Rethinking John Snow's South London study: A Bayesian evaluation and recalculation

Thomas Koch[*], Kenneth Denike

University British Columbia, Vancouver BC, Canada V6K 2S1

Abstract

Famously, John Snow attempted to convince a critical professional audience that public water supplied to South London residents by private companies was a principal vector for the transmission of cholera. The result has been called the sine qua non of the "epidemiological imagination," a landmark study still taught today. In fact, Snow twice attempted to prove public water supplies spread cholera to the South London population. His first, published in 1855, suffered from an incomplete data set that limited its descriptive and predictive import. In 1856, armed with new data, Snow published a more definitive study. 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. The result hopefully revitalizes Snow's research as a teaching case through the application of a contemporary statistical approach.
© 2006 Elsevier Ltd. All rights reserved.

Keywords: Cholera; Empirical Bayes estimation; Medical history; Modifiable area unit problem; John snow; Small area unit problem

Introduction

Famously, John Snow attempted to convince a critical professional audience that public water supplied to South London residents by private companies was a principal vector for the transmission of cholera. He did this in two separate studies, the first of a neighborhood cholera outbreak in the Broad Street area of Soho (Snow, 1855, pp. 38–56; Smith, 2002) and the second a larger scale study of cholera in South London (Snow, 1855, pp. 71–92). In the latter Snow attempted to assign variable rates of cholera in the population to private water

companies serving central South London (the Southwark–Vauxhall and Lambeth Waterworks companies) to prove cholera was far more intense in one company's service area than the others.[1] The result is the sine quo non of what Ashton (1974) called "the epidemiological imagination." and a textbook example for generations of students (Vandenbroucke, Rooda, & Beukers, 1991; pp. 971–972) in epidemiology (Rothman, 2002), medical geography (Melnick, 2002; Robinson

---

[1]William Farr, 1852 was the first to attempt to compare mortality rates from cholera on the basis of South London water company service areas at the scale of the registration district. John Simon (1856), a physician at the London Board of Health, was equally aware of the unique qualities of the area and its potential for experimentation (Eyler, 1973, pp. 117–118).

*Corresponding author. Tel.: +1 604 714 0348; fax: +1 604 822 6150.

E-mail address: tomkoch@shaw.ca (T. Koch).

1982, p. 179) and public health (Carvalho, Lima, & Kriebel, 2004).

Geographers have focused their attention on Snow's Broad Street neighborhood study and, McLeod argues, its "dot map that makes him a hero in medical geography" (Koch, 2005, chapter 6; McLeod, 2000, p. 923). For epidemiologists and public health officials, as well as some historical cartographers (Robinson 1982, p. 179), it was the South London study (and its map) that deserved heroic status. In this study, Snow faced two daunting technical problems in an attempt to prove a causal relationship between water supply and cholera. First, his data set of deaths due to cholera, compiled from field records reported by local registrars and collected by London's General Registrar Office (GRO), were incomplete. Second, available population figures for each water company were not easily transposed to the scale of the registration district or sub-district in a manner that would permit precise spatial assignment. The result inhibited even a general calculation of local mortality ratios on the basis of water supply areas and prevented completion of the natural experiment Snow promised in his landmark work, the second edition of *On the Mode of Communication of Cholera* (Snow, 1855), hereafter referred to as MCC-2.

In 1856, armed with additional data (Simon, 1856), 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 to water companies according to the best estimate then available. Snow then attempted to improve his finding's resolution by applying the resulting mortality rates at the sub-district level. The result, he boasted, "supplies a greater amount of statistical evidence than was ever brought to bear on a medical subject" (Snow, 1856, p. 248).

We demonstrate that in this second attempt Snow made not merely minor arithmetic errors but more importantly critical, conceptual mistakes that adversely affected his results. While his findings appear to conform to mortality ratios based on the GRO's records of cholera deaths—a signal proof of accuracy for Snow (Snow, 1856, p. 10)—a comparison of variance for the two distributions is unimpressive, and in a few sub-districts Snow's calculated mortality is wildly different from those

actually reported (Vinten-Johansen, Brody, Paneth, Rachman, & Rip, 2003, pp. 275–276).

The result is not merely a statistician's quibble. Snow's goal in the 1856 paper was to define a statistical process that would serve to predict the incidence of cholera on the basis of local water supply. His intent was to prove through this methodology that cholera was a waterborne disease. We demonstrate here that Snow's statistical process, one central to then evolving disease studies, was flawed. Both identifying its problems and demonstrating their correctives serves both an historical evaluation of Snow's work and contemporary studies in which predictive statistical models, of which Snow's was an early example, are frequently employed.

Here we first correct for Snow's arithmetic errors and then employ an empirical Bayes estimation (EBE) (Balsted, 2004), "method of moments" approach whose a priori perspective offers a critical corrective to Snow's system of calculation (Bailey & Gatrell, 1995, pp. 303–307; Martuzzi & Elliott, 1996). We then reconstruct Snow's findings, demonstrating a more robust comparison of variance that shrinks unacceptably large differences between Snow's conclusions and the GRO's mortality records in individual registration sub-districts.

The result offers important insights into both Snow's thinking and the limits of the calculations he presented in his works. In addition, it appears to present an unusually clear example of the benefits of the EBE approach to a class of "modifiable area unit problems" (Openshaw, 1984) in which area unit size employed in an analysis, and the relationship between different area units, effect the result (Cromley & McLafferty, 2002, p. 110).

Cholera

Cholera is a bacterial disease causing intense diarrhea that has swept the globe in a series of global pandemics ending with the recent, sixth pandemic, in the latter half of the twentieth century (CDC, 2000). The first outbreak began in India early in the 19th century and spread to England in 1831 in the first of four 19th century epidemics occurring in 1831–1833, 1848–1850, 1853–1854, and 1866 (Morris, 1976, p. 23). In part as a result of cholera's mortality rate of between 20 and 25 percent (Morris, 1976, p. 13), the nature of cholera—was it air or waterborne—and the means of its diffusion were subjects of intense professional

debate (for a review see Koch, 2005). Indeed, it is no exaggeration to state that in the second-half of the 19th century, cholera was the clinical focus of different theories of disease generation and diffusion. The problem, as an analysis of Snow's study of cholera in London in the 1850s makes clear, was conceptual and statistical at once.

