LANCASTER UNIVERSITY

Rivers, rainfall, and risk factors: geostatistical and epidemiological approaches to disentangle potential transmission routes of typhoid fever

by

Jillian Sarah Karklin Gauld

A thesis submitted in partial fulfilment for the degree of Doctor of Philosophy

in the

Faculty of Health and Medicine

Lancaster Medical School

February 2020

Declaration

This thesis has not been submitted in support of an application for another degree at this or any other university. It is the result of my own work and includes nothing that is the outcome of work done in collaboration except where specifically indicated. Many of the ideas in this thesis were the product of discussion with my supervisors, Dr. Jonathan M. Read, Professor Nicholas A. Feasey, and Professor Peter J. Diggle.

Chapter 2 of this thesis has been published in the following academic publication:

Jillian S. Gauld, Franziska Olgemoeller, Rose Nkhata, Chao Li, Angeziwa Chirambo, Tracy Morse, Melita A. Gordon, Jonathan M. Read, Robert S. Heyderman, Neil Kennedy, Peter J. Diggle, Nicholas A. Feasey, *Domestic River Water Use and Risk of Typhoid Fever: Results from a Case-control Study in Blantyre*, Malawi, Clinical Infectious Diseases, ciz405, https://doi.org/10.1093/cid/ciz405

Jillian S. Gauld, BScH, MSc

Lancaster University, UK

February 2020

Abstract

Typhoid fever, caused by the bacterium *Salmonella* Typhi, is a severe febrile illness, with over 20 million cases and 100 thousand deaths occurring annually. In 2011, Blantyre, Malawi experienced a sharp increase in the incidence of typhoid fever, and transmission continues today. Although the disease is generally known to spread through the fecal-oral route, the precise mechanisms of transmission in endemic locations are not well characterized. Therefore, a challenge exists in determining which water and sanitation interventions may be the most important for control of typhoid fever. This thesis attempts to identify risk factors for typhoid fever in this setting, and employs geostatistical, epidemiological, and genomic approaches to data collected as part of routine disease surveillance as well as typhoid-specific epidemiological studies. The findings from this thesis indicate that transmission of typhoid fever in Blantyre is complex, with both environmental and social factors important components. Evidence of environmental transmission as found, through the use of non-drinking water from local rivers identified as a risk factor. This finding was used to generate hypotheses: testing whether river catchments are predictors of genomic patterns, and exploring rainfall anomalies as time-dependent predictors of incidence. Both investigations yielded significant results: river catchments were predictors of genomic patterns, and rainfall anomalies were found to be protective, further bolstering the hypothesized environmental component of transmission. Typhoid fever can also lead to severe clinical complications, and a methodological contribution was included that enabled the attribution of intestinal perforations to typhoid fever, independent of microbiological testing. Although new vaccines for typhoid offer a promising tool for control, investment in non-vaccine interventions will likely be critical for elimination, and the work

presented suggests possible opportunities for interventions focused around hydrological systems and water usage.

Acknowledgements

I'd like to thank my supervisors, Dr. Jonathan Read, Professor Peter Diggle, and Professor Nicholas Feasey for their support throughout this thesis. The many late-afternoon (or evening) Skype calls, marathon email threads, continuous time-zone conversions, and positive encouragement were very much appreciated. I'd also like to thank Peter and Jon for making themselves so available for meetings when visiting Lancaster (both planned and impromptu), allowing me to make the most of my trips. I appreciate being so readily included in the goings-on of the group while in town: seminars, coffee time, and hikes (walks). The students and staff at CHICAS make a fantastic group, and I hope to be able to continue collaborating after my studies conclude.

Thank you to Nick for initially supporting me in my idea to pursue a PhD, connecting me to my eventual supervisors at Lancaster University, as well as providing the data for this thesis. I am additionally grateful for the travel funding support, through his grant with the Bill and Melinda Gates Foundation. This resulted in a unique variety of supervisory meeting locales, including Kampala, Hanoi, Blantyre, Seattle, Annecy, Atlanta, New Orleans, London, Liverpool and Lancaster.

I am additionally very grateful for the funding provided by the Institute for Disease Modeling. I am specifically grateful to Philip Welkhoff, Hao Hu, and Mike Famulare for enabling and encouraging this pursuit.

I am very thankful for the hospitality of Annabelle Edwards, who shared her home of Stoneybeck Cottage with me while staying in Lancaster. Having a comfortable place to live made the longer (as long as 3 month) stays possible and productive. Finally, thank you to my family and friends for their support and patience.

Contents

List of Figures

List of Tables

List of Papers

This thesis contains the following appended papers.

Paper 1: *Domestic River Water Use and Risk of Typhoid Fever: Results from a Case-control Study in Blantyre, Malawi.* **Jillian S. Gauld**, Franziska Olgemoeller, Rose Nkhata, Chao Li, Angeziwa Chirambo, Tracy Morse, Melita A. Gordon, Jonathan M. Read, Robert S. Heyderman, Neil Kennedy, Peter J. Diggle, Nicholas A. Feasey

- Published in *Clinical Infectious Diseases*, doi: 10.1093/cid/ciz405.
- Contribution: co-design of statistical analyses, implementation and interpretation of analyses, drafting of manuscript, incorporation of co-author comments, corresponding author
    - Figures: JG: 2.1; FO: 2.2
    - Tables: All tables
    - Supplementary Material: FO: S2.1; PJD: S2.2; JG: S2.3

Paper 2: *Integrating spatial and genomic data to identify transmission patterns of typhoid fever in Blantyre, Malawi.* **Jillian S. Gauld,** Franziska Olgemoeller, Eva Heinz, Rose Nkhata, Chao Li, Sithembile Bilima, Alexander M. Wailan, Neil Kennedy, Jane Mallewa, Melita A. Gordon, Jonathan M. Read, Robert S. Heyderman, Nicholas R. Thomson, Peter J. Diggle, Nicholas A. Feasey

- Manuscript circulated to co-authors in preparation for submission.
- Contribution: co-design of statistical analyses, implementation and interpretation of analyses, drafting of manuscript
    - Figures: JG: 3.1, 3.2, 3.3b,c; EH: 3.3a
    - Tables: JG: 3.2, 3.3; FO: 3.1
    - Supplementary Material: JG: S3.1, S3.2, S3.4; EH: S3.3

Paper 3: *Rainfall anomalies and typhoid fever in Blantyre, Malawi.* **Jillian S. Gauld**, Peter J. Diggle, Nicholas A. Feasey, Jonathan M. Read

- Manuscript circulated to co-authors in preparation for submission.
- Contribution: co-design of statistical analyses, implementation and interpretation of analyses, drafting of manuscript and creation of all figures and tables.

Paper 4: *Intestinal perforations associated with a high mortality and frequent complications during an epidemic of multidrug-resistant typhoid fever in Blantyre, Malawi.* Franziska Olgemoeller, Jonathan J. Waluza, Dalitso Zeka, **Jillian S. Gauld**, Peter J. Diggle, Jonathan M. Read, Thomas Edwards, Chisomo L. Msefula, Angeziwa Chirambo, Melita Gordon, Emma Thomson, Tiya Chilunjika, Robert S. Heyderman, Eric Borgstein, Nicholas A. Feasey

- Under review at *Clinical Infectious Diseases.*
- Contribution: Implemented time series analysis attributing intestinal perforations to typhoid fever, drafted section 5.3.5, contributed to discussion, corresponding author
    - Drafting of manuscript: FO and NF
    - Figures: JG: 5.2; FO: 5.1
    - Tables: JG: 5.2; FO: 5.1
    - Supplementary material: JG: S5.1

Acronyms

BSI	bloodstream infection
CI	confidence interval
DEM	digital elevation model
EA	enumeration area
GBD	global burden of disease
IQR	interquartile range
LL	log-likelihood
LMIC	low- and middle- income countries
MCET	morbidity, carriage, and genomic epidemiology of typhoid
MCMC	Markov chain Monte Carlo
MCML	Monte Carlo maximum likelihood
MDS	multidimensional scaling
MLW	Malawi-Liverpool Wellcome Trust Clinical Research Programme
OR	odds ratio
PC	principal coordinate
PCA	principal components analysis
PCR	polymerase chain reaction
QECH	Queen Elizabeth Central Hospital
SNP	single nucleotide polymorphisms
SRTM	Shuttle Radar Thematic Mapper
TCVs	typhoid conjugate vaccines
USGS	United States Geological Survey
VBNC	viable but non-culturable
WASH	water, sanitation, and hygiene
WGS	whole genome sequence
XDR	extensively drug resistant

1 Introduction

1.1 Thesis overview

The overall focus of this thesis is on the epidemiology of typhoid fever in Blantyre, Malawi. This involves a number of statistical analysis approaches: incorporating genomic, spatial, and risk factor data to better understand transmission in this setting.

1.2 Typhoid fever

1.2.1 Epidemiological overview and burden

Typhoid fever, caused by the bacterium *Salmonella enterica* serovar Typhi (*S.* Typhi), remains a significant cause of morbidity and mortality in the developing world. Estimated global burden is approximated at 10-20 million cases per year [1–4], with the majority of cases occurring in Africa and south Asia. The disease is primarily characterized by a high, sustained fever.

As high fever may indicate any number of diseases, including, but not limited to malaria, tuberculosis, influenza, or other bacterial bloodstream infections, symptom-based diagnosis is not possible. The gold standard diagnostic test for typhoid fever is by culture of blood or a bone marrow sample, which requires diagnostic infrastructure that does not exist in many clinics in resource-limited settings. Rapid diagnostics or alternative burden evaluation methods either perform poorly or do not exist. Typhoid fever surveillance is therefore often limited to passive surveillance from a few major hospitals in these regions. Because of these challenges, estimates of burden are limited to few reporting sites, with as few as 13-14 low and middle-income countries contributing data to global burden estimates [2,3]. Where blood culture diagnostics do exist, sensitivity of this method for detection of *S.* Typhi is

limited, with sensitivity estimates at approximately 59% [5]. Therefore, under-reporting of cases is an accepted challenge in typhoid fever epidemiology and estimates of disease burden, affecting the targeting of control strategies, and general advocacy of typhoid as a global health problem.

Without effective treatment, typhoid fever can be a fatal disease. Case-fatality is currently estimated to be 1-2% in endemic countries [4,6]. The most extreme complication is intestinal perforation, requiring surgical intervention to resolve. However, data on "surgical typhoid" are largely absent from global burden of disease estimates [4]. Antibiotics are the primary tool of treatment for typhoid fever, but resistance is an ongoing concern. Resistance to chloramphenicol, the antibiotic originally used for treatment of typhoid fever, was reported as early as the 1950s [7], and outbreaks of these resistant strains were reported in the 1970s [8]. Since then, multi-drug resistant (MDR) typhoid fever, or typhoid fever resistant to three or more first-line antibiotics, including chloramphenicol, has become a global concern [9]. More recently, resistance to fluoroquinolones has emerged, and has become widespread in Asia [10]. Most recently, the appearance of extensively drug resistant (XDR) typhoid fever, which extends resistance to third generation cephalosporins, was reported in 2018 [11], further limiting the treatment of typhoid fever. For XDR typhoid, azithromycin is the only widely available/ affordable antibiotic remaining for treatment of the disease. We are therefore threatened by typhoid that is impossible to treat. In the pre-antibiotic era, case-fatality was estimated to be 15% [12], meaning deaths from typhoid fever could increase an order of magnitude without effective control measures.

1.2.2 Control strategies and challenges

Vaccines for typhoid fever have been in development since the late 1800s [13]. Field trials of candidate vaccines such as the live oral vaccine Ty21a vaccine

occurred beginning in the 1970s, and demonstrated long-lasting efficacy in school age children [14]. However, formulas that were safe and immunogenic in young children, had long lasting immunogenicity, and were feasible to distribute in developing countries were lacking. The development of new conjugate vaccines (TCVs) for typhoid has provided alternatives that are safe, efficacious, and likely to provide lasting immunity in young children [15]. World health organization (WHO) prequalification of the first conjugate vaccines occurred in January of 2018 [16]. As a result, there has been an increased effort to prevent typhoid through vaccination in areas of highest burden, with field trials underway in a number of locations [17].

Although vaccines for typhoid fever have occasionally been demonstrated to be effective in reducing incidence in settings with large-scale field trials [18], elimination of local transmission of the disease has not yet occurred without water or sanitation interventions. In the United States, typhoid incidence declined after widespread sanitation and water improvements were implemented throughout the early 1900s, a common theme across most developed countries. In Santiago, Chile, a ban on the irrigation of produce with wastewater in the 1990s helped expedite control to elimination as a public health problem, and was aided, but not driven by, use of the vaccine [19]. Mass vaccination initiatives for typhoid fever have not previously occurred, but the conjugate vaccine offers an opportunity to do so.

There is some concern that even widespread use of TCVs may not be able to eliminate typhoid fever as a public health problem. Human challenge models suggest a high level of clinical protection conferred with the vaccine using a syndrome-based definition, 87% [15]. However, protection from any microbiological outcome was only estimated at 55%, suggesting the presence of the bacteria in the body, and subsequent shedding of the disease, still may occur.

If these observations hold in endemic settings, although the conjugate vaccine may be an excellent vaccine for reduction in mortality outcomes and clinical disease, its usefulness as a tool for reduction of ongoing transmission, and therefore control, may be limited.

Because of this uncertainty, exploration of non-vaccine interventions is still necessary. Barring widespread sanitation and water improvements, understanding specifically how *S.* Typhi is transmitted is a critical step in this process.

1.2.3 Transmission routes

S. Typhi, the aetiological agent of typhoid fever, is a human restricted pathogen whose only known reservoir is human hosts. It is believed that typhoid fever is transmitted through the fecal-oral route, in which individuals are exposed through ingestion of *S.* Typhi that has previously been excreted through the feces of an individual shedding the disease. How this chain of exposure occurs can vary between locations, however at a high level, transmission is often categorized into two modes: "short-cycle" and "long-cycle".

