properly applied. Second, and much more seriously, they seem to misunderstand or misinterpret Snow's data and analysis and as a

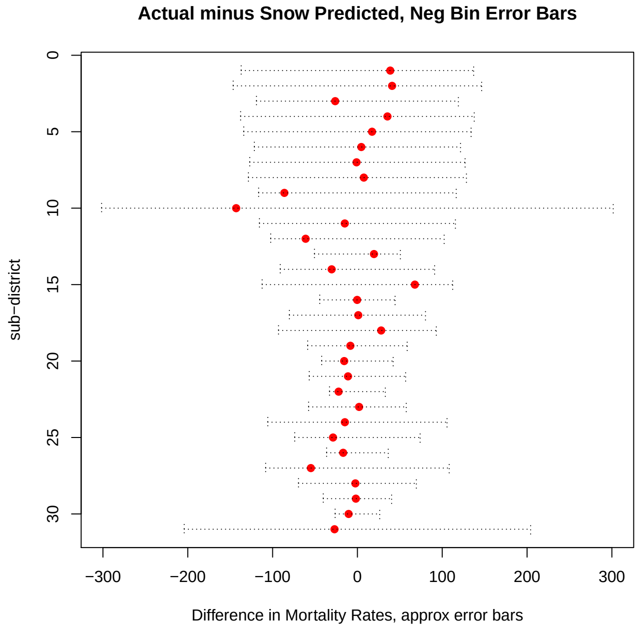

Figure 9: Difference in Mortality Rates (per 10,000) Between Actual and Snow's Predicted for 1854 for the 31 Sub-districts shown in Snow [1856] Table VI (my Table 13)

The red circles are the actual sub-district mortality rate minus Snow's predicted rate based on Southwark and Lambeth populations and Snow's calculated rates (160 per 10,000 for Southwark & Vauxhall and 27 for Lambeth). For consistency with other figures, sub-districts are sorted as in Snow [1855], the "first 12" Southwark-only followed by the "next 16" jointly-supplied sub-districts. The approximate 95% error bars assume the counts are generated by an underlying Negative Binomial process (Poisson mixture process) with different rates and variances for each sub-district and for the calculated versus actual mortality rates: $SE(r1-r2) = \sqrt{\left(r1/n1 + r1^2/\theta\right) + \left(r2/n1 + r2^2/\theta\right)}$, with the Gamma mixing parameter $\theta = 12.8$.

result they inappropriately alter the mortality data from the Registrar-General (used by Snow and others). Although not ill-intentioned this altering of the original mortality data does invalidate their analysis and conclusions. Finally, in my judgment the problem they seek to address is not truly a methodological or conceptual problem in Snow's 1856 argument. Snow indeed lacked the statistical tools to test many of his hypotheses (as Koch and Denike point out); indeed this monograph is an extended attempt to apply current statistical thinking to Snow's data. But even more importantly I argue that Snow's arguments, even without modern hypothesis-testing, are a convincing and persuasive demonstration of water as the causal source of cholera infection.[46]

Turning first to statistical testing, Koch and Denike apply a paired t-test to the differences in counts (deaths) by sub-district. For each sub-district they calculate the difference in mortality. The null hypothesis is that mortality is the same in each sub-district or that the mean of the differences are zero.

In comparing the actual versus calculated mortality we need to use mortality rates per 10,000 population rather than the raw counts. This is particularly necessary here because of the sometimes very different

[46] One additional (somewhat nitpicking) criticism of Koch and Denike is the many typos in their transcription of Snow's tables and their paper overall. For example their Figure 4 (Snow's 1856 Table VI) has errors in the 1851 population for Christchurch, St. Saviour, St. Mary Magdelen, and London Road, to name a few. Regarding their analysis, on page 280 they write "The results returned, $\gamma = 411/4177$ (equalling 0.09983) ...". The figure 4177 should read 4117 (the sum of 3,706 + 411, the figures from Snow's Table V for the assigned Southwark & Vauxhall and Lambeth deaths). The phrase should read "$\gamma = 411/4177$" which is indeed 0.09983. As far as I can determine their typos do not carry through to their later calculations.

populations between actual versus calculated. With different populations the counts will differ even when the underlying mortality rates are the same.[47] Running a paired t-test on the mortality rates produces:

	Mean	S. Dev	Std Err	t-test	Sig (2-tailed)
Difference in mortality rates (per 10,000)	-10.1428	39.5349	7.1007	-1.4284	0.1635

Table 14: Paired t-test for mortality rates from Snow [1856] Table VI (my Table 13)

This supports Snow's assertion of that the actual and calculated mortality "bear a close relation", contrary to Koch and Denike's claim.

Table 15 shows the test performed by Koch & Denike using counts, which implies rejecting the equality of counts. But given the difference in populations, the test of counts would not be appropriate even *if* the paired t-test were an appropriate statistical approach.

	Mean	S. Dev	Std Err	t-test	Sig (2-tailed)
Difference in counts (deaths)	28.7609	47.0320	8.4472	3.405	0.002

Table 15: Paired t-test for death counts from Snow [1856] Table VI (my Table 13)

More fundamentally the paired t-test comparison of sub-district mortality (e.g. Table 14) is not an appropriate test. The primary reason is that it does not incorporate the stochastic nature of the counts and rates for the individual sub-districts, displayed in the right panel of Figure 8. A second reason is that it does not incorporate the uncertainty in the estimated Southwark and Lambeth mortality rates used to predict counts and rates. A third reason is that the actual mortality includes *three* sources for water: Southwark & Vauxhall Company, Lambeth Company, and "Other" (pump-wells, Thames, ditches). We should incorporate all three sources when comparing an actual versus predicted mortality.

Turning from statistical testing to the data, the more serious problem with Koch and Denike [2006] is misunderstanding Snow's original data, which leads them to (unintentionally) alter and thus contaminate the original mortality data. The problem Koch & Denike set out to address is that for the 1854 cholera outbreak there were 623 deaths in the South London districts that were not assigned to water company supplier (Southwark & Vauxhall Company versus Lambeth Company). They state "there were 623 houses in which cholera occurred that could not be assigned reflexively to any single district nor to either of the two water supplier areas." (p. 275) The first part of this statement is simply incorrect while the second part is slightly confusing. There were 623 deaths in houses where the house could not be assigned to water *supplier*, but all those death (and the houses in which they occurred) were clearly assigned, in the original reports from the Registrar-General, to Registration District and sub-district. The assignments to Registration District are clearly shown in Snow's Table V (Koch and Denike's Figure 3, my Table 16). There has never (to my knowledge) been any question about the reliability of the Registrar-General's assignment of deaths to

[47] Snow is comparing the overall sub-district mortality with a prediction based on only Southwark and Lambeth supplied customers, a population that may be substantially lower. For Putney the overall population is 5,280 while the combined Southwark plus Lambeth population is only 74: many houses were supplied by pump-wells or the Thames. For Kennington 1st the total population is 24,261while the combined Southwark and Lambeth population is only 18,483. The observed count (total population 24,261) is 305, giving an estimated rate of 125.7 per 10,000. If we apply the rate of 125.7 to the Southwark and Lambeth population (18,483) we would have an expected count of only 232.2. Comparing 305 versus 232.2 is a difference of 72.7 but this reflects only differing population size and not underlying mortality or infection.

District or Sub-District – in contrast to assignment to water *supplier* within sub-district.[48][49]

Koch & Denike proceed to re-allocate those 623 deaths across water supplier *and* Registration Districts (see for example their Figure 6, my Table 17). In doing so they move deaths across Districts, deaths that were reliably located – assigned to District by the Registrar-General. The re-assignment across Districts is neither necessary nor justified and corrupts the original data. The re-assignment introduces substantive errors in mortality rates for those districts with either below-average or above-average unassigned deaths. For example the mortality rate for St. Saviour, Southwark is increased from 137.4 to 156.8 (per 10,000), while Lambeth is reduced from 66.5 to 56.5. All of these changes in overall mortality at the Registration District level are arbitrary and counter to the observed data.[50]

The original problem in Snow's analysis of the 623 deaths not assigned to supplier can, I think, be handled more simply by allocating at the District or sub-district level in proportion to the reported deaths for the two suppliers. This is what Snow proposes:

> The instances in which the water supply was not specified, or not ascertained, in the returns made by the district registrars must evidently nearly all have been cases in which the house was supplied by one or other of the water companies, for, if the persons received no such supply, and obtained water from a pump well, canal, or ditch, there could be no difficulty in knowing the fact. Moreover, as the two water companies are guided by precisely the same regulations, the difficulty in ascertaining the supply is exactly the same with regard to one as the other; I, therefore, concluded that I could not be wrong in dividing the non-ascertained cases between the two companies in the same proportion as those which were ascertained, and I have done so at the foot of table V (Snow [1856] p. 247)

As Snow argues this is reasonable because a) there appears to be little difference between customers of the

[48] Throughout, Koch & Denike imply, incorrectly, that the problem with the 623 deaths is spatial location. On page 272 they state: "Snow attempted to extend the incomplete field using a then recently compiled but still incomplete inventory of deaths from cholera at registration district and sub-district levels through the simple expedient of allocating cases that *could not otherwise be spatially located* [emphasis added] to water companies according to the best estimate then available." All the 623 deaths were correctly *geographically* or spatially located within District and sub-district. (See, for example, Snow's 1856 Table V. Even more instructive is the appendix to Snow [1855] where Snow lists *all* deaths for the four weeks ending 5th August 1854. Snow had the address of each death and all deaths were assigned to sub-districts, as is obvious from perusing the list. Snow visited the households to ascertain the water source. In some cases Snow could not ascertain the water source; for example on p. 141 "At St. Thomas's Hospital, supposed from Red Cross Street, Southwark, on 31st July, a charwoman, aged 50, 'cholera'". In all cases, however, the location was known and recorded.)

[49] There are other instances where Koch & Denike's statements are not good representations of Snow's analysis and methods. Koch and Denike [2006] p. 273 state: "'All that was required,' [Snow] wrote, 'was to learn the supply of water to each individual house where a fatal attack of cholera might occur' (Snow, 1855, p. 75). *Data permitting assignment of cholera deaths to either of the water suppliers was unavailable, however*. [emphasis added]" Data on assignments *is* available because Snow himself collected the data through August 26 and convinced the Registrar-General to collect data following August 26. Snow was stating the problem but he then proposed *and implemented* the solution (see Snow [1855] p. 76 ff). Another example is page 278 where Koch and Denike state "Required was some form of smoothing permitting a better appreciation of the reliability of risk of cholera in registration districts and thus a better assignment of the 623 unassigned cases (*561 from Southwark-Vauxhall, 62 from Lambeth Company*) [emphasis added]." This seems to be a mis-understanding (or poor explanation) of Snow's data. The fundamental problem is that we do not know how many of those 623 unassigned cases are Southwark & Vauxhall and how many Lambeth. The figures of 561 and 62 are in fact only Snow's estimates, his allocation of the unascertained cases between Southwark and Lambeth based on the reported (known) assignments of 3,706 and 411 (see Snow's Table V). The problem is not (as Koch and Denike seem to imply) the location or *geographic* assignment of deaths, but the assignment to water supplier for deaths that are well-identified as to location.

[50] I have focused on Koch & Denike's Figure 6 which shows results at the District level. Their Figure 7 re-assigns the 623 deaths at the sub-district level and some of those results alter mortality even more dramatically. Consider the District of Newington, comprised of three sub-districts and an overall population of 64,816. Koch and Denike's Figure 7 show deaths from both companies (Southwark and Lambeth) across the three Newington sub-districts totaling to 501.14. Adding the 2 pump-wells deaths gives a total of 503.14 versus the true total of 694 deaths. Dividing by the population gives a Newington District mortality rate of 77.6 per 10,000, dramatically below the true rate of 107.1.

Table 16: Snow [1856] Table V – Deaths and Mortality by Water Supply and Registration District for 1854 Outbreak

Registration District	Pop 1851	Southwark	Lambeth	Pump-well	Unasc	Total	Mortality per 10,000
1, St. Saviour, Southwark	35,731	406	72	10	3	491	137.4
2, St. Olave, Southwark	19,375	277	0	8	28	313	161.5
3, Bermondsey	48,128	821	0	25	0	846	175.8
4, St. George, Southwark	51,824	388	99	0	56	543	104.8
5, Newington	64,816	458	58	2	176	694	107.1
6, Lambeth	139,325	525	138	24	240	927	66.5
7, Wandsworth	50,764	268	7	106	40	421	82.9
8, Camberwell	54,667	352	33	115	49	549	100.4
9, Rotherhithe	17,805	207	0	46	30	283	158.9
10, Greenwich & sub-districts, Sydenham		4	4	2	1	11	–
TOTAL	482,435	3,706	411	338	623	5,078	105.3

The 1851 population for Camberwell is reported by Snow in his Table V as 54,607 when in fact it should be 54,667 (according to cross-checks by summing and comparing with his 1856 Table VI and 1855 Table VIII). This is one of the few errors or typos in Snow's publications that I have found.

Southwark & Vauxhall Company versus the Lambeth Company (this is the whole basis for Snow using the South London data); b) thus unassigned houses should be random relative to supplier; c) so the probability of an unassigned house belonging to one or the other supplier should be proportional to the probability of death by supplier at the sub-district level.

7.2 Extending Difference-in-Differences and Quasi-Randomized Trials Using Detailed Population Data

Snow [1856] and Simon [1856] published estimates collected by the Registrar-General of density, houses, and population served by the water companies (Southwark & Vauxhall Company versus the Lambeth Company) by sub-district. Although there are various errors in the estimates (discussed more in the appendix) this provides the underlying population-at-risk data needed to extend both the difference-in-differences and the quasi-randomized trial analyses.

There are three sets of analyses we can perform, determined by the details of the count (death) data that we have:

1849 vs 1854 by Sub-district: Snow [1855] Table XII has data on overall sub-district counts for 1849 and 1854 and these data form the basis for the 1849 versus 1854 difference-in-differences analysis. The overall count includes individuals supplied by three sources: Southwark & Vauxhall, Lambeth, and "Other" (pump-wells, Thames, ditches, etc.). We now have the population for each source. We can expand the original "Lambeth" 1854 fixed effect to include effects for those three sources, measuring the population fraction supplied by each source. This probably comes the closest to what Snow was trying to accomplish in Snow [1856] Table V.

7 wks Southwark vs Lambeth by Sub-district Snow [1855] Table VIII gives deaths by supplier (Southwark & Vauxhall, Lambeth, Other) for the first seven weeks (ending 26th August 1854). This formed the basis of Snow [1855] Table IX, measuring a treatment effect in a quasi-randomized trial. With population by supplier by sub-district we can examine each sub-district and test for the difference between

Table 17: Koch and Denike [2006] Figure 6 – Deaths and Mortality by Water Supply and Registration District, Re-assigned *Across Districts* with Empirical Bayes Estimation Process

Registration District	Pop 1851	Southw	Lamb	Pump-well	Unasc	Total	Mortality per 10,000 K&D Calc	Original
1, St. Saviour, Southwark	35,731	467.96	82.37	10	0	560	156.8	137.4
2, St. Olave, Southwark	19,375	317.63	1.28	8	0	327	168.7	161.5
3, Bermondsey	48,128	943.92	1.32	25	0	970	201.6	175.8
4, St. George, Southwark	51,824	447.88	112.82	0	0	561	108.2	104.8
5, Newington	64,816	527.36	66.72	2	0	596	92.0	107.1
6, Lambeth	139,325	605.61	157.72	24	0	787	56.5	66.5
7, Wandsworth	50,764	307.59	9.03	106	0	423	83.3	82.9
8, Camberwell	54,667	405.02	38.24	115	0	558	102.1	100.4
9, Rotherhithe	17,805	237.06	1.26	46	0	284	159.7	158.9
10, Greenwich & sub-districts, Sydenham		6.78	2.43	2	0	11	–	–
TOTAL	482,375	4,267	473	338	0	5,078	105.3	105.3

The "Southwark" and "Lambeth" columns are from Koch and Denike [2006] Figure 6, "assignments using the empirical Bayes estimation process". Pump-well are from Snow [1856], the "Total" is the sum of the other columns. The 623 unassigned deaths from Snow [1856] Table V are (incorrectly) re-assigned across sub-district.

Southwark versus Lambeth much more carefully. This only covers the first seven weeks (through 26th August) and we also have to recognize the problems with the population estimates (discussed more in the appendix) which are more serious at the sub-district than District level.