Grand experiment I: 1855

In 1852 William Farr, then chief statistician of the GRO in London, published a detailed 400-page study of cholera in the years 1848–1849 that concluded public water supplies contributed to the spread of what Farr believed was a fundamentally airborne disease (Eyler, 1979; Farr, 1852). While he believed water contributed to the disease's diffusion Farr also argued that it would be difficult or impossible to prove its complicity at the scale of Metropolitan London where many private companies competing for customers, often at the level of the neighborhood street. In the second edition of MCC-2, Snow promised a grand experiment based on data from the 1854 London epidemic that would "thoroughly test the effect of water supply on the progress of cholera" (Snow, 1855, p. 75).

Using records collected by Farr's GRO registrars, and his own research, Snow developed a list of cholera deaths in 1854 in central South London supplied by two competing companies, South-wark–Vauxhall Company and the Lambeth Waterworks Company (Snow, 1855, p. 76). This data, which was to serve as numerator in his calculation of mortality rates, was based on data not only from registrars reporting to Farr but also on Snow's investigation of a limited set of sub-districts in the early weeks of the epidemic (Snow, 1854). To be useful, however, it needed a denominator based on service populations for the two companies. "All that

was required," he 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. While Snow was able to construct precise mortality ratios at these levels for earlier epidemics, for example, those of 1849 and 1853, he could not do the same for the 1854 epidemic that was his focus (Snow, 1855, p. 73). "I was unable at the time to show the relation between the supply of houses in which fatal attacks took place and the entire supply of each district and subdistrict [sic], on account of the latter circumstance not being known" (Snow, 1856, p. 7).

What Snow did have was a return to Parliament by water suppliers reporting the total number of houses they respectively supplied in Metropolitan London (Snow, 1855, p. 72). But because these returns did not specify the location of those houses Snow had only a *general* total of South London households supplied by each of the water companies as the denominator for the mortality ratios he sought to construct (Rothman, 2002, p. 61). Snow therefore could not link the homes of cholera victims to the water service areas. Therefore, while he could show a remarkable difference in mortality between Southwark–Vauxhall Company and Lambeth Company service populations he could not locate precise mortality ratios based on available data. Still, the result of his 1855 study is perhaps the most frequently reproduced table (Fig. 1) in epidemiology and public health (Carvalho et al., 2004), and with Snow's smaller scale Broad Street study, the basis for Snow's enduring fame (Vinten-Johansen et al., 2002, pp. 392–396).

While suggestive, the results were considered definitive neither by Snow nor his critics. They were too coarse to permit precise water supply assignments to homes of decedents, preventing the

	Number of houses	Deaths from Cholera
Southwark-Vauxhall Company	40,046	315
Lambeth Company	26,107	37
Rest of London	256,423	59

Snow, J. 1855. Table IX. *On the Mode of Communication of Cholera*, 86).

Fig. 1. Proportion of deaths to 10,000 houses during the first 7 weeks of the 1854 epidemic comparing two South London water supply companies and the mortality by household in greater London.

construction of cholera mortality ratios in the registration sub-districts of South London where "the mixing of the [water] supply is of the most intimate kind. The pipes of each company go down all the street, and into nearly all the courts and alleys" (Snow, 1855, p. 74). Snow lacked a mechanism to apply his analysis at a scale or resolution permitting the transformation of gross mortality into precise mortality ratios based on population for the two water companies.

Snow's map in MCC-2, best seen today in a digital version online (http://www.ph.ucla.edu/epi/snow/snowmap2_1854.html) presents the problem in graphic terms. While he could delineate the general service areas of the two water companies, and the area they served jointly, he could not join the mortality of the registration district and sub-district populations occurring within those jurisdictions in the map. Simply, there was no way to assign cholera mortality on the basis of population in the
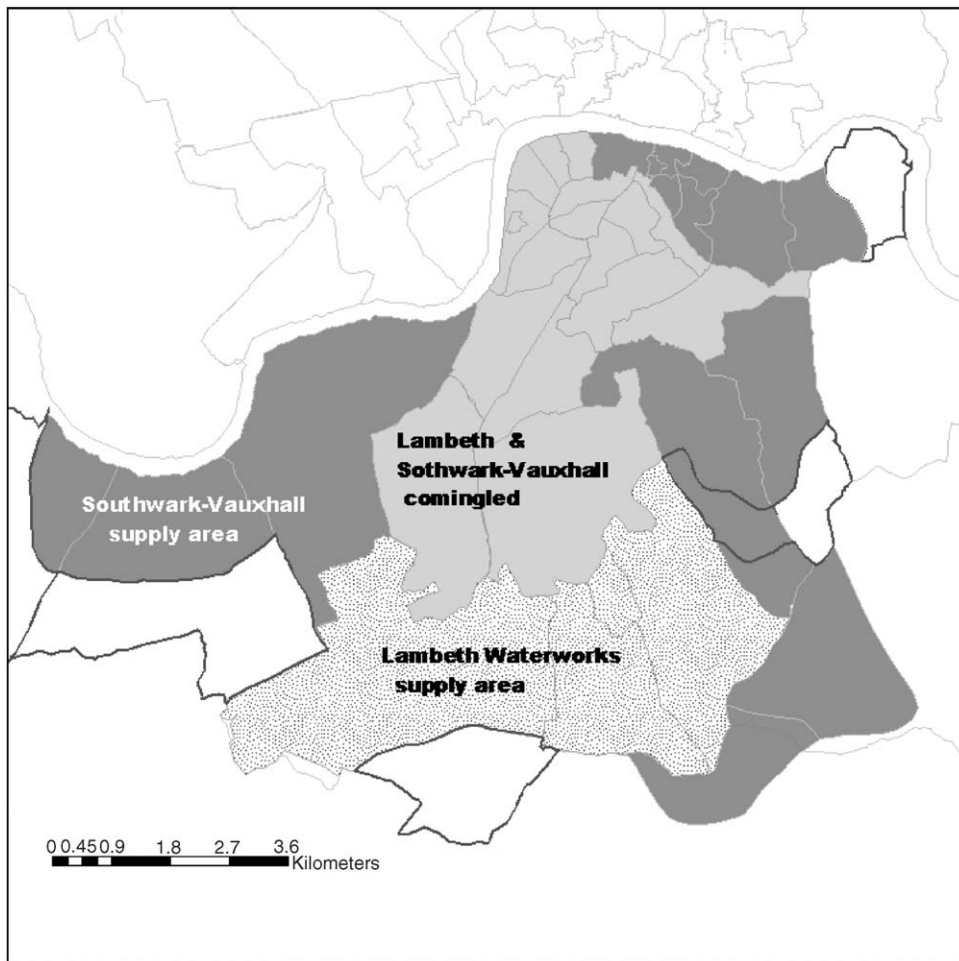
Water Company Supply Areas