Short-cycle transmission represents infections spread from person-to-person through proximate contaminated food vehicles, a route made famous by the case of "Typhoid Mary" [20]. If an infectious individual does not have access to or use appropriate hand hygiene, *S.* Typhi on their hands is capable of infecting others, through food vehicles. This transmission route has been well documented in reviews of outbreaks [21]. After the decline of typhoid fever occurred through improvements in sanitation systems and water treatment, outbreaks of typhoid fever can still occur through food handlers. Today, short-cycle outbreaks of typhoid fever still occur in the United States, mostly through long-term carriers of the disease [22,23].

Long-cycle transmission represents infections transmitted through environmental mediators. Examples of these mediators include drinking water, open sewers, and food crops irrigated with wastewater. The dominant environmental mediators can vary between locations: In Kathmandu, Nepal, the contamination of drinking water from stone taps with sewage has been implicated as a likely transmission route for typhoid [24], while in Santiago, Chile, in the 1980s, high endemic levels of typhoid were maintained through the irrigation of produce with contaminated wastewater [19]. It is critical to note that, in the majority of locations, what happens to *S.* Typhi in between fecal excretion into the environment and exposure of the next individual is uncertain; we do not know what the ecological niches are, or if there is an environmental reservoir, as is the case with *V. cholerae* [25].

The relative importance of intermediate ecological niches to typhoid transmission is unknown. Risk factor studies have attempted to identify potential exposures (a review of these exposures is contained in section 2.1), yet identifying the *contaminant* of these exposures is often difficult. For example, it is currently unclear how to disentangle whether lettuce, if identified as a risk factor, was contaminated by irrigation of produce through the long-cycle, or through a food handler directly. These unknowns lead to general uncertainty in which environmental and/or behavioral interventions are most important for control.

1.2.4 Presence and persistence of *Salmonella* Typhi in the environment

The environmentally mediated component of typhoid transmission is enabled through the extended shedding of infected individuals, and persistence of *S.* Typhi in the environment.

Individuals with typhoid fever may demonstrate prolonged periods of shedding. Those who are acutely infected with typhoid typically shed *S.* Typhi from 2 to 4 weeks [26,27]. Further, approximately 2 percent of infected individuals become chronic carriers, defined as individuals who shed *S.* Typhi for more than a year. This is due to the colonization of *S.* Typhi in the gallbladder [28], and these chronic carriers are capable of shedding *S.* Typhi for a lifetime. However effective treatment with antibiotics, particularly fluorquinolones, is known to curtail shedding of the disease [29].

S. Typhi is additionally capable of surviving in the environment, and has been experimentally demonstrated in the laboratory in both water and food. A study in 1999 explored the survival of *Salmonella* Typhi in water [30]. By marking *S.* Typhi with a green fluorescence protein, researchers were able to explore both the decline of culturable cells in water, but also propose that survival and conversion to viable-but non-culturable (VBNC) cells occurs. The study found that the decline of culturable cells in all types of water was rapid over the 27-day period (0.75 day^{-1} and 1.3 day^{-1} exponential decay in groundwater and pondwater, respectively), with viable cell counts declining at a lower decay rate (0.25 day^{-1} and 0.35 day^{-1} exponential decay) for groundwater and pondwater, respectively). Survival is better in groundwater than pond water, a conclusion attributed to the greater presence of protozoa in the latter medium.

S. Typhi's survival in food vehicles is under-studied compared to *Salmonella* strains that cause outbreaks in developing countries today (*Salmonella* Typhimurium, *Salmonella* Enteritidis). A study in 1976 contaminated young lettuce plants with *Salmonella* Typhi, and demonstrated survival of the bacteria declined over time, but persisted to the age of maturity of the plant [31]. One study showed survival of typhi on sprouts up to 10 days,

and even demonstrated growth in experiments with inoculation at the seed germination stage [32]. The book *Microbial Survival in the Environment* (1984) offers a comprehensive view of laboratory studies from mostly the 1920s through the 1940s: sewage sludge (2- 83 days), radish surfaces (60 days), soil (10-70 days) were some of the mediums tested, with the primary take-away being that there is evidence of heterogeneity of *S.* Typhi survival in the environment, and there is a possibility on certain vehicles for it to persist for weeks to months [33].

1.2.5 Detection of *Salmonella* Typhi in the environment

Few studies have attempted to detect *S.* Typhi in the environment in endemic settings. Successful identification of environmental *S.* Typhi through culture was demonstrated in Santiago, Chile in the 1970s. Moore swabs, passive filtration tools made of sterile fabric or gauze [34], were used in the to isolate and culture live *S.* Typhi from both from sewage drainage outside the homes of chronic carriers [35] as well as from the irrigation water [36]. However, even with the swabs placed directly outside the homes of known shedders of *S.* Typhi, sensitivity was low at 25% [37]. Culturing of live S. Typhi from environmental samples in present day has not yielded positive results [24,38], and therefore live culture is not currently accepted as a sensitive detection method.

One potential reason for the lack of sensitivity of culture methods in the detection of live *Salmonella* Typhi is that *S.* Typhi may enter a viable-but-non-culturable state (VNBC). VNBC cells of *S.* Typhi have been shown to be able to be epidemiologically relevant, however: a laboratory study in 1996 resuscitated VNBC cells with broth, and successfully colonized mice [39].

Quantitative PCR is an alternative to culture methods, as VBNC cells could be detected without resuscitation. qPCR was used to detect *S.* Typhi in drinking water samples in Kathmandu, Nepal [24], which supported conclusions

that the contamination of stone aquifers are a likely driver of typhoid in the city. These methods have additionally been applied in Dhaka and Mirzapur, Bangladesh [40], yielding contrasting detection rates in drinking water samples between the two cities.

Some challenges exist in the interpretation of these results, however. New evidence suggests that when using qPCR for environmental samples to detect *S.* Typhi, the commonly used primers used are subject to specificity issues. This specifically relates to the ability to differentiate between *S.* Typhi and non-typhoidal *Salmonella* in samples spiked with both [41]. Therefore, without either improved culture methods for growing live *S.* Typhi in a laboratory setting, or investment in qPCR methods that prioritize specificity, which require moving beyond a single gene target, our ability to conclusively detect and measure *S.* Typhi in the environment is limited.

1.3 Study population: Blantyre, Malawi

1.3.1 Geographic and demographic context

The city of Blantyre is located in the Southern Region of Malawi (Figure 2.1, Chapter 2). The city is geographically diverse: the area surrounding the city is mountainous, and the city itself contains a number of small mountains and hills. The city additionally contains ten major rivers, which all drain out of the city. The city has experienced substantial population growth, with the estimated population growing from 649,000 in 2008 to 800,000 in 2018 according to census [42]. Much of the population lives in unplanned areas or informal settlements, the largest of which is Ndirande, with 118,000 individuals living in this area as of 2007 [43].

In these informal settlements, sewage infrastructure and safe water availability is not reliable. The National Statistics Office of Malawi conducts a

Welfare Monitoring Survey, most recently in 2011 [44], and reflects the reality of informal settlements and lack of infrastructure. It was estimated that only 3.6% of the city's population has a toilet that flushes to the sewage system, and 59.6% of the population use a 'basic latrine', which, out of the choices (flush or pour flush to sewer system, flush to a septic tank, an improved latrine, VIP, eco-san), likely indicates a pit latrine. Although this was not surveyed, it has been noted that the rocky soil in Blantyre often prevents the digging of pit latrines further than three meters deep [45], possibly adding a challenge to sewage management.

In 2012, the Millennium Cities Initiative additionally conducted a case study in Blantyre to assess the city's health, water and sanitation, education, gender, and infrastructure needs according to the Millennium Development Goals. An assessment of wastewater treatment revealed that, although there are five treatment facilities, three are not functioning. According to this document, the wastewater and sewage that is supposed to be treated at these facilities is often diverted to the rivers, untreated [45].

Drinking water is an additional challenge in the city: From the Welfare Monitoring Survey [44], only 23.8% of the population has water piped into their property, indicating a need to travel outside the home (at least short distances) for water. 65.5% of the population retrieved water from a public tap or standpipe, while 7.1% used a bore hole. The quality of these sources was not assessed. Almost all residents surveyed (98.7%) purchased food from a local market, while 22% additionally grew their own produce.

1.3.2 Disease surveillance in Blantyre

Blantyre is the site of Queen Elizabeth Central Hospital (QECH), the largest government hospital in Malawi. The hospital provides free healthcare to

residents of Blantyre, and is a referral center for the surrounding Southern region of the country. The Malawi-Liverpool Wellcome Trust Clinical Research Programme (MLW) was established in 1995, and has conducted continuous routine blood culture surveillance at QECH, beginning in 1998.

Between 1998 and 2010, an average of 14 cases of typhoid fever per year were diagnosed at QECH, and only 6.8% were resistant to ampicillin, chloramphenicol, and cotrimoxazole. However, by 2014 there were 782 cases per year, 97% of which were resistant to all three of these antibiotics [46]. Potential drivers of the emergence of typhoid fever were explored in a mathematical modelling study, including increased duration of infectiousness and increased transmission rate [47]. It was found that an increased duration of shedding in the population, possibly caused by multi-drug resistance preventing the ability of first-line antibiotics to be effective in treating the disease, may be responsible for the emergence seen. This model was based on simplified assumptions of transmission and population mixing, and despite transmission still occurring today, the mechanisms of typhoid fever transmission are unknown.

1.3.3 Morbidity, carriage, and genomic epidemiology of typhoid (MCET) study

The previous sections demonstrate both a need for non-vaccine interventions for control of typhoid fever, as well as methodological limitations in our ability to rapidly determine them through current epidemiological and microbiological methods. Given advances in sequencing technology and ease of geo-location of cases, opportunities exist to better understand typhoid transmission in combination with statistical analysis techniques, and doing so in a single site may offer a framework for assessment of these methods and their conclusions in parallel.

The morbidity, carriage, and genomic epidemiology of typhoid (MCET) study was initialized in 2015, after the confirmed emergence of typhoid fever, in order to better understand transmission of typhoid fever in Blantyre, Malawi. This project harnessed the routine blood culture surveillance ongoing at QECH to recruit hospital cases of typhoid fever into a cohort study. The households of cases were geo-located, and demographic information was recorded. Within this cohort, a nested case-control study of children 9 years of age and under was conducted to identify risk factors for typhoid fever, using additional controls recruited from the community. Further, a subset of isolates was whole-genome sequenced.

1.4 Aims and structure of the thesis

The goals of my thesis were to utilize MCET data, along with data collected through routine blood culture surveillance at MLW, to better understand the epidemiology and transmission of typhoid fever in this setting. This included utilizing multiple methodological approaches and specific questions nested within each. The specific aims are listed below, with methodological rationale following:

1. Case-control analysis to identify risk factors for typhoid fever in Blantyre, Malawi (Chapter 2)
2. Incidence mapping to identify areas of highest incidence and unexpected hot-spots (Chapter 3)
3. Spatial-genomic analysis to explore spatial scales of genetic relatedness of case isolates as a proxy for the transmission process (Chapter 3)
4. Analysis of temporal trends to explore the relationship between patterns of rainfall and typhoid incidence (Chapter 4)

5. (analysis contribution) Predicting the impact of the increase in typhoid fever on reported surgical intestinal perforations, a severe complication of typhoid fever (Chapter 5)

1.5 Methods overview

The following sections outline methodological approaches and motivation for employing these methods.

1.5.1 Spatial statistical methods for epidemiological data

Within the field of infectious disease epidemiology, the ability to geolocate cases in a population is becoming increasingly efficient and inexpensive due to advances in GPS technology and decreasing costs of using these tools. Therefore, spatial analysis methods are becoming more widely used alongside epidemiological data.

In descriptive studies of spatial patterns of cases, the goal of the analysis is often limited to the detection (or not) of clustering of cases, usually by testing the null hypothesis that the observed spatial pattern is completely random. The formal definition of completely random depends on the data-format: for point data, a completely random pattern is a realization of a Poisson process; for small-area count data, a completely random pattern is one for which the counts follow independent Poisson distributions. For a review of clustering methods see, for example, Alexander and Cuzick (1992) [48].

Previous examples of the use of spatial statistical methods for the analysis of typhoid data are mostly limited to this type of cluster detection. Spatiotemporal cluster detection was used for typhoid in rural Cambodia, which noted heterogeneity in clusters of disease across both space and time in the study site [49]. Similar clustering was noted for typhoid fever in Dhaka, Bangladesh [50].

With the understanding from these clustering investigations that cases of typhoid fever may occur non-randomly across space, some studies have incorporated this clustering into risk factor outcomes. In Kibera, Kenya, a spatial case-control study was conducted in order to understand the drivers of incidence in the setting [51]. Spline smoothing was used to account for spatial autocorrelation.

The spatial statistical methods that are used in this thesis have the more ambitious goal of estimating the strength and scale of spatial dependence within a geostatistical modelling framework, as described below.

1.5.2 Tools for geostatistical modelling

Given the increase in availability of spatial data, tools for geostatistical modelling have become more widely available. The primary tool used for spatial analyses in this thesis is PrevMap, a package in R, which is based on a generalized linear mixed modelling framework [52].

Briefly, generalized linear mixed models are statistical models that allow for random effects. Random effects are, essentially, unobserved random processes that account for variation in the outcome that cannot be explained by measured explanatory variables or sampling variation. For example, if we model the growth of a child over time, we may need to include a random effect on the level of individual child, to account for unexplained differences between children if their size at the beginning of the observations is not measured. These random effects are most often assumed to be normally distributed.

When measurements are taken across space, we can similarly assume that there will be some random differences in measurement that vary across space. Often in practice, it has been observed that this variability, however, is spatially correlated. That is, observations that are closer together have more similar

outcomes than those that are further apart. Modelling this type of spatially correlated random effect and model structure is often referred to as model-based geostatistics [53].

