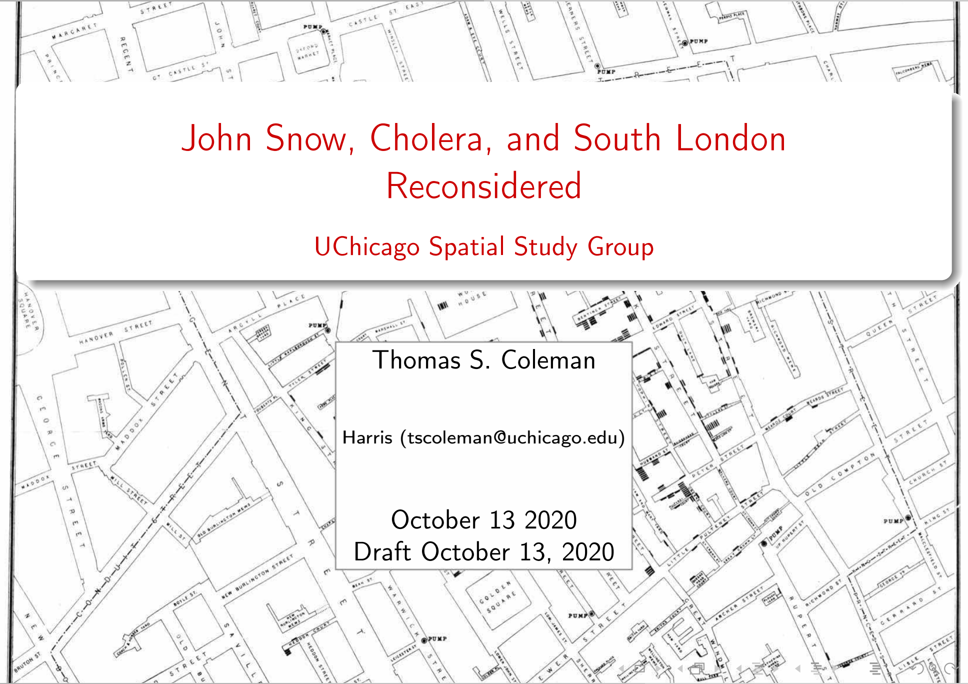


John Snow, Cholera, and South London Reconsidered

UChicago Spatial Study Group

Thomas S. Coleman

Harris (tscoleman@uchicago.edu)

October 13 2020
Draft October 13, 2020

What Causes Cholera? Hugely Important in 1850s London

Horrendous way to die – dehydration, convulsions, blue skin, die within hours

Scourge of mid-1800s London – 1831-32 6,526 dead; 1849 14,137; 1853-54 10,738

Massive uncertainty as to cause

- Bad air (miasma); bad breeding (poverty); bad ground (plague pits)

Huge public health & policy question – and one man knew the answer:

- John Snow & bad water – effort to prove contaminated water as causal agent

Why John Snow and 1850s Cholera?

This work (joint with Peter Vinten-Johansen) is to re-examine and reconsider Snow's analysis of South London

Why Snow? Three reasons:

1. **Rollicking Good Tale** – full of heroism, death, and statistics
2. **Causal Inference** – template for how to marshal evidence in support of a causal explanation
3. **Statistics & Instruction** – The data are simple but the analysis demonstrates multiple data analytic tools we use today
   - Snow cited as first instance of both Difference-in-Differences (Angrist & Pischke) and IV / randomization (Green 2018)
   - combining maps and data (GIS or geographic information systems)
   - regression and error analysis

Snow's cholera work is also a humbling reminder of the sometimes meandering path towards truth: even with overwhelming evidence and strong analysis Snow failed to convince the medical establishment, the public, or the authorities

Broadly, Three Strands or "Blocks" of Evidence

Snow marshaled evidence in 1855 & 1856 to convince skeptics

1. Isolated events, e.g. Albion Terrace
   - 1849, Discovery of waterborne theory
   - single event, 17 houses
2. Broad Street Outbreak
   - Aug-Sep 1854, 700 deaths over roughly 2 weeks, 10 square blocks
3. South London "Grand Experiment"
   - Summer & Fall 1854, customers supplied by two water companies
   - large scale, 400k mixed (quasi-random) subjects

Data or Evidence Blocks

Albion Terr	Broad St	South London
17 houses single outbreak	~10 sq blocks 2wks, 700 deaths	summer/fall 1854 ~400k subjects mixed treated & untreated

Preview: Hypotheses, Snow's Analysis, Modern Re-Analysis

Hypothesis	Challenge & Alternatives	Snow's Analysis	Problems	Modern Approach
Water Causal	Challenge – Confound Alternatives – Elevation – Crowding – Social class – Miasma – others	**1855 Table IX** – Direct comparison of mixed populations (quasi-randomized trial)	– 1855: all subdistricts together (no detailed population-by-supplier) – 1856: unbalanced	– Randomization limited to jointly-supplied (16) subdistricts – Regression & subdistrict FE control for balance – Count regression
		1855 p. 89 (Table XII) – Difference-in-Differences	– 1855: marginal significance (substantial heterogeneity)	Combine DiD & Prediction – DiD with continuous treatment effect (population fractions) – 3 suppliers – DiD controls for confounders – Residual Deviance & R^2 measures % explained – Count regression
Water Most Important	Challenges – Confound – "How important?" Alternatives – As above	**1856 Table VI** – Prediction from supplier population fractions	– Only 2 suppliers – No way to measure quality of fit – Ignores change 1849 to 1854	

Outline

Cholera – Disease of Poor Sanitation

What is Cholera?