1854 Southwark vs Lambeth by District Snow [1856] Table V gives deaths by supplier by nine Districts for the full 1854 period (ending October 1854).[51] These data allow testing for the Southwark versus Lambeth effect in a quasi-randomized trial framework at a somewhat more aggregated level than the sub-district data above, but covering the full 1854 cholera outbreak

7.2.1 1849 versus 1854 Difference-in-Differences

With population by supplier and sub-district we can estimate a mortality rate specific to each of the two water companies. Snow [1855] Table XII (my Tables 32 and 33) have variation in mortality and Snow [1856] Table VI (my Table 13) shows the variation in population. We can now estimate how much of the mortality variation is due to differences in mortality rates and population variation.

We would like to estimate an equation of the form:

$$R_{subdis,yr} = R_S \cdot P_S + R_L \cdot P_L + R_O \cdot P_O + R_L^{54} \cdot P_L \cdot I_{yr=1854} + \delta_{54} \cdot I_{yr=1854} + \varepsilon \quad (2)$$

δ_{54} Year effect – increase in rate for 1854 versus 1849

R_S Mortality rate by supplier: for Southwark, Lambeth, Other – estimated parameter

P_S Proportion of sub-district population supplied by Southwark, Lambeth, Other – observation (input data)

[51] Totals across supplier tie against the sub-district data in Table VI which in turn mostly match the 1854 data in Snow Snow [1855] Table XII. There are small differences which I assume reflect revisions in the Registrar-General's tabulations.

$I_{yr=1854}$ *etc.* Indicator or dummy variables that are 1 for year=1854, region=joint, etc.

R_L^{54} Change in mortality rate for Lambeth in 1854

This says that the overall rate for each sub-district is a population-weighted average of the rates for Southwark, Lambeth, and "other" customers. The coefficient R_L^{54} measures the decrease in mortality in 1854 for Lambeth customers, which is presumably due to clean water.

This provides a reasonable conceptual and statistical framework for examining Snow's contention (Snow [1856] p. 248) that the actual versus predicted mortality "bear a close relation". The mortality rates are estimated and the regression allows us to test the statistical significance of those estimates.

Table 18 shows the results for running such a linear regression, which extends and refines Table 7. We find a reduction in mortality by a factor on the order of 5-8 (for example 8.5=173.6/20.4). This is larger than estimated in Table 7 but it should be. In Table 7 we did not have data on the *proportion* of Lambeth customers, only that some customers were supplied by Lambeth. From Table 13 or Table 34 we can see that some sub-districts had a high proportion of Lambeth customers (e.g. Christchurch) while others had a low proportion (e.g. Kennington 1st). Furthermore, accounting for the proportion of "other" supplies is important since some sub-districts (e.g. Wandsworth and Putney) had almost no supply from either water company.

Table 18: Regressions for Sub-District Difference-in-Differences, Population Proportions

	Linear (levels)	Linear, house density	Linear, Fixed Effects	Neg Binom, house density
Lambeth effect	**-126.5**	**-126.5**	**-126.5**	**-1.389**
standard error	**27.01**	**26.23**	**20.16**	**0.269**
z value	-4.7	-4.8	-6.3	-5.2
p-value	2.1E-05	1.4E-05	1.2E-06	2.5E-07
Housing density	NO	11.68	NO	0.066
z value		2.0		1.1
R^2 / Resid Deviance (df, p-value)	65.3%	67.3%	80.7%	61.3 (50, p=15.1%)
Southwark mortality rate (1854)	173.6	90.5	141.7	117.2
Lambeth mortality rate (1854)	20.4	-68.3	22.6	23.4

Deaths by sub-district from 1849 and 1854 for the 28 sub-districts ("first 12" Southwark-only and "next 16" jointly-supplied) shown in Snow [1855] Table XII and my Table 32, with population from Snow [1856] Table VI. Total 56 observations. The linear regressions are fit in levels, the Negative Binomial in logs. The Negative Binomial regressions is fitted with the R function glm.nb function from the MASS package.

The adjusted R-squared shows that a large fraction of the variation across sub-districts is accounted for, supporting Snow's claim of a close relation, while the large z-values (low p-values) show that the Lambeth effect (the difference in mortality between the Southwark & Vauxhall Company versus the Lambeth Company) is statistically significant.

We can include additional variables, such as the housing density, to test whether sub-district characteristics such as crowding are important. The second column of the table shows that including density has not effect on the estimate coefficient and itself is marginally statistically significant. An even stronger test is to include sub-district fixed effects. This allows us to control for any and all differences across sub-districts. Doing so does not alter the Lambeth effect.

Equation 2 is linear with normal error term and ignores that rates are based on counts with possibly sample size (population) differing by sub-district. We should use a count (Poisson or Negative Binomial) regression as discussed in the appendix. The final column of Table 18 shows a Negative Binomial regression with essentially the same results: a large and statistically significant Lambeth effect; no large effect for crowding (housing density); and a large portion of the overall variation explained by the estimated difference in mortality rates (the low Residual Deviance). Further count regressions are discussed in the appendix.[52]

In summary, extending Snow's difference-in-differences with the water company population proportions provides strong statistical support for Snow's argument that variation in the supply of water was a primary driver both in the reduction in mortality from 1849 to 1854 and in variation across sub-districts.

7.2.2 Seven weeks, by Sub-districts: Quasi-Randomized Trial

Snow [1855] Table VIII provides data on deaths reported separately by water source by sub-district. Table IX aggregates those data to directly compare mortality rates for Southwark-supplied versus Lambeth-supplier customers, but at the aggregate level (overall south London). Sub-district data is more powerful – because of the effective mixing at the street level sub-district data provide a reasonable approximation to a randomized trial. For his 1855 publication, however, Snow did not have population by sub-district so could not compare on a sub-district basis (thus the aggregate data in Table IX).

The population by water supplier and sub-district published in Simon [1856] and then Snow [1856] provides the population-at-risk with which we can perform a comparison at the sub-district level. To introduce the idea, Table 19 mirrors the format of Snow's Table IX but for a particular sub-district – St. Peter, Walworth in this case. Comparing within a sub-district should produce mixing across a more homogeneous population with common characteristics (elevation, housing stock, weather characteristics), ensuring closer to true randomization. For St. Peter, Walworth the difference in mortality rates is large: the difference is 55.1 expressed as a rate per 10,000 and a factor of almost 16 when while expressed as a ratio. Using the standard t-test for comparing the difference in rates gives an approximate 95% confidence band of ± 13.1, implying that the difference in rates is far from zero (the rates are very different).[53]

Table 19: Deaths and Mortality per 10,000 Population, First Seven Weeks of 1854 Cholera Epidemic, St. Peter, Walworth

Water Supplier	Population	Deaths from Cholera	Deaths per 10,000
Southwark and Vauxhall	14,274	84	58.8
Lambeth Company	10,724	4	3.7
Both Cos Together	24,998	90	36.0

Following the format of Snow [1855] Table IX. Population from Snow [1856] Table I & II or VI, Deaths (counts) from Snow [1855] Table VIII.

In fact we have sixteen sub-districts (the "next 16" jointly-supplied sub-districts) where customers from the Southwark & Vauxhall Company are mixed with those from the Lambeth Company. The counts are displayed in Table 35 and the mortality rates in Table 36. We want to test these jointly rather than individually (as done in Table 19). One feasible and appropriate statistical framework is count regression (Poisson and

[52] One issue with the Negative Binomial regression is that the equation is logarithmic (as in Equation 1) rather than linear. The software for fitting count regressions usually assumes a logarithmic rather than linear function for the rate.

[53] The usual formula for the standard error of the difference in rates is $(SE(r1-r2) = \sqrt{r1(1-r1)/n1 + r2(1-r2)/n2}$, discussed more fully in the Appendix Section 10.3.1. Applying this formula gives a standard error of 6.7 (rate per 10,000), and multiplied by 1.96 gives 95% confidence bands of ± 13.1.

Negative Binomial regression) discussed above for the difference-in-differences analysis (and discussed in detail in Appendix Section 10.1).

Particularly important when testing across multiple sub-districts is the potential for "overdispersion" – more variation than implied by a Poisson process. Overdispersion implies that a Poisson process (and thus a Bernoulli process with fixed rates) may not be appropriate, and that the usual standard error just discussed may not be appropriate.

As shown in the appendix these data do exhibit overdispersion. Table 20 shows results for Negative Binomial regressions that allow for overdispersion. Even when we incorporate the observed variation across sub-districts, the effect of clean water (the "Lambeth effect" in Table 20) is very large and significant, both economically and statistically. Furthermore, including housing density does not change the Lambeth effect and is itself statistically and economically insignificant, supporting Snow's contention that the effect of clean water overrides other observable factors.

Table 20: Poisson and Negative Binomial Regressions for Sub-District Quasi Randomization, Seven Weeks Ending 26th August

	Neg Binom	Neg Binom, house density
Lambeth effect (ln)	**-1.888**	**-1.870**
standard error	**0.158**	**0.158**
z value	-11.9	-11.8
p-value	1.0E-32	2.0e-32
effect (ratio)	6.60	6.49
theta ("size" from Gamma mixing)	11.27	11.77
Housing density	NO	-0.068
standard error		0.079
p-value		39.1%
Residual Deviance (df, p-value)	38.8 (30, p=13.0%)	38.7 (29, p=10.8%)
Southwark pred mort (per 10,000)	45.7	46.0
Lambeth pred mort (per 10,000)	6.92	6.89

Data on deaths by sub-district and by supplier (Southwark & Vauxhall Co versus Lambeth Co) for the "next 16" sub-districts jointly supplied by the two companies, where the mixing by supplier effectively produces quasi-random assignment. Data from Snow [1856] Table I and II which show a) Deaths by water source (Southwark & Vauxhall Co, Lambeth Co, pump-wells, Thames, unascertained), b) houses and density in 1851, c) houses and population by supplier (Southwark versus Lambeth Cos). Death counts are for the seven weeks ending 26th August (as in Snow [1855] Table VIII). Total of 32 observations. The Negative Binomial regressions are fitted with the R function glm.nb function from the MASS package. The parameter "theta" is the size or θ for a "parametrization (1)" Negative Binomial (see Section 10.1). The "Residual Deviance" is approximately chi-squared and the p-value is the probability of observing a chi-squared RV that large or larger – a small p-value indicates the regression does not fit the data.

Although the data from Snow [1855] Table VIII (used in these tables) only covers the first seven weeks of the outbreak (through August 26th 1854) they are good quality data, the assignment to water company having been performed by Snow himself. The conclusion from this analysis (as for all the data examined by Snow) is that the effect of water is large and unambiguous in the transmission of cholera.

7.2.3 1854, Full Outbreak, by Registration District: Quasi-Randomized Trial

With the publication of population data and deaths assigned to water source in Snow [1856] Table V we can compare District-by-District for the full 1854 outbreak. This extends the analysis of the prior section to the full outbreak, but limits the comparison to aggregate Districts rather than sub-districts: The Registrar-General apparently did not publish the assignment to water supplier by sub-district, only at the District level.

Table 21: Poisson and Negative Binomial Regressions for District Quasi Randomization, Full 1854 Outbreak

	Poisson	Poisson, fixed effects	Neg Binom	Neg Binom, house density
Lambeth effect (ln)	**-1.714**	**-1.700**	**-1.725**	**-1.727**
standard error	**0.048**	**0.049**	**0.339**	**0.338**
z value	-36.0	-34.5	-5.1	-5.1
p-value	1.7E-284	1.7E-260	3.5E-07	3.1E-07
robust standard error	0.259	0.201	0.225	0.226
z value	-6.6	-8.5	-7.7	-7.6
effect (ratio)	5.55	5.47	5.61	5.62
theta ("size" from Gamma mixing)	0.00	0.00	2.35	2.38
Residual Deviance (df, p-value)	316.6 (20, <1E-6)	200.6 (12, <1E-6)	30.5 (20, 6.2%)	30.7 (19, 4.4%)
Housing density				0.077
standard error				0.183
z-value				0.4
Southwark pred mort (per 10,000)	159.9	159.9	167.0	167.0
Lambeth pred mort (per 10,000)	28.8	28.8	29.7	29.6
"Other" pred mort (per 10,000)	53.0	53.0	109.6	104.0

Data on deaths by sub-district and by supplier (Southwark & Vauxhall Co versus Lambeth Co versus pump-wells) for the nine Registration Districts shown in Snow [1856] Table V, where the mixing by supplier effectively produces quasi-random assignment. Data from Snow [1856] Table V shows a) houses and density in 1851, d) houses and population by supplier (Southwark versus Lambeth Cos), c) Deaths by water source (Southwark & Vauxhall Co, Lambeth Co, pump-wells, Thames, unascertained). Death counts are for the full 1854 outbreak (ending October). The 623 "unascertained" deaths are allocated at the District level in proportion to the observed deaths assigned to Southwark versus Lambeth, following Snow [1856] p. 247. Total of 23 observations: 9 Districts supplied by Southwark, 7 by Lambeth (St. Olave Southwark and Rotherhithe had no Lambeth population), and 7 "other" (Bermondsey and Newington have implied negative population supplied by "other"). The Poisson and Negative Binomial regressions are fitted with the R function glm (family=poisson) and the glm.nb function from the MASS package. The parameter "theta" is the size or θ for a "parametrization (1)" Negative Binomial (see Section 10.1). Robust standard errors are calculated with the R "sandwich" package, using the default "HC3" for the adjusted variance-covariance matrix (see Zeileis [2004] and the R "sandwich" manual). The "Residual Deviance" is approximately chi-squared and the p-value is the probability of observing a chi-squared RV that large or larger – a small p-value indicates the regression does not fit the data.

Table 21 shows the Poisson and Negative Binomial regressions. The first column is the Poisson regression. The residual deviance is so large that we can reject the process as Poisson, meaning that the reported standard errors are substantially too small. The second column includes fixed effects, and again the residual deviance is large enough to reject the hypothesis that the count process is Poisson. What is particularly important, however, is that the "Lambeth effect" is essentially unchanged, which means that even when we allow the mortality rate to differ by District, the Lambeth effect remain.

The final two columns allow variation across Districts in the underlying mortality. We can have more confidence in these estimated standard errors because the residual deviance is much smaller (indicating this statistical model fits reasonably well). We still have a large and statistically significant Lambeth effect. Including housing density does not change the size or statistical significance of the Lambeth effect, and housing density itself has no effect (insignificant economically and statistically). This supports Snow's assertion that water quality was predominant and crowding (housing density) had little effect.