For illustrative purposes, equation 1.1 is an example of the parameterization of a linear geostatistical model, as implemented in PrevMap and proposed in [53]:

$$Y_i = \alpha + \beta d(x_i) + S(x_i) + Z_i \qquad [1.1]$$

The random effects in these models are comprised of two components, Z_i and $S(x_i)$. Z_i are independent zero-mean Gaussian random variables, essentially the spatially-uncorrelated random effects. Also known as the 'nugget effect,' this component is named for the variation in presence of gold nuggets at a single mining site. In an epidemiological setting it can represent either variance in measurements at a single spatial point, or spatial correlation that is occurring at a smaller scale than what is sampled. $S(x)$ is the spatially structured random effect, or the 'spatial random effect'. This is parameterized by a multivariate normal distribution and a covariance matrix. The values making up the covariance matrix can be obtained from the variogram as defined in equation 1.2:

$$\mathrm{V(u)} = \tau^2 + \sigma^2\{1 - \rho(u)\} \qquad [1.2]$$

In equation 1.2, τ^2 is the aforementioned nugget effect, σ^2 is the partial sill, and $\rho(u)$ represents the correlation at distance u. The variance of Y is $\tau^2 + \sigma^2$. The covariance between any two values of Y at locations distance u apart is $\sigma^2\rho(u)$. In PrevMap, the Matérn correlation function [54] is used for $\rho(u)$. A schematic representation of a typical *variogram*, is shown in Figure 1.1.

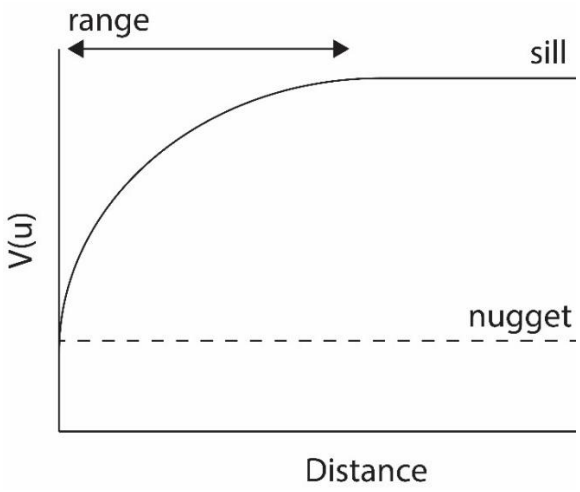


Figure 1.1 Theoretical variogram illustrating the parameterization of the spatial random effect.

In addition to the sill, $\tau^2 + \sigma^2$, and nugget, τ^2, represented in Figure 1.1, the shape and effective distance of spatial correlation complete the specification of the variogram. The smoothness and physical range of spatial correlation are contained in the $\rho(u)$ function of equation 1.2. For the Matérn correlation function, these are components represented by two parameters, κ and ϕ, respectively [equation 1.3].

$$\rho(u) = \{2K_{\kappa}^{-1}\Gamma(\kappa)\}^{-1}(h/\phi)^{\kappa}K_{\kappa}(h/\phi) \qquad [1.3]$$

Where K_{κ} is the modified Bessel function of the third kind of order κ.

In addition to any predetermined covariates included in a geostatistical model, we must fit the parameters determining the spatial random effect. Fitting typically occurs for the parameters τ^2, σ^2, and ϕ, while κ remains fixed, as it is difficult to estimate [55]. Fitting these geostatistical models occurs through Maximum Likelihood Estimation for linear models, and Markov-Chain Monte Carlo methods for generalized linear geostatistical models. Both of these processes are automated through PrevMap.

PrevMap has successfully been applied across a number of studies. For example, mapping the incidence of Snakebites in Sri Lanka [56], and exploring the impact of changing sample size on the ability to accurately detect Malaria hotspots [57].

1.5.3 Genetic analysis approaches and multidimensional scaling

Genomics have been a useful tool for better understanding the epidemiology of infectious diseases. With the increased availability and decreasing costs of obtaining whole genome sequence (WGS) data, applications have been diverse; WGS data have been used to understand the dynamics of an outbreak of *Mycobacterium tuberculosis* [58], and trace the origins of the 2014 Ebola epidemic [59].

Sequencing of *Salmonella* Typhi offers distinct challenges. The highly clonal nature of *S.* Typhi was originally revealed through analysis of 200 gene fragments (88,739 base-pairs). Only 88 single nucleotide polymorphisms (SNPs) were identified after sequencing 105 Typhi isolates, confirming the highly clonal nature of *S.* Typhi and identifying the drug resistant haplotype H58 [60]. Given the genome is over 4 million base-pairs long, further evolutionary signal can be revealed analysis of the entire genome [61]. The first multiple-genome study sequenced 17 isolates from the previous gene fragment analysis, identified 1,964 SNPs, excluding repetitive regions [62], however there are now thousands of genomes available which has led to a revision of the haplotyping scheme [63].

Some studies have previously explored the spatial patterns of *S.* Typhi genomics. SNP typing for 1,500 previously identified SNPs was performed on pediatric *S.* Typhi isolated in Kathmandu, Nepal [64], and isolates were categorized into 28 distinct genotypes. Geospatial analysis of these patterns [62]

found no household clustering. However, these 28 genotypes were not linked together in any way via genetic relatedness, so these were treated as independent categorical variables. This makes any in-depth analysis challenging, and may mask more subtle patterns of relatedness. On a more macro-scale, the spatiotemporal distribution of genomic lineages, defined as groups containing a combination of SNPs identified through whole-genome sequencing, were explored in Siem Reap province in Cambodia [49], and revealed distinct spatial-genetic clusters. However, this study is similarly limited in that the analysis was restricted to using the categorical representation of these genomic lineages.

Multidimensional scaling (MDS) and other multivariate methods such as principal components analysis (PCA) have been useful for extracting continuous summary measures of genetic relatedness since the 1960s [65]. These methods reduce complex data containing relational measures of genetic-relatedness, to a small number of synthetic variables that can approximate components of this relatedness in a pre-specified number of dimensions. This approach is agnostic to assumptions of evolutionary models or common ancestors. Its aim is simply to describe relationships between isolates. This can be particularly useful in exploratory scenarios, and allows us to retain a continuous representation of genetic relatedness across samples, as opposed to categorical haplotypes or lineages, as described above. These multivariate methods of summarizing genetic relatedness have been used widely, for example, in the exploration of spatial transmission in malaria genetics [66] and global population structures in human population genetics [67].

1.5.4 Weather patterns and disease

In addition to spatial and genomic analyses methods, for diseases that are possibly mediated by the environment, characterizing the temporal trends of infectious disease in relation to weather patterns may offer complementary

insights into transmission patterns of disease. For well-characterized transmission routes, such as malaria, incorporating weather into statistical models can allow for evaluation of variables for early warning systems [68]. For diseases without established transmission routes, weather patterns can offer unique insight into mechanisms at play, such as pathogen accumulation during the dry seasons, as suggested for diarrheal illness [69].

Because many diseases exhibit seasonal dynamics, and so do weather patterns, cross-correlation of disease and weather patterns is to be expected regardless of whether a mechanistic link occurs. This makes establishing a causal link challenging. Typhoid is known to be a seasonal disease across most settings [70]. Therefore, unsurprisingly, rainfall is correlated with typhoid incidence. In Dhaka, it was found that a 3-5 week lag in rainfall was associated with an increase in typhoid cases [50]. In a multi-site investigation, it was observed that rainfall often precedes the disease, and a positive association with temperature is frequent [70], however this was not a universal finding across sites.

In analysis scenarios where cross-correlation is a given due to seasonality of both the weather and disease pattern, a method of more convincingly establishing causality from a time series analysis can be helpful. Working with extreme disease or weather events in this context, rather than the raw data, offers a potential solution. As extreme events do not tend to have seasonal regularity, there is no expected cross-correlation, as long as the extreme events can be effectively identified. For example, more or less rainfall than expected in a given season predicting more or fewer cases than expected may be more convincing than a seasonal lag, when establishing a causal link. This approach has been used in establishing the link between rainfall events and diarrhea [69]. Methods of identifying these extreme events and incorporating them into a

model vary but can include using pre-specified thresholds, or by extracting residuals from a smoothed (de-seasonalized) model.

2 Domestic river water use and risk of typhoid fever: results from a case-control study in Blantyre, Malawi

Clinical Infectious Diseases

Jillian S. Gauld[1,2], Franziska Olgemoeller[3,4], Rose Nkhata[4], Chao Li[3,5], Angeziwa Chirambo[4,6], Tracy Morse[7, 11], Melita A. Gordon[4,6] , Jonathan M. Read[2], Robert S. Heyderman[8], Neil Kennedy[9,10], Peter J. Diggle[2], Nicholas A. Feasey[3,4]

1. Institute for Disease Modelling, Bellevue, USA
2. Centre for Health Informatics, Computing, and Statistics, Lancaster University, Lancaster, UK
3. Department of Clinical Sciences, Liverpool School of Tropical Medicine, Liverpool, UK
4. Malawi-Liverpool Wellcome Trust Clinical Research Programme, Blantyre, Malawi
5. Xi'an Jiaotong University Health Science Center, Shaanxi, China
6. Institute of Infection and Global Health, The University of Liverpool, Liverpool, UK
7. Centre for Water, Environment, Sustainability and Public Health, University of Strathclyde, Glasgow, UK
8. Division of Infection and Immunity, University College London, London, UK
9. Department of Paediatrics, University of Malawi the College of Medicine, Blantyre, Malawi
10. School of Medicine, Dentistry and Biomedical Sciences, Queen's University Belfast, UK
11. Centre for Water, Sanitation, Health and Appropriate Technology Development, University of Malawi – Polytechnic, Blantyre, Malawi

Abstract

Typhoid fever remains a major cause of morbidity and mortality in low and middle-income settings. In the last 10 years, several reports have described the re-emergence of typhoid fever in southern and eastern Africa, associated with multidrug-resistant H58 *Salmonella* Typhi. Here, we identify risk factors for pediatric typhoid fever in a large epidemic in Blantyre, Malawi.

A case-control study was conducted between April 2015 and November 2016. Cases were recruited at a large teaching hospital, while controls were recruited from the community, matched by residential ward. Stepwise variable selection and likelihood ratio testing were used to select candidate risk factors for a final logistic regression model.

Use of river water for cooking and cleaning was highly associated with risk of typhoid fever (OR 4.6 [CI: 1.6-12.5]). Additional risk factors included protective effects of soap in the household (OR 0.6 [CI: 0.4-0.98]) and more than one water sources used in the previous 3 weeks (OR 3.2 [CI: 1.6-6.2]). Attendance at school or other daycare was also identified as a risk factor (OR 2.7 [CI: 1.4-5.3]) and was associated with the highest attributable risk (51.3%).

These results highlight diverse risk factors for typhoid fever in Malawi, with implications for control in addition to the provision of safe drinking water. There is an urgent need to improve our understanding of transmission pathways of typhoid fever, both to develop tools for detecting *S.* Typhi in the environment, and inform water, sanitation, and hygiene interventions.

2.1 Introduction

Typhoid fever continues to be a major cause of morbidity and mortality in low and middle-income settings, with an estimated 10-20 million cases occurring annually, and approximately 200,000 deaths [1–3]. In south and southeast Asia, *Salmonella* Typhi was identified as the most common bacterial pathogen

associated with bloodstream infection (BSI) among hospitalized patients between 1990 and 2010 [71]. In contrast, *S.* Typhi was not described as a major cause of BSI in southern and eastern African countries during the same period, even in centers with long term bacteremia surveillance [72]. Instead, nontyphoidal serovars of *Salmonella* were much more prominent causes of BSI. Since 2012, the picture has changed dramatically, with multiple reports describing the emergence of typhoid as a major cause of BSI in southern and eastern Africa [46,73–75]. Though the drivers of this recent emergence remain unclear, typhoid is now acknowledged as a significant public health problem in both Africa and Asia [76].

S. Typhi is a human-restricted pathogen, and transmission occurs via the fecal-oral route. Its ecological niche after excretion remains poorly described, but there is evidence for heterogeneity in pathways of environmental exposure. For example, typhoid transmission has been linked to contamination of the water supply in Kathmandu, Nepal [24], whereas in Santiago, Chile, endemicity was maintained until the early 1990s through irrigation of salad crops with wastewater [19]. These contrasting data suggest that the critical intervention points at which typhoid transmission may be interrupted in the environment may be context-specific. In addition to transmission through an ecological niche, *S.* Typhi may also be transmitted within the household, most often through direct contamination of food by an infected individual. This type of transmission is not only present in endemic settings, but has led to outbreaks of typhoid after endemicity has been interrupted through widespread sanitation improvements [21]. This poses an additional challenge for control.

Both transmission pathways are important in the spread of *S.* Typhi, but their relative importance in endemic settings is poorly understood. Risk factor studies have been conducted in a variety of locations, including both endemic

and outbreak settings, to better understand the dominant drivers of transmission. Previously identified risk factors for typhoid include recent contact with individuals diagnosed with typhoid or enteric fever [77–79], food, including consuming flavored ices [80] and ice cream [81] or ice cubes [82], buying lunch at school [80] or eating roadside or outdoor vended food [81–83], and drinking unsafe or untreated water at home [79,82,84] or drinking water at work [81]. Exposure to water used for purposes other than drinking has also been identified as a risk for typhoid, such as bathing and brushing teeth [82]. Findings on sanitation show lack of soap in the household and limited handwashing are associated with typhoid [78,82,83,85,86], while having a latrine in the household has been found to be protective in Indonesia [78], but a risk factor in Nepal [87]. In endemic locations, the majority of work has been done in Asian, Oceanian, and South American countries, and has so far been limited on the African continent [77–86].

These findings implicate a variety of water, sanitation, and hygiene (WASH) factors, but the heterogeneity among locations indicates a need for site-specific investigation, particularly in regions that have been under-studied or where typhoid is re-emerging. Furthermore, although many food and water exposures have been previously identified, detailed studies describing where in the food preparation or production cycle, or through which aspect of water usage *S.* Typhi is entering and amplifying are lacking. This hampers the planning of effective intervention strategies at the source of contamination. Understanding the complexity of water, sanitation and hygiene factors in transmission has assumed greater importance following the emergence of cephalosporin resistant typhoid in Pakistan [11], which threatens the role of antimicrobials in typhoid control. Whilst the typhoid conjugate vaccine offers a promising tool for control, targeted water and sanitation interventions are likely to be necessary too.

Blantyre is the second-largest city in the country of Malawi, located in the Southern region (Figure 2.1). WASH-related interventions over the last 10 years in Blantyre have focused on water access, with an increase in kiosks, trials of delivery systems, and protection of open sources, but interventions on household water treatment and improved sanitation have been limited. Blantyre has experienced a sharp increase in typhoid, increasing from an average of 14 cases per year between 1998 and 2010, to over 700 in 2013 [46]. Typhoid has remained endemic in Blantyre, and the mechanism of this sustained transmission is currently unknown. We therefore conducted a case-control study to investigate risk factors for typhoid in this setting.