Lambeth Waterworks Company	Snow's 1855 water supply areas based on South London registration sub-district data (Table VI). Registration disctricts outside the study area are left blank.
S-V and Lambeth Companies	
Southwark-Vauxhall Company	
Registration districts	

Fig. 2. A contemporary rendering of the water supply areas distinguished in Snow's map. Without data on population per registration district intensity of cholera could not be precisely located. Map by authors.

large area in which both companies vied for customers. Without that he could not precisely allocate deaths to either water company or effectively calculate the mortality ratios for registration districts or sub-districts, jurisdictional levels at which mortality data was recorded by the GRO. His map, therefore, was of service areas but not of cholera incidence. Snow therefore lacked a definitive argument—cartographic or statistical—that water was the principal variable explaining the wide variation in mortality ratios between water suppliers, a point not lost on either his contemporary critics (for example, Parks, 1855; Farr, 1853) or Snow's later admirers (Frost, 1936, p. 179) (Fig. 2).

Grand experiment II: 1856

In 1856, London Board of Health medical officer John Simon published a paper that provided the necessary data to translate Snow's remarkable difference in mortality between customers of the two companies into comparable *rates* of mortality among their customers at a relatively fine resolution. Simply, registration district and registration sub-district populations developed by the GRO were added to the picture, permitting a denominator to be added and precise mortality ratios to be constructed for the two water companies. "In the Report on the Cholera Epidemics of London as affected by the Consumption of Impure Water, lately written by Mr. Simon, and published by the General Board of Health, there is a statement of the number of houses supplied by each of the water companies respectively in each district and sub-district" (Snow, 1856, p. 7). The former were areas for which civil statistics (births and deaths) were by law reported (Registration of Births, 1835) on a weekly basis to the GRO while the latter were registration district sub-divisions, each with a single registrar charged with collection of birth and mortality data (Eyler, 1979, p. 43).

Simon effectively completed Snow's grand experiment, calculating mortality ratios of 13 deaths per 10,000 persons in the Southwark–Vauxhall service area compared to 3.7 per 10,000 persons for Lambeth Water Company customers in the 1853–1854 South London epidemic. "Of the 3476 tenants of the Southwark and Vauxhall Company who died of cholera in 1853–1854, two-thirds would have escaped if their water supply had been like their neighbors" he concluded, and "of the much larger number—tenants of both companies—who

died in 1848–1849, also two-thirds would have escaped…" (Simon, 1856, p. 9). Not to be outdone, Snow then used Simon's data (after correcting what he believed were jurisdictional errors (Snow, 1856, p. 7) to construct a statistical model of mortality at first registration district and then registration sub-district levels that would generate mortality results similar to those that actually occurred and were reported by local registrars to the GRO (Vinten-Johansen et al., 2003, p. 274). Snow sought in this way a statistical argument at least as compelling as one earlier constructed by Farr (1852) to demonstrate a clear, inverse relationship between increasing altitude and decreasing rates of cholera per 10,000 persons.

Snow: registration districts

The heart of Snow's approach, and the problems inherent in it, were distilled in his Table V (Fig. 3). In it (Snow, 1856, p. 17) Snow attempted first to calculate the number of deaths for each of ten registration districts for each of the two water suppliers, and then to calculate their respective mortality per 10,000 persons in those registration districts, based on 1851 population data, in a manner that returned a general mortality ratio for each of the two water supply areas. To do this Snow divided the water supply of all houses in each registration district in which fatal attacks of cholera occurred by the estimated population of each district. He then multiplied the result by 10,000.

Thus for homes supplied by Southwark–Vauxhall Company in St. Savior, Southwark, registration district, Snow divided 406 (houses in which fatal attacks occurred) by the estimated registration district population served by the company (19,617) to return a mortality ratio, after multiplying by 10,000 of 207 (206.963). Finally, Snow calculated general mortality rates from cholera for both water companies (160 per 10,000 and 27 per 10,000) by the simple, global expedient of dividing total deaths per water company by total estimated population for each service area. These ratios are given in the table in the final row in the two columns for mortality per 10,000 persons per water supply area.

A problem in this approach was how Snow addressed data lacking spatial assignment: 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. Snow assigned them on the basis of an a priori

Registration Districts	Number of inhabited houses in 1851	Population in 1851	Estimated constant population per house	"Number of houses and estimated number of persons supplied in 1854 with water as under"				Water supply of houses in which fatal attacks cholera took place				Deaths from cholera in the epidemic of 1854	Mortality per 10,000 supplied with water	
				By the Southwark and Vauxhall Co.		By the Lambeth Company		Southwark and Vauxhall Co.	Lambeth Co.	Pumps, wells, & other sources	Supply not ascertained		Southwark and Vauxhall Co.	Lambeth Co.
				No. of houses	Estimated population	No. of houses	Estimated population							
St. Saviour, Southwark..................	4,000	35,731	7.8	2,631	19,617	1,689	14,201	406	72	10	3	491	207	50
St. Olave, Southwark.....................	2,960	19,735	8.2	2,193	18,638	0	0	277	0	8	28	313	148	-
Bermondsey.................................	7,007	48,128	6.9	8,402	57,884	268	1,785	821	0	25	0	846	142	-
St. George, Southwark..................	6,992	51,824	7.4	3,419	25,039	3,183	23,712	388	99	0	56	543	155	41
Newington..................................	10,458	64,816	6.2	5,224	31,940	5,473	33,531	458	58	2	176	694	143	17
Lamberth....................................	20,447	139,325	6.8	8,077	54,982	11,763	83,786	525	138	24	240	927	96	16
Wandsworth................................	8,276	50,764	6.1	3,028	18,390	618	3,870	268	7	106	40	421	145	18
Camberwell.................................	9,412	54,607	5.8	4,005	23,472	1,835	10,478	352	33	115	49	549	150	31
Rotherhithe.................................	2,792	17,805	6.4	2,336	14,951	0	0	207	0	46	30	283	138	-
Greenwich & sub-dis. Sydenham...	-	-	-	-	-	-	-	4	4	2	1	11	-	-
Houses not identified...................	-	-	6.6	411	2712	25	165	-	-	-	-	-	-	-
Totals.................................	72,344	482,435	0.7	39,726	267,625	24,854	171,528	37,706	411	338	623	5,078	138	23
Non-ascertained cases distributed in proportion to others.................	-	-	-	-	-	-	-	561	62	-	-	-	-	-
Population (Registrar-General)......	-	-	-	-	266,516	-	173,748	4,267	473	338	-	5,078	160	27

Fig. 3. Snow's Table V calculates mortality per 10,000 persons during the 1854 cholera epidemic for South London residents in 10 registrations districts based upon water supplier (1856, p. 16).

summary location parameter, the global mean ratio of deaths per company. "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, in order to obtain a complete view of the influence of the water supply during the whole epidemic of 1854" (Snow, 1856, p. 9).