7.3 Error Analysis for Randomized Control Trial Incorporating Overdispersion

CURRENTLY IN THE APPENDIX, NEEDS TO BE REVISED AND MOVED HERE

8 Causal Assessment Procedure – (based on Katz and Singer [2007] – preliminary & incomplete)

Katz and Singer [2007] propose an "Attribution Assessment Procedure" to weigh the disparate evidence and conflicting explanations associated with reports of a chemical weapons attack. Such an exercise has many similarities with efforts to determine causal effects in social sciences generally, and cholera in 1850s London in particular. Katz and Singer propose seven steps, which I modify slightly:

1. Divide evidence into blocks or types of evidence
2. Assess the type and quality of each block
3. Develop groups of hypotheses
4. Assess each evidence block for strength of rejection for each hypothesis
    - Consider *rejection* of hypotheses (refute, neutral, consistent) rather than strength of association (support of hypotheses)
    - Consider statistical tests used for each evidence block
    - Consider what covariates and confounders each evidence block and statistical test rules out
5. Organize evidence blocks by hypothesis into matrix
6. Choose hypothesis not contradicted
7. Strongest hypothesis checked

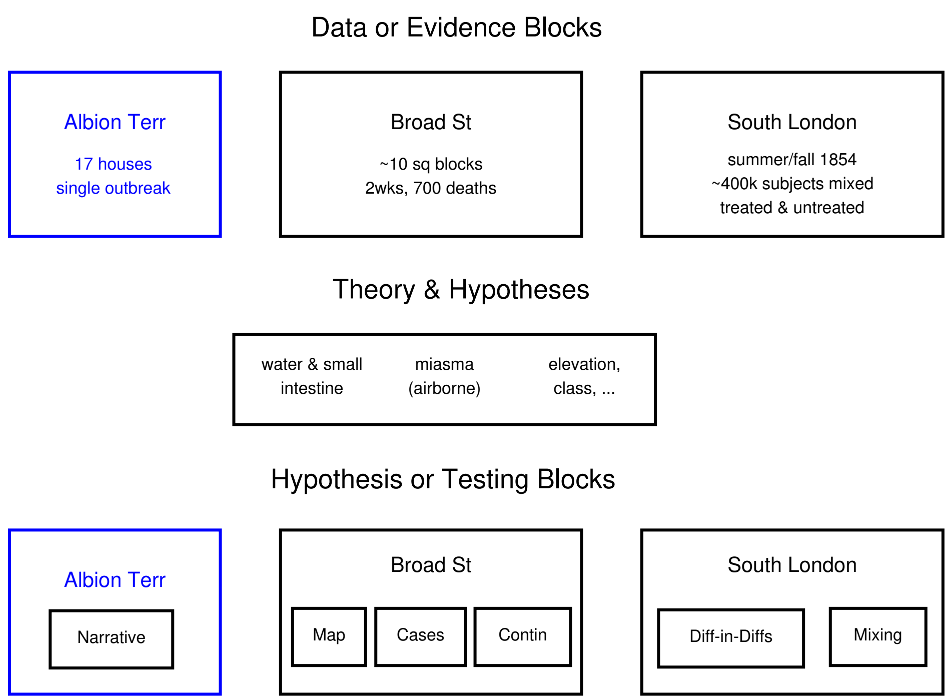

Figure 10: Graphical Overview of Causal Assessment Procedure

Table 22: Theory & Hypotheses by Evidence Block

	T1: Water	T2: Miasma	T3: Class, Elevation, ...	Comment
Albion	Contradict: no Strength: na	Contradict: yes Strength: strong	Contradict: neut Strength: na	Narrative, showing 1) no infection in nearby houses; 2) contamination of water supply. Lays out mechanism.
Broad 1 – mapping	Contradict: no Strength: med	Contradict: no Strength: med	Contradict: yes Strength: med	Identifies location of outbreak (Broad St pump) but does not immediately distinguish between water & miasma. Rejects class because deaths not distinguished by class.
Broad 2 – cases	Contradict: no Strength: strong	Contradict: yes Strength: strong	Contradict: yes Strength: na	Sharply distinguishes water from miasma, rejects miasma: 1) Those who should have died but escaped; 2) Those who should have escaped but died. Rejects class because deaths roughly equal by class. Lays out mechanism (index case, baby girl Frances).
Broad 3 – contin table	Contradict: no Strength: strong	Contradict: yes Strength: med	Contradict: yes Strength: med	Rules out "survivorship bias": shows strong association pump water & illness, both among drinkers and non-drinkers. "medium" for T2&T3: maybe could produce correlation between water & miasma
S London 1 – DiDs	Contradict: no Strength: strong	Contradict: yes Strength: med	Contradict: yes Strength: med	Moderately rules out confounders, strengthens water causality. "medium" for T2&T3: maybe could produce correlation between water & miasma
S London 2 – Randomization	Contradict: no Strength: strong	Contradict: yes Strength: strong	Contradict: yes Strength: strong	Strongly rules out confounders, strengthens water causality

Table 23: Statistical Test Used and Water Theory Covariates & Confounders Ruled Out – by Evidence Block xx1

	Evidence For	Statistical Test	Covariates ruled out	Assumptions needed – strength of ruling out
Broad 2 – cases				
Broad 3 – contin table	Strong association between water & illness	2x2 contingency table, χ^2 & Fisher Exact. p-level: $\ll .01\%$		
S London 1 – DiDs	Large & significant mortality difference, dirty vs clean water.	Count (Neg Binom) regression. p-level: 0.14%	Most: miasma; class & income; age; sex; elevation; crowding; location (by Thames)	Fixed effects (differences across regions and time) are linear and stable. It is possible, although convoluted & artificial, to construct non-water effects that produce observed data.
S London 2 – Randomization, Snow 1855	Large mortality difference, dirty vs clean water.	No formal statistical test or error analysis	Most: miasma; class & income; age; sex; elevation; crowding; location (by Thames)	
S London 2 – Randomization, Snow 1856	Large & significant mortality difference, dirty vs clean water.	Count (Neg Binom) regression. p-level: $\ll .01\%$	Most: miasma; class & income; age; sex; elevation; crowding; location (by Thames)	Not many since this is close to random mixing (RCT)

Table 24: Close to pump but did not die

	Water 1	Water 2	Miasma 1	Miasma 2
Core	Drinking	Drinking	Breathing	Breathing
Auxiliary	P[drink~ distance]	P[drink~ in-house wells]	P[breath~ distance]	P[breath~ ??]
Implication	deaths~ distance	deaths~ distance & wells	deaths~ distance	??
Core Refuted?	YES	NO	YES	??

These are cases where individuals lived or worked close to the Broad Street pump but did not die at the rate of others living close to the pump. The two most important cases are the St. James workhouse at 50 Poland Street and the Lion Brewery at 50 Broad Street. At the workhouse there were five deaths among 535 inmates but if mortality had been at the same rate as surrounding buildings there should have been roughly 100. The Brewery employed more than 70 workers and there were no deaths among them. (Snow [1855] p 42, but these cases are discussed widely - see Johnson [2007], Hempel [2007], Tufte [1997a])

Table 25: Far from pump but did die

	Water 1	Water 2	Miasma 1	Miasma 2
Core	Drinking	Drinking	Breathing	Breathing
Auxiliary	P[drink~ distance]	People travel to Broad St	P[breath~ distance]	Water infected by air
Implication	deaths~ distance	deaths~ taste for Broad St	deaths~ distance	deaths~ taste for Broad St
Core Refuted?	YES	NO	YES	NO

These are cases where individuals lived far from the pump but did die. The most striking case was Susannah Eley from Hampstead, but also includes higher mortality around the Little Marlborough Street pump and school-girls from south of the pump (nearer the Bridle Street pump) – see Snow [1855] pp 44-45, Westminster and London School of Hygiene and Tropical Medicine [1855] pp 112-113), also discussed in Tufte [1997b], Johnson [2007], Hempel [2007] p 217.

9 Conclusion

John Snow's 1855 monograph *On the mode of communication of cholera* provides a valuable example and guide for modern-day researchers in the social sciences, a guide for assembling persuasive evidence of a causal effect. The power of Snow's argument derives from employing the following components, although the final was not really available to Snow at the time:

- Well-articulated theory
- Testing predictions against evidence – Consistent consideration and rejection of alternatives
- Multiple tests and sources of data
- Careful and honest assessment of the statistical reliability of the evidence

Well-articulated theory

Snow proposed his waterborne theory of cholera in the 1849 pamphlet *On the mode of communication of cholera* (Snow [1849]; an earlier and shorter version of the 1855 monograph with none of the 1854 evidence discussed in this essay). Without the benefit of the germ theory of disease or any evidence on the bacterium *Vibrio cholerae* Snow nonetheless proposed a consistent (and correct) theory of the infection and transmission of cholera.

> The excretions of the sick at once suggest themselves as containing some material which, being accidentally swallowed, might attach itself to the mucous membrane of the small intestines, and there multiply itself by the appropriation of surrounding matter ... That a portion of the ejections or dejections [of an infected individual] must often be swallowed by healthy persons is, however, a matter of necessity. ... The views here explained open up to consideration a most important way in which the cholera may be widely disseminated, viz., by the emptying of sewers into the drinking water of the community (Snow [1849] pp 8-11)

As Johnson [2007] points out, the power of this theory is that it addresses phenomena at different scales, from the individual to the city:

> The strength of his model derived from its ability to use observed phenomena on one scale to make predictions about behavior on other scales up and down the chain. ... If cholera were waterborne then the patterns of infection must correlate with the patterns of water distribution in London's neighborhoods. Snow's theory was like a ladder; each individual rung was impressive enough, but the power of it lay in ascending from bottom to top, from the membrane of the small intestine all the way up to the city itself. (Johnson [2007] p 148)

This theory provided not only a consistent framework to guide Snow in his collection and examination of data, but also predictions at a range of scales, predictions that Snow (either explicitly or implicitly) was able to test and place in contrast with predictions from alternate theories and hypotheses.

I should emphasize that this theory was well-articulated but by no means a formal mathematical model. Snow was not estimating or testing parameters but rather predictions about how "patterns of infection must correlate with the patterns of water distribution". Most importantly he was contrasting (again sometimes explicitly and sometimes implicitly) predictions of the waterborne theory versus predictions of alternate theories.

Testing predictions against evidence – Consistent consideration and rejection of alternatives

Snow wrote his 1855 monograph to convince skeptics and throughout he was responding to criticisms of his theory. This skepticism pushed him to consider (and reject) a wide range of alternatives. This is the appropriate method for building a causal argument – consider all alternatives that can be proposed – and Snow had the benefit of having at hand a wide range of alternative hypotheses.

Important examples of confronting predictions with evidence show up in both the Broad Street outbreak and the South London "Grand Experiment". In the Broad Street outbreak, the three most important predictions were about the spatial pattern of deaths (concentrated around the pump); anomalous cases ("Those who should have died but escaped" and "Those who should have escaped but died"); and the pattern of death among those who drank from versus did not drink from the pump. In all cases the waterborne theory and the mechanism of transmission are central.

The simplest version of the waterborne theory (the pump is infected and the likelihood people drink from the pump is proportional to distance from residence to the pump) predicts that infection (and mortality) should be concentrated around the pump and fall off with distance. This pattern is also predicted by an airborne transmission (miasma) theory that posits emanations from the pump or nearby sewers. The pattern is not predicted by many alternatives – for example that infection depends on what floor individuals live on or social status. Thus the broad spatial pattern of deaths, the concentration around the pump shown in Snow's famous map, rules out a variety of hypotheses (say social class: "The mortality appears to have fallen pretty equally amongst all classes, in proportion to their numbers." Snow [1855] p. 48) but does not sharply distinguish between waterborne versus airborne transmission (miasma).

The waterborne and miasma theory do differ, however, in some of the details of spatial patterns. The waterborne theory, together with specific circumstances of water supply or drinking patterns would predict either fewer deaths close to the pump ("Those who should have died but did not") or more further from the pump ("Those who should have escaped but died"), and these patterns would not be consistent with miasma. Snow (and the Reverend Whitehead) documented such patterns. "Those who should have died but did not" included groups who were close to the pump but did not die at the expected rates. Inmates of the St. James workhouse and workers at the Lion Brewery were quite close to the pump but drank from alternate sources and had low mortality rates, observations that were consistent with Snow's theory but not miasma.

"Those who should have escaped but died" were individuals who lived further from the pump but nonetheless died at high rates, higher than expected given the distance from the pump. Once again the waterborne theory (together with auxiliary details of individual circumstances or behavior) could predicted these detailed patterns but the miasma could not. Again Snow and Whitehead documented specific cases and found a connection with the Broad Street pump, with Susannah Eley, the widow from Hampstead, being the most striking example.

One final and critical prediction that distinguishes the waterborne from the miasma theory is that infection should differ dramatically between those who drank from the pump versus those who did not. Snow himself did not present evidence on this question around the Broad Street outbreak (although it is central to the South London analysis), but the Reverend Whitehead did in his contribution to the Vestry report (Westminster and London School of Hygiene and Tropical Medicine [1855], a report produced in concert with Snow). Data collected by the Reverend Whitehead showed that the rate of infection differed substantially between those who drank versus did not drink from the water. Although Whitehead's presentation of this data is somewhat anecdotal by today's standards, it is important in failing to contradict the waterborne theory and contradicting the miasma theory.

For the South London "Grand Experiment" the differing mortality among drinkers of contaminated versus clean water is the central issue. In 1854 the Southwark and Vauxhall Company provided contaminated water while the Lambeth Company provided clean (or cleaner) water. Customers of these two companies were mixed across and within sub-districts in a geographically compact region of south London. The waterborne theory predicts a large difference in mortality that follow the pattern of water supply by company. Customers were mixed within and across sub-districts on all characteristics *including* water supplier. This mixing should control for and average out the effect of all other factors. If water were *not* the causal agent (if it were miasma or elevation or social standing or any other cause) then the mixing would leave no observed difference in mortality between customers of the two companies. If water were the causal agent, there would be observed differences between customers of the two companies. Finding a large effect would imply that water supply was a causal agent because all other possible factors were matched, controlled for, or averaged out.[54]

[54] Individuals would be effectively matched on location-related characteristics such as elevation or weather since the sub-districts were geographically similar and compact. Other characteristics, such as age, sex, social standing, would be controlled for by averaging over the large number of customers who were effectively randomly assigned to water supplier.

Snow did find a substantial difference across customers, following the pattern of water supply company. Since all other factors were controlled for it is hard to interpret this observation as anything but ruling out all (or at least a very wide range) of alternatives.

Multiple Tests and Sources of Data

To convince skeptics Snow employed a variety of evidence from multiple sources. The Broad Street outbreak is a case study of an outbreak limited in time and spatial extent, with detailed attention to the circumstances of individual cases. Snow provided evidence of multiple sorts. He mapped the outbreak to show the concentration of deaths around the Broad Street pump. He provided narratives for individual case after case, showing the role of water in deaths close to and far from the pump. Snow argues convincingly that water from the Broad Street pump figured in the vast preponderance of deaths during the period from August 31st to September 18th. He builds a convincing argument that other factors – miasma, social standing, crowding, and others – are poor explanations for the observed mortality.

The south London "Grand Experiment" was very different – a large-scale natural experiment independent of the Broad Street episode. The mixing of customers, some supplied with contaminated water from the Southwark and Vauxhall Company and some with clean water from the Lambeth Company, meant that all other factors were matched or averaged out. In the language of modern statistics randomization controlled for all the factors that others proposed as causal factors for cholera – among them elevation, weather, housing, social status , age, gender. The comparison was as close to a controlled scientific experiment as one is likely to find in the social sciences. Although Snow did not provide formal or structured tests, his comparisons map to two more modern approaches: difference-in-differences regression and randomized control trials.

Role of Statistics

Freedman [1991] p 291 is correct in saying that the true power of Snow's work derives from "good design, relevant data, and testing predictions against reality in a variety of settings." But the application of formal statistics and regression tools enhances and strengthens Snow's conclusions. Translating his verbal and tabular comparisons into a formal regression or other statistical framework allows us to weigh the strength of the individual pieces of evidence. Comparing the drop in mortality from 1849 to 1854 for the jointly-supplied (Vauxhall and Lambeth) regions using a difference-in-differences regression allows us to answer important question: Precisely how much did the mortality drop? Was this truly large relative to the range of the variation observed across sub-districts and across time? In other words was the estimated coefficient statistically significant?

I believe that Snow's work, when embedded in a the statistical framework, provides a rebuttal to the claim by Freedman [1991] (p 304) that "regression models are not a particularly good way of doing empirical work in the social sciences today", and also provides an example of using regression models fruitfully.

10 Appendix

10.1 Statistical Framework for Count Data in Difference-in-Differences Analysis – Poisson and Negative Binomial Regression

The generalized difference-in-differences Equation 1 provides a framework that allows us to extend Snow's analysis and ask questions that Snow and his peers were simply not equipped to answer. There is an important statistical issue, however, that we glossed over by writing Equation 1 in terms of a linear regression in rates with a normal error term. We do indeed want to think of the rate as a linear function of our variables, but the observed data are in terms of deaths (counts) and population (exposure). St. Saviour has more deaths in 1849 than St. Olave (283 versus 157) but a lower rate (144 versus 196 per 10,000 people) because of St. Saviour's larger population. The random variable we should use is the actual count (of deaths), which we can think of as generated from the underlying rate by a statistical process determined by the rates in Equation 1.

The simplest statistical process we could think of would be a binomial process, assuming that each individual has the same rate of infection and death, making the observed count a binomial random variable. Cholera infection was a relatively rare event, even in the 1850s London epidemics, with rates on the order of 100 per 10,000 persons or 1%. For such low rates a Poisson process is a good approximation to the binomial, and the Poisson is ideal for various mathematical reasons that will become apparent shortly.

> The Poisson process describes the arrival of multiple independent events. For the low intensities we have here the Poisson is a good approximation to the Binomial because the probability of multiple events is low. For an intensity (rate) of μ the probability of one Poisson event is $\exp(-\mu) \cdot \mu$ and the probability of two is $\exp(-\mu) \cdot \mu^2/2$. For a 1% rate the probability of one event is 0.99% and the probability of two is 0.005%. Using the Poisson for approximating the Binomial has a long history. We can also see that the distributions are very similar by noting that the mean and variances for the count for a binomial and Poisson will be close. For a Binomial with population n the mean and variance are *np* and *np(1-p)* while for a Poisson with intensity μ they are both $n \cdot \mu$. When p is small then *1-p* will be close to 1 and the variances will be very close.