2.2 Methods

2.2.1 Data collection and study site

Queen Elizabeth Central Hospital (QECH) in Blantyre, Malawi, provides free healthcare to urban Blantyre and the surrounding district, and tertiary care to the Southern region of Malawi. Laboratory surveillance for BSI has been routine since 1998, and is conducted through the Malawi-Liverpool-Wellcome Trust Clinical Research Programme (MLW), based at QECH [88]. Pediatric patients are eligible for routine blood culture if they present to the hospital with non-specific febrile illness and test negative for malaria, have persistent febrile illness after treatment for malaria, or are severely ill with suspected sepsis. Blood was drawn for each eligible patient (2-4mL), followed by automated culture (BacT/ALERT, Biomerieux) and serotyping for identification of *Salmonella* Typhi [88].

Cases were defined as children under 9 years of age with blood culture confirmed *S.* Typhi infection diagnosed between April 2015 and November 2016 at QECH, and who originated from the Blantyre urban area. Eligible controls were healthy children under 9 years of age and were recruited at a 4:1 ratio

throughout the study period. Children under 9 years of age were prioritized for the study because of the known frequency of typhoid in this age group [46]. A high-resolution census sub-divided urban Blantyre into 393 enumeration areas (EAs), each with an estimated population size (Figure 2.1) [42]. To avoid spatial over-matching, which would have made it impossible to identify small-scale spatial heterogeneity of risk, controls were matched by larger residential wards rather than by EA. To approximate the random selection of controls within a ward, we selected EAs with probability proportional to population size. Within each sampled EA, households were approached along a random path until an eligible control was identified and consent was taken from legal guardians.

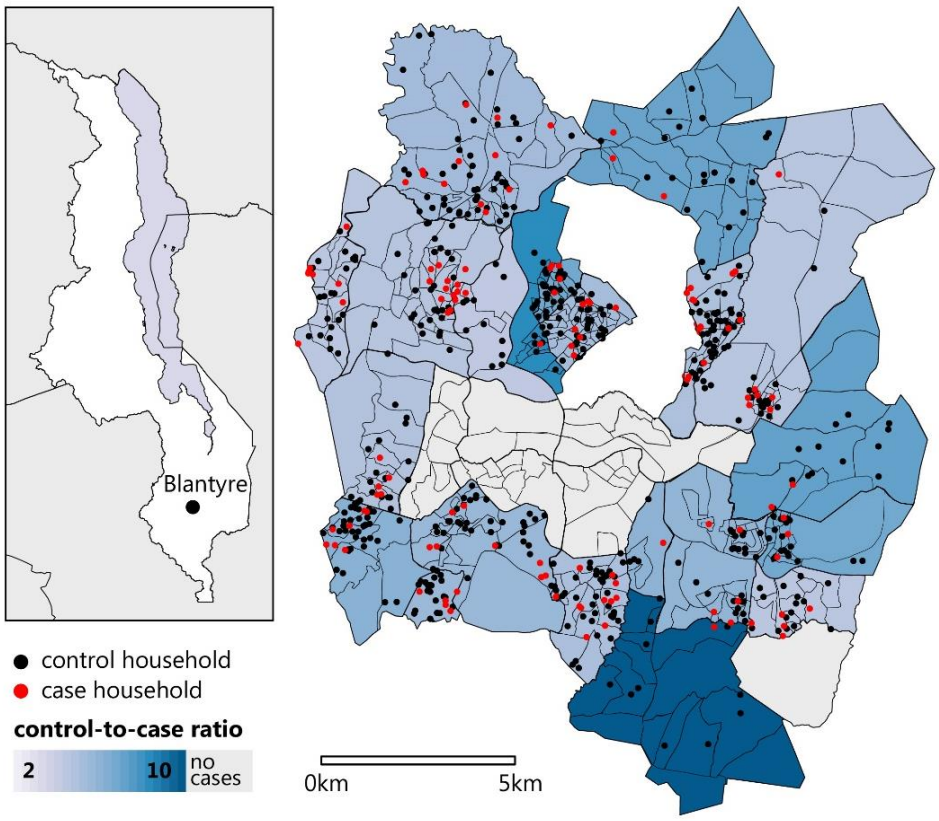


Figure 2.1 Location of Blantyre within the country of Malawi (inset), and the Blantyre study boundaries. Enumeration areas are represented by the smaller polygons, while residential wards are indicated in the larger, shaded by the ratio of controls to cases. Households of cases (red) and controls (black) are plotted as points, with precise locations masked by randomization.

Exclusion criteria specific to cases after initial recruitment was living outside EA boundaries. For both cases and controls, individuals were excluded if they had a household member who previously was diagnosed with typhoid during the period of the study.

A standardized questionnaire was administered to the guardians of participants, where guardian was defined as a caregiver for the child, above 18 years of age. The questionnaire recorded both demographic and socioeconomic indicators, as well as potential risk factors for typhoid. The incubation period for typhoid in outbreak settings can be highly variable, but is not known to frequently extend longer than three weeks [21]. Therefore, questions distinguished exposure in the last three weeks from exposure in the last year. Sources of water for drinking and water used for cooking and cleaning were separately surveyed. The location and altitude of households and identified water sources were collected using Garmin Etrex 30 GPS devices.

Controls were requested to provide a stool sample to describe asymptomatic shedding of *S.* Typhi, described in Supplementary Material 2.1.

2.2.2 Statistical analysis

Logistic regression was used to assess potential risk factors in the study. Residential ward was included as a fixed effect for all analyses, to take account of the stratified sampling design for controls. The majority of predictor variables were assessed directly from the questionnaire, while distance to hospital, distance to primary water source, and elevation change between the household and water source were calculated for each individual, using the recorded household locations, water source locations, and ascertained GPS coordinates of QECH. Due to the large number of questions in the initial survey, stepwise forward variable selection was conducted to reduce the number to an interpretable size. This process began with the base model, defined as the fixed

effect of residential ward, plus intercept. At each iteration, likelihood ratio tests were conducted to compare the base model with each potential variable addition. The variable addition resulting in the lowest p-value from the likelihood ratio test was then added to the model. The process was repeated with the base model now updated with the added variable. The process stopped when no variable addition improved the model at a significance level of $p<0.05$.

The final logistic regression model was fitted using the resulting selected variables. Odds ratios with 95% confidence intervals were calculated using coefficients and standard errors estimated from the fitted model. Unadjusted individual odds ratios were also calculated for each selected variable to assess dependence of multivariate model findings on the combination of included parameters. To enable comparison between continuous variables in the study, we rescaled each so an increase in scaled value is equal to one standard deviation increase in the unscaled value. Due to only one individual reporting more than one febrile family member, and one individual reporting more than two water sources, for the final model fit these continuous variables were converted to categorical variables.

Finally, we extend the multivariate logistic regression model to estimate the potential percentage reduction in cases in our population attributed to removing reported exposures. Detailed methods are described (Supplementary Material 2.2). Because we do not know the null exposure value of continuous variables, these calculations were only made for variables that were binary, and those were estimated to be significant in the model.

To investigate spatial correlation in risk within residential wards, we assessed the residuals of the fitted logistic regression model [53]. All statistical analyses were conducted using R statistical software, version 3.5.1 [89].

This study was approved by the University of Malawi, College of Medicine Research and Ethics Committee [P.08/14/1617], the Liverpool School of Tropical Medicine Research Ethics Committee [14.042] and the Lancaster University Faculty of Health and Medicine Ethics Committee [FHMREC17014].

2.3 Results

During the study period, 189 children were diagnosed with blood culture confirmed typhoid (Figure 2.2). There were no cases of *Salmonella* Paratyphi A. 125 cases were included in the study, with a median age of 5 (IQR 3-7); 60 patients were not recruited, amongst whom 35 declined participation, 24 could not be reached after diagnosis, and 1 patient died from complications of perforation prior to recruitment. After recruitment, two patients were excluded because they were secondary cases in households that had previously been surveyed, and two cases were excluded from the analysis because their household location fell outside the study boundary. One control was excluded, due to another household member having culture-confirmed typhoid during the week of recruitment.

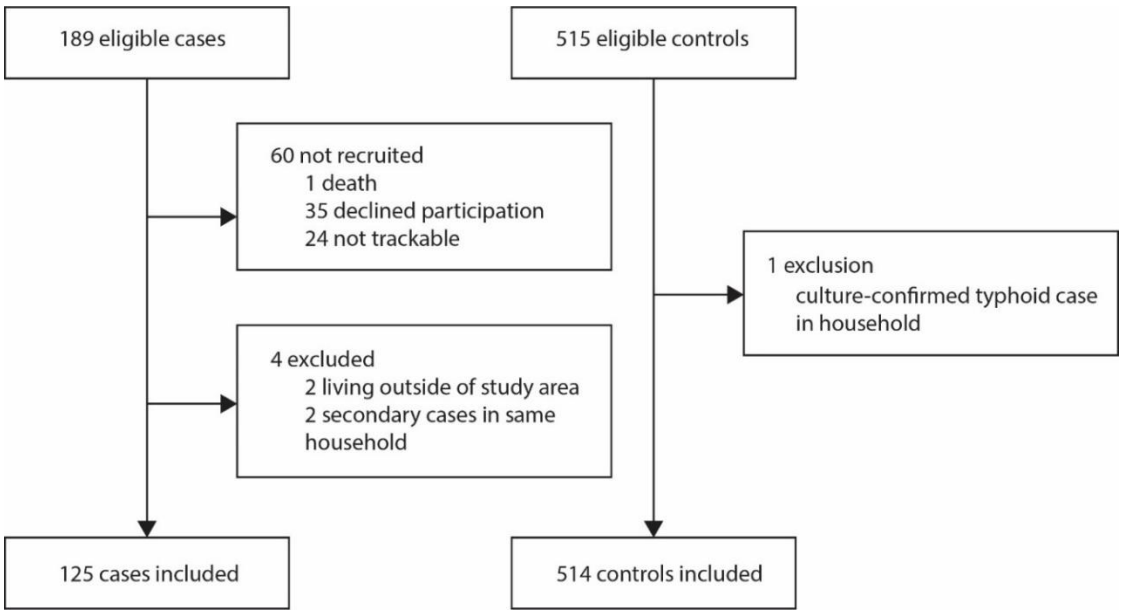


Figure 2.2 Consort chart for cases and controls in the study.

Cases tended to be older than controls (Table 2.1) but were similar in distribution of gender. Though the overall ratio of controls to cases was 4.2:1, control ascertainment resulted in a heterogeneity of the ratio of controls to cases between residential wards (Figure 2.1). Six residential wards did not contain any cases. Amongst the 123 controls tested, 0/123(0%) were stool culture positive for *S.* Typhi, therefore no further action was taken, however, 3/123(2.4%) were PCR positive (95% CI:0.8-6.9%).

Table 2.1 Baseline characteristics of cases and controls enrolled in the study.

Characteristic	**Cases (n=125)**		**Controls (n=514)**		**p-value**
Age (years)	n	%	n	(%)	< 0.005
≤ 2	28	(22)	185	(37)	
3-5	35	(28)	209	(40)	
6-8	61	(49)	120	(23)	
Gender					0.38
Male	61	(49)	294	(51)	
Female	64	(51)	278	(49)	

Variable selection reduced the 97 initial variables to 14 (Table 2.2, Supplementary Material 2.3). The 125 cases and 514 controls were reduced to 122 and 507, respectively, due to missing data in the final variable set. Out of

the 14 final variables selected in the model, 8 were directly related to water exposures.

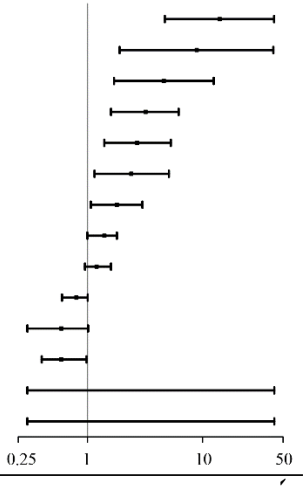
Logistic regression identified several significant risk factors for typhoid in children (Table 2.2). Factors suggestion environmental exposure included cooking and cleaning with river water (OR 4.6 [CI:1.7-12.5]) and water from an open dug well (OR 2.4 [CI:1.1-5.1]), having more than one drinking water source (OR 3.2 [CI:1.6-6.2]), and being from a household growing crops (OR 1.8 [CI:1.1-3.0]). Conversely, availability of soap to wash hands after the toilet (OR 0.6 [CI:0.4-0.98]), was protective. Risk factors suggesting the importance of social interaction patterns were identified, including spending the day at school or in child care (OR 2.7 [CI:1.4-5.3]) and having one or more household members admitted to the hospital with febrile illness in the last four weeks (OR 8.9 [CI:1.9-41.2]. Seeking care for severe illness at QECH was selected for in the model, adjusting for differential case-ascertainment through the hospital between cases and controls. Estimates of attributable risk are summarized in Table 2.2. The highest attributable risk percentage was spending the day at school or daycare (51.3%), followed by growing crops by the household (17.4%). Attributable risk percentages were lower, and similar, for cooking and cleaning with river water (10.3%) and water from an open dug well (8.3%).

There was no significant spatial correlation of residuals from the analysis (Supplementary Material 2.3), indicating that the variables in the questionnaire and/or spatially matching on residential ward sufficiently accounted for unexplained spatial variation in risk.

`

Table 2.2 Estimated odds ratios for univariate models and selected multivariate model. Numeric variables are scaled for presentation of estimates; thus odds ratios are presented as increased risk per 1 standard deviation increase in the value.