While Snow's strong conviction that unassigned cases mirrored general pattern of cases was reasonable a priori, the variations in distribution argue for smoothing the allocation of unassigned deaths according to occurrence by district and company. Snow might equally have assumed that unassigned cases reflected mistakes by individual registrars and that the non-ascertained cases would be better distributed by local means by district. Further, any and all variations in density of homes per district (6.4–7.8 persons per house), location, population (17,805–140,000 persons), and socio-economic status of the registration district might argue for the eccentric assignment of these cases.

No less critically, Snow's final figures of relative morality by water supplier for each registration district estimated directly from a priori through his reflexive use of a global mean, 160 deaths per 10,000 for Southwark–Vauxhall and 27 deaths per 10,000 for Lambeth Waterworks Company. This in effect negated, or at least diminished the resolution

returned through his district-by-district analysis. These problems make Snow's mortality ratios suspect and their application to the registration sub-district level problematic.

Registration sub-districts

Snow then attempted to improve the resolution of his findings by transposing his conclusions from the level of the registration district to that of the registration districts' 31 constituent sub-districts, the scale required if the results were to be precisely mapped.[2] In this manner, a moveable area unit problem was fully engaged. The effect of Snow's a priori assumptions calculated at the level of the registration district in his Table V (here Fig. 3) thus were folded into his final calculations at the level of the 31 sub-districts in his Table VI (here Fig. 4), the level at which his model was to be completed.

Minor arithmetic errors with a real but minimal effect on the table's results can be found in Snow's calculations at the level of both registration district and sub-district analysis, not surprising in the work

---

[2] The construction of Snow's map of the South London study is one that has never been fully explained. A careful examination suggests the boundaries of the water supplier areas are defined by registration sub-district rather than registration district data. A definitive proof of this, and a detailed analysis of the map, is forthcoming in a subsequent publication.

Registration Districts	Registration Sub-Districts	Population in 1851	Estimated population supplied with water			Deaths from cholera in 1854		Calculated mortality in the population, supplied with water			
			Southwark and Vauxhall Co.	Lambeth Co.	Both companies together	Total deaths	Deaths per 10,000 living	Southwark and Vauxhall Co. at 100 per 10,000	Lambeth Co. at 27 per 10,000	The two Companies	Calculated deaths per 10,000 supplied by the two Companies
St. Saviour, Southw.........	1. Christchurch	10,022	2,915	13,234	16,149	113	71	46	36	82	57
	2. St. Saviour	10,700	16,337	898	17,235	378	192	261	2	263	158
St. Olave........................	1. St. Olave	8,015	8,745	0	8,745	161	201	140	0	140	160
	2. St. John, Horselydown	11,360	9,360	0	9,360	152	134	150	0	150	160
Bermondsey...................	1. St. James	18,899	23,173	693	23,866	362	192	370	2	372	156
	2. St. Mary Magdalen	13,034	17,258	0	17,258	247	177	276	0	276	160
	3. Leather Market	15,295	14,003	1,092	15,095	237	155	224	3	227	150
St. George, Southw.........	1. Kent Road	18,126	12,630	3,997	16,627	177	98	202	11	213	134
	2. Borough Road	15,862	8,937	6,672	15,609	271	171	143	18	161	104
	3. London Road	15,836	2,872	11,497	14,309	95	53	46	31	79	55
Newington....................	1. Trinity	20,922	10,132	8,370	18,502	211	101	162	22	184	99
	2. St. Peter, Walworth	29,861	14,274	10,724	24,998	391	131	228	29	257	103
	3. St. Mary	14,033	2,983	5,484	8,467	92	66	48	15	63	74
Lambeth.........................	1. Waterloo, part 1	14,088	3,548	11,939	15,487	59	42	57	31	86	55
	2. Waterloo part 2	18,348	7,171	12,533	19,704	118	64	115	34	149	76
	3. Lambeth church, pt. 1	18,409	3,113	15,878	18,991	49	27	50	43	93	49
	4. Lambeth church pt. 2	26,784	7,868	16,023	23,891	195	73	126	43	167	71
	5. Kennington, part 1	24,261	15,775	2,708	18,483	305	126	253	7	260	146
	6. Kennington, part 2	18,848	7,874	5,620	13,494	143	75	126	15	141	105
	7. Brixon	14,610	1,922	9,356	11,278	48	33	31	25	50	49
	8. Norwood	3,977	0	1,060	1,060	10	25	0	3	3	28
Wandsworth...................	1. Clapham	16,290	6,747	134	6,881	167	103	108	0	108	158
	2. Battersea	10,560	6,276	276	6,552	171	162	100	1	101	152
	3. Wandsworth	9,611	907	94	1,001	59	61	15	0	15	149
	4. Putney	5,280	74	0	74	9	17	1	0	1	160
	5. Streatham	9,023	0	3,244	3,244	15	17	0	9	9	27
Camberwell....................	1. Dulwich	1,632	0	25	25	0	0	0	0	0	0
	2. Camberwell	17,742	9,139	639	9,778	242	136	146	2	148	151
	3. Peckham	19,444	5,438	392	5,830	175	90	87	1	88	151
	4. St. George	15,849	4,295	5,437	9,732	132	83	69	15	84	86
Rotherbithe....................	Rotherbithe	17,805	12,218	0	12,218	283	159	196	0	196	160
Houses supplied in streets where no death occurred		-	28,929	23,338	52,267	-	-	-	-	-	-
Houses not identified...		-	2,712	165	2,877	-	-	-	-	-	-
Totals..		482,435	267,625	171,528	439,153	5,067	105	4,282	462	4,744	108
Population as estimated by the Registrar-General......		-	266,516	173,748	440,264	-	-	4,267	473	4,740	108

Fig. 4. Snow's Table VI: registration sub-distribct mortality based upon district level calculations (Snow, 1856, p. 19). A digital version of this table is available at http://www.epi.msu.edu/johnsnow/illustrationchapters/illustrations%20chp10.htm.

of a physician who carried out all his computations with paper and pencil after a full day of medical practice.[3] For example, Snow miscalculated deaths per 10,000 in Christchurch registration sub-district as 57 rather than, correctly, 50 deaths per 10,000 persons ((82/16149) × 10,000). Nor do all his columns necessarily sum correctly.