We can move from the *Rate* in Equation 1 to counts by simply noting the the rate is the count divided by population:

$$\ln(Rate) = \ln(count/population) = \ln(count) - \ln(population)$$

$$\Rightarrow \ln(count) = \ln(Rate) + \ln(population)$$

We can now write a proper regression equation with a proper error term and *count* being a proper random variable:

$$\begin{aligned} \ln(count_{sub-dist,yr}) &= \mu + \delta_{54} \cdot I_{yr=1854} + \gamma_J \cdot I_{region=joint} \\ &+\beta \cdot I_{region=joint} \cdot I_{yr=1854} + \alpha \cdot Covariates + \ln(population_{sub-dist,yr}) + \varepsilon \end{aligned} \tag{3}$$

In terms of Equation 1 this is just

$$\ln(count_{sub-dist,yr}) = \ln(R_{sub-dist,yr}) + \ln(population_{sub-dist,yr}) + \varepsilon$$

The term $\ln\left(population_{sub-dist,yr}\right)$ is commonly called an "offset".

The difference between Equation 3 and a standard linear regression equation is the error term, which is *not* normally-distributed. We want the count itself to be Poisson-distributed ($count \sim Poisson$), so we want to re-write Equation 3 in the form:

$$\begin{aligned} count_{sub-dist,yr} &= \exp\left(R_{sub-dist,yr}\right) \cdot \exp\left(population_{sub-dist,yr}\right) \cdot \exp\left(\varepsilon\right) \\ &= \exp\left(R_{sub-dist,yr}\right) \cdot \exp\left(population_{sub-dist,yr}\right) \cdot \eta \end{aligned} \tag{4}$$

This implies that η (or $\exp(\varepsilon)$) is Poisson-distributed with mean 1 (in other words $\eta = \exp(\varepsilon) \sim Poisson(mean = variance = 1)$). The additive error term ε (additive in Equation 3) is some unknown distribution but the distribution for ε does not really matter. Starting with the assumption that $\eta = \exp(\varepsilon) \sim Poisson(mean = 1)$ we can use the observed counts and maximum likelihood techniques to fit the regression, Equation 3 or 4.[55]

Although the statistical framework around Equation 3 seems more complicated than simple linear regression (with normal errors) it is not, with the only difference being that the error term $\eta = \exp(\varepsilon)$ is not normally-distributed. Here the data are Poisson-distributed counts rather than (say) normally-distributed income, but the ideas are the same and we can use maximum likelihood to estimate the parameters.

With the statistical regression framework of Equation 3 we now have the tools to start examining the observed number of deaths (counts) and rates for the 28 sub-districts for 1849 and 1854 shown in Tables 32 and 33. First let us simply estimate the aggregate data in Tables 4, 5, and 6 by (Poisson) regression. This will provide what Snow's analysis did not and those tables cannot: measures of statistical significance and confidence bands.

The first column of Table 26 shows the results of fitting Equation 3 (or 1). The estimated coefficients are the same as Table 5[56] but we now have tools with which to judge the precision of and our confidence in the coefficients.

There are issues, however, regarding the raw or un-adjusted standard errors from the Poisson regression. (For this reason I report robust standard errors in the table − for the column 1 estimate the un-adjusted Poisson standard error is 0.039.) The problem is that the assumption that deaths are Poisson-Distributed is not realistic, for a variety of reasons: first formal statistics reject the simple Poisson model (the Residual Deviance is very large); second we see considerable internal variation in the data (mortality rates shown in Table 33 differ substantially across sub-districts); third intuition tells us that rates are unlikely to be the same for everyone (the probability of infection will differ across people and location).

The Residual Deviance – a measure of quality of fit comparing actual with predicted counts – asymptotically should be distributed as chi-squared. For the simple Poisson regression in column 1 of Table 5 it is very

[55] Discussions of Poisson and Negative Binomial Regression usually explain them in the context of general linear models or GLM (see the software manuals and discussion: Zeileis et al. [2017], stata.com [2017], ncs [2017]). They start by assuming the *expected* count for sub-district i and year t is a liner function of observables:

$$\begin{aligned} \ln\left[E\left(count_{it} \mid data, parms\right)\right] = \mu + \delta_{54} \cdot I_{yr=1854} + \gamma_s \cdot I_{sub-dist} \\ + \beta_{treat} \cdot I_{treatment-degree} \cdot I_{yr=1854} + \ln\left(population_{it}\right) \end{aligned} \tag{5}$$

with the auxiliary assumption that the observed count is distributed as a Poisson variable with mean given by Equation 5:

$$count_{it} \mid data, parms \sim Poisson\left[mean = \mu + \delta_{54} \cdot I_{yr=1854} + \cdots\right]$$

This is of course equivalent to Equations 3 or 4.

[56] The coefficients should be the same because the Poisson regression is effectively weighting the average by sub-district population.

Table 26: Poisson and Negative Binomial Regressions for Sub-District Difference-in-Differences 1849 vs 1854

	Poisson, Single treatment effect	Poisson, Single treatment effect, sub-district FE	Neg Binom, Single treatment effect	Neg Binom, two treatment effects
Treatment – "less Lambeth" (ln)	**-0.511**	**-0.511**	**-0.500**	**-0.338***
standard error	**0.039**	**0.039**	**0.246**	**0.248**
z value (coeff/SE)	-13.2	-13.2	-2.0	-1.4
p-value	0.0%	0.0%	4.2%	17.3%
robust standard error	0.211	0.234	0.230	0.242
z value (coeff/SE)	-2.4	-2.2	-2.2	-1.4
treatment (ratio)	1.67	1.67	1.65	1.40
Treatment – "more Lambeth" (ln)				**-1.132**
standard error				**0.353**
z value (coeff/SE)				-3.2
p-value				0.14%
robust SE				0.295
z value (coeff/SE)				-3.8
treatment (ratio)				3.10
theta ("size" from Gamma mixing)			4.96	5.57
Residual Deviance (df, p-value)	1542 (52, p<1e-10)	456.8 (26, p<1e-10)	59.8 (52, p=22%)	60.0 (51, p=18%)
sub-district fixed effects	NO	YES	NO	NO
Single region / less Lambeth fixed effect	-0.036*		-0.032*	-0.064*
More Lambeth fixed effect				0.059*
Time fixed effect	0.084*	0.084*	0.057*	0.057*

Deaths by sub-district from 1849 and 1854 for the 28 sub-districts ("first 12" Southwark-only and "next 16" jointly-supplied) shown in Snow [1855] Table XII and my Table 32, with population from Snow's Table VIII. Total 56 observations. The "Residual Deviance" is approximately chi-squared and the p-value is the probability of observing a chi-squared RV that large or larger – a small p-value indicates the regression does not fit the data. * = ***not*** significant at the 10% level (robust errors for Poisson regression). The Poisson and Negative Binomial regressions are fitted with the R function glm (family=poisson) and the glm.nb function from the MASS package. The parameter "theta" is the size or θ for a "parametrization (1)" Negative Binomial (see appendix). Robust standard errors are calculated with the R "sandwich" package, using the default "HC3" for the adjusted variance-covariance matrix (see Zeileis [2004] and the R "sandwich" manual).

large: 1,565 while a chi-square variable with 52 degrees of freedom has a 0.1% quantile of only 89.3. We can pretty conclusively reject this version of our Poisson regression. Rejection of this constant-rate Poisson assumption implies that the un-adjusted standard error is too small.

Figure 14 shows why this restricted Poisson regression does not fit the observations and also why the (un-adjusted) standard error is too small. The graphic shows the observed and predicted mortality rates for sub-districts, with approximate 95% confidence bands drawn around the predicted rates.[57] The confidence bands are based on a constant-rate Poisson and are far too small given the observed mortality, which varies considerably both across sub-districts and across time. A majority of the observations are outside the calculated bands, implying that a constant-rate Poisson random variable is simply not able to account for the observed variation.[58]

It should come as no surprise that the assumption behind Table 1 and the simple Poisson in the first column

[57]The 95% error bands are calculated for a Poisson distribution with mean equal to the predicted 1849 count - in R the command for the 2.5% quantile is *10000 * qpois(.025,lambda=predcount1849) / pop1851*). The 1854 "actual" is adjusted for the year effect and treatment effect, so that it is comparable to the 1849 predicted. The alternative would be to display both 1849 and 1854 predicted, producing a more cluttered graphic.

[58]Remember that the mean and variance for a Poisson random variable are the same, both equal to the rate.

of Table 26 (all individuals across all sub-districts have the same probability of death) is strongly at variance with observation. Underlying conditions, and thus the probability of death, will vary by sub-district and within sub-district across individuals. The data are telling us that a simple constant rate Poisson does not apply, and by implication the un-adjusted standard errors for the Poisson regression are too low.

We can relax the assumption of constant mortality in two directions. First, we can allow mortality rates to vary across sub-district by incorporating fixed effects in Equation 1; replacing the single indicator $I_{region=joint}$ with multiple indicators $I_{sub-dist}$, one for each sub-district. This will account for unobserved sub-district differences or heterogeneity. Second, we can relax the constant-rate Poisson assumption – same mortality for every person – for example by using the negative binomial.

Consider sub-district fixed effects first. The second column of Table 26 and Figure 15 show the results for a sub-district fixed-effect Poisson regression. The Residual Deviance shows this model still does not account for the observed variation (a chi-squared of 456.8 is still extremely large), and Figure 15 shows the reason: although the fixed effects account for variation between sub-districts, many of the observations still fall outside the 95% confidence bands. The across-time variation in rates (within sub-districts) is still too large to be accounted for by a Poisson random variable – the variance of the Poisson is just too small.[59]

Again, we should not be surprised that the Poisson does not fit the observed data – the Poisson assumes the same mortality rate for everyone within sub-districts, while in reality there will be differences across people. Both logic and the data are telling us that we need to allow for unobserved differences or heterogeneity across individuals. One standard way to do this is to model the counts as a negative binomial. The negative binomial is a Gamma mixture of Poissons, meaning that each individual's risk of death is Poisson-distributed with rate $\tilde{\mu}$ (as an approximation to binomial, with some overall mean m) but the rate $\tilde{\mu}$ itself is multiplied by a Gamma($\theta, 1/\theta$) random variable across the population. The Gamma distribution models the heterogeneity in individual-level mortality rates.

Parametrizations for the Gamma and Negative Binomial Distributions

There are multiple ways to parametrize both the Negative Binomial and the Gamma distributions, so notation can be very confusing.

Negative Binomial Distribution

1. Size (θ) & Mean (m): NB(m, θ); Mean $= m$, Variance $= m + m^2/\theta$
   - *Size* = shape parameter inherited from the Gamma mixing distribution, aka size, sometimes called dispersion because this is what provides the higher variance ("overdispersion") relative to the Poisson. But for the term "dispersion" also see parametrization (3) below.
   - *Mean* = mean of the negative binomial, the overall mean count

[59] This is called "overdispersion" in the literature and is frequently encountered in practice. It is almost inevitable that the Poisson will predict too little variation for large populations. Remember that the mean and the variance of the Poisson are the same so as the population increases (for a constant mortality rate) the mean count and the variance will increase proportionally with population, but the standard deviation will increase only with the square root of population. This implies that, when rates are considered rather than counts, the observed standard deviation of the rate will go *down* as population increases.

- Parametrization used by R in the MASS package glm.nb function. One of the alternative parametrizations allowed by R. Poisson limit is $\theta \to \infty$

2. Size (θ) & Prob (p): NB(θ, p_{sucess}); Mean $= \frac{\theta p}{1-p}$, Variance $= \frac{\theta p}{(1-p)^2}$
   - *Size* = shape parameter of the Gamma mixing distribution, aka target number of successful trials
   - *Probability* = probability of success in each trial
   - Parametrization allowed by R
   - Parametrization used by Mathematica **but** with probability of *failure* $p_{failure} = 1 - p_{success}$

3. Dispersion (α) & Mean (m): NB(m, α); Mean $= m$, Variance $= m + m^2 \cdot \alpha$
   - *Dispersion* = shape parameter inherited from the Gamma mixing distribution, this is 1/size of parametrization (1). Again, this parameter is what provides the higher variance ("overdispersion") relative to the Poisson.
   - Parametrization used by STATA. Poisson limit is $\alpha \to 0$

Parametrizations (1) and (2) are related by $m = \frac{\theta p}{1-p}$ and $p_{success} = \frac{m}{\theta+m}$.

Gamma Distribution

1. Shape (a) & Scale (b): Gamma(a, b); Mean $= ab$, Variance $= a \cdot b^2$
   - Parametrization used by Mathematica, allowed by R
2. Shape (c) & Rate (d): Gamma(c, d); Mean $= c/d$, Variance $= c/d^2$
   - Parametrization allowed by R

Moving from Gamma mixture of Poissons to Negative Binomial

- A Poisson($\tilde{\lambda}$) mixed with a parametrization (1) $\lambda\tilde{}\sim$Gamma($\theta, m/\theta$) produces a parametrization (1) Negative Binomial NB(m, θ). This is the parametrization used by R.
  - A Gamma($\theta, m/\theta$) has Mean $= m$, Variance $= m^2/\theta$.
  - A NB(m, θ) has Mean $= m$, Variance $= m + m^2/\theta$
- STATA mixes a mean m Poisson with a parametrization (1) Gamma($1/\alpha, \alpha$) to obtain a parametrization (3) NB(m, α); Mean $= m$, Variance $= m + m^2 \cdot \alpha$
- A Poisson(m) or Poisson($pop \cdot \mu$) will have mean and variance $m = pop \cdot \mu$, standard deviation $\sqrt{m} = \sqrt{pop \cdot \mu}$. The mean scales with the population so the rate (intensity) remains the same, but the standard deviation scales with $\sqrt{pop}$ so measured in rate terms the standard deviation decreases as population increases.
- For the Gamma($\theta, m/\theta$) both the mean and the standard deviation scale (asymptotically) with population, so in rate terms both remain constant as population changes. The behavior of the Gamma induces similar scaling for the mixed Poisson, i.e. the Negative Binomial.

The third column of Table 26 shows Negative Binomial regression results (no sub-district fixed effects). The Residual Deviance tells us that this model does a reasonable job accounting for the observed variation in counts; the Residual Deviance is 59.7 which is not high for a chi-squared with 52 degrees of freedom (p-value 22%). Figure 16 shows why the Negative Binomial works better than the Poisson regressions. The error bars (the 95% confidence bands for a Negative Binomial) are now wide enough to capture all except three out of the 56 observations (5.4%, close to the 5% we expect for 95% confidence bands). The unobserved heterogeneity modeled by the Negative Binomial captures the wide variation in observed mortality across sub-districts and across time, after controlling for average differences between 1849 and 1854 and between treated and untreated sub-districts.

The important question now is whether there remains a "treatment effect" – a reduction in mortality in 1854 for sub-districts supplied partially by the Lambeth company. And the answer is "yes, but the evidence so far is not overwhelming." The estimated coefficient is -0.505, essentially the same as the -0.511 from Table 5 (and the Poisson regressions). Now, however, we have a statistical model and thus standard error that reasonably captures the variation in the date shown in Table 33. This standard error (0.248) tell us that the -0.505 estimate is not estimated with extreme precision and we should not have exceedingly high confidence that it is different from zero: the p-value is 4.1%.[60]

But we do have more information than incorporated in Table 5 and the estimate for the third column of Table 26. Snow remarks that four sub-districts had more Lambeth-supplied customers than the others:

> In certain sub-districts, where I know that the supply of the Lambeth Water Company is more general than elsewhere, as Christchurch, London Road, Waterloo Road 1st, and Lambeth Church 1st, the decrease of mortality in 1854 as compared with 1849 is greatest, as might be expected. (Snow [1855] p 89)

We can incorporate this by estimating two treatment effects, one "less-Lambeth" and the other "more-Lambeth" (for the four identified sub-districts). The fourth column of Table 26 shows the Negative Binomial regression for these two treatment effects. As should be the case if Snow's assertion and the waterborne hypothesis are correct, the reduction in mortality is greater (and by a substantial amount) for these four sub-districts: -1.137 for "more-Lambeth" versus -0.344 for "less-Lambeth". Furthermore, the "more-Lambeth" estimated coefficient has a small standard error relative to the estimate (the p-value is 0.14%), which implies high confidence in the existence and magnitude of the treatment effect.

Figure 17 shows why the "more-Lambeth" treatment effect is statistically significant but the "less Lambeth" is not.[61] When we overlay the 1854 observed mortality rates (the blue triangles) on top of the 1849 rates (red circles) and draw error bands, we see that all the 1854 rates are within the bands *except* for sub-districts 16, 20, 22, and 23. Three of the four (16, 20, 22) are "more-Lambeth" sub-districts. The fact that so many of the "more-Lambeth" sub-districts have exceptionally low 1854 rates is the strong evidence for a treatment effect that shows up in the small p-value of 0.14% in the fourth column of Table 26. For 10 out of 12 of

[60]The Negative Binomial regression fits the data reasonably well, and the estimated robust standard errors are not very different from the un-adjusted: 0.232 in this case.