Variable	Cases (122)	Controls (507)	Unadjusted OR	p value	Adjusted OR (95% CI)	p value	Attributable Risk (%)	Adjusted OR (95% CI)
Seeking care at QECH if child is severely ill, no. (%)	118 (97)	370 (73)	10.9 (4.0, 30.2)	<0.001	14.1 (4.7, 41.8)	<0.001	–	
One or more household members admitted to hospital for febrile illness in last four weeks, no. (%)	9 (7)	7 (1)	5.7 (2.1, 15.6)	<0.001	8.9 (1.9, 41.2)	0.006	–	
Cooking and cleaning with river water in the previous three weeks, no. (%)	15 (12)	16 (3)	4.3 (2.1, 9.0)	<0.001	4.6 (1.7, 12.5)	0.002	10.3	
More than one drinking water sources used last three weeks, no. (%)	28 (23)	38 (7)	3.7 (2.2, 6.3)	<0.001	3.2 (1.6, 6.2)	<0.001	15.4	
Child spends the day at school, preschool, nursery or any other daycare, no. (%)	99 (81)	312 (62)	2.7 (1.7, 4.4)	<0.001	2.7 (1.4, 5.3)	0.005	51.3	
Cooking and cleaning using water from an open dug well in the previous three weeks, no. (%)	20 (16)	35 (7)	2.6 (1.5, 4.8)	0.001	2.4 (1.1, 5.1)	0.020	8.3	
Family grows crops, no. (%)	47 (38)	137 (27)	1.7 (1.1, 2.6)	0.127	1.8 (1.1, 3.0)	0.027	17.4	
Age (years), median (range)	5 (0– 8)	3 (0– 8)	1.7 (1.4, 2.1)	<0.001	1.4 (1.0, 1.8)	0.053	–	
Distance to from household to primary water source (meters), median (range)	78 (1– 738)	52 (0– 748)	1.2 (1.0, 1.5)	0.013	1.2 (1.0, 1.6)	0.118	–	
Number of days water is stored, median (range)	2 (1– 7)	2 (1– 20)	0.74 (0.6, 0.96)	0.024	0.8 (0.6, 1.0)	0.054	–	
Experienced water shortage in the house or surrounding area in the past two weeks, no. (%)	38 (31)	172 (31)	1.0 (0.7, 1.6)	0.897	0.6 (0.3, 1.0)	0.056	–	
Soap available to wash hands after the toilet in the previous three weeks, no. (%)	70 (57)	360 (71)	0.5 (0.4, 0.8)	0.002	0.6 (0.4, 0. 98)	0.042	–	
Stores drinking water in drum, no. (%)	0 (0)	20 (4)	2.6 e−07 (0, inf)	0.977	1.2 e −7 (0, inf)	0.984	–	
Used stream or river water for drinking in the last three weeks, no. (%)	0 (0)	4 (1)	7.2 e−07 (0, inf)	0.984	1.1 e −8 (0, inf)	0.992	–	



2.4 Discussion

This study provides detailed insight into the risk factors for pediatric typhoid in an urban African setting. Our findings point to complex and varied risks for typhoid in Blantyre, including water sources, household indicators of sanitation and hygiene, and social interaction patterns such as school attendance.

In multivariate analysis, cooking and cleaning with river water was the principal environmental exposure identified in the study. Cooking and cleaning with water from an open dug well was additionally identified as a risk factor. No sources of drinking water were associated with typhoid, contrasting with other studies that implicate drinking water sources as risk factors [79,81,82,84]. Potential explanations include that communities are aware of the risks associated with drinking unclean water, but less aware of the risks of indirect exposure, such as through pans or other items that may come into contact with food. Alternatively, people may prioritize safe water for drinking, but cannot afford to purchase or transport the volume of safe water needed for use in other household tasks. It is estimated that less than 5% of the population is connected to the sewage network, with the majority of the population utilizing pit latrines [44]. Open dug wells and nearby rivers used for cooking and cleaning water may become contaminated with runoff from pit latrines, particularly during rain events, providing a plausible epidemiological link.

Our findings indicate that individuals are at a higher risk for typhoid when using multiple drinking water sources. Previous work examining water access in urban Malawi identified limited access hours, tariffs, low water pressure, and too few water kiosks as structural barriers to adequate potable water for household activities [90]. These water access challenges are likely to influence the number and type of water sources used, and may necessitate the

use of unsafe sources. In other studies, distance, access, and behavioral factors have been found to influence decisions around accessing potable water [91–93].

We also identify risk factors where exposure could occur through either interaction with contaminated environments, infected individuals, or both. Having household members hospitalized for febrile illness was identified as a risk factor; as was attending school or other day care. In the context of schools, however, it is uncertain whether the key exposure is direct contact with a contaminated environment [94], food handlers contaminating meals [94,95], or transmission routes such as contact with infectious children. The presence of soap in the household was found to be protective, consistent with findings in other locations [78,82,83,85,86], further supporting a tool that interrupts exposure.

Coming from a household that grows crops is a risk factor for typhoid in Blantyre, consistent with the experience in Santiago, Chile, where irrigation of crops with wastewater was a driver of typhoid transmission [19]. Neither irrigation with human nor animal waste was found to be a significant risk, however fecal contamination of food crops still may be possible in Blantyre through runoff from latrines, or irrigation with fecally contaminated river water.

Calculation of attributable risk has enabled us to estimate frequency of exposure to these risk factors in the population. Spending the day in school or day-care was associated with the highest attributable risk, highlighting the importance of this common exposure among children in our study and associated challenges with WASH in schools [96]. A small percentage of cases and controls reported cooking and cleaning with river water/water from an open dug and thus these factors were associated with lower estimated attributable risks, however such behaviours are commonly described in qualitative and observational research in Malawi [97,98]. There is therefore a possibility of

under-reporting these types of exposures, and further research on quantifying these patterns would be useful. The study has some limitations; the extended incubation period of typhoid necessitated a 2-3 week window for assessing potential exposures, and recall bias cannot be excluded. Controls were recruited throughout the study period, and not matched over time, limiting our ability to control for seasonality. We focused on young children with the goal of capturing household-related risk factors, assuming younger children move around the city less than adults and are therefore less likely to become exposed outside of the household. Regardless, the potential for differential risk factors for older children and adults may limit the generalizability of these findings to older age groups. We assessed WASH risk factors through a questionnaire, rather than by direct observation in or transect walks around participant households. Lastly, by basing our study on sentinel surveillance of patients presenting to QECH, we have selected for more severe disease, and have not captured minimally symptomatic or sub-clinical typhoid, which may be associated with differential risk factors.

We provide new insights into risk factors for typhoid in an urban African context, challenging the dogma that transmission of *S.* Typhi can be interrupted solely by the provision of safe drinking water. Instead, we highlight the importance of usage of water for purposes other than drinking, of hand hygiene, and of preschool/daycare attendance in the transmission of typhoid in this setting. Future work should confirm our findings by direct assessment of *S.* Typhi in the environment. Developing novel tools for the identification of *S.* Typhi in the environment will help to identify transmission routes rapidly, and without in-depth risk factor analyses for each epidemic or endemic location.

Funding

This work was supported by Bill and Melinda Gates Foundation Investment [OPP1128444]. The Malawi Liverpool Wellcome Trust Clinical Research Programme is supported by Wellcome Trust Major Overseas Programme [206545/Z/17/Z].

Acknowledgements

The authors would like to thank the staff and patients of Queen Elizabeth Central Hospital and the University of Malawi, the College of Medicine and the control participants for their support.

3 Integrating spatial and genomic data to identify transmission patterns of typhoid fever in Blantyre, Malawi

Jillian S. Gauld[1,2], Franziska Olgemoeller[3,4], Eva Heinz[7,3], Rose Nkhata[4], Chao Li[3,5], Sithembile Bilima[4], Alexander M Wailan[7], Neil Kennedy[9,10], Jane Mallewa[11] Melita A. Gordon[4,6], Jonathan M. Read[2], Robert S. Heyderman[8,] Nicholas R Thomson[7,12], Peter J. Diggle[2], Nicholas A. Feasey†[3,4]

1. Institute for Disease Modelling, Bellevue, USA
2. Centre for Health Informatics, Computing, and Statistics, Lancaster University, Lancaster, UK
3. Department of Clinical Sciences, Liverpool School of Tropical Medicine, Liverpool, UK
4. Malawi-Liverpool Wellcome Trust Clinical Research Programme, Blantyre, Malawi
5. Xi'an Jiaotong University Health Science Center, Shaanxi, China
6. Institute of Infection and Global Health, The University of Liverpool, Liverpool, UK
7. Wellcome Sanger Institute, Cambridge, UK
8. Division of Infection and Immunity, University College London, London, UK
9. Department of Paediatrics, University of Malawi the College of Medicine, Blantyre, Malawi
10. School of Medicine, Dentistry and Biomedical Sciences, Queen's University Belfast, UK
11. Adult Medicine, University of Malawi the College of Medicine, Blantyre, Malawi
12. Department of Pathogen Molecular Biology, London School of Hygiene and Tropical Medicine

Abstract

A sharp increase in cases of multidrug resistant typhoid fever in Blantyre, Malawi was observed in 2011. Transmission continues today, but the key environmental niches and dominant transmission routes remain unknown. This poses a challenge for targeting water and sanitation interventions.

Between March 2015 and January 2017, 549 patients presenting to Queen's Hospital, Blantyre, with blood culture confirmed typhoid fever were recruited to a cohort. For a subset of these patients, households were geo-located, and *S.* Typhi isolates were whole genome sequenced (WGS). Pairwise SNP distances were converted into informative variables using multidimensional scaling and incorporated into a geostatistical modelling framework.

Spatial risk analyses revealed a heterogenous distribution of *Salmonella* Typhi isolates across the city. Pairwise SNP distance and physical household distances were significantly correlated. We evaluated the ability of river catchment to explain the spatial patterns of genomics observed, and found that river catchment significantly improved the fit of the model. We also found small scale spatial correlation of the genetic signatures amongst households living up to 50 meters apart.

Our findings support epidemiological evidence that river systems play a key role in the transmission of *S.* Typhi in Blantyre. These findings will help inform targeted environmental surveillance, to confirm the presence of *S.* Typhi in rivers and understand heterogeneity in exposure. We present compelling evidence of the value of integrating complex data to understand the transmission of environmentally dependent pathogens, which in this case can be used to inform the deployment of control measures.

3.1 Introduction

Typhoid fever remains a major cause of morbidity and mortality in developing countries, with an estimated 11 million cases occurring annually [99]. In March 2018, the typhoid conjugate vaccine (TCV) was recommended by WHO for control of typhoid fever, providing momentum for global initiatives to combat this disease [76]. Although this vaccine offers a high level of clinical protection, it is less certain whether it will prevent shedding of the disease [15]. Therefore, interventions using TCVs alone may not be sufficient for control in endemic locations. Multi-faceted initiatives pairing vaccine with water, sanitation and hygiene (WASH) interventions may be necessary.

Prioritizing WASH interventions for typhoid is challenging, as transmission routes do not appear to be consistent across locations [19,24]. Risk factor investigations are useful for identifying specific exposures, however achieving a broad understanding of the pathways of transmission for the purposes of intervention remains a challenge. This is further complicated due to the difficulty of detecting *Salmonella* Typhi from the environment [41]. This limits our understanding of where best to intervene to interrupt transmission in settings with inadequate water, sanitation and hygiene infrastructure.

Spatial and genomic data may offer insight into transmission patterns of typhoid fever. Geo-locating cases as a part of routine surveillance has become increasingly common. Spatially informed disease control interventions and investigations are being developed and utilized, such as for the targeting Polio vaccines [100], and investigating hot-spots and transmission routes of Ebola [101]. Geospatial analyses for typhoid fever to-date have revealed the spatially heterogenous nature of the disease at both municipal and national scales [50,102], but consistent spatial predictors of disease have yet to be identified [3].

As costs have declined, whole genome sequencing (WGS) has become a valuable component of infectious disease research, and WGS of bacterial genomes offers the potential to type bacteria with high discrimination and reproducibility [103,104]. Individuals closely linked along a transmission chain have closely related genomes, so WGS offers the resolution to confirm or refute the existence of a transmission chain. WGS data can inform and enhance spatial analyses, providing further insight into pathogen transmission. For example, a genomic investigation of typhoid in rural Cambodia revealed that genetic groups of *S.* Typhi are distinctly spatially clustered [49].

Blantyre, Malawi experienced rapid emergence of multi-drug resistant (MDR) typhoid fever beginning in 2011 [46]. Despite ongoing transmission, the dominant transmission pathways remain unknown. A case-control study revealed complex risk factors related to both WASH (i.e. river water usage) and social exposures (i.e. school/ daycare attendance) [105]. In this study, we explore the spatial and genomic patterns of typhoid transmission in Blantyre through a cohort study, in order to further characterize incidence patterns, and disentangle transmission occurring in the city.

3.2 Methods

3.2.1 Setting and case ascertainment

Any patient diagnosed with culture-confirmed typhoid fever presenting to QECH, Blantyre, Malawi was recruited to the prospective observational cohort between April 2015 and January 2017. Informed written consent was sought from adult participants and from the legal guardians of children. Questionnaires were used to record age, residential area, HIV status, in- vs outpatient treatment, clinical presentation, complications and outcome using Open Data Kit (https://opendatakit.org/). Residential location and location of any

household water source used within three weeks prior to diagnosis was recorded for children under the age of 9 beginning in April 2015, and for all cases beginning in August 2015. Garmin Etrex GPS devices were used for geo-location of households of cases that were enrolled in the nested case-control study [105], while the electronic PArticipant Locator application (ePAL) was used to geo-locate the households for the remainder of the cohort [106].

3.2.2 Incidence mapping

We aimed to describe the incidence of culture confirmed typhoid fever associated with presentation to QECH. The denominator was derived from a 2016 census of Blantyre and surrounding areas [107], dividing the urban catchment area of the ePAL system into 275 enumeration areas (EAs). This census included population structure by age band, and number of households, along with shapefiles for each EA. There are now numerous approaches to adjusting incidence of typhoid fever based on sensitivity of diagnostic and health care utilization [3,108], however as none are standardized, no adjustment has been used in this study.

All statistical analyses were conducted using R statistical software, version 3.5.1 [89]. In order to model incidence across the city, a Poisson log-linear model with a spatial random effect was fitted using the PrevMap package [12]. Rates were estimated for each EA and age band, with estimated population size in each age band as an offset. Covariate effects were explored, including distance from the centroid of each EA to QECH, average household size, population density per square kilometer, elevation at the centroid, ascertained from digital elevation model (DEM) data [110], and hydrological catchment, extracted using ArcGIS (Supplementary Material 3.1). The statistical model is further described in Supplementary Material 3.2.