In his Table VI (Snow, 1856, p. 18) Snow calculated sub-district deaths per 10,000 persons on the basis of population and recorded deaths, using the previous table's general mortality ratios

(160 per 10,000 and 27 per 10,000) for each of the two water suppliers in each registration sub-district. Snow again calculated a final mortality ratio per 10,000 persons globally—from the totals of population and deaths in the "Totals" column taken from his totals in Table V of 4740 deaths to population of 440,264 served by the two companies rather than by summing his calculated findings row-by-row in Table VI (yielding only 4175 and a ratio of 94.83). Snow then compared the result with the mortality reported by GRO data. "It will be observed that the calculated mortality bears a very close relation to the real mortality in each subdistrict. This relation exists with regard to the real mortality in each subdistrict" (Snow, 1856, p. 10). His result appeared

---

[3] A minor error in Snow's 1855 calculation that does not effect Snow's conclusions has been recently reported by Carvalho et al. (2004).

sufficiently close to the deaths recorded by the official registrars (108 versus 105 deaths per 10,000) to serve, Snow believed, as conclusive proof of his model.

Results

Basic statistical tests, available to us but not to Snow, suggest his conclusion was less than convincing (Fig. 5). While paired samples correlations suggest a high degree of correlation between Snow's outcome and one based on the Registrar-General's findings (.907)—his hypothesis—the real story is told by a paired samples test. At a 95% confidence interval the lower and upper bounds of Snow's calculations of population mortality for both companies are unacceptably high. Most critically, the lower bound is well above 0.0, a clear sign of problems (Norušis, 1999, pp. 223–224). The .002 says at that level of significance (95%) that Snow's hypothesis that his projected calculations based on mortality by water supplier agreed with the Registrar-General's realized mortality is unlikely.

A quick visual check of the data adds substance to these statistical cautions. In several registration sub-districts in Table VI Snow's predictions are clearly inaccurate (Vinten-Johansen et al., 2003, pp. 275–276). In Putney, for example, Snow predicted 160 deaths per 10,000 persons compared to the 17 per 10,000 reported by the Registrar-General's figures in Snow's work. In the more populated St. George, Southwark registration district, the sub-district of Borough Road reported 171 deaths per 10,000 where Snow's method predicted 104. While Snow was content with the "very close" general relation observed in his tables (Snow, 1856, p. 10) it would not serve, today, as sufficiently robust.

Combined mortality for both companies, Snow's penultimate column, gives an erroneous sum of 4744 persons rather than, correctly, 4175 persons. Total population for all sub-districts in 1851 is not 482,435, as Snow's totals showed, but 482,399 persons. More critically, his totals of deaths per

10,000 living resulted from global calculations (Fig. 3) that ignored the detail of the registration districts he had worked so hard to create. Thus, for example, his calculations generate 4175 deaths if summed by sub-district, omitting 569 cases in his Table VI, and resulting in an actual ratio of 95.08 deaths per 10,000 persons served by the two water companies.

Snow calculated mortality at the registration sub-district level based on the ratio of cholera deaths (160 per 10,000 persons for Southwark–Vauxhall and 27 per 10,000 persons for Lambeth Waterworks Company) returned in his Table V (Fig. 3). Snow "lost" in his calculations several thousand registration district residents living in registration districts (and sub-districts) on streets in which no deaths occurred. He was aware of this problem, an apparent artifact of Simon's data set, but had no way to adjust his population figures. "Instead of being able to compare, as I could wish, the mortality in the houses supplied by each company with the exact number of houses supplied, I have only been able to compare it with the number of houses in the streets in which deaths occurred" (Snow, 1856, p. 8). The result adversely affected the resulting ratios based on population in his attempt to calculate mortality at the sub-district level.

Bayesian analysis

First stage

Snow's minor arithmetic errors are insufficient to explain the problems suggested by the two-tailed significance test or the lower bounding, however. For that it is necessary to turn again to the a priori assumptions of Table V and Snow's general approach to the unascertained cases. 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). The EBE

	Mean	S. Dev.	St. Error MEAN	95% Confidence Interval of the difference		T-test	Sig. (2-tailed)
				Lower	Upper		
Deaths from cholera in 1854: both companies	28.7609	47.0320	8.4472	11.5094	46.0124	3.405	0.002

Fig. 5. Confidence and two-tailed significance tests reject Snow's assumption that his projected calculations agreed with the figures returned by the Registrar-General.

Registration District	Cholera cases Southwark-V (Snow)	cholera cases Lambeth (Snow)	Cholera cases Southwark-V (EBE)	cholera cases Lambeth (EBE)
1. St. Saviour, Southwark	406	72	467.96	82.37
2. St. Olave, Southwark	277	0	317.63	1.28
3. Bermondsey	821	0	943.92	1.32
4. St. George	388	99	447.88	112.82
5. Newington	458	58	527.36	66.72
6. Lambeth	525	138	605.61	157.72
7. Wandsworth	268	7	307.59	9.03
8. Camberwell	52	33	405.02	38.24
9 Rotherhithe	207	0	237.06	1.26
10. Greenwich& Sydenham	04	04	6.78	2.43

Fig. 6. Snow's data on deaths per registration district compared with assignments using the empirical Bayes estimation process. The difference reflects the allocation of the 623 previously unassigned cases.

approach (Press, 2003, p. 212) offered a useful strategy to combine Snow's belief with Snow's registration district data set in a manner permitting a better allocation of the problematic cases. EBE provided reasonable starting estimates, allowing for their refinement in the process, and significantly, a careful test of Snow's a priori, locative decisions. This identified as problematic Snow's decision to allocate unascertained cases on the basis of prior distribution.