[61]The 1849 observations are represented by the red circles, and given the fitted parameters for the Negative Binomial distribution we can infer 95% confidence bands around each observation. If we have a count of m in 1849 (say Waterloo Rd (1st) with 193 deaths and rate of 137 per 10,000) we want the 2.5% and 97.5% quantiles of a Negative Binomial with mean m (the observed count) and size θ equal to the fitted "theta" parameter. (The mean count will be m and the standard deviation will be $\sqrt{m + m^2/\theta}$.) We want to convert the quantiles from counts to rates, so we divide by the population. The R command will be 10000 * qnbinom(.025,size=theta,mu=deaths) / pop

the remaining joint sub-districts the 1854 rate is lower, and while this is suggestive of a treatment effect, it does not rise to the level of definitive evidence. Further note that for the Southwark-only sub-districts with all contaminated water (the left panel in Figure 17) the 1854 rate is sometimes lower (six out of 12) and sometimes higher (three out of 12).

10.2 Differences Between Snow's Tables VII & VIII (Deaths by Sub-District & Supplier for 4 weeks ending 6th August versus 7 weeks ending 26th August)

Snow separately reported counts for the four weeks ending 5th August in Table VII (p 84) and the seven weeks ending 26th August in Table VII (p 85). In the body of the text I discuss results for the seven weeks (counts from Table VIII, shown in Tables 9 and 11). One can of course perform the same analysis for the four weeks with data from Table VII, shown here in Tables 27 and 28. The results for the four weeks are similar except they show an even larger difference in mortality between Southwark-supplied and Lambeth-supplied households: Lambeth mortality lower by a factor of 11 (66.59/5.96) instead of the 7.6 shown by the ratio 304/40 in Table 11. As Snow remarks "As the epidemic advanced [during August], the disproportion between the number of cases in houses supplied by the Southwark and Vauxhall Company and those supplied by the Lambeth Company, became not quite so great, although it continued very striking." (Snow [1855] p 82)

Table 27: Deaths from Cholera in the First Four Weeks of 1854 Cholera Epidemic – Supplier along rows, Region along columns

Water Supplier	Southwark & Vauxhall "first 12" sub-districts	Southwark and Lambeth jointly-supplied "next 16"	Lambeth only "final 4"	All sub-districts total
Southwark and Vauxhall	171	115	–	286
Lambeth Company	–	14	0	14
Both Suppliers	171	129	0	300

From Snow [1855] Table VII

Table 28: Mortality per 10,000 Households, First Four Weeks of 1854 Cholera Epidemic

Water Supplier	Southwark & Vauxhall "first 12" sub-districts	Southwark and Lambeth jointly-supplied "next 16"	Lambeth only "final 4"	All sub-districts total
Southwark & Vauxhall	75.08	**66.59**	–	71.42
Lambeth Company	–	**5.96**	0.0	5.36
Both Suppliers	75.08	31.64	0.0	45.35

Number of Deaths from Snow [1855] Table VII, number of houses my Table 10, imputed from Snow's Table IX and Population as discussed in the main text.

Snow reports the first four weeks separately from the total seven weeks because of a small difference in data collection between the first four and last three weeks. For the full period for the jointly-supplied "next 16" sub-districts Snow himself determined the supplier – Southwark & Vauxhall versus Lambeth – for each household in which a death occurred ("In table VIII, showing the mortality in the first seven weeks of the epidemic, the water supply [whether Southwark & Vauxhall or Lambeth] is the result of my own personal inquiry, in every case, in all the sub-districts to which the supply of the Lambeth Company extends [i.e. the "next 16" jointly-supplied and the "last 4" Lambeth only]." Snow [1855] p 83).

In the "first 12" Southwark & Vauxhall only sub-districts there was no supply by the Lambeth Company but some households obtained their water from a well, ditch-water, etc. To compare Southwark & Vauxhall versus Lambeth it is necessary to count deaths from Southwark-supplied households and exclude households supplied by well. For the four weeks ending August 6th another doctor, Mr. John Joseph Whiting, L.A.C., assisted Snow and inquired about the water source "in Bermondsey, Rotherwaite, Wandsworth, and certain other districts, which are supplied only by the Southwark and Vauxhall Company." (Snow [1855] p 79, see also Snow [1856] Table II as reproduced at http://www.ph.ucla.edu/epi/snow/cholerawatersouthlondon.html).

For the final three weeks for the Southwark-only "first 12" sub-districts Snow allocated the total deaths (from the official "Weekly Returns of Births and Deaths in London") between the Southwark & Vauxhall company versus other sources (wells, etc.) in the same proportion as measured by Dr. Whiting in the first four weeks: "the water supply of the last three weeks is calculated to have been in the same proportion by the Company [Southwark & Vauxhall], or by pump wells, etc., as in the first four weeks ... [in] sub-districts ... marked with an asterisk." (Snow [1855] p 83)

In any case, mortality for Lambeth-supplied households is dramatically lower, whether we look at the full seven weeks (Table 11) or the first four (Table 28).

10.3 Error Analysis for Randomized Control Trial Incorporating Overdispersion – NEEDS TO BE REVISED

Snow, in Snow [1856], used data on houses and housing density by Registration District and sub-district to more carefully test the difference between supply by the Southwark and Vauxhall Company versus the Lambeth Company.[62] These data effectively resolve the problem (discussed in the main text) that Snow faced in producing Snow [1855] Table IX – that he did not know the population served by the Southwark & Vauxhall Company versus the Lambeth Company at anything except a very aggregate level.

These data also provide a case study for error analysis of data from a randomized control trial. It is true that the South London data was not randomized by Snow himself, but the competition for customers between the two companies in the 1840s, followed by the choice of the Lambeth Company to move its water source in 1852, did effectively randomize customers. The study is instructive in highlighting the importance of heterogeneity and mixture distributions as the proper framework for statistical testing and error analysis, a lesson which carries over to current analysis of clinical trials and randomized trials.

Table 29 shows the population by region and supplier, and Table 30shows the mortality for the first seven weeks of the outbreak (using the deaths from Table 9).

The mortality for Southwark-supplied versus Lambeth-supplied customers is dramatically different, by a factor of 6.90-times. Even thought this difference is large we need to measure the precision of the estimated effects, and specifically test whether this difference – either in level terms (46.4 – 6.73) or in ratio terms (46.4 / 6.73) – is larger than we should expect from random variation.

[62] Data published in Simon [1856] provided the number of houses supplied by the two companies in each district and sub-district, albeit with error (as discussed by Snow).

Table 29: Population (1851) by Water Supplier and by Sub-Districts from Snow (1856)

Water Supplier	Southwark & Vauxhall "first 12" sub-districts	Southwark and Lambeth jointly-supplied "next 16"	Lambeth only "final 4"	All sub-districts total (pop assigned to sub-district)
Southwark & Vauxhall	122,928	113,056	–	235,984
Lambeth Company	4,084	139,606	4,335	148,025
Both Suppliers	127,012	252,662	4,335	384,009

Population by sub-district and by supplier (Southwark & Vauxhall Co versus Lambeth Co) from Snow [1856] Table I and II which show a) Deaths by water source (Southwark & Vauxhall Co, Lambeth Co, pump-wells, Thames, unascertained), b) houses and density in 1851, c) houses and population by supplier (Southwark versus Lambeth Cos). Note, however, that the assignment of houses (and population) to supplier has errors, as discussed in Snow [1856] p. 245 ff and in the following sub-section.

Table 30: Mortality per 10,000 Persons, First Seven Weeks of 1854 Cholera Epidemic, Using Population Data from Snow (1856)

Water Supplier	Southwark & Vauxhall "first 12" sub-districts	Southwark and Lambeth jointly-supplied "next 16"	Lambeth only "final 4"	All sub-districts total
Southwark & Vauxhall	60.0	**46.4**	–	53.5
Lambeth Company	0.00	**6.73**	9.23	6.62
Both Suppliers	58.1	24.5	9.23	35.4

Deaths from Table 9 and population from Table 29.

The standard approach for testing randomized clinical trials, discussed in the next sub-section, gives an incomplete and potentially wrong answer. The reason is instructive and provides another case study based on Snow's data for how to pursue empirical analysis. Specifically, the 16 jointly-supplied sub-districts show substantial variation in mortality, more variation than can be accounted for simply by random variation due to finite populations. The observations are telling us that there is additional variation across and within sub-districts. Such variation contradicts the assumption of a single Bernoulli rate across the whole population. We must incorporate some form of mixing of rates, either through mixing across sub-districts (sub-district fixed effects) or individual-level heterogeneity. This approach is discussed in a following sub-section.

10.3.1 Standard Error Analysis for Randomized Control (Clinical) Trials

The standard approach for estimating and testing incidence or response rates such as the mortality shown in Table 30 would be to assume that death is the outcome of a Bernoulli trial:

- Assume outcome or response to treatment is Bernoulli. In Snow's case outcome is win (survive) or lose (die) and treatment is drinking clean water.
- The number of responses is the sum of Bernoulli random variables and so is Binomial. For a large sample this will go to normal.
- The rate is the number of responses (death in this case) divided by the population (the population at risk). This will be a mean and will go to normal by the central limit theorem.

- Because the count is binomial, if the rate is r the count is $r \cdot n$ and the variance of the count will be $r(1-r) \cdot n$. The variance of the rate (dividing the count by sample or population n so the variance by n^2) will be $^{r(1-r)}/_{n}$. The standard error (the standard deviation of the observed rate) will be $\sqrt{^{r(1-r)}/_{n}}$.
- We will generally have two samples (control and treatment) with rates r_1 and r_2. Each rate will go to normal, and so we can test for equality of the rates with a standard t-test for equality of means. We will want to allow for different variances, which implies the standard error for the difference in rates will be $SE(r1-r2) = \sqrt{^{r1(1-r1)}/_{n1} + {}^{r2(1-r2)}/_{n2}}$. (The expression for degrees-of-freedom is given below.)
- For mathematical tractability we may work with Poisson distributions rather than Bernoulli and Binomial. For rates that are small the difference is small (Binomial variance $^{r(1-r)}/_{n}$ and Poisson variance $^{r}/_{n}$) and Poisson random variables have the nice property that the sum of Poissons is still Poisson (with rates averaged).

This Bernoulli / Poisson statistical framework seems (and in many respects is) both reasonable and innocuous. However, for both empirical and logical reasons (detailed shortly) the simple Bernoulli / Poisson assumption is not appropriate for Snow's data and may not be for many clinical trials. The alternative we are pushed to adopt is a mixture model: the distribution of the observed (estimated) rate will be a mixture of Bernoullis (or, as a more convenient approximation, a mixture of Poissons).

The mixture framework has a number of important implications. First, the distribution of the rate will not go to normal but will generally be skewed. Second, the standard deviation of the rate (standard error) will not go to zero as sample size increases: there may remain an irreducible level of variation of observed rates due to the variation of probabilities in the population. Third, as a result of the non-normality of the the distribution of the observed rate, the simple statistical testing framework outlined above must be modified. One relatively simple approach is to use count (Poisson or Negative Binomial) regression. Finally, because of the irreducible variation, we should expect variation from one trial to another, even for large samples.

Applying this standard error analysis to Snow's data (Tables 9, 29, 30) we have:

r_1&n_1 Mortality rate and population for Southwark & Vauxhall customers, controls drinking dirty water: $r_1 = .004643,\ n_1 = 113,056$

r_2&n_2 Mortality rate and population for Lambeth customers, treated with clean water: $r_2 = .000673,\ n_2 = 139,606$

SE r_1 Standard Error (standard deviation of distribution) of rate 1: $SE1 = \sqrt{^{r1(1-r1)}/_{n1}}$; $SE1$=.0002021

SE r_2 Standard Error (standard deviation of distribution) of rate 2: $SE2 = \sqrt{^{r2(1-r2)}/_{n2}}$; $V2$= .00006940

$r_1 - r_2$ Difference in rates: .00464-.000673=.00397

SE($r_1 - r_2$) Standard Error of the difference in rates: $SE(r1-r2) = \sqrt{^{r1(1-r1)}/_{n1} + {}^{r2(1-r2)}/_{n2}}$. This assumes that variances are different. SE($r_1 - r_2$)=.000214

t-ratio Test statistic: $\frac{r1-r2}{SE(r1-r2)}$; t-ratio = 18.6, p-value = 1.6E-77 df degrees-of-freedom $= \frac{(V1+V2)^2}{V1^2/(n1-1)+V2^2/(n2-1)}$ (with $V1 = SE1^2$, $V2 = SE2^2$). With the large sample size here the test statistic the degrees of freedom is so large (139,716) that the ratio will be normally distributed rather than Student-t.

For South London's huge sample size (252,662 total population) the standard analysis would imply very low standard errors and precise estimates. The problem posed by Snow's data is that the observed variation of sub-district rates belies this precision and requires an extension of the simple statistical framework just described. (See also Jakobsen et al. [2015] for a discussion of using count data in randomized trials.)

10.3.2 Error Analysis Incorporating Overdispersion

The Bernoulli / Poisson assumption appears (and in many respects is) innocuous and imminently reasonable. But it is incomplete. It presumes that the only source of variation in the counts and rates at the group level is from random variation due to finite sample size. As the sample grows bigger the distribution narrows. The standard deviation is $\sqrt{r(1-r)/n}$ and as n grows the distribution shrinks around the rate r. This is shown in Figure 11.

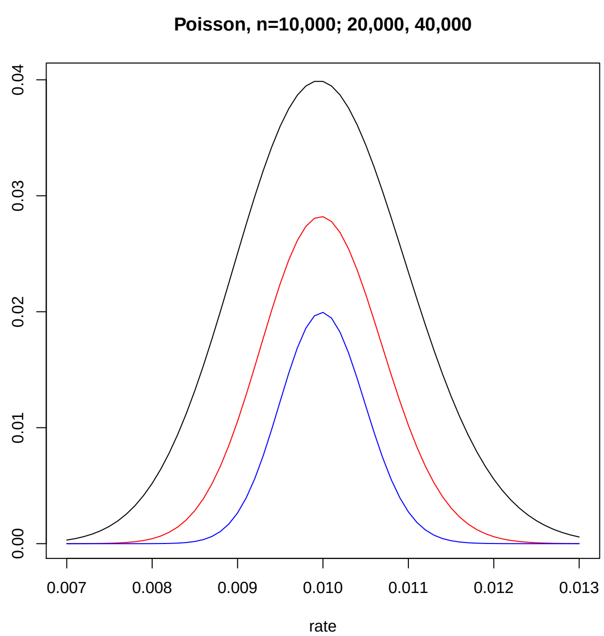

Figure 11: Binomial / Poisson Density for Fixed Rate (.010) and Increasing Group (Population) Size

Focusing on only the aggregate data in Table 30 we cannot examine or test the Bernoulli / Poisson assumption. In contrast the sub-district data in Tables 33 and 36 do allow us to examine the assumption, and the data show that mortality by sub-districts varies substantially. We can compare the actual rates with the aggregate calculated rates from Table 30 (the same rates as the Poisson regression in Table 20). Figure 12 shows the actual (solid) and predicted (circle) mortality rates.

We should, of course, expect variation in sub-district rates simply from random variation in counts. The question is whether the observed variation is more than we would expect, given the population (sample size) for the sub-districts. Figure 12 shows approximate 95% confidence bands for this random variation. Six observations are outside the bands, 17% of the sub-districts, a clear indication that there is more variation than accounted for simply by finite size of the sub-districts.