3.2.3 Sequencing & Phylogenetic analysis

In order to investigate the genetic patterns of *S.* Typhi in this study, isolates of MDR *S.* Typhi were whole genome sequenced on Illumina HiSeq2500 machines generating 150 bp paired-end reads. Reads were mapped against the high-quality reference genome of *S.* Typhi 1036491 isolated in Blantyre, Malawi in 2012. An alignment was generated selecting only sites containing ACGT (no gaps or Ns) using snp_sites [111]. A pairwise SNP matrix was generated from this alignment and was used for spatial modelling. For further phylogenetic analyses, recombinant sites and mobile elements were removed following analysis of the mapping-based alignment and phage characterization.

The phylogeny was reconstructed using iq-tree [112] under the general time-reversible model. Ascertainment correction was done for the SNP-only alignment, and support was assessed using 1000 bootstrap replicates. The resulting tree was assessed for phylogenetic signal using tempest (v1.5.1;[113]) and root-to-tip correlation was calculated. The phylogenetic tree was reconstructed into a joint ancestral tree, and rPinecone [114] was used to further group the isolates based on this tree, using 2 and 4 as relevant SNP cutoffs for minor and major clusters, respectively. Detailed protocols for genomic analyses are found in Supplementary Material 3.3.

3.2.4 Spatio-genetic modelling

Firstly, we tested for correlation between SNP distance and spatial distance by comparing the correlation coefficient calculated from the data with those calculated after random permutations of the household locations. Next, the pairwise distance matrix of all absolute differences of SNPs was mapped to two dimensions using multidimensional scaling (MDS), with the principal coordinate values henceforth referred to as genetic scores. We used a linear model with a

spatial random effect to predict genetic score across the city. Finally, based on results of a risk factor study from a subset of this cohort that points to river water as a potential exposure [105], we explored the ability of river catchment to predict genetic score. Analyses were conducted using R statistical software, version 3.5.1 [89] and the PrevMap package [12] (Supplementary Material 3.4).

3.3 Results

3.3.1 Characteristics of cohort

S. Typhi was isolated from 658 blood cultures between March 28, 2015 and January 12, 2017 (Figure 3.1), with an additional 2 isolates obtained from cerebrospinal fluid. 641/660 (97%) of all isolates were multidrug-resistant to ampicillin, chloramphenicol and cotrimoxazole. 12 isolates (1.8%) were susceptible to all tested drugs. 542 cases were recruited to the cohort study, of whom 314 cases consented to provide their household locations, and 256 MDR isolates were whole genome sequenced (Figure 3.1).

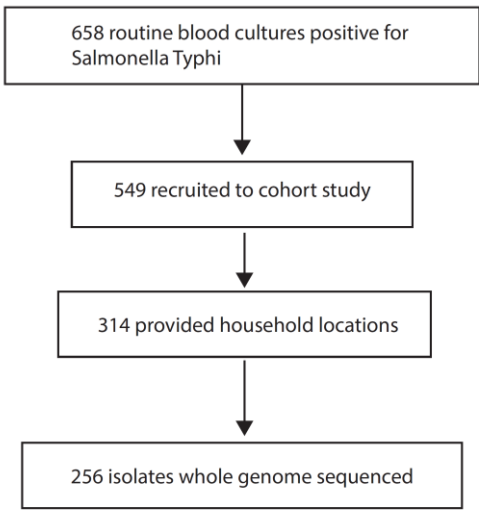


Figure 3.1 Consort chart showing individuals recruited to study.

The characteristics of the cohort are summarized in Table 3.1. The median age was 11 years (IQR: 6-19), and the HIV seropositivity rate was 10.7% (37/346). 391/542 (73%) patients were hospitalized, and hospital records were retrieved for 326. The most frequently reported modes of presentation were fever and gastrointestinal complaint (abdominal pain and/or diarrhea and vomiting) in 45%, and non-focal febrile illness in 43%. The fatality rate of the cohort was 1.5% (8/520).

Table 3.1 Characteristics of the recruited cohort.

Characteristic	**Value**
Age, median years (range)	11 (6-19)
Female, n (%)	256/542 (47.2)
HIV reactive or exposed, n (%)	37/346 (10.7)
Malaria test positive, n (%)	7/533 (1.3)
Living in urban Blantyre, n (%)	484/542 (89)
Death, n (%)	8/520 (1.5)
Admitted, n (%)	391/542 (72.1)
Length of hospital stay, median (IQR)	4 (3-7)

3.3.2 Incidence mapping

314 cases provided household locations, with 17 cases occurred outside of enumeration area bounds. This resulted in 297 of the 658 blood-culture confirmed cases recruited to the cohort being included in the geostatistical incidence model. The sensitivity of blood culture as a diagnostic test is known

to be approximately 60% [5], therefore our modeled estimates represented estimated minimum incidence rates only. Total predicted incidence in each enumeration area is plotted in Figure 3.2. The model predicts the highest risk in the <5 age band, followed by the 5-14 age band (Table 3.2). Of the evaluated covariates, average household size was a significant predictor of incidence (Table 3.2), while other measured covariates did not significantly improve the model (Supplementary Material 3.2). Distance from hospital was not a significant predictor of incidence, suggesting that distance from care may not affect health-seeking behavior across the city for these severe cases.

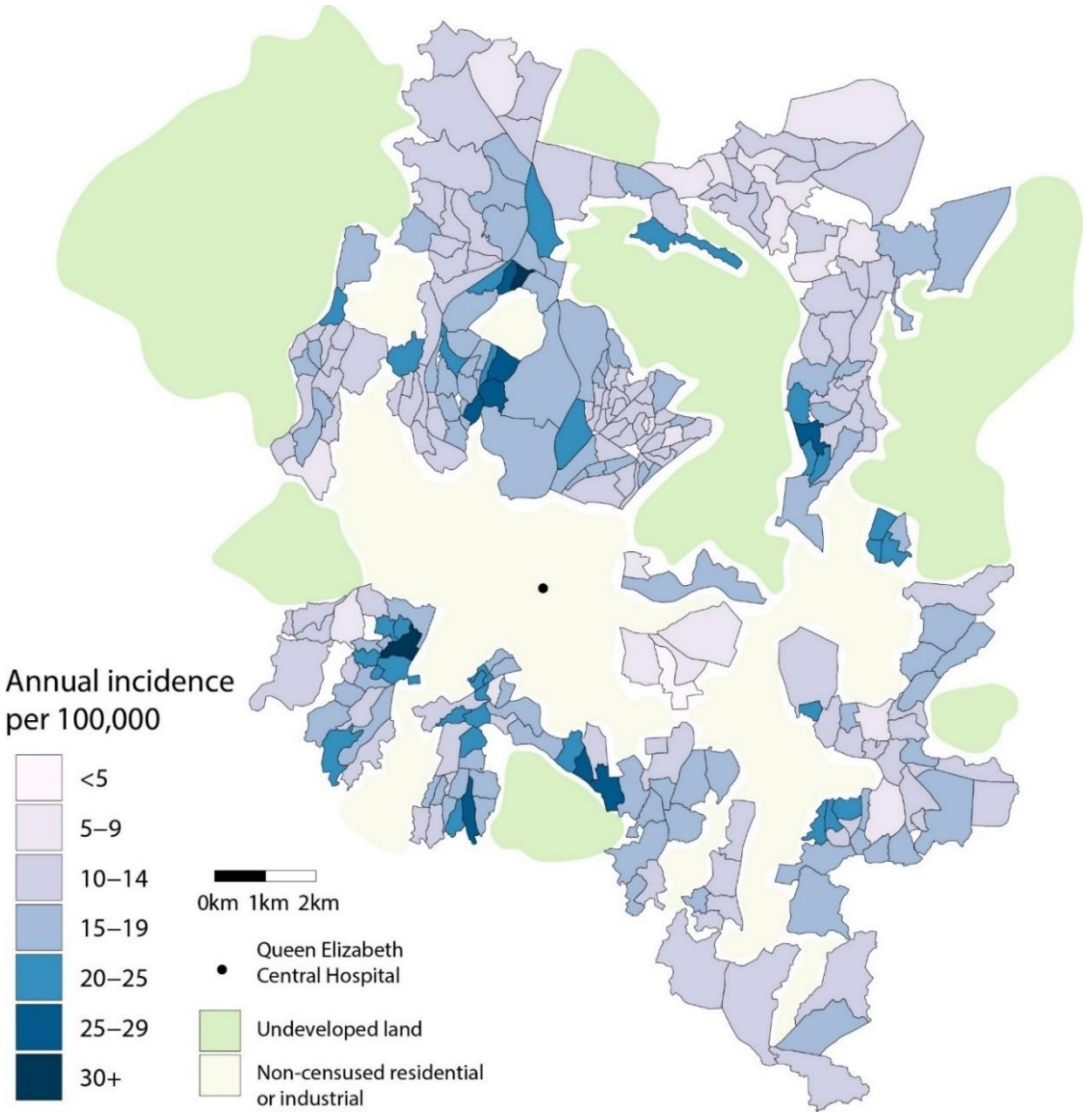


Figure 3.2 Estimated incidence rate for enumeration areas in Blantyre.

Incidence across Blantyre is geographically heterogenous, with an overall estimated incidence rate of 14.7 per 100,000 per year, reflecting a likely lower bound on incidence in the region. High incidence regions exist within the city, with 11 enumeration areas predicted to have an incidence greater than 30 per 100,000 whilst 4 enumeration areas have an incidence rate of less than 5 per 100,000. Small-scale spatial correlation is estimated to occur in the region, with the estimated range of spatial correlation (phi) at 528 meters (Table 3.2). This indicates a practical range of spatial correlation (>5% correlation) reaching approximately 1600 meters. Areas of higher or lower risk that expected given their covariate values are identified (Supplementary Material 3.2).

Table 3.2 Parameter estimates for geostatistical incidence model.

Parameter	**Estimate**	**Standard error**	**P-value**
Intercept	-5.25	0.560	<0.001
Average household size	-0.829	0.129	<0.001
Age 5-14	1.108	0.076	<0.001
Age <5	1.168	0.075	<0.001
log(sigma^2)	-1.797	0.243	-
log(phi)	6.269	0.331	-
log(tau^2)	-0.251	0.510	-

3.3.3 Genomic epidemiology

MDR isolates of 256 patients were whole-genome sequenced and all were H58 haplotype. Root-to-tip correlation (0.07) indicated insufficient temporal signal to allow a temporal analysis (Supplementary Material 3.3). Whilst no visual

association with river catchment can be observed, six distinct genomic clusters of isolates were identified (Fig 3.3A).

Significant correlation between SNP distance of isolates and physical distance of household locations was observed in the data ($p = 0.001$) (Supplementary Material 3.4). The first two principal coordinates (PCs) resulting from the multivariate analysis account for 38% of the variation in the SNP matrix (Figure 3.3B). Scores along primary axis (PC 1) are similar for the majority of the cohort, with the exception of 11 isolates whose genetic score is approximately -15, suggesting a distinct genetic group. These individuals were also observed to be clade 6 in the tree (Figure 3.3A). Although genetically distinct, these 11 individuals do not appear to form a distinct spatial or temporal cluster compared to the rest of the cohort (Supplementary Material 3.4), so it is unclear how they are related.

PC 1 shows no evidence of spatial correlation (Supplementary Material 3.4), indicating that, although this axis reflects an aspect of genetic relatedness, it may not be relevant to the spatial component of the genetics we are aiming to observe. Genetic scores are more evenly distributed along the secondary axis (Figure 3.3B). Further, there appears to be spatial correlation of these scores approaching 2500 meters, as seen from the empirical variogram (Figure 3.3C), and confirmed by statistical test (Supplementary Material 3.4). We therefore fit the linear geostatistical model to PC 2.

Blantyre has complex river network (Figure 3.4A), and we have previously identified use of river water in the household as a risk factor for typhoid [105]. Using river catchment as a categorical predictor in the linear model significantly improves the model's fit to the genomic patterns observed compared to an intercept-only model (LL -301.9 vs. -289.4, $p = 0.003$).

Parameter estimates indicate similar genetic scores for individuals in catchments 2 and 8 (Table 3.3), distinct from the rest of the catchments. To confirm this observation, we conducted a contrast test to compare the mean coefficient values between catchments 2 and 8, and the rest of the river catchments. The difference was significantly different from zero as evaluated using a t-test ($p<0.001$).

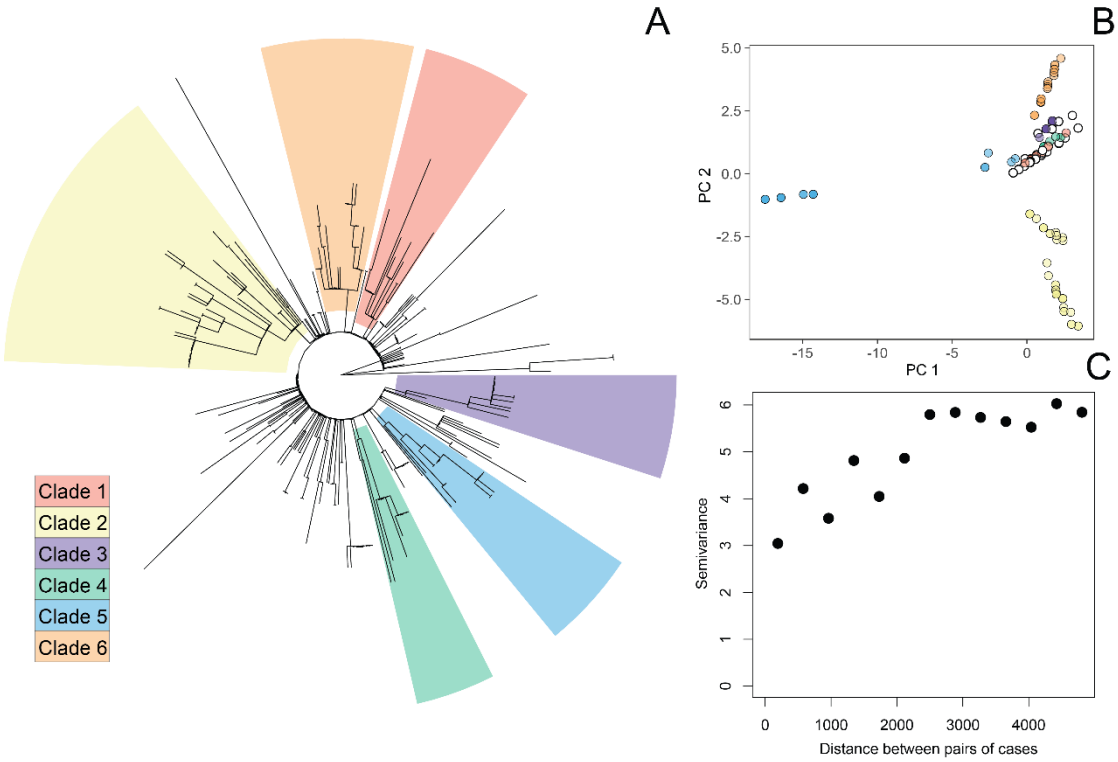


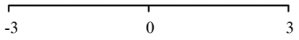
Figure 3.3 A. Tree showing major clades, B. Decomposition of SNP matrix into the first two 2 principal coordinates of the multidimensional scale, points colored by membership of major clades corresponding to the tree C. Empirical variogram of PC 2 of SNP distance matrix.