In general, a Bayesian approach provides a means by which existing but incomplete or missing data sets can be reviewed and the assumptions based upon their use critiqued. The strength of the EBE version of Bayesian analysis is its utility in estimating parameters of distribution without the necessity of assessing parameters of a priori distribution as would be done in conventional Bayesian statistics. Once estimated, in this paper these parameters are used in a hierarchical way as priors for conventional Bayesian prediction. That is, the first stage of this analysis uses Snow's estimates based on his analysis of Simon's data and the subsequent stage involves Snow's comparison of the resulting calculations with the GRO's data on cholera incidence.

Here we summarize our application of the EBE "method of moments," approach to Snow's data (Bailey & Gatrell, 1995, pp. 306, 329; Martuzzi & Elliott, 1996). It is necessary to note parenthetically the complex and robust literature on EBE methodologies whose relation to this problem, and by extension others of its type, will require a separate, more technical paper now in preparation. In our current application Snow's belief in the complicity of the Southwark–Vauxhall company's water in the evidence, and the data returned by the Registrar-General, set the stage for a two-stage analysis. First, we allocate the unknown cases in Snow's Table V using an a priori distribution reflecting Snow's belief in the complicity of Southwark–Vauxhall Company water. In the second stage we estimate distribution of deaths by sub-district conditional upon the Registrar-General's records in the final row of Snow's Table VI. For simplicity's sake, we describe here the general procedure using only the Lambeth Waterworks Company registration districts although our final analysis obviously required the approach be calculated for each water supplier in all registration districts.

In general terms, the first stage EBE methodology permits direct estimation of two parameters, sample variance (ϕ) and a pooled mean (γ), critical to the weighting formula. We followed Snow's a priori assumptions and similarly used the global mean (total cholera deaths in Lambeth/total deaths for both companies in each registration district) to calculate variance (ϕ) and the pooled mean (γ). We did so, however, in a manner that took full account of district variations and assured integrity of population and case data across all districts. This permitted us to employ not Snow's rough global totals but those in which registration district figures were summed.

The prior mean (γ),[4] defined here as cholera per district i for each water company (here we use Lambeth Waterworks for ease of description), was summed and then divided by the total number of cholera deaths. Next, a weighted sample of var-

---

[4]Symbolically, this can be simply stated as $\gamma = \Sigma y_i / \Sigma n_i$ where y_i is the sum of cholera deaths per Lambeth district (i), and n is the number of cholera deaths for both companies in district i.

iances (ϕ) was calculated.[5] The results returned, $\gamma =$ 411/4177 (equalling .09983) and ϕ equals .007333, were then used to in a weighting factor ($wi = \phi/(\phi + \gamma/ni)$) where ni is the number of deaths from cholera in a district. Estimation of the posterior distribution of unassigned cases to each registration district was then carried out using the Bayes estimation formula ($\theta = \gamma + wi(ri - \gamma)$). The table in Fig. 6 compares Snow's original data for both water suppliers and those returned by the EBE approach allocating "non-ascertained" and original cases by district.

Bayesian analysis

Second stage

The results from the first stage at the district level provide estimates of parameters of the a priori at the sub-district level then used to estimate deaths associated with the two companies. The assumption here is that "equivalent past experience" should prevail and the best estimate for a posteriori is merely allocation of deaths according to the registration districts re-cast by population in streets where deaths occurred. The resulting revised figures for the registration districts (Table V) then were used as a priori data to calculate new estimates at the registration sub-district level conditional upon registrar's records.

For the sub-districts in St. Saviour, and Christchurch in Southwark registration district (pop. 19,252 person), for example, the 467.96 cholera deaths attributed to the Southwark–Vauxhall Company in the reworked Table V were allocated to the two constituent sub-districts, Christchurch (pop. 2,915) and St. Savior (pop. 16,337), in effect, on the basis of the percentage of registration district population for each (.1584 versus .8486). In the final iteration of the model this assigned 70.86 deaths to Christchurch and 397.135 to St. Savior, a change reflecting the addition of previously unassigned cases.

Fig. 7 presents a summary of the conclusions of this approach comparing (a) 105 deaths per 10,000 persons based on actual deaths reported by local registrars (b) Snow's 108 deaths per 10,000 persons

based on his calculations and (c) those returned by the EBE approach. The last three columns present the deaths per sub-district calculated on the basis of those reported and the 623 unassigned deaths allocated, sub-district-by-sub-district, through the EBE process.

Because Snow calculated deaths per 10,000 persons using a global mean both his and the Registrar-General's mortality rate, Snow's benchmark, are here recalculated by summing deaths per 10,000 persons per district. The effect is observable in the estimated mortality, reflecting both the skewed nature of the data and the problem of the unassigned cases.

Clearly, the EBE approach more precisely agrees with Snow's interpretation of the observed result based on the global mean. Confidence and significance testing improved dramatically with the EBE calculations. At a 95% confidence interval the range between lower and upper bounds changed from 11.5094 (lower) and 46.0124 (upper) where Snow's findings and the registrars' reported deaths were compared to −6.64908 and 27.7457, for the Bayesian recalculation. In the latter, the upper and lower bounds were comfortably set around the 0.0 point, where they should be. T-test results changed from 3.405 based on Snow's calculation of mortality for both water suppliers to 1.253 and the registrar's results, a significant improvement that now argues the likelihood of his model's efficacy in predicting realized mortality on the basis of water supplier assignments. In its proven congruence between his sub-district calculations and the reported findings of mortality during the epidemic by the local registrars this approach in effect proves Snow's thesis, albeit 150 years after the fact.

Graphically, the result permits the addition of cholera mortality ratios to Snow's map of water supply areas (Fig. 2). In effect, the map gains depth, represented in Fig. 8 through a map in which district mortality rates are combined in the water supply areas of Lambeth Water Company, Southwark–Vauxhall Water Company, and the service area they shared. It was this depth of analysis Snow originally sought in his 1855 study but, without population data, was unable to achieve.

Discussion

Snow attempted to develop a statistical model with predictive capabilities that stands today as a landmark event in disease studies. While his overall goal was to prove the waterborne nature of cholera

---

[5]EBE method of moments formula calculates variance in the following manner: is $\phi = (\sum ni(ri - \gamma)2/\sum ni) - \gamma/\bar{n}$ where ri is the ratio of cholera deaths for Lambeth in district i and $\bar{n}$ is the typical number of cholera deaths by either Southwark and Vauxall companies.