- Vibrio Cholerae – bacterium that infects the small intestine of humans
- Causes severe diarrhea (& vomiting) that drains fluids
- Death from dehydration & organ failure
- Oral Rehydration Therapy highly succesfull (roughly 1960s)
  - In case you ever need it, here's the recipe – 1 liter boiled water, 1/2 teaspoon salt, 6 teaspoons sugar, mashed banana (potassium)

Cholera thrives in crowded cities with poor sanitation

- Transmitted through recycling (drinking) sewage
- When cholera exits one victim, needs to find a way into gut of others
- Victorian London was an ideal playground for cholera to thrive

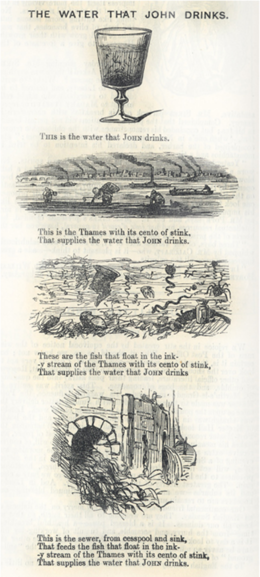
Cholera Loved Victorian London

Victorian London was an ideal playground for cholera

- Mid-1800s London was dirty, smelly place with no organized sewage treatment
- Efforts to improve sanitation made things worse
    - cesspools relatively safe – did not provide access to thousands of guts
- Public Health Act of 1848 accelerated the connection of houses to sewage lines
    - helped clean up streets, flushed filth to Thames
- By mid-1800s, cholera had easy access from the gut of one to thousands of victims

Contemporaries were aware of dirty water (*Punch* 1849)

- But water not recognized as vector for cholera



Solution – Construction of Bazalgette "Outfall Sewers"

Sewers that sloped towards outfalls (discharge points) lower on the Thames

- Construction started (under Bazalgette) 1859, response to 1858 "Great Stink"
- Embankments along Thames – what we see today
  - Embedded discharge pipes – still used today (?)
  - Decreased width, increased flow – scouring effect
- Moved sewage downstream, below London & water in-take

One final outbreak, 1866, limited to east London, last area unserved by sewers

John Snow's Research & Publications

Doctor – pioneer in anesthesia & medical hygiene

- Provided Queen Victoria with anesthesia during childbirth

Research and writing on Cholera

- 1849: "On the Mode of Communication of Cholera"
  - Laid out theory and evidence for waterborne transmission
- 1855: "On the Mode of Communication of Cholera"
  - Substantially expanded, additional evidence and argument
- 1856: "Cholera and the water supply in the south district of London in 1854"
  - Refined randomized analysis

John Snow's 1849 Theory & 1855 Evidence

1849: Snow developed theory of infection & transmission

- Based on medical knowledge and study of single events – Horsleydown & Albion Terrace

Fully-developed & modern theory of disease

- Infects & reproduces in the small intestine
- Exits from victim, into water supply
- Infects new victims through drinking dirty water

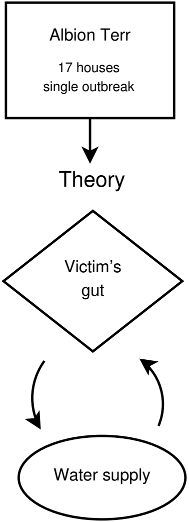
Implications for patterns of infection, across scales

- "from the membrane of the small intestine all the way up to the city itself" (Johnson)

Snow's work grounded by theory

> *Snow had a good idea – a causal theory about how the disease spread – that guided the gathering and assessment of evidence. (Tufte)*

1855: evidence & argument to convince skeptics



Alternative Theories

Miasma (Smells & Airborne)

- Cholera infectious & transmitted through the air
- Generally accepted in mid-1800s

Elevation, Crowding & Class, Others

- Elevation: lower elevation $\rightarrow$ more infection
- Crowding & Class: lower class & crowding $\rightarrow$ more infection

None of these absolutely crazy – correlated with cholera (and dirty water)

- Raw sewage associated with bad smells & dirty drinking water
- Lower class associated with crowding & poor sanitation

Other non-infectious theories (I won't seriously consider)

- Emanations from the ground
- Plague burying-pit near Broad Street pump

Modify Katz & Singer as "Causal Assessment Procedure"

Still tentative, based on Katz & Singer's analysis of possible Chemical & Biological Weapons attacks, 1970s-80s, "Can an Attribution Assessment Be Made for Yellow Rain?"

1. Divide evidence into blocks or types of evidence
2. Assign to each block a *veritas* rating – quality of data
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)
5. Organize evidence blocks by hypothesis into matrix
6. Choose hypothesis not contradicted
7. Strongest hypothesis checked

Theory, Data, Hypothesis Testing

Data or Evidence Blocks

Albion Terr
17 houses
single outbreak

Broad St
~10 sq blocks
2wks, 700 deaths

South London
summer/fall 1854
~400k subjects mixed
treated & untreated

Theory & Hypotheses

theory: water & small intestine

H1: water a causal factor

H2: water primary causal factor

alternatives: elevation, crowding, class, ...

Hypothesis or Testing Blocks

Albion Terr
- Narrative

Broad St
- Map
- Cases
- Contin

South London 1855
- Mixing
- DiD
- Act v Pred

modern re-analysis

- Mixing
- DiD w cont treatment eff

"Grand Experiment" – Water Supply Changes

Two water companies served south London – Southwark & Vauxhall Co and Lambeth Co. – 486,936 customers, 300,000 **intimately mixed**

- In 1830s & 1840s companies competed for customers, often on same street

  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)

1849 epidemic

- Both companies drew water from low in the Thames – near Vauxhall bridge

1852

- Lambeth Company moved source to Thames Ditton (upstream of London)
- In response to Act of Parliament, requiring move (by 1855)

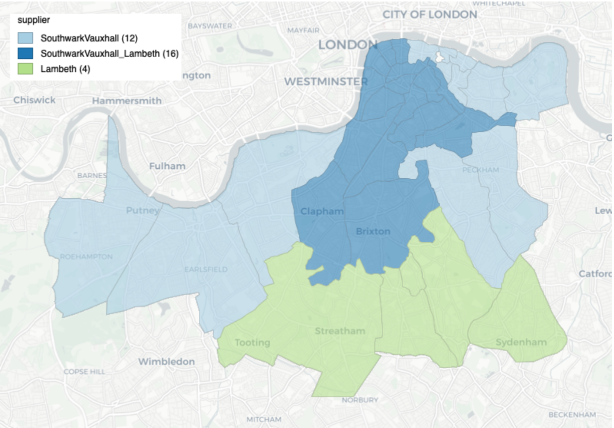
1854 epidemic

- Southwark & Vauxhall Co supplied dirty water
- Lambeth Co supplied cleaner water