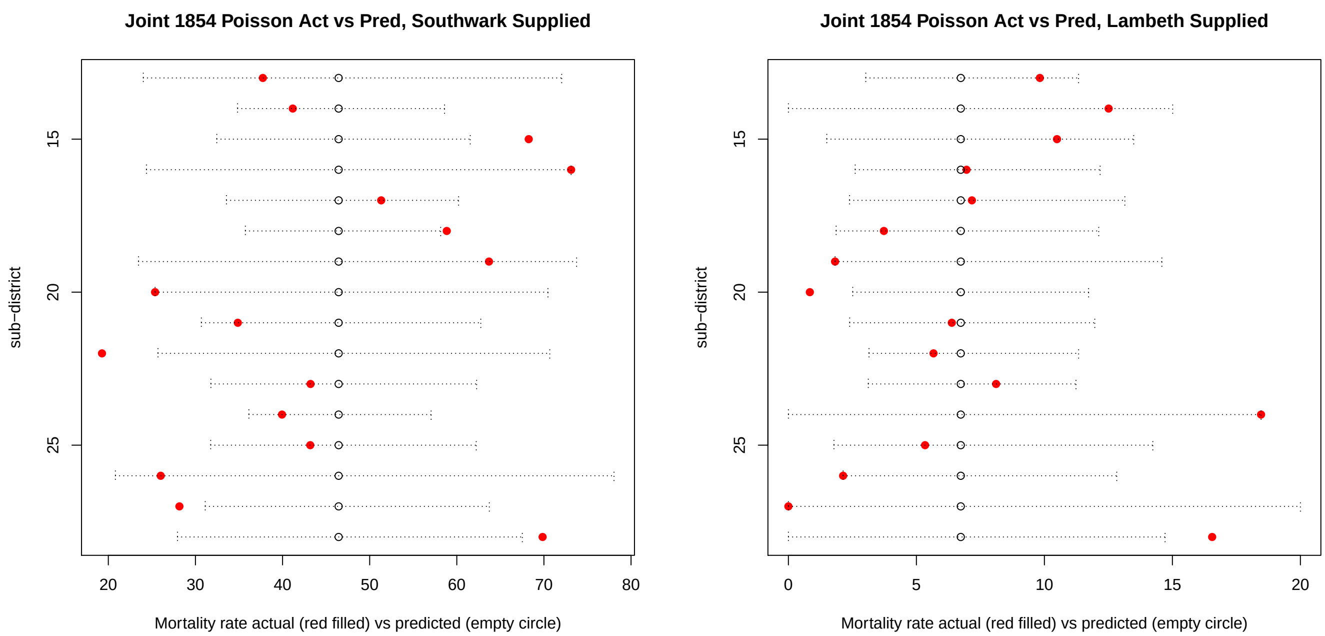

Figure 12: Mortality per 10,000 for "Next 16" Jointly-Supplied Sub-Districts, Separately for Customers Supplied by Southwark & Vauxhall Co versus Lambeth Co. Actual (for the First Seven Weeks, ending 26th August 1854) and Predicted from Poisson Count Model (with 95% confidence bands)

"Actual" is the mortality per 10,000 population calculated using the counts and population from Snow [1856] Tables I and II (population from the Registrar-General, published in Simon [1856], counts by supplier collected by Snow and Mr. Whiting for the seven weeks ending 26th August 1854, published in Snow [1855] Table VIII). "Predicted" is calculated from the estimated Poisson model in the first column of Table 20, with intensity (probability of death) estimated for Southwark customers and Lambeth customers.

A more formal statistical measure, shown in Table 20, is the residual deviance for the Poisson regression. The residual deviance will be (approximately) chi-squared distributed with a large value and low p-value indicating that the residual deviance after fitting the model is too large. This similarly rejects the hypothesis that the sub-district variation is consistent with the Bernoulli / Poisson assumption

There are two possible explanations for overdispersion and large residual deviance seen in Figure 12. The first is that sub-districts have intrinsically different (but fixed) Poisson infection rates. This can easily be accommodated by adding sub-district fixed effects, as shown in the second column of Table 20 and in Figure 13. Sub-district variation by itself does not invalidate the Bernoulli / Poisson framework, *if* we are willing to assume that the sub-district differences are fixed and unchanging. The second possible explanation is that there is random variation in the Poisson infection rates across sub-districts, and that the differences we observe are (at least partially) the result of random variation in the underlying Poisson rates.

Simply comparing across sub-districts cannot distinguish between fixed effects versus random effects. We do, however, have additional observations and variation – we observe rates for customers supplied by the Southwark & Vauxhall Company versus Lambeth Company.[63] We can test whether the residual variation

[63] In the analysis of the difference-in-differences data, Tables 26, raw data shown in Table 33, we have observations for 1849

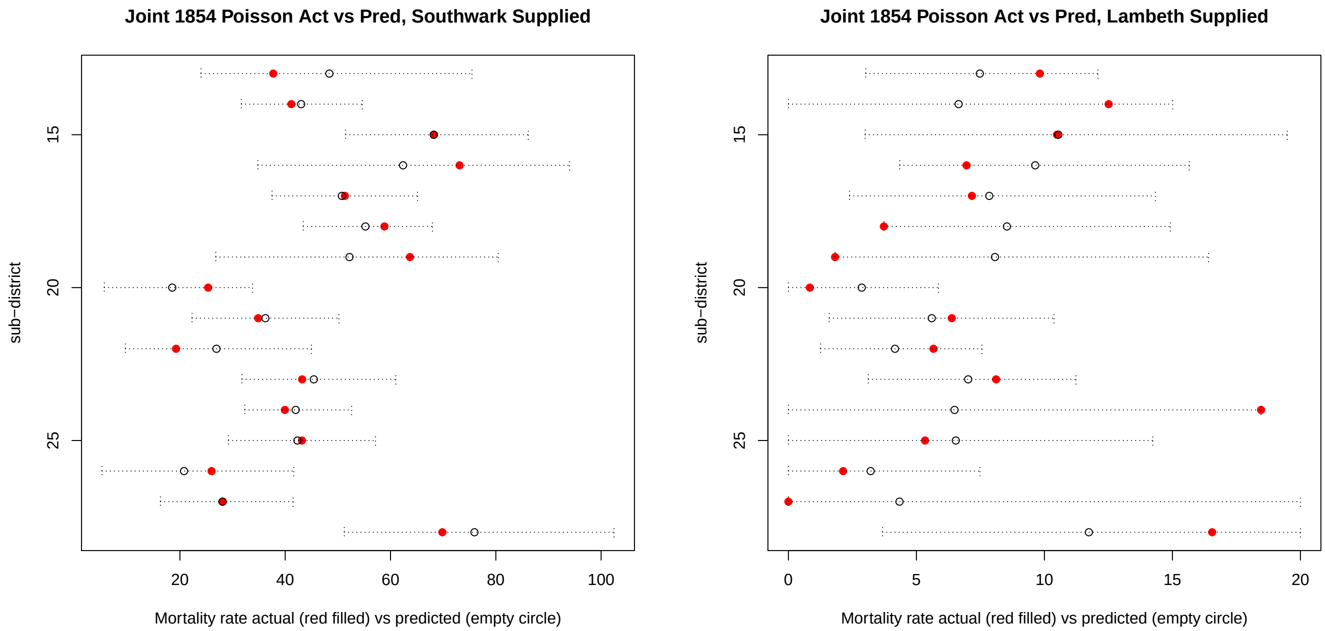

Figure 13: Mortality per 10,000 for "Next 16" Jointly-Supplied Sub-Districts, Separately for Customers Supplied by Southwark & Vauxhall Co versus Lambeth Co. Actual (for the First Seven Weeks, ending 26th August 1854) and Predicted from Poisson Count Model (with 95% confidence bands)

"Actual" is the mortality per 10,000 population calculated using the counts and population from Snow [1856] Tables I and II (population from the Registrar-General, published in Simon [1856], counts by supplier collected by Snow and Mr. Whiting for the seven weeks ending 26th August 1854, published in Snow [1855] Table VIII). "Predicted" is calculated from the estimated Poisson model in the first column of Table 20, with intensity (probability of death) estimated for Southwark customers and Lambeth customers.

across supplier (after accounting for a clean-water treatment effect and the sub-district fixed effects) is larger than would be expected given purely variation in Poisson counts. Figure 13 shows the fitted fixed effects regression and we can see that most observations are within the 95% confidence bands, but there are three (maybe four) observations right at the edge. The residual deviance in the second column of Table 20 provides a formal statistical test with p-value 5.1%, showing that we can reject the fixed effects model at the 10% but not the 5% level.

The question here is not the estimated effect itself but the standard error, the precision with which we estimate the effect. If there is random variation in the underlying rates, across sub-districts or across control versus treated, then the standard error estimated from the Binomial or the Poisson assumption will be too small.

Mixing and random variation in the rates could be modeled by a mixture of Poissons. The Poisson intensity for sub-districts or control versus treated are drawn from a random distribution. The gamma is a convenient distribution to use for the distribution of Poisson intensities, giving as it does a Negative Binomial distribution for the resulting counts.[64]

A negative binomial will have variance of counts of $r \cdot n + {}^{r^2 n^2}/_{\theta}$ and thus standard deviation of the rate of $\sqrt{{}^{r}/_{n} + {}^{r^2}/_{\theta}}$. This will *not* go to zero as n increases and the reason is simple: there is an underlying

versus 1854.

[64]Mixing with other distributions could be equally plausible. The gamma is often used because it is mathematically convenient: a gamma-mixture of Poissons is negative binomial.

Table 31: Poisson and Negative Binomial Regressions for Sub-District Quasi Randomization, Seven Weeks Ending 26th August

	Poisson	Poisson, fixed effects	Neg Binom	Neg Binom, house density
Lambeth effect (ln)	**-1.931**	**-1.867**	**-1.888**	**-1.870**
standard error	**0.112**	**0.119**	**0.158**	**0.158**
z value	-17.2	-15.7	-11.9	-11.8
p-value	1.3E-66	6.9E-56	1.0E-32	2.0e-32
robust standard error	0.165	0.213	0.176	0.187
z value	-11.7	-8.8	-10.7	-10.0
effect (ratio)	6.90	6.47	6.60	6.49
theta ("size" from Gamma mixing)			11.27	11.77
sub-district fixed effects	NO	YES	NO	NO
Housing density	NO	NO	NO	-0.068
standard error				0.079
p-value				39.1%
Residual Deviance (df, p-value)	79.4 (30, p<1e-5)	24.9 (15, p=5.1%)	38.8 (30, p=13.0%)	38.7 (29, p=10.8%)
Southwark pred mort (per 10,000)	46.4	46.4	45.7	46.0
Lambeth pred mort (per 10,000)	6.73	6.73	6.92	6.89

Data on deaths by sub-district and by supplier (Southwark & Vauxhall Co versus Lambeth Co) for the "next 16" sub-districts jointly supplied by the two companies, where the mixing by supplier effectively produces quasi-random assignment. Data from Snow [1856] Table I and II which show a) Deaths by water source (Southwark & Vauxhall Co, Lambeth Co, pump-wells, Thames, unascertained), b) houses and density in 1851, c) houses and population by supplier (Southwark versus Lambeth Cos). Death counts are for the seven weeks ending 26th August (as in Snow [1855] Table VIII). Total of 32 observations. The Poisson and Negative Binomial regressions are fitted with the R function glm (family=poisson) and the glm.nb function from the MASS package. The parameter "theta" is the size or θ for a "parametrization (1)" Negative Binomial (see Section 10.1). Robust standard errors are calculated with the R "sandwich" package, using the default "HC3" for the adjusted variance-covariance matrix (see Zeileis [2004] and the R "sandwich" manual). The "Residual Deviance" is approximately chi-squared and the p-value is the probability of observing a chi-squared RV that large or larger – a small p-value indicates the regression does not fit the data.

distribution among individuals of the rate (incidence of cholera) and this distribution will induce variation in rates even for a large population. The third column of Table 20 shows a Negative Binomial regression, where we assume that the sub-district effects are random. We cannot reject that model even at the 10% level.

What is unavoidable when we examine the sub-district data is that a *fixed and constant* Bernoulli rate (or Poisson intensity) across the whole population is rejected by the data. We should not consider this a nuisance – it is telling us something important about our world. The difficulty is discerning exactly what it is telling us. It seems reasonable that sub-districts could be inherently different, so that fixed effects may be reasonable. But are those differences fixed or are they random? Would the variation across sub-districts be different were we to draw another sample? For the 1849 versus 1854 comparison discussed in Section 10.1 we do have two samples drawn from two time periods, and they do show more variation than we should see from fixed Poisson rates. For the 1854 comparison discussed in this section we have less data and cannot reliably determine whether the sub-district variation is fixed or random.

Whether the differences are fixed or random makes little difference to the size of the estimated effect, but does matter for the implied precision of the estimate. Ignoring such variability or treating it as fixed may lead to over-confidence in a single sample and will bias our statistical testing. As an example, for the aggregate data (Table 30) the constant-Bernoulli assumption implies a z-value (ratio to standard error) of 18.6, while a Negative Binomial regression that assumes the sub-district variation is random gives a z-value of 11.9. This is still very large and implies a highly-significant effect, but is much lower than for the constant-Bernoulli assumption.

Practical application of heterogeneity and mixing distributions to the analysis of randomized trials is relatively straightforward. For small sample sizes the heterogeneity (mixture) will be less important than stochastic variation. For a Poisson with sample size n the standard error of the rate will be $\sqrt{r/n}$ while for a Negative Binomial (gamma-mixed Poisson) it will be $\sqrt{r/n + r^2/\theta}$. For small sample size and low rate r, the first term will dominate and the standard error will behave like that for a Poisson (or constant-rate Bernoulli). For example, with n=50 and reasonable values (say r=.05, $\theta = 5$) the difference is small: 0.032 for Poisson, 0.039 for Negative Binomial. Heterogeneity (the r^2/θ term) dominates when sample size is large: for n=5,000 the standard error would be 0.0032 for Poisson and 0.023 for Negative Binomial.

For large samples, however, we can actually measure the within-sample variation. We can split our overall sample into sub-samples and use count regressions (Poisson and Negative Binomial) to test for and estimate possible heterogeneity and mixing. When heterogeneity (a non-constant Bernoulli process) is present this will show up as more variation across sub-samples than predicted by the Bernoulli or Poisson. We can estimate the heterogeneity via an assumed gamma (or other) mixing distribution and apply our statistics tests based on the mixed Poisson.[65]

10.4 Explanation / List of Tables for Snow [1855]

- Table III: 1849, by District. With "Annual value of House & Shop room to each person in £"
- Table V: 1853, by District, 17wks Aug 21 - Dec 17, 1853
- Table VI p 73: 1853, by Sub-District, August 1853 - January 1854
- Table VII: 1854 p 84: by Sub-District, four weeks ending 5th August. Categorized by source (Southwark & Vauxhall; Lambeth; Pump-wells; River Thames, ditches, etc.; Unascertained) categorization carefully performed by Snow
- Table VIII p 85: by Sub-District, seven weeks ending 26th August. Categorized by source (Southwark & Vauxhall; Lambeth; Pump-wells; River Thames, ditches, etc.; Unascertained) categorization carefully performed by Snow
- Table IX p. 86: Summary of Table VIII (plus additional for "Rest of London"), and rates per household
- Table X p. 87: By week Sept 2 - October 14, for South London, from Registrar-General. Categorization of water source but not as thoroughly done as by Snow for July & August.
- Table XI p. 88: 1854, population and death rates for all London and aggregate districts ("West", "North", ...) and for houses supplied by Southwark & Vauxhall vs Lambeth. (Rates to population, not to houses as for Table IX)

[65] One important point – the simple testing based on the t-test for equality of two means cannot be applied in the presence of heterogeneity (a mixture distribution). The sum of Bernoullis is Binomial which goes to normal in the limit, but the sum of mixed Bernoullis or Poissons will not go to normal in the limit. For example, the Negative Binomial can be decidedly non-normal. Testing must be based on the Negative Binomial or other count regression.

- Table XII p. 90: Deaths 1849 & 1854 compared (not rates). By sub-district. cf Table VII. For 1854 through October 21 (cf p 89) "It is necessary to observe, however, that the supply of the Lambeth Company has been extended to Streatham, Norwood, and Sydenham, since 1849, in which year these places were not supplied by any water company."