District	Sub-district	Total cholera deaths	Reg. Gen. Deaths per 10,000	Snow Deaths per 10,000	Bayes Deaths per 10,000	Bayes deaths Southwark-V	Bayes deaths Lambeth	Bayes deaths both companies
St. Saviour, S.	Christchurch	113	71	51	90.46	69.58	76.51	146.08
	St. Saviour	378	192	153	229.27	389.95	5.19	395.14
St. Olave	St. Olave	161	201	160	170.28	148.91	0.00	148.91
	St. John Horsleydown	152	134	160	170.28	159.38	0.00	159.38
Bermondsey	St. James	362	192	156	158.55	377.77	0.62	378.38
	St. Mary Magdalen	247	177	160	163.02	281.34	0.00	281.34
	Leather Market	237	155	150	151.87	228.28	0.97	229.25
St. George, S.	Kent Road	177	98	128	147.41	226.15	18.94	245.09
	Borough Road	271	171	103	122.77	160.03	31.61	191.64
	London Road	95	53	54	73.70	51.43	54.47	105.90
Newington	Trinity	211	101	100	99.42	167.33	16.62	183.95
	St. Peter, Walworth	391	131	103	102.82	235.74	21.30	257.03
	St. Mary Magdalen	92	66	74	71.05	49.26	10.89	60.16
Lambeth	Waterloo, Pt. 1	59	42	57	39.72	39.12	22.40	61.51
	Waterloo, Pt. 2	118	64	75	52.06	79.06	23.51	102.57
	Lambeth church Pt. 1	49	27	49	33.76	34.32	29.79	64.11
	Lambeth church Pt.2.	195	73	71	48.89	86.74	30.06	116.80
	Kensington, Pt. 1	305	126	141	96.85	173.92	5.08	179.00
	Kensington, Pt. 2	143	75	105	72.15	86.81	10.54	97.35
	Brixton	48	33	50	34.35	21.19	17.55	38.74
	Norwood	10	25	27	18.76	0.00	2.00	2.00
Wandsworth	Clapham	167	103	157	164.37	112.78	0.32	113.10
	Battersea	171	162	154	161.12	104.91	0.66	105.57
	Wandsworth	59	61	148	153.70	15.16	0.22	15.39
	Putney	9	17	160	0.02	1.24	0.00	1.24
	Streatham	15	17	27	23.80	0.00	7.72	7.72
Camberwell	Dulwich	0	0	27	36.00	0.00	0.09	0.09
	Camberwell	242	136	151	163.68	157.71	2.33	160.04
	Peckham	175	90	151	163.42	93.84	1.43	95.27
	St. George	132	83	86	96.53	74.12	19.83	93.95
Rotherhithe	Rotherhithe	283	159	160	158.39	193.52	0.00	193.52 (includes unassigned cases)
	TOTALS	5,067	105	108	105.434			
Totals using sub-district mean			97.830	107.988	105.434			

Fig. 7. Comparison of mortality ratios based on observed deaths, Snow's model, and the EBE approach advanced in this paper. "Bayes" deaths for Southwark–Vauxhall and Lambeth companies include allocation of the 623 previously unassigned deaths to sub-districts. These are contrasted with "Total cholera deaths" assigned by Snow.

his methodological goal was a type of modeling that in the mid-1850s was in its infancy. We have attempted to demonstrate in this paper not simply that his methodology was limited but more importantly that its limits are (or should be) evident. Using the global mean in his creation of the rates in Snow's Table V was an error even then, one that gave the appearance of congruence with the GRO's data but in the process lost the specificity of the data he was considering. Here we demonstrated that problems inherent in Snow's methodology, and especially in his transposition between registration district and sub-district levels, could be corrected using a modern EBE approach.

One cannot criticize Snow for a failure to have the statistical expertise of modern researchers. His work was, as he claimed, among the most rigorous statistical treatments of its day. Nor is it hard to understand the failure of more modern epidemiologists and spatial analysts to perceive the problems
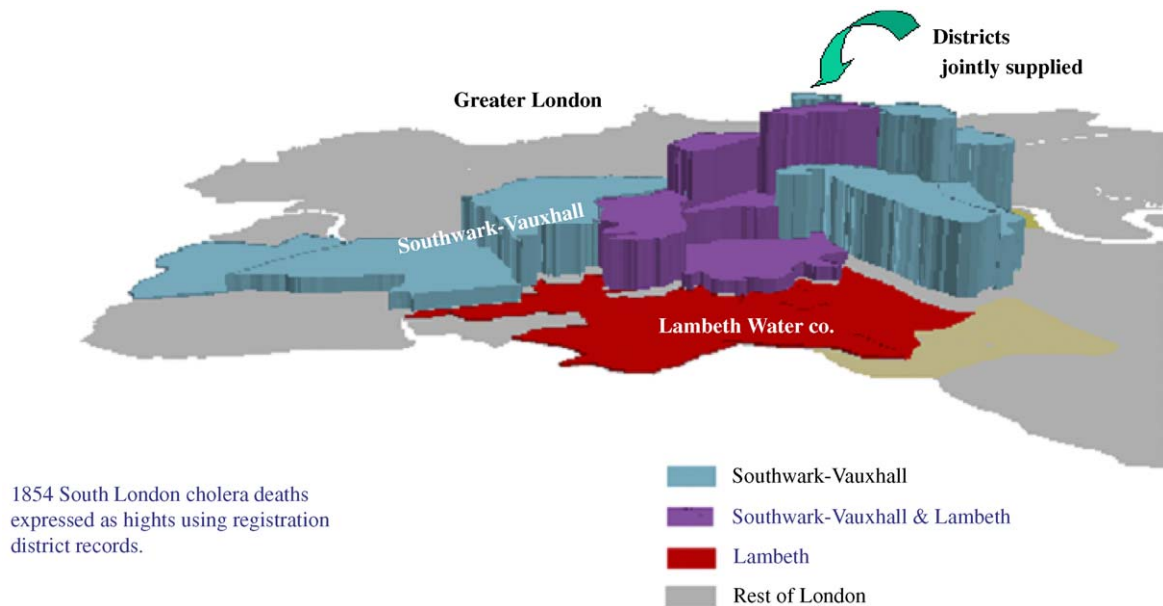
Fig. 8. This map has added mortality per 10,000 persons for all registration districts added to the water supply areas mapped by Snow in the South London study area. Map by authors.

inherent in Snow's response to the problems he faced. Because Snow's study is a foundation piece of the modern epidemiological research paradigm (Carvalho et al., 2004, p. 422), attention has focused on the conclusions in Snow's Table VI, not the problems embedded in his Table V (Vinten-Johansen et al., 2003, pp. 273–277). And, perhaps, Snow's fame has inhibited an unbiased review of his method and results.