Estimates of spatial correlation of the geostatistical model highlight the multiple scales of spatio-genetic clustering. The range parameter, phi, indicates that the practical range of spatial correlation (the distance at which the spatial correlation decays to 0.05) is approximately 192 meters, indicating the model's

spatial random effect is capturing small geographic-scale spatial correlation. Though the city's geographical range spans approximately 20 kilometers, households in the cohort are clustered, with 59% of the cohort having another cohort member within a distance of 192 meters. We conducted a sensitivity analysis using geo-located water sources instead of household locations. As the majority of individuals lived within close proximity of their water sources, the results were consistent with the findings using household location (Supplementary Material 3.4).

Table 3.3 Estimated parameters for geostatistical genetic model.

Catchment	Genomic score	95% CI
1	0.091	
2	−1.240	
3	1.301	
4	0.308	
5	0.496	
6	0.353	
7	1.084	
8	−1.092	
9	0.807	
10	0.640	



Parameter	Description	Estimate	Std. error
$sigma^2$	spatially correlated variance	4.116	1.106
phi	range of spatial correlation	40.496	1.119
tau^2	nugget (nonspatial) variance	0.165	1.859

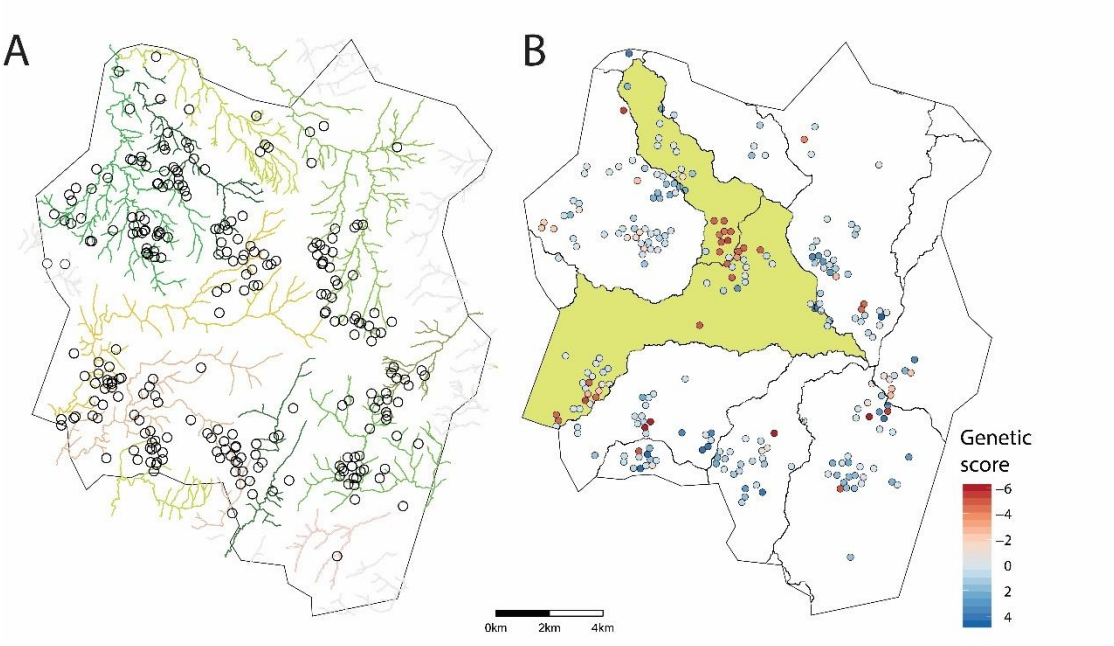


Figure 3.4 A. Major rives in Blantyre with household locations of cases, B. Genetic score and river catchments delineated, with catchments 2 and 8 highlighted in yellow. Precise locations of households are masked by randomization and overlapping points have been jittered for visualization.

3.4 Discussion

We describe a multi-faceted epidemiological investigation, enhanced by high resolution genotyping and geostatistical modelling. Mapping the cohort first enabled us to explore spatial patterns of incidence. The results reveal unexpected heterogeneity in incidence of typhoid fever in the context of a generalized epidemic across the city of Blantyre. A single incidence estimate for Blantyre would mask areas of both high and low incidence rates within enumeration areas across the city. This is relevant for considerations of control with the conjugate vaccine: in the case of limited resources for vaccination, these high-incidence regions can be prioritized. Additionally, the cost-efficiency of the vaccine decreases with lower incidence rates [115], and providing data indicating

higher incidence regions within city may help motivate policy-makers to support a vaccination strategy.

Useful and generalizable incidence covariates are not commonly known for typhoid [3]. Out of the covariates tested in our study, only average household size was significantly predictive of risk. This contrasts with other studies that have found increased risk in lower elevation areas [51]. We do not have evidence of decreasing incidence with distance to the hospital (within the city), which is consistent with the severity of the cases captured in this cohort of individuals presenting to the hospital. The increased incidence with decreasing household size, after accounting for population density and controlling for differential age distributions of EAs, may be a consequence of younger families living in greater socioeconomic precarity. Further work into exposure-related factors that would be associated with household composition is needed.

It is important to note the incidence rates presented are a lower-bound of the underlying burden for a number of reasons. Case ascertainment for typhoid is highly dependent on the surveillance framework: Our hospital-based surveillance is likely only capturing the most severe of cases, while an active surveillance framework would be more efficient in detecting mild or sub-clinical illness in the community, leading to higher estimates [116]. Further, we do not adjust rates by the sensitivity of blood culture, which is approximately 60%, but varies based on blood culture volume [5]. Data on sample volumes were not available in this study. Finally, the differential time periods for mapping the recruited cohort vs. cases under the age of 9 is leading to an under-estimate of incidence rates, however is not likely to affect the spatial patterns observed.

While reconstruction of phylogeny was successful in identifying discrete clades of *S.* Typhi, much information is masked by reducing the genetic data to

categories in this way. Instead, we incorporated the full spectrum of genetic variation by starting with all-against-all SNP distance followed by multidimensional scaling, which created a continuous variable for further analysis. Modelling the principal coordinates extracted from the SNP- matrix enabled us to view a continuous representation of genetic relatedness spatially, as well as test the predictive power of spatial covariates. Results from the spatio-genetic modelling show a high correlation of spatial distance with genetic relatedness. This contrasts with data from Kathmandu, Nepal, where haplotypes were not distinctly clustered [62]. However, SNP-typing, not WGS, was performed in this previous analysis, possibly masking more subtle clustering at non-sequenced locations in the genome.

The spatial-genomic clustering identified in the current study supports the results from a recent case-control analysis in Blantyre, which evaluated risk factors from a subset of the current cohort [105], and identified cooking and cleaning with river water as a potential exposure pathway. This case control study provided the empiric evidence for our current analysis, in which adding river catchment to our spatial-genomic model resulted in a significantly improved fit. This offers evidence that transmission within hydrological catchments may be occurring, and is consistent with emerging data from Fiji, showing heterogeneity of disease between hydrological catchments [117].

However, small-scale spatial correlation existed in the model even accounting for river catchment, which reflects the complexity of typhoid transmission. This may relate to social exposure factors such as attendance at school or daycare, additionally identified in the risk factor study in Blantyre [105], but the data needed to test this hypothesis do not currently exist.

There are limitations to the spatial-genomic analyses. Our single tested spatial covariate in the spatial-genomic analysis was hydrological catchment. Although our hypothesis that these were critical to transmission was grounded in previous work [105], we are missing predictors reflecting broader social interactions such as school attendance or food market, which might also have explained the spatial clustering seen, and should be investigated in future studies. The modelling of 'genetic score,' extracted through principal coordinates analysis, has allowed a flexible framework to explore the spatial patterns of genomics. However, some of the genetic signals present may be missed through this analysis due to the reduction of the genetic data to a two-dimensional space. We are additionally only sequencing a subset of the most severe cases; to add further granularity to our data and greater depth to our understanding of typhoid transmission, we would need representative geo-located isolates from mild cases and asymptomatic carriers, as well as data from environmental surveillance.

Currently, typhoid conjugate vaccines are being introduced in areas with known typhoid transmission, including Blantyre. However, additional interventions are likely to be necessary for sustained control, and identifying precise intervention points for water and sanitation interventions is often a challenge without detailed risk factor studies. Pairing spatial and genomic data has helped to identify that the rivers of Blantyre play a role in typhoid transmission, and this finding can help inform targeted inventions for typhoid control in this setting. The development of methods for rapid detection of *S.* Typhi in the environment will be critical to support the planning of public health interventions to interrupt transmission in the future.

4 Rainfall anomalies and typhoid fever in Blantyre, Malawi

Jillian S. Gauld[1,2], Peter J. Diggle[2], Nicholas A. Feasey[3,4], Jonathan M. Read[2]

1. Institute for Disease Modelling, Bellevue, USA
2. Centre for Health Informatics, Computing, and Statistics, Lancaster University, Lancaster, UK
3. Department of Clinical Sciences, Liverpool School of Tropical Medicine, Liverpool, UK
4. Malawi-Liverpool Wellcome Trust Clinical Research Programme, Blantyre, Malawi

Abstract

Typhoid fever is a major cause of febrile illness in developing countries. The final stage of transmission is via the fecal-oral route, however intermediate environmental pathways are poorly understood. The interaction between weather events such as rainfall and typhoid fever may offer insight into these roles. We investigate this relationship in Blantyre, Malawi, where multi-drug resistant typhoid fever has been transmitting since 2011.

We examined cross-correlations of rainfall and detrended typhoid fever, and utilized a quasi-Poisson generalized linear modelling framework to explore the predictive power of rainfall anomalies on typhoid fever. We found that the peak in rainfall precedes the peak in typhoid fever by approximately 15 weeks, a lag that does not indicate a direct biological link. However, when exploring anomalies in rainfall (either more or less rain than expected), we found a significant protective effect of anomalous rainfall on typhoid fever, at a two-week lag.

The extended lag between rainfall and typhoid fever, far exceeding the incubation period of the infection, indicates the existence of an unknown intermediate step in the transmission pathway. The significant protective effect of rainfall anomalies at a two-week lag suggests inordinate rainfall may cleanse the environment, while less than usual rainfall may prevent fecal material from washing into either environmental systems such as rivers, or directly into households.

In summary, rainfall anomalies may be protective in Blantyre. However, this relationship may change from location-to-location depending on the sewage infrastructure and drinking water quality. These results can help to better

understand the environmental mediation of typhoid transmission, and offer insights for future water and sanitation intervention strategies.

4.1 Introduction

Typhoid fever, caused by the bacterium *Salmonella* Typhi is a major cause of febrile illness in developing countries, with 10-20 million cases occurring annually [4]. Individuals can be exposed through direct interaction with infected individuals, via food handling or contamination of other fomites. Additionally, exposure to *S.* Typhi can be mediated by the environment, i.e. contaminated drinking or household water, sewage, or food. Although *Salmonella enterica* serovar Typhi is a human restricted pathogen, its behavior in the environment after fecal excretion remains obscure.

In many locations with ongoing transmission of *S.* Typhi, the specific mechanisms of environmentally mediated, or 'long-cycle,' transmission are unknown. In Chile, irrigation of crops with wastewater was identified as risk factor for typhoid. After this practice was banned, typhoid incidence declined to near-elimination levels [19]. In Nepal, transmission through drinking water was posited, and further bolstered by environmental sampling [24]. Understanding these pathways is important for designing non-vaccine control measures.

As climate is a key determinant of environmental conditions, the impact of weather events, such as rainfall on typhoid, could help to identify the environmental drivers of transmission in endemic locations. Further, if a link to a weather pattern is established, this may help to predict fluctuations in disease incidence. Currently, however, we do not understand the impact of rainfall on typhoid. Flooding or extreme rainfall events may overwhelm pit latrines or other forms of waste management, whilst drought conditions may offer opportunities

for contamination of drinking water through negative pressure facilitating leaks into water pipes.

Because many diseases exhibit seasonal dynamics, cross-correlation of disease incidence and weather variables is frequently assumed regardless of whether a mechanistic link occurs, thus establishing a causal or mechanistic link is challenging. Typhoid is known to be seasonal [70], therefore it is unsurprising that seasonal rainfall is also correlated with disease incidence. In Dhaka, a 3-5 week lag in rainfall was associated with an increase in typhoid cases [50]. In a multi-site investigation, it was observed that rainfall often precedes the disease, and a positive association with temperature is frequent [70], however this was not a universal finding across the evaluated study sites.

Where cross-correlation is a given, time series analysis can be helpful in establishing causality, in particular by considering the relationship between disease and weather anomalies. Because weather anomalies are not predictable in an annual/ seasonal way, this removes the issue of an expected cross-correlation, as long as the extreme events are identified effectively and well characterized. For example, more or less rainfall than expected in a given season predicting more or fewer cases than expected may be more convincing than a seasonal lag, when establishing a causal link. This approach has been used in establishing a link between rainfall events and diarrhea [69]. Methods of identifying these extreme events and incorporating them into a model vary, but can include exceedance of a pre-specified threshold, or by extracting residuals from a smoothed (de-seasonalized) model.