32 Subdistricts, 12 S&V only, 16 joint, 4 Lambeth

Registration Districts & Sub-Districts – Need to keep straight

- Deaths collected weekly by Registrar-General, by District & Subdistrict
- In this region of South London, 32 sub-districts
- Snow viewed "mixed" as crucial experiment



- "First 12" – Southwark & Vauxhall Water Co only – dirty water 1849 & 1854
- "Next 16" Mixed or Joint – Southwark & Vauxhall Co *and* Lambeth Water Co – 1849 dirty water, 1854 part dirty (S&V) & part clean (Lambeth)
- "Final 4" – Lambeth Water Co only – not relevant, not supplied in 1849

Deaths: Combined (All Suppliers) vs Direct (By Supplier)

Data available in 1855

- Deaths (combined all suppliers) 1849 & 1854, full epidemic
- Population (combined all suppliers)
- Deaths by supplier, first 7 weeks of epidemic (collected by Snow)

Data available in 1856 (originally published by Simon)

- Population by supplier (only S&V shown here)

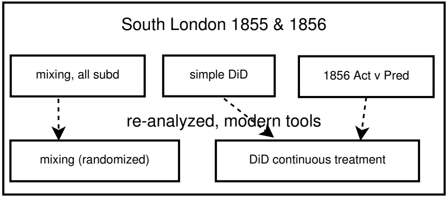
						1854, first 7 wks		
	subdistricts	Deaths 1849	Deaths 1854	Supplier	Population 1851	Deaths S&V	Deaths Lam	Pop S&V
1	St. Saviour	283	371	SV	19,709	115	0	16,337
2	St. Olave	157	161	SV	8,015	43	0	8,745
13	Christchurch	256	113	SV & Lambeth	16,022	11	13	2,915
14	Kent Road	267	174	SV & Lambeth	18,126	52	5	12,630
29	Norwood	2	10	Lambeth	3,977	0	2	0
	TOTAL	6,328	5,042		486,936	1,263	98	266,516

Combined (all suppliers)	Direct (by supplier)
$D_{subdist} = D_{S\&V} + D_{Lam} + D_{Other}$	$\{D_{S\&V}, D_{Lam}, D_{Other}\}$

Snow's Analysis – 3 Approaches

Mixed or quasi-random direct comparison

- Snow determined supplier – by bill or chloride test
- Visited all houses (deaths) for 7 weeks ending Aug 26



Diff-in-Diffs comparison of combined (all suppliers) mortality rates

- For each subdistrict, observe combined deaths all suppliers
- Compare 1849 vs 1854 and Treated (clean) vs untreated (dirty) subdistricts

Actual v Predicted across subdistricts for 1854 only

- Predicting mortality based on fraction of S&V vs Lambeth customers

Snow's Analysis – 3 Approaches

Hypothesis	Challenge & Alternatives	Snow's Analysis	Problems	Modern Approach
Water Causal	Challenge – Confound Alternatives – Elevation – Crowding – Social class – Miasma – others	**1855 Table IX** – Direct comparison of mixed populations (quasi-randomized trial)	– 1855: all subdistricts together (no detailed population-by-supplier) – 1856: unbalanced	– Randomization limited to jointly-supplied (16) subdistricts – Regression & subdistrict FE control for balance – Count regression
		1855 p. 89 (Table XII) – Difference-in-Differences	– 1855: marginal significance (substantial heterogeneity)	Combine DiD & Prediction – DiD with continuous treatment effect (population fractions) – 3 suppliers – DiD controls for confounders – Residual Deviance & R^2 measures % explained – Count regression
Water Most Important	Challenges – Confound – "How important?" Alternatives – As above	**1856 Table VI** – Prediction from supplier population fractions	– Only 2 suppliers – No way to measure quality of fit – Ignores change 1849 to 1854	

Snow Modern in View of Mixing (Randomization)

Recognized that mixing (randomization) would average out differences

> *As there is no difference whatever, either in the houses or the people receiving the supply of the two Water Companies, or in any of the physical conditions with which they are surrounded, it is obvious that no experiment could have been devised which would' more thoroughly test the effect of water supply on the progress of cholera than this. (1855 p. 75)*

Cited as first instance of Randomization and Instrumental Variables (Greene 2018, also Deaton, others)

Direct Comparison of Mixed or Randomized Population

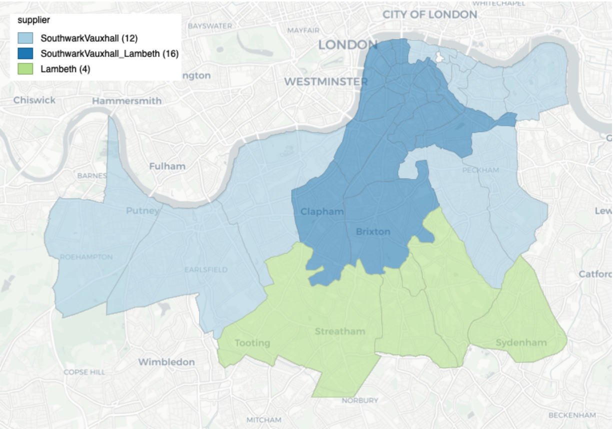
Table: Houses, Deaths, and Mortality Rates per 10,000 Households, First Seven Weeks of 1854 Cholera Epidemic – Table IX

Water Supplier	Number of houses	Deaths from Cholera	Deaths in each 10,000 houses
Southwark & Vauxhall Co supply	40,046	1,263	315.4
Lambeth Co supply	26,107	98	37.5
Rest of London	256,423	1,422	59
Ratio Effect: Southwark & Vauxhall vs Lambeth			8.40