It was through a review of Snow's South London study as a potential example for a computerized mapping class that the problem was first identified and later confirmed through confidence tests. A resolution using an EBE approach seemed to be the logical answer (Martuzzi & Elliott, 1996). Certainly, that choice offered a range of benefits that extends beyond correction of an historical data set. Pedagogically, epidemiologists in the past have lamented Snow's failure to create a real case cohort study permitting the construction of risk ratios based on water supplier or other factors (Rothman, 2002, p. 86). That lament assumed Snow went no further than the 1855 study, however. The 1856 paper considered here provides the potential for rigorous risk analysis among cohorts at not one but two levels of data. In addition, Snow's data offers modern instructors a critical example of the difficulties that arise when data collected at one level is transposed to another.

Finally, we believe the problems Snow faced not simply in his registration district calculations but in transposing them to the level of the registration sub-district provide a critical example of the small area unit problem and the difficulties invoked as data are transposed from level to level. In an era of increasingly digital recording and storage of data at different scales (enumeration district, census district, county, regional, state, etc.) the problem is one that increasingly confronts contemporary health researchers (Kirby, 1996; Wakefield & Elliott, 1999). From this perspective Snow's work remains a model of imaginative thinking, and in this study, a practical caution of the problems in transposition from one data level to another. For us, the results not only suggest the potential of the EBE approach but also and as importantly give to Snow's historical study a contemporary methodological standing.

Acknowledgements

References

Ashton, J. (1974). *The epidemiological imagination*. Philadelphia: Open University Press.

Bailey, T. C., & Gatrell, A. C. (1995). *Interactive spatial data analysis*. Essex, UK: Longman Group Ltd.

Balsted, W. M. (2004). *Introduction to Baysian statistics*. Hoboken, NY: John Wiley & Sons.

Carvalho, F. M., Lima, F., & Kriebel, D. (2004). Re: On John Snow's unquestioned long division. *Amercian Journal of Epidemiology, 159, 159*, 422.

CDC (2000). *Weekly Epidemiological Record 31* (4. Aug.), *Cholera*, 1999 (pp. 249–256). Geneva: World Health Organization.

Cromley, E. K., & McLafferty, S. L. (2002). *GIS and public health*. NY: Guilford.

Eyler, J. M. (1973). William Farr on the Cholera: The Sanitarian's disease theory and the statistician's model. *Journal of the History of Medicine and Allied Sciences, 28*(2), 79–100.

Eyler, J. M. (1979). *Victorian social medicine: The ideas and methods of William Farr*. Baltimore: Johns Hopkins University Press.

Farr, W. (1852). *Report on the mortality of cholera in England*, 1848–1849. London: W. Clowes.

Farr, W. (1853). Supplement to the weekly return: Cholera and the London water supply. *Registrar General's weekly return of births and deaths in London #14* (19 November) pp. 401–406.

Frost, W. H. (1936). Appendix. In W. H. Frost (Ed.), *Snow on cholera* (pp. 179–186). New York: The Commonwealth Fund.

Kirby, R. S. (1996). Toward congruence between theory and practice in small area analysis and local public health data. *Statistics in Medicine, 15*, 1859–1866.

Koch, T. (2005). *Cartographies of disease: Maps, mapping, and medicine*. Redlands, CA: ESRI Press.

Martuzzi, M., & Elliott, P. (1996). Empirical Bayes estimation of small area prevalence of non-rare conditions. *Statistics in Medicine, 15*, 1867–1873.

McLeod, K. S. (2000). Our sense of Snow: The myth of John Snow in medical geography. *Social Science & Medicine, 50*, 923–935.

Melnick, A. L. (2002). *Introduction to geographic information systems in public health*. Gaithersburg, MD: Aspen Pub.

Morris, R. J. (1976). *Cholera 1832: The social response to an epidemic*. London: Croom Helm.

Norušis, M. J. (1999). *SPSS 9.0 guide to data analysis*. Upper Saddle River, NJ: Prentice-Hall.

Openshaw, S. (1984). *The modifiable area unit problem* (Concepts and Techniques in Modern Geography, No. 39). Norwich, UK: Geo Books.

Parks, E. A. (1855). Review: Mode of communication of cholera by John Snow MD. *British and Foreign Medical Review, 15*, 449–463.

Press, J. S. (2003). *Subjective and objective Bayesian statistics: Principles, models and applications*. NY: John Wiley and Sons.

Registration of Births, Deaths, and Marriages in England Act of 1835 (1835). (6 & 7 Will IV c.86).

Robinson, A. K. (1982). *Early thematic mapping in the history of cartography*. Chicago: University of Chicago Press.

Rothman, K. J. (2002). *Epidemiology: An introduction*. NY: Oxford University Press.

Simon, J. (1856). *Report of the last two cholera epidemics of London, as affected by the consumption of impure water*. London: HMSO.

Smith, G. D. (2002). Behind the broad street pump: aetiology, epidemiology, and prevention of cholera in mid-19th century Britain. *International Journal of Epidemiology, 31*, 920–932.

Snow, J. (1854). Communication of cholera by Thames Water. *The Medical Times and Gazette*, 9 (2 Sept), 247–248.

Snow, J. (1855). On the mode of communication of cholera. In W.H. Frost (Ed.). *Snow on cholera* (2nd ed.). New York: The Commonwealth Fund, 1936.

Snow, J. (1856). Cholera and the water supply in the south districts of London in 1854. *Journal of Public Health, October*. Published by: Queen St., London: T. Richards.

Vandenbroucke, J. P., Rooda, H. M. E., & Beukers, H. (1991). Who made John Snow a hero? *American Journal of Epidemiology, 133*(10), 967–973.

Vinten-Johansen, P., Brody, H., Paneth, N., Rachman, S., & Rip, M. (2003). *Cholera, chloroform, and the science of medicine: A life of John Snow*. NY: Oxford University Press.

Wakefield, J., & Elliott, P. (1999). Issues in the statistical analysis of small area health data. *Statistics in Medicine, 19*, 2377–2399.