10.5 Detailed Tables and Figures for South London "Grand Experiment"

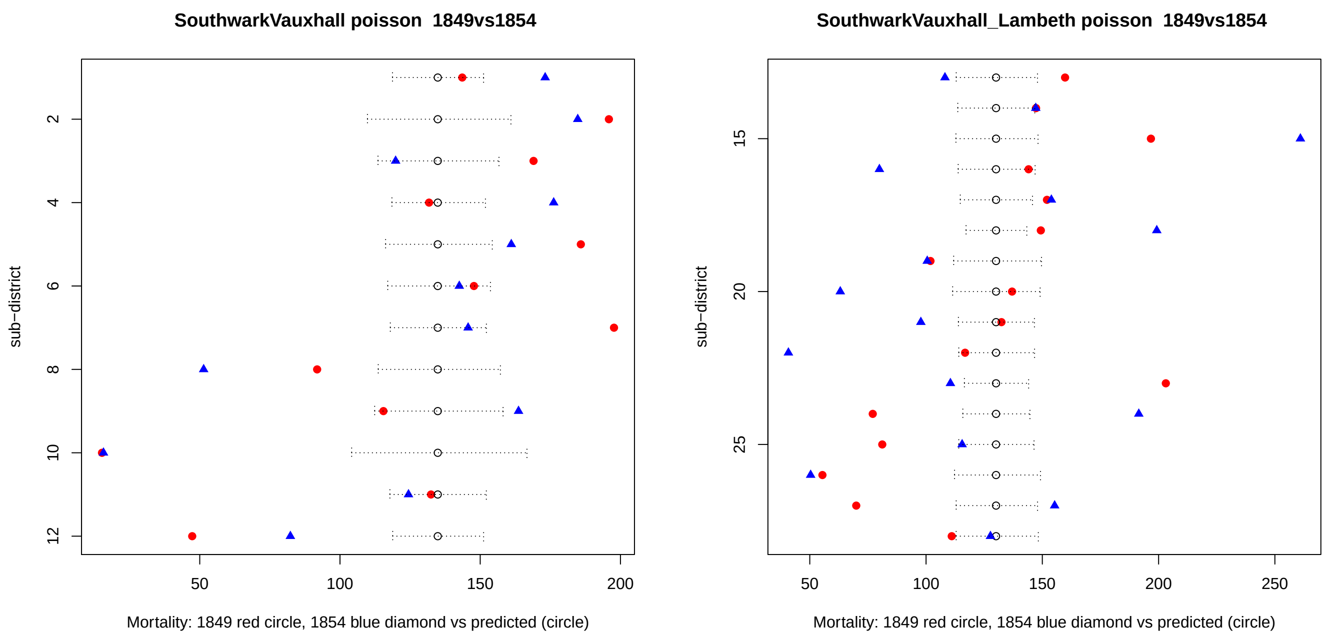

Figure 14: Mortality per 10,000, Poisson Count Model, Same Rate All Sub-Districts, Predicted (with 95% confidence bands) and Actual 1849 & 1854 (Adjusted for Time and Single Treatment Effect)

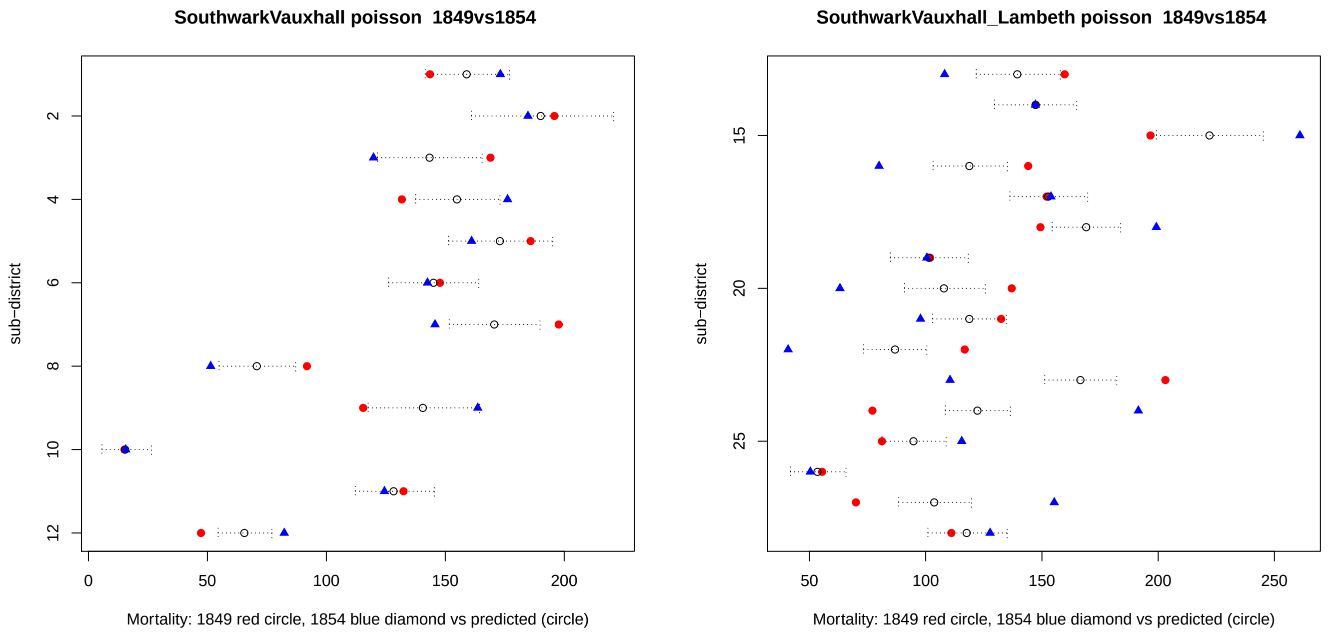

Figure 15: Mortality per 10,000, Poisson Count Model, Different Rates for Sub-Districts, Predicted (with 95% confidence bands) and Actual 1849 & 1854 (Adjusted for Time and Single Treatment Effect)

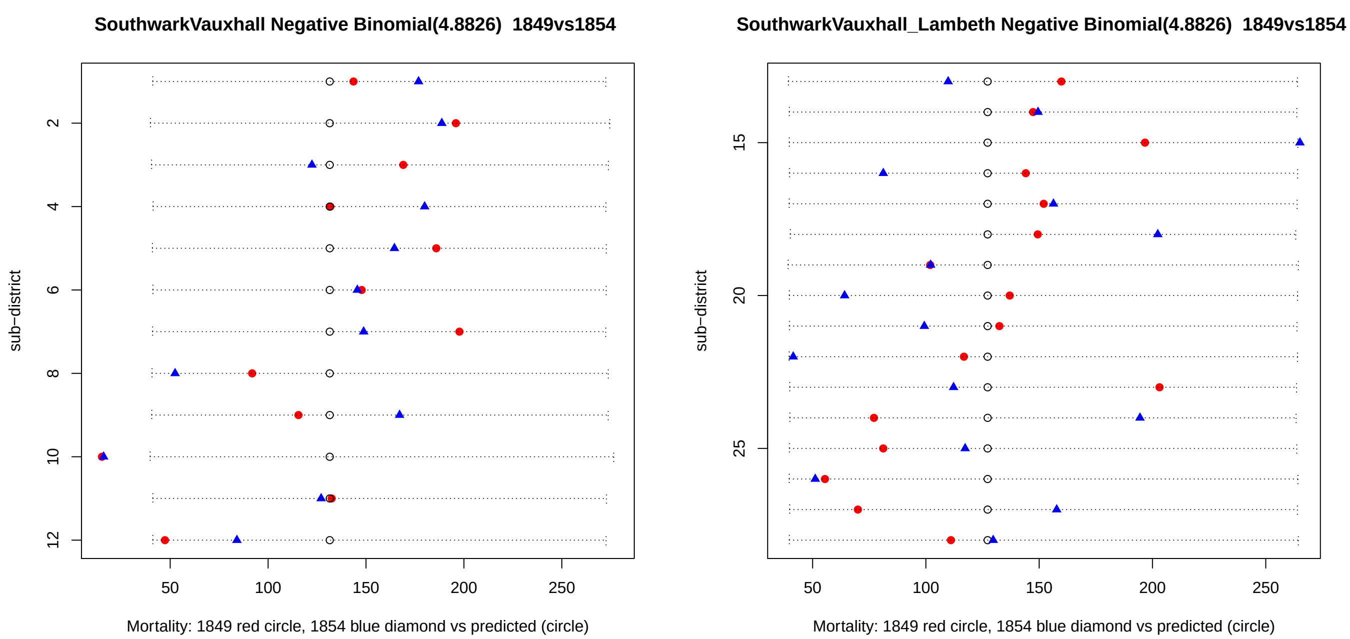

Figure 16: Mortality per 10,000, Negative Binomial, Same Rates for Sub-Districts & Single Treatment Effect, Predicted (with 95% confidence bands) and Actual 1849 & 1854 (Adjusted for Time and Single Treatment Effect)

Table 32: Snow's Table XII – Deaths from Cholera in 1849 & 1854 (Snow [1855] p 90) with Population in 1851 from Table VIII (p 85)

	Sub-Districts	Deaths from Cholera in 1849	Deaths from Cholera in 1854	Water Supplier	Population in 1851
1	St. Saviour, Southwark	283	371	SouthwarkVauxhall	19,709
2	St. Olave, Southwark	157	161	SouthwarkVauxhall	8,015
3	St. John, Horsleydown	192	148	SouthwarkVauxhall	11,360
4	St. James, Bermondsey	249	362	SouthwarkVauxhall	18,899
5	St. Mary Magdalen	259	244	SouthwarkVauxhall	13,934
6	Leather Market	226	237	SouthwarkVauxhall	15,295
7	Rotherhithe	352	282	SouthwarkVauxhall	17,805
8	Battersea	111	171	SouthwarkVauxhall	10,560
9	Wandsworth	97	59	SouthwarkVauxhall	9,611
10	Putney	8	9	SouthwarkVauxhall	5,280
11	Camberwell	235	240	SouthwarkVauxhall	17,742
12	Peckham	92	174	SouthwarkVauxhall	19,444
13	Christchurch, Southwark	256	113	SouthwarkVauxhall & Lambeth	16,022
14	Kent Road	267	174	SouthwarkVauxhall & Lambeth	18,126
15	Borough Road	312	270	SouthwarkVauxhall & Lambeth	15,862
16	London Road	257	93	SouthwarkVauxhall & Lambeth	17,836
17	Trinity, Newington	318	210	SouthwarkVauxhall & Lambeth	20,922
18	St. Peter, Walworth	446	388	SouthwarkVauxhall & Lambeth	29,861
19	St. Mary, Newington	143	92	SouthwarkVauxhall & Lambeth	14,033
20	Waterloo Road (1st)	193	58	SouthwarkVauxhall & Lambeth	14,088
21	Waterloo Road (2nd)	243	117	SouthwarkVauxhall & Lambeth	18,348
22	Lambeth Church (1st)	215	49	SouthwarkVauxhall & Lambeth	18,409
23	Lambeth Church (2nd)	544	193	SouthwarkVauxhall & Lambeth	26,784
24	Kennington (1st)	187	303	SouthwarkVauxhall & Lambeth	24,261
25	Kennington (2nd)	153	142	SouthwarkVauxhall & Lambeth	18,848
26	Brixton	81	48	SouthwarkVauxhall & Lambeth	14,610
27	Clapham	114	165	SouthwarkVauxhall & Lambeth	16,290
28	St. George, Camberwell	176	132	SouthwarkVauxhall & Lambeth	15,849
29	Norwood	2	10	Lambeth	3,977
30	Streatham	154	15	Lambeth	9,023
31	Dulwich	1	0	Lambeth	1,632
32	Sydenham	5	12	Lambeth	4,501
	First 12 sub-districts	2261	2458	first12	167,654
	Next 16 sub-districts	3905	2547	next16	300,149
	Last 4 sub-districts	162	37	last4	19,133
	TOTAL	6,328	5,042		486,936

Table 33: Mortality Rates from Cholera per 10,000 Persons in 1849 & 1854 (from Snow Table XII & using population in 1851 from Table VIII)

	Sub-Districts	Deaths rates from Cholera in 1849, per 10,000	Deaths rates from Cholera in 1854, per 10,000	Water Supplier	Degree of Lambeth Supply
1	St. Saviour, Southwark	144	188	SouthwarkVauxhall	none
2	St. Olave, Southwark	196	201	SouthwarkVauxhall	none
3	St. John, Horsleydown	169	130	SouthwarkVauxhall	none
4	St. James, Bermondsey	132	192	SouthwarkVauxhall	none
5	St. Mary Magdalen	186	175	SouthwarkVauxhall	none
6	Leather Market	148	155	SouthwarkVauxhall	none
7	Rotherhithe	198	158	SouthwarkVauxhall	none
8	Battersea	115	178	SouthwarkVauxhall	none
9	Wandsworth	92	56	SouthwarkVauxhall	none
10	Putney	15	17	SouthwarkVauxhall	none
11	Camberwell	132	135	SouthwarkVauxhall	none
12	Peckham	47	89	SouthwarkVauxhall	none
13	Christchurch, Southwark	160	71	SouthwarkVauxhall & Lambeth	more_Lambeth
14	Kent Road	147	96	SouthwarkVauxhall & Lambeth	less_Lambeth
15	Borough Road	197	170	SouthwarkVauxhall & Lambeth	less_Lambeth
16	London Road	144	52	SouthwarkVauxhall & Lambeth	more_Lambeth
17	Trinity, Newington	152	100	SouthwarkVauxhall & Lambeth	less_Lambeth
18	St. Peter, Walworth	149	130	SouthwarkVauxhall & Lambeth	less_Lambeth
19	St. Mary, Newington	102	66	SouthwarkVauxhall & Lambeth	less_Lambeth
20	Waterloo Road (1st)	137	41	SouthwarkVauxhall & Lambeth	more_Lambeth
21	Waterloo Road (2nd)	132	64	SouthwarkVauxhall & Lambeth	less_Lambeth
22	Lambeth Church (1st)	117	27	SouthwarkVauxhall & Lambeth	more_Lambeth
23	Lambeth Church (2nd)	203	72	SouthwarkVauxhall & Lambeth	less_Lambeth
24	Kennington (1st)	77	125	SouthwarkVauxhall & Lambeth	less_Lambeth
25	Kennington (2nd)	81	75	SouthwarkVauxhall & Lambeth	less_Lambeth
26	Brixton	55	33	SouthwarkVauxhall & Lambeth	less_Lambeth
27	Clapham	70	101	SouthwarkVauxhall & Lambeth	less_Lambeth
28	St. George, Camberwell	111	83	SouthwarkVauxhall & Lambeth	less_Lambeth
29	Norwood	5	25	Lambeth	all
30	Streatham	171	17	Lambeth	all
31	Dulwich	6	0	Lambeth	all
32	Sydenham	11	27	Lambeth	all
	First 12 sub-districts	135	147	first12	none
	Next 16 sub-districts	130	85	next16	some
	Last 4 sub-districts	85	19	last4	all
	TOTAL	130	104		some

Table 34: Houses & Population by Sub-District from (Snow [1856] Tables I & II, from Simon [1856])

	Sub-Districts	Houses 1851	Pop per house	Southwark Houses	Southwark Pop	Lambeth Houses	Lambeth Pop
1	St. Saviour, Southwark	2,713	7.3	2,238	16,337	123	898
2	St. Olave, Southwark	880	9.1	961	8,745	0	0
3	St. John, Horsleydown	1,480	7.7	1,170	9,360	0	0
4	St. James, Bermondsey	2,863	6.6	3,511	23,173	105	693
5	St. Mary Magdalen	1,865	7.5	2,301	17,258	0	0
6	Leather Market	2,279	6.7	2,090	14,003	163	1,092
7	Rotherhithe	2,792	6.4	1,909	12,218	0	0
8	Battersea	1,760	6	1,046	6,276	46	276
9	Wandsworth	1,522	6.3	144	907	15	94
10	Putney	918	5.7	13	74	0	0
11	Camberwell	2,851	6.2	1,474	9,139	103	639
12	Peckham	3,457	5.6	971	5,438	70	392
13	Christchurch, Southwark	1,887	8.5	343	2,915	1,557	13,234
14	Kent Road	2,558	7.1	1,779	12,630	563	3,997
15	Borough Road	2,069	7.7	1,176	8,937	878	6,672
16	London Road	2,365	7.5	383	2,872	1,533	11,497
17	Trinity, Newington	3,224	6.5	1,661	10,132	1,372	8,370
18	St. Peter, Walworth	4,925	6.1	2,340	14,274	1,758	10,724
19	St. Mary, Newington	2,309	6.1	489	2,983	899	5,484
20	Waterloo Road (1st)	1,729	8.1	438	3,548	1,474	11,939
21	Waterloo Road (2nd)	2,191	8.4	864	7,171	1,510	12,533
22	Lambeth Church (1st)	2,451	7.5	415	3,113	2,117	15,878
23	Lambeth Church (2nd)	3,849	7	1,124	7,868	2,289	16,023
24	Kennington (1st)	3,977	6.1	2,586	15,775	444	2,708
25	Kennington (2nd)	3,288	5.7	1,206	7,874	986	5,620
26	Brixton	2,362	6.1	310	1,922	1,509	9,356
27	Clapham	2,657	6.1	1,106	6,747	22	134
28	St. George, Camberwell	2,845	5.6	767	4,295	971	5,437
29	Norwood	600	6.6			160	1,066
30	Streatham	1,419	6.4			515	3,244
31	Dulwich	259	6.3			4	25
32	Sydenham	801	5.6				
	TOTAL	73,145	na	39,726	267,625	24,854	171,528
	Houses in streets with no deaths			4,500	28,929	3,643	23,338
	Not identified			411	2,712	25	165

Table 35: Deaths by Water Source for Seven Weeks ending 26th August 1854, Snow [1855] Table VIII

	Sub-Districts	Overall	Southwark & Vauxhall	Lambeth	Pump-wells	River Thames & ditches	Un-ascertained
1	St. Saviour, Southwark	125	115	0	0	10	0
2	St. Olave, Southwark	53	43	0	0	5	5
3	St. John, Horsleydown	51	48	0	0	3	0
4	St. James, Bermondsey	123	102	0	0	21	0
5	St. Mary Magdalen	87	83	0	0	4	0
6	Leather Market	81	81	0	0	0	0
7	Rotherhithe	103	68	0	0	35	0
8	Battersea	54	42	0	4	8	0
9	Wandsworth	11	1	0	2	8	0
10	Putney	1	0	0	1	0	0
11	Camberwell	96	96	0	0	0	0
12	Peckham	59	59	0	0	0	0
13	Christchurch, Southwark	25	11	13	0	0	1
14	Kent Road	57	52	5	0	0	0
15	Borough Road	71	61	7	0	0	3
16	London Road	29	21	8	0	0	0
17	Trinity, Newington	58	52	6	0	0	0
18	St. Peter, Walworth	90	84	4	0	0	2
19	St. Mary, Newington	21	19	1	1	0	0
20	Waterloo Road (1st)	10	9	1	0	0	0
21	Waterloo Road (2nd)	36	25	8	1	2	0
22	Lambeth Church (1st)	18	6	9	0	1	2
23	Lambeth Church (2nd)	53	34	13	1	0	5
24	Kennington (1st)	71	63	5	3	0	0
25	Kennington (2nd)	38	34	3	1	0	0
26	Brixton	9	5	2	0	0	2
27	Clapham	24	19	0	5	0	0
28	St. George, Camberwell	42	30	9	2	0	1
29	Norwood	8	0	2	1	5	0
30	Streatham	6	0	1	5	0	0
31	Dulwich	0	0	0	0	0	0
32	Sydenham	4	0	1	2	0	1
	TOTAL	1,514	1,263	98	29	102	22

Total total counts are as reported by the Registrar-General. Allocation to water source by Snow and Mr. Whiting, as described in Snow [1855] p. 76 ff.