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

- Found LARGE Lambeth effect
- But suffered from potential confounding – includes **all** subdistricts

Problem: Direct Comparison Uses *All* Subdistricts



- Snow wanted to use Joint (Mixed) subdistricts
- Population (houses) by supplier for overall region only
- Potential for confounding (bias if S&V-only subdistricts different than joint)

Second Approach – Diff-in-Diffs

Hypothesis	Challenge & Alternatives	Snow's Analysis	Problems	Modern Approach
Water Causal	Challenge – Confound Alternatives – Elevation – Crowding – Social class – Miasma – others	**1855 Table IX** – Direct comparison of mixed populations (quasi-randomized trial)	– 1855: all subdistricts together (no detailed population-by-supplier) – 1856: unbalanced	– Randomization limited to jointly-supplied (16) subdistricts – Regression & subdistrict FE control for balance – Count regression
		1855 p. 89 (Table XII) – Difference-in-Differences	– 1855: marginal significance (substantial heterogeneity)	Combine DiD & Prediction – DiD with continuous treatment effect (population fractions) – 3 suppliers – DiD controls for confounders – Residual Deviance & R^2 measures % explained – Count regression
Water Most Important	Challenges – Confound – "How important?" Alternatives – As above	**1856 Table VI** – Prediction from supplier population fractions	– Only 2 suppliers – No way to measure quality of fit – Ignores change 1849 to 1854	

Simple Diff-in-Diffs – Before vs After, Treated vs Control

Table: Mortality Rates 1849 & 1854, Summary Snow 1855 Table XII

	1849 Deaths per 10,000	1854 Deaths per 10,000	Ratio 1849 - 1854
Always Dirty – Southwark & Vauxhall Water Company Only ("First 12" subdistricts)	134.9 dirty, S&V only	146.6 dirty, S&V only	0.92 diff in time
Dirty / Clean – Joint Southwark & Vauxhall and Lambeth Companies ("Next 16" subdistricts)	130.1 dirty, joint	84.9 (partial) clean	1.53 diff in time & treatment
Ratio: Next 16 less First 12	0.96 diff in region	1.73 diff in region & treatment	1.67 (partial) treatment

Comparing the S&V-only subdistricts vs the Jointly-supplied subdistricts

- Interestingly, Snow did not convert deaths to rates – missed an opportunity
- Large treatment effect, but need to evaluate statistical significance

Problem: treatment effect only marginally significant

DiD as Regression

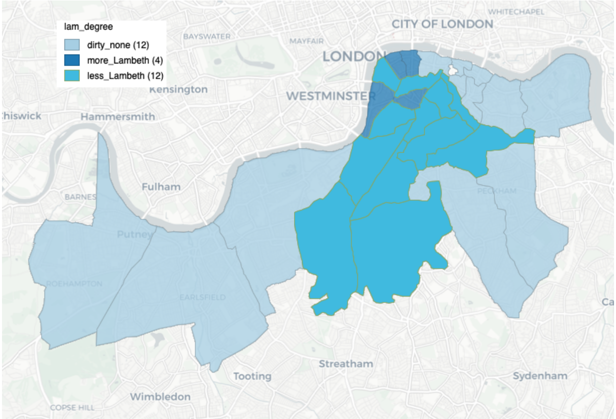
$$\ln\left(Rate_{subdist,yr}\right) = \ln\left(count_{subdist,yr}/population_{subdist,yr}\right) = \hat{\mu} + \hat{\delta}_{54} \cdot I_{yr=1854} + \hat{\gamma}_J \cdot I_{subdist=joint} + \hat{\beta} \cdot I_{subdist=joint} \cdot I_{yr=1854} + \varepsilon_{s,y}$$

Region or Sub-Districts – Supplied by	1849 Death Rate (log)	1854 Death Rate (log)	Diff 1854 less 1849
First 12 – Southwark Only	μ	$\mu + \delta_{54}$	δ_{54}
Next 16 – Joint Southwark and Lambeth	$\mu + \gamma_J$	$\mu + \delta_{54} + \beta + \gamma_J$	$\delta_{54} + \beta$
Diff Joint less Southwark	γ_J	$\beta + \gamma_J$	β

Regression framework allows us to

- Use subdistrict detail, and additional regressors (if available)
- Test for statistical significance (both for finite population and "within-sample" variation)
- Extend the DiD framework to continuous treatment and actual-vs-predicted

Snow Highlighted Difference in "Lambeth Degree"



- Four subdistricts where "the supply of the Lambeth Water Company is more general than elsewhere"

Snow Highlighted Difference in "Lambeth Degree"

	1849 Deaths per 10,000	1854 Deaths per 10,000	Ratio 1849 - 1854
Always Dirty – Southwark & Vauxhall Water Company Only ("First 12" subdistricts)	134.9 dirty, S&V only	146.6 dirty, S&V only	0.92 diff in time
Dirty / Clean – "More Lambeth" in Joint (4 subdistricts)	138.8 dirty, more	47.2 more clean	2.94 time & more
Dirty / Clean – "Less Lambeth" in Joint (12 subdistricts)	127.6 dirty, less	95.6 less clean	1.34 time & less
Ratio: "More Lambeth" vs Dirty	0.97 diff in region	3.11 region & more	3.20 more treatment
Ratio: "Less Lambeth" vs Dirty	1.06 diff in region	1.53 region & less	1.45 less treatment

Larger effect for "More Lambeth"

- Now, treatment effect is highly significant (see below)

Third Approach – in 1856, Actual-vs-Predicted

Hypothesis	Challenge & Alternatives	Snow's Analysis	Problems	Modern Approach
Water Causal	Challenge – Confound Alternatives – Elevation – Crowding – Social class – Miasma – others	**1855 Table IX** – Direct comparison of mixed populations (quasi-randomized trial)	– 1855: all subdistricts together (no detailed population-by-supplier) – 1856: unbalanced	– Randomization limited to jointly-supplied (16) subdistricts – Regression & subdistrict FE control for balance – Count regression
		1855 p. 89 (Table XII) – Difference-in-Differences	– 1855: marginal significance (substantial heterogeneity)	Combine DiD & Prediction – DiD with continuous treatment effect (population fractions) – 3 suppliers – DiD controls for confounders – Residual Deviance & R^2 measures % explained – Count regression
Water Most Important	Challenges – Confound – "How important?" Alternatives – As above	**1856 Table VI** – Prediction from supplier population fractions	– Only 2 suppliers – No way to measure quality of fit – Ignores change 1849 to 1854	

Actual-vs-Predicted by Subdistrict

	1	2	3	4	6	10
Subdistrict	Pop 1851	Population Estimates by Supplier South-wark	Lambeth	Both Cos	Actual Rate	Predicted Rate
Christchurch, Southwark	16,022	2,915	13,234	16,149	***70.5***	***51.0***
St. Saviour, Southwark	19,709	16,337	898	17,235	***191.8***	***153.1***
St. Olave, Southwark	8,015	8,745	0	8,745	***200.9***	***160.0***
St. John, Horsleydown	11,360	9,360	0	9,360	***133.8***	***160.0***
St. James, Bermondsey	18,899	23,173	693	23,866	***191.5***	***156.1***
...	...	...				

$$\hat{R}_{subdis,both} = \frac{N_{subdis,Southwark} \cdot \hat{R}_{Southwark} + N_{subdis,Lambeth} \cdot \hat{R}_{Lambeth}}{N_{subdis,Southwark} + N_{subdis,Lambeth}} = F_S \cdot \hat{R}_S + F_L \cdot \hat{R}_L$$

$\hat{R}_{Southwark}$, $\hat{R}_{Lambeth}$ *Overall* rate for supplier (average across all subdistricts) $\hat{R}_{Southwark} = 160$ and $\hat{R}_{Lambeth} = 27$

- Correlation is 0.75 ($R^2 = 56\%$), but Snow had no statistics, could only argue individual cases
- Problems: a) Ignores "Other" (pumps, wells); b) Population not accurate (see St. James)

Conditional Map – Highlights High S&V Mortality

Vertical: Fraction of population Lambeth

Horizontal: Fraction of population Southwark & Vauxhall

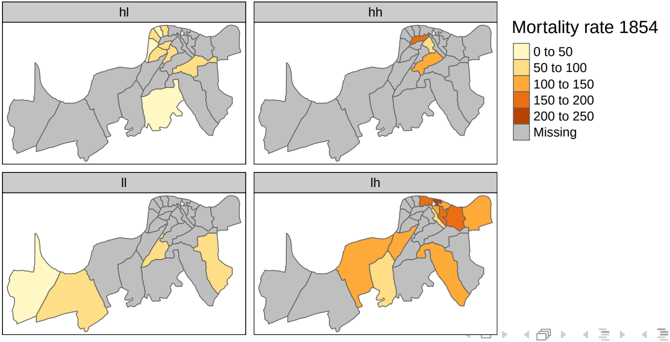
- Fraction Lambeth Low & High: Low mortality (along vertical)
- Fraction S&V Low: Low mortality; Fraction S&V high: High mortality (along horizontal)

Conditional Map – Highlights High S&V Mortality

Vertical: Fraction of population Lambeth

Horizontal: Fraction of population Southwark & Vauxhall

- Fraction Lambeth Low & High: Low mortality (along vertical)
- Fraction S&V Low: Low mortality; Fraction S&V high: High mortality (along horizontal)



Overview – Problems and Modern Re-Analysis

<table>
<tr><th>Hypothesis</th><th>Challenge & Alternatives</th><th>Snow's Analysis</th><th>Problems</th><th>Modern Approach</th></tr>
<tr><td rowspan="2">Water Causal</td><td rowspan="2">Challenge
– Confound
Alternatives
– Elevation
– Crowding
– Social class
– Miasma
– others</td><td>1855 Table IX
– Direct comparison of mixed populations (quasi-randomized trial)</td><td>– 1855: all subdistricts together (no detailed population-by-supplier)
– 1856: unbalanced</td><td>– Randomization limited to jointly-supplied (16) subdistricts
– Regression & subdistrict FE control for balance
– Count regression</td></tr>
<tr><td>1855 p. 89 (Table XII)
– Difference-in-Differences</td><td>– 1855: marginal significance (substantial heterogeneity)</td><td rowspan="2">Combine DiD & Prediction
– DiD with continuous treatment effect (population fractions)
– 3 suppliers
– DiD controls for confounders
– Residual Deviance & R^2 measures % explained
– Count regression</td></tr>
<tr><td>Water Most Important</td><td>Challenges
– Confound
– "How important?"
Alternatives
– As above</td><td>1856 Table VI
– Prediction from supplier population fractions</td><td>– Only 2 suppliers
– No way to measure quality of fit
– Ignores change 1849 to 1854</td></tr>
</table>

Population-by-Supplier & Direct Comparison (Mixed, Randomized)

Publication (in 1856) of population-by-supplier allowed direct comparison of mixed subdistricts

- Snow published the Actual-vs-Predicted, but not re-analysis of direct comparison (quasi-randomized) – not sure why
- Re-analysis shows same result – large Lambeth effect

Water Supplier	Number of houses	Deaths from Cholera	Deaths in each 10,000 houses	Approx Deaths per 10,000 persons		Population by Supplier	Deaths	Deaths per 10,000 persons
Southwark & Vauxhall Co supply	40,046	1,263	315.4	100.9	S&V only supply, "First 12"	119,603	738	132.2
					S&V only supply, "Next 16" joint	118,888	525	94.6
Lambeth Co supply	26,107	98	37.54	12.01	Lambeth supply, "Next 16" joint	147,961	94	13.6
Ratio Effect: S&V vs Lambeth			8.40	8.40	Ratio Effect, joint subdistricts			6.95