Table 36: Mortality per 10,000 population by Water Source for Seven Weeks ending 26th August 1854, Snow [1855] Table VIII and Snow [1856] Tables I & II

	Sub-Districts	Overall	Southwark & Lambeth combined	Southwark & Vauxhall	Lambeth
1	St. Saviour, Southwark	63.4	66.7	70.4	–
2	St. Olave, Southwark	66.1	49.2	49.2	–
3	St. John, Horsleydown	44.9	51.3	51.3	–
4	St. James, Bermondsey	65.1	42.7	44.0	–
5	St. Mary Magdalen	62.4	48.1	48.1	–
6	Leather Market	53.0	53.7	57.8	–
7	Rotherhithe	57.8	55.7	55.7	–
8	Battersea	51.1	64.1	66.9	–
9	Wandsworth	11.4	10.0	11.0	–
10	Putney	1.9	0.0	0.0	–
11	Camberwell	54.1	98.2	105.0	–
12	Peckham	30.3	101.2	108.5	–
13	Christchurch, Southwark	15.6	14.9	37.7	9.82
14	Kent Road	31.4	34.3	41.2	12.51
15	Borough Road	44.8	43.6	68.3	10.49
16	London Road	16.3	20.2	73.1	6.96
17	Trinity, Newington	27.7	31.3	51.3	7.17
18	St. Peter, Walworth	30.1	35.2	58.8	3.73
19	St. Mary, Newington	15.0	23.6	63.7	1.82
20	Waterloo Road (1st)	7.1	6.5	25.4	0.84
21	Waterloo Road (2nd)	19.6	16.7	34.9	6.38
22	Lambeth Church (1st)	9.8	7.9	19.3	5.67
23	Lambeth Church (2nd)	19.8	19.7	43.2	8.11
24	Kennington (1st)	29.3	36.8	39.9	18.46
25	Kennington (2nd)	20.2	27.4	43.2	5.34
26	Brixton	6.2	6.2	26.0	2.14
27	Clapham	14.7	27.6	28.2	0.00
28	St. George, Camberwell	26.5	40.1	69.8	16.55
29	Norwood	20.1	18.8	–	18.76
30	Streatham	6.6	3.1	–	3.08
31	Dulwich	0.0	0.0	–	–
32	Sydenham	8.9	–	–	–
	TOTAL	31.1	31.0	47.2	5.71

Counts are from Snow [1855] Table VII (reproduced in Snow [1856] Tables I & II). "Overall" mortality rates are the total counts divided by 1851 population. "Southwark & Lambeth combined" are the counts by supplier divided by population by supplier. This will differ from the "Overall" because it excludes pump-wellls, Thames water, and unascertained. (Also because of errors in the population ascribed to Vauxhall versus Lambeth, as described by Snow.) "Southwark & Vauxhall" and "Lambeth" mortality are counts divided by the population from Snow [1856] Tables I & II (originally from Simon [1856]).

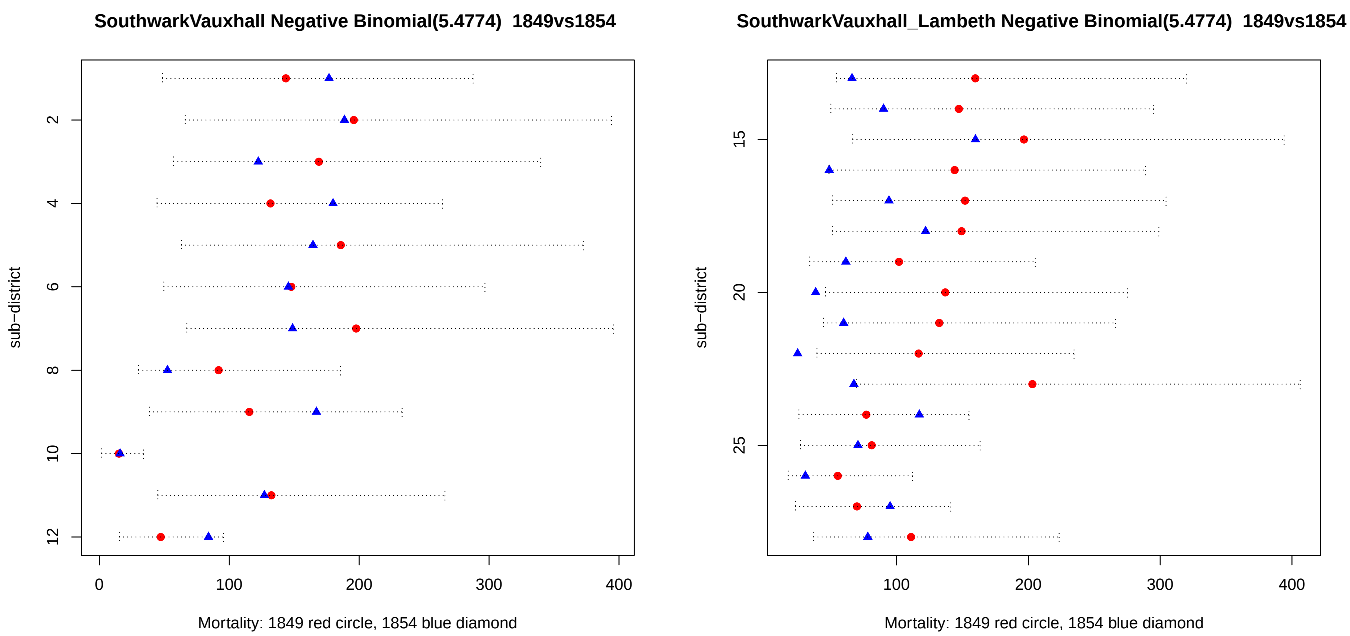

Figure 17: Mortality per 10,000, Negative Binomial, Same Rates for Sub-Districts & Two Treatment Effects, Actual 1849 (with 95% confidence bands) vs 1854 (Adjusted for Time Effect but not Treatment)

References

Poisson Regression. Technical report, NCSS Statistical Software, 2017. https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Poisson_Regression.pdf.

Joshua D. Angrist and Jorn-Steffen Pischke. *Mostly Harmless Econometrics: An Empiricist's Companion.* Princeton University Press, 1 edition edition, December 2008.

BMJ. 13. Study design and choosing a statistical test | The BMJ. URL https://www.bmj.com/about-bmj/resources-readers/publications/statistics-square-one/13-study-design-and

William Casselman. The Legend of Abraham Wald, 2016. URL http://www.ams.org/publicoutreach/feature-column/fc-2016-06.

Robert A. Dahl. Cause and Effect in the Study of Politics. In Daniel Lerner, editor, *Cause and effect: Hayden colloquium on scientific method and concept.* Free Press, New York, 1965.

Charles Dimaggio. A Brief Introduction to Spatial Analysis in R. URL http://www.columbia.edu/ cjd11/charles$_d imaggio/DIRE/styled-4/code-16/(20)$.

A Evans. Cuasation and Disease: A chronological journey. *Journal of Epidemiology*, 108(4):249, 1978.

David Freedman. Statistical Models and Shoe Leather. *Sociological Methodology*, 21:291–313, 1991. ISSN 0883-4237, 2168-8745. doi: 10.2307/270939. URL https://www-jstor-org.proxy.uchicago.edu/stable/270939.

David Freedman. From association to causation: some remarks on the history of statistics. *Statistical Science*, 14(3):243–258, August 1999. ISSN 0883-4237, 2168-8745. doi: 10.1214/ss/1009212409. URL https://projecteuclid.org/euclid.ss/1009212409.

Ralph R. Frerichs. John Snow Site, a. URL http://www.ph.ucla.edu/epi/snow.html.

Ralph R. Frerichs. John Snow maps, b. URL http://www.ph.ucla.edu/epi/snow/highressnowmap.html.

Sandra Hempel. *The Strange Case of the Broad Street Pump: John Snow and the Mystery of Cholera.* University of California Press, Berkeley, first edition edition, January 2007. ISBN 978-0-520-25049-9.

JC Jakobsen, M Tamborrino, P Winkel, N Haase, A Perner, J Wetterslev, and C Gluud. Count Data Analysis in Randomised Clinical Trials. *Journal of Biometrics & Biostatistics*, 06(01), 2015. ISSN 21556180. doi: 10.4172/2155-6180.1000227. URL https://www.omicsonline.org/open-access/count-data-analysis-in-randomised-clinical-trials-2155-6180-10

Steven Johnson. *The Ghost Map: The Story of London's Most Terrifying Epidemic–and How It Changed Science, Cities, and the Modern World.* Riverhead Books, New York, reprint edition edition, October 2007. ISBN 978-1-59448-269-4.

Rebecca Katz and Burton Singer. Can an Attribution Assessment Be Made for Yellow Rain? Systematic Reanalysis in a Chemical-and-Biological-Weapons Use Investigation. *Politics and the Life Sciences*, 26(1): 24–42, March 2007.

Thomas Koch and Kenneth Denike. Rethinking John Snow's South London study: A Bayesian evaluation and recalculation. *Social Science & Medicine*, 63(1):271–283, July 2006. ISSN 0277-9536. doi: 10.1016/j.socscimed.2005.12.006. URL `http://www.sciencedirect.com/science/article/pii/S0277953605006933`.

Imre Lakatos. *The Methodology of Scientific Research Programmes: Volume 1: Philosophical Papers*. Cambridge University Press, Cambridge, November 1980. ISBN 978-0-521-28031-0.

Peter Li. R Package for Analyzing John Snow's 1854 Cholera Map. URL `https://github.com/lindbrook/cholera`.

John Mackenzie. Mapping the 1854 London Cholera Outbreak, 2010. URL `https://www1.udel.edu/johnmack/frec682/cholera/`.

Marc Mangel and Francisco J. Samaniego. Abraham Wald's Work on Aircraft Survivability. *Journal of the American Statistical Association*, 79(386):259–267, June 1984. ISSN 0162-1459. doi: 10.1080/01621459.1984.10478038. URL `https://www.tandfonline.com/doi/abs/10.1080/01621459.1984.10478038`.

K. S. McLeod. Our sense of Snow: the myth of John Snow in medical geography. *Social Science & Medicine (1982)*, 50(7-8):923–935, April 2000. ISSN 0277-9536.

Karl R. Popper. *Popper Selections*. Princeton University Press, Princeton, N.J, February 1985. ISBN 978-0-691-02031-0.

Penguin Press. Abraham Wald and the Missing Bullet Holes, July 2016. URL `https://medium.com/@penguinpress/an-excerpt-from-how-not-to-be-wrong-by-jordan-ellenberg-664e708cfc3d`.

Simon Rogers. John Snow's data journalism: the cholera map that changed the world. *The Guardian*, March 2013. ISSN 0261-3077. URL `http://www.theguardian.com/news/datablog/2013/mar/15/john-snow-cholera-map`.

Kenneth J. Rothman. *Epidemiology: an introduction*. Oxford University Press, New York, N.Y., 2002. ISBN 978-0-19-513553-4.

Narushige Shiode, Shino Shiode, Elodie Rod-Thatcher, Sanjay Rana, and Peter Vinten-Johansen. The mortality rates and the space-time patterns of John Snow's cholera epidemic map. *International Journal of Health Geographics*, 14, June 2015. ISSN 1476-072X. doi: 10.1186/s12942-015-0011-y. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4521606/`.

Shino Shiode. Revisiting John Snow's map: network-based spatial demarcation of cholera area. *International Journal of Geographical Information Science*, 26(1):133–150, January 2012. ISSN 1365-8816. doi: 10.1080/13658816.2011.577433. URL `https://doi.org/10.1080/13658816.2011.577433`.

John Simon, editor. *Report on the last two cholera-epidemics of London: as affected by the consumption of impure water addressed to the Rt. Hon. The President of the General Board of Health, by the Medical Officer of the Board*. Printed by Eyre and Spottiswoode, for HMSO, London, 1856. URL `https://collections-nlm-nih-gov.proxy.uchicago.edu/catalog/nlm:nlmuid-0260772-bk`. OCLC: 14531255.

John Snow. *On the mode of communication of cholera.* John Churchill, London, 1849. OCLC: 14550757.

John Snow. *On the mode of communication of cholera.* London : John Churchill, 2nd edition, 1855. URL `http://archive.org/details/b28985266`.

John Snow. Cholera and the water supply in the south district of London in 1854. *Journal of Public Health and Sanitary Review*, 2:239–257, October 1856. URL `http://www.ph.ucla.edu/epi/snow/cholerawatersouthlondon.html`.

stata.com. poisson - Poisson regression. Technical report, 2017. URL `https://www.stata.com/manuals13/rpoisson.pdf`.

Stephen M. Stigler. *The seven pillars of statistical wisdom.* Harvard University Press, Cambridge, Massachusetts, 2016. ISBN 9780674088917 (pbk.: alk. paper).

Edward R. Tufte. *Visual and Statistical Thinking: Displays of Evidence for Making Decisions.* Graphics Press, Cheshire, Conn, unknown edition edition, April 1997a. ISBN 978-0-9613921-3-0. https://www.edwardtufte.com/tufte/books_textb.

Edward R. Tufte. *Visual Explanations: Images and Quantities, Evidence and Narrative.* Graphics Press, 1st edition edition edition, February 1997b. ISBN 978-1-930824-15-7. https://www.edwardtufte.com/tufte/books_visex.

Peter Vinten-Johansen, Howard Brody, Nigel Paneth, Stephen Rachman, and Michael Russell Rip. *Cholera, Chloroform and the Science of Medicine: A Life of John Snow.* Oxford University Press, Oxford ; New York, 1 edition edition, May 2003. ISBN 978-0-19-513544-2.

Abraham Wald. A Reprint of 'A Method of Estimating Plane Vulnerability Based on Damage of Survivors. Technical Report CRC-432, CENTER FOR NAVAL ANALYSES ALEXANDRIA VA OPERATIONS EVALUATION GROUP, CENTER FOR NAVAL ANALYSES ALEXANDRIA VA OPERATIONS EVALUATION GROUP, July 1980. URL `http://www.dtic.mil/docs/citations/ADA091073`.

Westminster and London School of Hygiene and Tropical Medicine, editors. *Report on the cholera outbreak in the parish of St. james,Westminster, during the autumn of 1854.* J. Churchill, London, 1855.

City of Westminster. Cholera and the Thames, 2018. URL `‘`.

Wilson. John Snow's Cholera data in more formats - Robin's Blog. URL `http://blog.rtwilson.com/john-snows-cholera-data-in-more-formats/`.

Achim Zeileis. Econometric Computing with HC and HAC Covariance Matrix Estimators | Zeileis | Journal of Statistical Software. *Journal of Statistical Software*, 11, November 2004. doi: 10.18637/jss.v011.i10. URL `https://www.jstatsoft.org/article/view/v011i10`.

Achim Zeileis, Christian Kleiber, and Simon Jackman. Regression Models for Count Data in R. Technical report, 2017. URL `https://cran.r-project.org/web/packages/pscl/vignettes/countreg.pdf`.