Randomization & Subdistrict Confounding

Still does not rule out confounding, because subdistricts vary and randomization is *within* but not *across* subdistricts

- Hypothetical Example: subregion 2 has low mortality and high proportion of Lambeth customers, produces overall low Lambeth mortality
- But the true effect of Lambeth (vs S&V) is small

	Southwark & Vauxhall			Lambeth			
	Pop	Deaths	Mortality Rate	Pop	Deaths	Mortality Rate	Ratio
Subregion 1	106,999	514	102.90	14,796	35	50.69	2.03
Subregion 2	11,889	11	19.83	133,165	59	9.49	2.09
Overall (total)	118,888	525	94.63	147,961	94	13.61	6.95
Ratio of Subregions			5.19			5.34	

Randomization as Averaging Over Confounders

Treatment effect is $\hat{\beta}_L$:

$$\ln\left(Rate_{supplier}\right) = \hat{\rho}_S + \hat{\beta}_L \cdot I_{supplier=Lam} + \varepsilon_{supplier}$$

Confounding: different unobservables for S&V vs Lambeth (correlated):
$E\left(\varepsilon_s \mid I_{s=L} = 0\right) \neq E\left(\varepsilon_s \mid I_{s=L} = 1\right) \neq 0$

- Violates standard regression conditions
- Randomization forces $E\left(\varepsilon_s \mid I_{s=L} = 0\right) = E\left(\varepsilon_s \mid I_{s=L} = 1\right) (\neq 0)$

Now, confounders drop out (the same for both S&V and Lambeth):

$$E\left(\ln\left(Rate_{supplier}\right) \mid I_{s=L} = 1\right) - E\left(\ln\left(Rate_{supplier}\right) \mid I_{s=L} = 0\right) = \hat{\beta}_L$$

BUT, averaging is within subdistricts, not across – still potential confounding

- Solution? Subdistrict Fixed Effects

Note in passing, randomization can also be viewed as Instrumental Variable (IV) that introduces variation orthogonal to observables and unobservables

- See Heckman (1996), Deaton (2018) for randomization as IV

Snow credited (by Greene, others) with first use of randomization as IV

Negative Binomial vs Poisson Model

We think about rates, but random variable is *count* (deaths)

$$R_{subdist,supplier} = count_{subdist,supplier} / population_{subdist,supplier}$$

First inclination to use Poisson model (counts Poisson)

- But does not fit the data

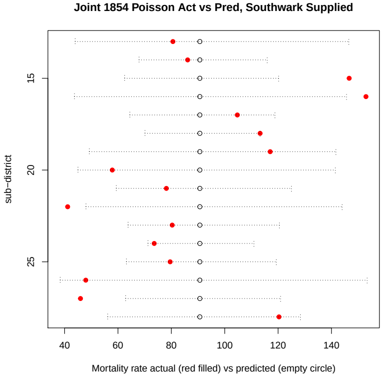
Explanation is simple

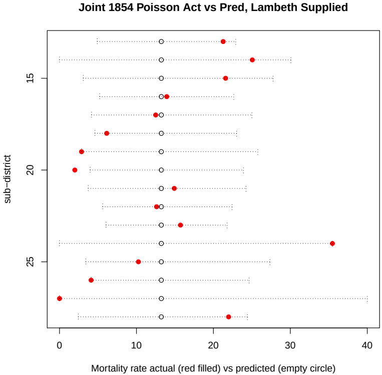
- Each subdistrict is Poisson (must be – approximation to sum of Bernoulli alive / dead)
- But random seeding of outbreak and propagation means that actual Poisson rate varies from subdistrict-to-subdistrict

My conclusion?

- Poisson is not the natural starting point
- Subdistrict is Poisson, but drawn from a distribution with some variance (distribution, not fixed number)
- Overall is mixture of Poissons – e.g. Negative Binomial as Gamma mixture of Poissons

Overdispersion – Fixed Poisson Does Not Fit





Mortality per 10,000 Separately for S&V and Lambeth Customers (for the First Seven Weeks, ending 26th August 1854). Actual and Predicted (with 95% confidence bands)

Note how many observations outside 95% bands: 5/32 (15%)

- Formal test: Residual Deviance (if low Resid Dev, high p-value, fail to reject)
- Use Negative Binomial count regression

Summary of Results

Table: Various Fitted Models

	1 Direct Southwark v Lambeth Joint Region 1855 Table VIII	2 Actual v Predicted DiD 1856 Table VI 1855 Table XII	3 DiD Single	4 DiD More Lambeth	5 DiD, Actual vs Predicted, Direct Southwark vs Lambeth
			1855 Table XII		
Raw Treatment	6.95		1.67	3.20	
Estimated Treat.	6.70	3.49	1.65	3.10	5.31
z-ratio	-12.04	-5.23	-2.03	-3.20	-7.08
p-value	2.3E-33	1.7E-07	4.21%	0.14%	1.5E-12
ResDev p-value	15.83%	14.58%	21.45%	15.74%	5.79%
Pseudo RSq	80.2%	69.3%	16.8%	25.1%	85.5%
Rate RSq	70.7%	72.1%	19.9%	24.5%	83.9%
No "Obs"	32	56	56	56	100
No Parameters	3	6	5	7	8
Notes	FE Poisson, p-value 2.5% FE $\hat{\beta}_L$=6.83, z=16				Treat Eff lower because 1849 Lam (1.50x); Late TE 3.1x

Conclusion from Mixed (Quasi-Randomized)

Strongly supports "Water as *a* causal factor"

- Strong Lambeth (treatment) effect (Lambeth 6.7-times lower, z-ratio 12.0)
- Evidence against confounding
  - Presumption that randomization averages over confounders within
  - Subdistrict Fixed Effects do not change result – reject confounding across

Randomization valuable, but randomization alone cannot answer all questions:

- Not strong evidence on "water as *primary* causal factor"
- No evidence on non-water differences between S&V vs Lambeth customers

For these, turn to DiD and comparison across subdistricts and across time

- Arguably "more difficult" test than mixed (randomized) direct comparison
- Example of the complementary nature of randomized and observational data

DiD Combined – Continuous Treatment

Simple (discrete treatment) DiD:

$$\ln\left(Rate_{subdist,yr}\right) = \hat{\mu} + \hat{\delta}_{54} \cdot I_{yr=1854} + \hat{\gamma}_J \cdot I_{subdist=joint} + \hat{\beta} \cdot I_{subdist=joint} \cdot I_{yr=1854} + \varepsilon_{s,y}$$

Region or Sub-Districts – Supplied by	1849 Death Rate (log)	1854 Death Rate (log)	Diff 1854 less 1849
First 12 – Southwark Only	μ	$\mu + \delta_{54}$	δ_{54}
Next 16 – Joint Southwark and Lambeth	$\mu + \gamma_J$	$\mu + \delta_{54} + \beta + \gamma_J$	$\delta_{54} + \beta$
Diff Joint less Southwark	γ_J	$\beta + \gamma_J$	β

But the treatment effect can be written as continuous based on fraction Lambeth:

$$\ln\left(Rate_{subdist,yr}\right) = \hat{\rho}_S + \hat{\delta}_{54} \cdot I_{yr=1854} + \hat{\rho}_L \cdot F_L + \hat{\rho}_O \cdot F_O + \hat{\beta}_L \cdot F_L \cdot I_{yr=1854} + \varepsilon_{s,y}$$

This now combines DiD *and* an extended Actual-vs-Predicted

- Act-v-Pred incorporates 3 suppliers: S&V, Lambeth, "Other"
- Underlying mortality rates ($\hat{\rho}_S, \ldots$) estimated from data to minimize SSQ
- Compares Act-v-Pred across subdistricts *and* time, comparing S&V vs Lambeth in 1849

DiD Continuous Treatment Tests Confounding & Importance

Continuous DiD allows testing for importance of treatment, as well as testing (and controlling for) confounding

$$\ln\left(Rate_{subdist,yr}\right) = \hat{\rho}_S + \hat{\delta}_{54} \cdot I_{yr=1854} + \hat{\rho}_L \cdot F_L + \hat{\rho}_O \cdot F_O + \hat{\beta}_L \cdot F_L \cdot I_{yr=1854} + \varepsilon_{s,y}$$

- $\hat{\rho}_L$ controls for non-water differences between S&V and Lambeth in 1849 (confounding)
- $\hat{\beta}_L$ measures treatment effect, purged of non-water differences ($\hat{\rho}_L$)
- R^2 (coefficient of determination for rates) measures how well prediction explains observations

Examining Hypothesis 2: Water as Primary Factor

$$\ln\left(Rate_{subdist,yr}\right) = \hat{\rho}_S + \hat{\delta}_{54} \cdot I_{yr=1854} + \hat{\rho}_L \cdot F_L + \hat{\rho}_O \cdot F_O + \hat{\beta}_L \cdot F_L \cdot I_{yr=1854} + \varepsilon_{s,y}$$

Hypothesis:	Base	Alt 1	Alt 2	Alt 3	Alt 4a	Alt 4b
	Water + Treatment Effect	Water No Treatment Effect	No Water, Subdistricts Only	Water + Treatment + Subdistricts	Water + Treatment + Housing Density	Housing Density Alone
Treat ratio	3.49	–	–	3.76	3.49	–
Res Dev p	14.6%	22.4%	0.08%	0.09%	12.3%	25.4%
Rate R^2	72.1%	56.9%	76.1%	89.7%	72.7%	24.5%
BIC	614.5	631.9	692.8	635.0	617.6	665.0
		remove $\hat{\beta}_L$	remove $\hat{\beta}_L$, $\hat{\rho}_S$, include FE	include FE	include density	density only

Conclusion: water matters, explains large proportion of rate variance, other factors (housing density, all subdistrict fixed characteristics) do not matter

Review so Far: Support Water, Reject Others

<table>
<tr><th>Hypothesis</th><th>Challenge & Alternatives</th><th>Snow’s Analysis</th><th>Problems</th><th>Modern Approach</th></tr>
<tr><td rowspan="2">Water Causal</td><td rowspan="2">Challenge
– Confound
Alternatives
– Elevation
– Crowding
– Social class
– Miasma
– others</td><td>1855 Table IX
– Direct comparison of mixed populations (quasi-randomized trial)</td><td>– 1855: all subdistricts together (no detailed population-by-supplier)
– 1856: unbalanced</td><td>– Randomization limited to jointly-supplied (16) subdistricts
– Regression & subdistrict FE control for balance
– Count regression</td></tr>
<tr><td>1855 p. 89 (Table XII)
– Difference-in-Differences</td><td>– 1855: marginal significance (substantial heterogeneity)</td><td rowspan="2">Combine DiD & Prediction
– DiD with continuous treatment effect (population fractions)
– 3 suppliers
– DiD controls for confounders
– Residual Deviance & R^2 measures % explained
– Count regression</td></tr>
<tr><td>Water Most Important</td><td>Challenges
– Confound
– “How important?”
Alternatives
– As above</td><td>1856 Table VI
– Prediction from supplier population fractions</td><td>– Only 2 suppliers
– No way to measure quality of fit
– Ignores change 1849 to 1854</td></tr>
</table>

Summary of Results

Table: Various Fitted Models

	1 Direct Southwark v Lambeth Joint Region 1855 Table VIII	2 Actual v Predicted DiD 1856 Table VI 1855 Table XII	3 DiD Single 1855 Table XII	4 DiD More Lambeth	5 DiD, Actual vs Predicted, Direct Southwark vs Lambeth
Raw Treatment	6.95		1.67	3.20	
Estimated Treat.	6.70	3.49	1.65	3.10	5.31
z-ratio	-12.04	-5.23	-2.03	-3.20	-7.08
p-value	2.3E-33	1.7E-07	4.21%	0.14%	1.5E-12
ResDev p-value	15.83%	14.58%	21.45%	15.74%	5.79%
Pseudo RSq	80.2%	69.3%	16.8%	25.1%	85.5%
Rate RSq	70.7%	72.1%	19.9%	24.5%	83.9%
No "Obs"	32	56	56	56	100
No Parameters	3	6	5	7	8
Notes	FE Poisson, p-value 2.5% FE $\hat{\beta}_L$=6.83, z=16				Treat Eff lower because 1849 Lam (1.50x); Late TE 3.1x

Summary

Direct (by supplier) Comparison using $\{D_{S\&V}, D_{Lam}, D_{Other}\}$

- Limited to joint subdistricts, same finding as Snow Table IX
- re Randomization, reject subdistrict confounding
- Strongly support water as a causal factor (not ruling out others)

Actual v Predicted DiD using $D_{subdist} = D_{S\&V} + D_{Lam} + D_{Other}$

- Combines Snow's idea of Actual vs Predicted and DiD
- Arguably tougher test than direct (compare 1849 vs 1854)
- Strongly supports water as *a* causal factor and water as *primary* causal factor

Simple DiD

- Using only data from 1855, still strongly supports water as a causal factor
- Statistics: shows importance of *correctly* measuring SEs, and NegBinom over Poisson

Combined & Direct using $D_{subdist} = D_{S\&V} + D_{Lam} + D_{Other}$ and $\{D_{S\&V}, D_{Lam}, D_{Other}\}$

- Reinforces other results
- Shows difference in early vs late part of epidemic

Supporting and Extending David Freedman's Comments

This detailed analysis of Snow's work supports Freedman's (1991) comments about Snow:

> *Snow's work is ... a success story for scientific reasoning based on nonexperimental data*
> *statistical technique can seldom be an adequate substitute for good design, relevant data, and testing predictions against reality in a variety of settings,*

But it modifies Freedman's skepticism about statistical arguments

> *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.*

to a more nuanced view: Snow's work proves the importance of marrying good design with good statistical analysis

Conclusion: Theory, Data, Hypothesis Testing

Data or Evidence Blocks

- **Albion Terr**: 17 houses, single outbreak
- **Broad St**: ~10 sq blocks, 2wks, 700 deaths
- **South London**: summer/fall 1854, ~400k subjects mixed, treated & untreated

Theory & Hypotheses

- water & small intestine
- miasma (airborne)
- elevation, class, ...

Hypothesis or Testing Blocks

- **Albion Terr**
  - Narrative
- **Broad St**
  - Map
  - Cases
  - Contin
- **South London**
  - Diff-in-Diffs
  - Mixing
  - No sub-district pop
  - With sub-district pop

Still Much to Learn From John Snow

This work (joint with Peter Vinten-Johansen) is to re-examine and reconsider Snow's analysis of South London

Why Snow? Three reasons:

1. **Rollicking Good Tale** – full of heroism, death, and statistics
2. **Causal Inference** – template for how to marshal evidence in support of a causal explanation
3. **Statistics & Instruction** – The data are simple but the analysis demonstrates multiple data analytic tools we use today
    - Snow cited as first instance of both Difference-in-Differences (Angrist & Pischke) and IV / randomization (Green 2018)
    - combining maps and data (GIS or geographic information systems)
    - regression and error analysis

Snow's cholera work is also a humbling reminder of the sometimes meandering path towards truth: even with overwhelming evidence and strong analysis Snow failed to convince the medical establishment, the public, or the authorities

Mortality Rates from Snow Table XII

	Sub-Districts	1849 per 10,000	1854 per 10,000	Water Supplier
1	St. Saviour, Southwark	144	188	SouthwarkVauxhall
2	St. Olave, Southwark	196	201	SouthwarkVauxhall
3	St. John, Horsleydown	169	130	SouthwarkVauxhall
4	St. James, Bermondsey	132	192	SouthwarkVauxhall
5	St. Mary Magdalen	186	175	SouthwarkVauxhall
6	Leather Market	148	155	SouthwarkVauxhall
7	Rotherhithe	198	158	SouthwarkVauxhall
8	Battersea	92	56	SouthwarkVauxhall
9	Wandsworth	115	178	SouthwarkVauxhall
10	Putney	15	17	SouthwarkVauxhall
11	Camberwell	132	135	SouthwarkVauxhall
12	Peckham	47	89	SouthwarkVauxhall
13	Christchurch, Southwark	160	71	Southwark&Lambeth
14	Kent Road	147	96	Southwark&Lambeth
15	Borough Road	197	170	Southwark&Lambeth
16	London Road	144	52	Southwark&Lambeth
17	Trinity, Newington	152	100	Southwark&Lambeth

Mortality Rates from Snow Table XII

	Sub-Districts	1849 per 10,000	1854 per 10,000	Water Supplier
18	St. Peter, Walworth	149	130	Southwark&Lambeth
19	St. Mary, Newington	102	66	Southwark&Lambeth
20	Waterloo Road (1st)	137	41	Southwark&Lambeth
21	Waterloo Road (2nd)	132	64	Southwark&Lambeth
22	Lambeth Church (1st)	117	27	Southwark&Lambeth
23	Lambeth Church (2nd)	203	72	Southwark&Lambeth
24	Kennington (1st)	77	125	Southwark&Lambeth
25	Kennington (2nd)	81	75	Southwark&Lambeth
26	Brixton	55	33	Southwark&Lambeth
27	Clapham	70	101	Southwark&Lambeth
28	St. George, Camberwell	111	83	Southwark&Lambeth
29	Norwood	5	25	Lambeth
30	Streatham	171	17	Lambeth
31	Dulwich	6	0	Lambeth
32	Sydenham	11	27	Lambeth
	First 12 sub-districts	135	147	first12
	Next 16 sub-districts	130	85	next16
	Last 4 sub-districts	85	19	last4
	TOTAL	130	104	