There is a distinction between (a) association between rainfall and incidence and (b) association between rainfall and incidence anomalies. For incidence and rainfall associations, we are attempting to explain the entire

seasonal pattern of cases with rainfall. For weather anomalies, we are only attempting to explain more or less than expected cases beyond an expected seasonal pattern, with the precise driver of the seasonal component left unknown. These different association types could lead to differing hypotheses regarding the impact of rainfall on transmission of the disease.

Since 1998, Queen Elizabeth Central Hospital in Blantyre, Malawi has conducted blood culture surveillance for typhoid fever. A sharp increase in reported cases occurred in 2011, the majority of which were multi-drug resistant [46]. Despite ongoing transmission, the mechanisms of transmission remain unknown. A risk factor study conducted in 2015 suggested complex interactions between environmental and common social exposures, including using river water for cooking and cleaning [105]. 59.6% of the population in Blantyre use non-flushing latrines, and it has been noted that the rocky soil in Blantyre often prevents the digging of pit latrines deeper than three meters [45], providing a hypothesis for a mechanistic link between rainfall events and subsequent contamination of river water or the surrounding environment. The goals of this study were twofold. First, we aimed to characterize the seasonality of typhoid with respect to rainfall. Second, in order to identify a possible causal pathway, we explored whether typhoid incidence could be predicted by rainfall.

4.2 Methods

4.2.1 Data and cleaning

Data was available between 1998-2017 from laboratory records from Queen Elizabeth Central Hospital. Anyone blood culture positive for *S.* Typhi, collected through routine hospital-based surveillance on both inpatients and outpatients, was recorded. We obtained weather data from the Malawi Meteorological Service, which included daily measurements of rainfall (mm).

Due to reporting and laboratory time lags based on the day of the week, we summarized the data into weekly counts of cases and weekly average rainfall. All data processing and subsequent analyses were conducted in R version 3.5.1 [89]. As there were limited cases of typhoid prior to 2012, to characterize the effect of rainfall on endemic transmission, analyses used information beginning January 1, 2012.

4.2.2 Modelling typhoid cases

We first modelled the time series of typhoid cases. Because we know typhoid cases are seasonal, and exhibited a large increase in 2011, we needed to incorporate both a seasonal term and a smooth time-trend. We did not attempt to explore any predictive drivers of the increase in 2011, as this has been explored previously through a dynamic modelling framework. That study attributed the rise in cases to an increase in shedding duration, possibly caused by multi-drug resistance [47]. We used a quasi-Poisson log-linear model [equation 4.1], which allows us to model typhoid case-counts over time while accounting for over-dispersion. We use the penalized regression spline (the default in *mgcv* package for the R statistical programming language) and an annual seasonal harmonic [equation 4.1].

$$\mathrm{E}(\mathrm{Y}_t) = \mu_t, \;\; \mathrm{Var}(\mathrm{Y}_t) = \phi\mu_t$$

$$\mu_t = \exp\left(\alpha + \beta_1 \cos\frac{2\pi t}{52} + \beta_2 \sin\frac{2\pi t}{52} + \mathrm{spline}(t)\right) \qquad [4.1]$$

4.2.3 Modelling weather and defining anomalies

In order to define weather anomalies, we needed to be able to predict an 'expected' amount of rainfall throughout our study period. We utilized a joint model with two components. First, we modeled the amount of rain on days with rainfall using a log-linear model with annual and six-month harmonic terms to describe the seasonal effect [equation 4.2].

$$m(t) = \log(rain_t) = \alpha + \beta_1 \cos\frac{2\pi t}{52} + \beta_2 \sin\frac{2\pi t}{52} + \beta_3 \cos\frac{4\pi t}{52} + \beta_4 \sin\frac{4\pi t}{52} + \epsilon \qquad [4.2]$$

The six-month harmonic terms in [4.2] were needed to capture the shape of the seasonal variation. Next, we modeled the probability of rainfall in any given week using logistic regression, including the same annual and six-month harmonic terms [equation 4.3].

$$f(t) = \log\frac{p}{(1-p)} = \alpha + \beta_1 \cos\frac{2\pi t}{52} + \beta_2 \sin\frac{2\pi t}{52} + \beta_3 \cos\frac{4\pi t}{52} + \beta_4 \sin\frac{4\pi t}{52} \qquad [4.3]$$

The expectation of total rainfall on any given day is therefore:

$$\mathrm{E}[rain(t)] = f(t)\ \exp\left(m(t) + \frac{\sigma^2}{2}\right) \qquad [4.4]$$

With σ^2 estimated from the fitted rainfall model [equation 4.2]. A rainfall anomaly was then defined, for each week in the study period, as the observed rainfall minus the expected rainfall.

4.2.4 Describing seasonal patterns

We examined cross-correlations of average weekly rainfall and typhoid fever cases, in order to characterize seasonal trends in relation to weather events in the raw data. Cross-correlations were generated between de-trended case counts, retaining the seasonal component, and average weekly rainfall, for lags spanning 0-24 weeks.

We then aimed to estimate the lag between the seasonal peaks of case incidence and rainfall. We generated 1000 realizations of model parameters using the multivariate normal sampling distribution of the parameter estimates for the fitted typhoid [equation 4.1], and rainfall [equation 4.2, 4.3] models. We then extracted the timing of the seasonal peaks for cases and rainfall from model predictions using these parameters. Finally, we took the difference in seasonal

peaks for each set of realizations to estimate the lag between cases and rainfall. We summarized the lag in terms of mean and 95% confidence intervals.

4.2.5 Predictive model

To explore the possibility of a causal relationship between rainfall and typhoid anomalies, similar to the case series, we used a quasi-Poisson log-linear model [equation 4.5].

$$\mathrm{E}(Y_t) = \mu_t, \;\; \mathrm{Var}(Y_t) = \phi\mu_t$$

$$\mu_t = (Offset_t) * \exp\left(\alpha + \beta_1 w_{t-1} + \beta_2 w_{t-2} + \beta_3 w_{t-3} + \beta_4 w_{t-4}\right), \qquad [4.5]$$

where w_s is the rainfall anomaly for week s.

This model accounted for the overall trend in cases by using the fitted values from the model described in equation 4.1, which includes both seasonal and time-trend components, as an offset term. Using this offset, we are only predicting case 'anomalies.'

We included terms for rainfall anomalies, as defined above, at lagged weeks 1 to 4. This range of lags was based on the known incubation period of typhoid [118], and allowing for potential delay in healthcare seeking and case identification. We explored potential relationships between rainfall anomalies and case anomalies using the model [equation 4.5], in which the rainfall anomaly effects are log-linear, and the following extension that allows anomaly effects to be log-quadratic [equation 4.6].

$$Y_t \sim \mathrm{Poisson}\left(\mu_t\right)$$

$$\mu_t = (Offset_t) * \exp\begin{pmatrix}\alpha + \beta_1 w_{t-1} + \beta_2 w_{t-2} + \beta_3 w_{t-3} + \beta_4 w_{t-4} + \\ \beta_5 {w_{t-1}}^2 + \beta_6 {w_{t-2}}^2 + \beta_7 {w_{t-3}}^2 + \beta_8 {w_{t-4}}^2\end{pmatrix} \qquad [4.6]$$

We evaluated the overall contribution of the rainfall to the incidence model using a Wald test [119], which provides an indication of whether the included model parameters are estimated to be significantly different from zero.

4.3 Results

4.3.1 Case series model

The case series model [equation 4.1] with and without seasonal components is shown in Figure 4.1A. The de-trended seasonal case-counts are shown Figure 4.1B, and the de-trended, de-seasonalized residuals are shown in Figure 4.1C, representing typhoid anomalies with and without the seasonal component, respectively. The fit to our joint model for the occurrence and amount of expected weekly rain is shown in Figure 4.2A. We used this model to generate the rainfall anomaly sequence as observed minus expected weekly rain (Figure 4.2B).

4.3.2 Seasonal comparisons

Correlations between detrended case counts (Figure 4.1B) and rainfall were calculated, and are shown in Figure 4.3A. Visually, it is apparent that rainfall is highly correlated with case counts at lags between approximately 10 and 20 weeks (Figure 4.3A). We can additionally observe the lag with the fitted rainfall and case model predictions over a single year (Figure 4.3B). The estimated lag between the peak rainfall and cases was 15.46 weeks [95% CI 13.28, 17.65].

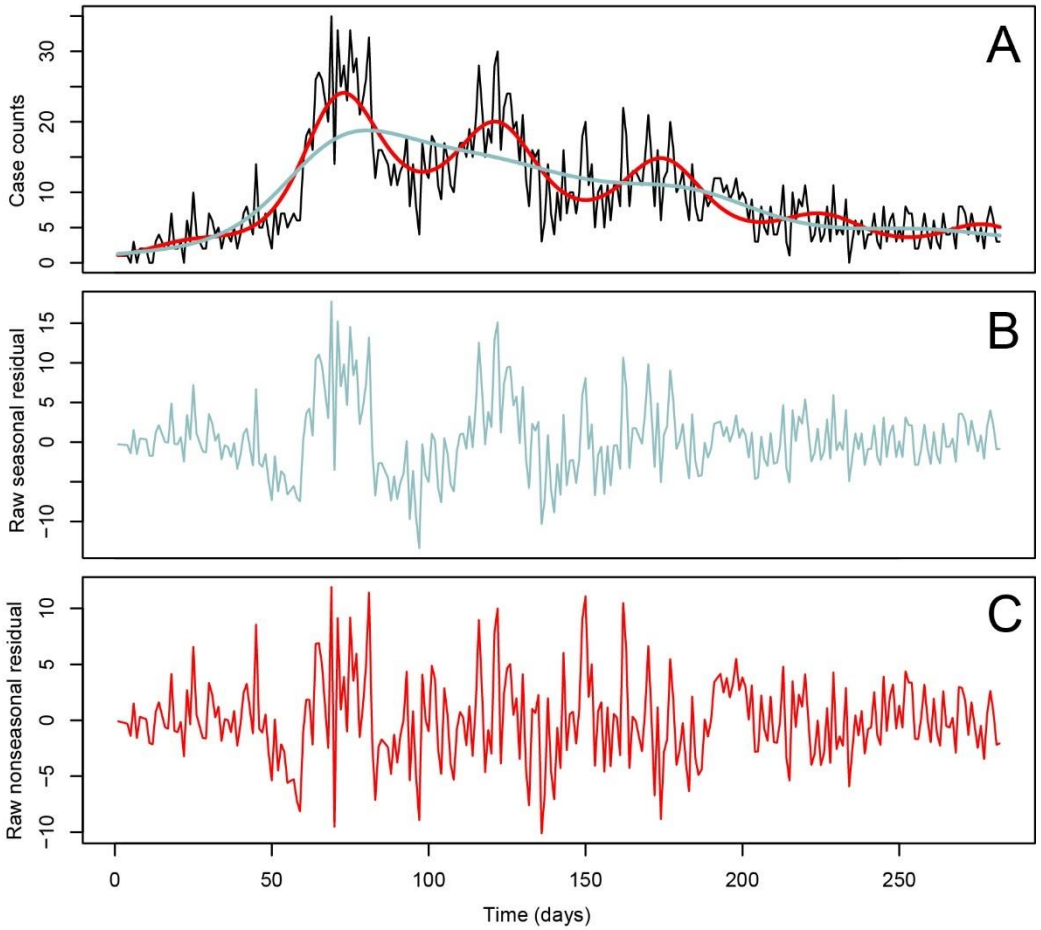


Figure 4.1 A. time series of case-counts (black), with long term trend (blue) and long term plus seasonal trend (red). B. Residuals from long-term trend model. C. Residuals from long term plus seasonal trend model.

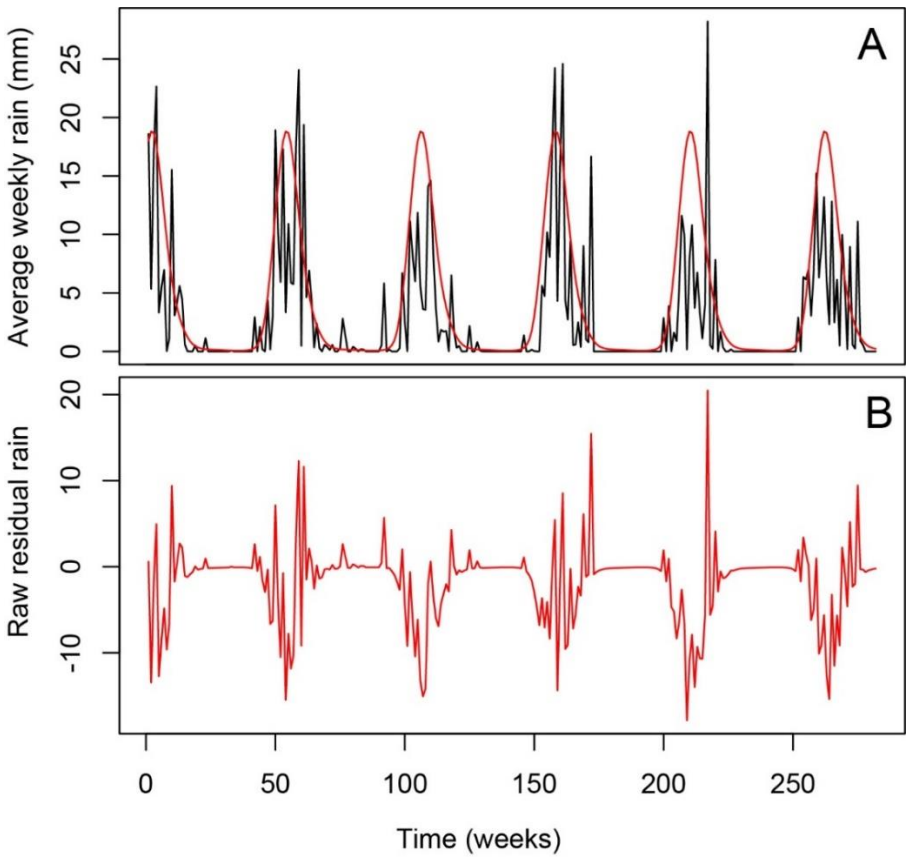


Figure 4.2 A. Average weekly rainfall (black), with fitted log-Gaussian model (red). B. Rainfall anomalies.

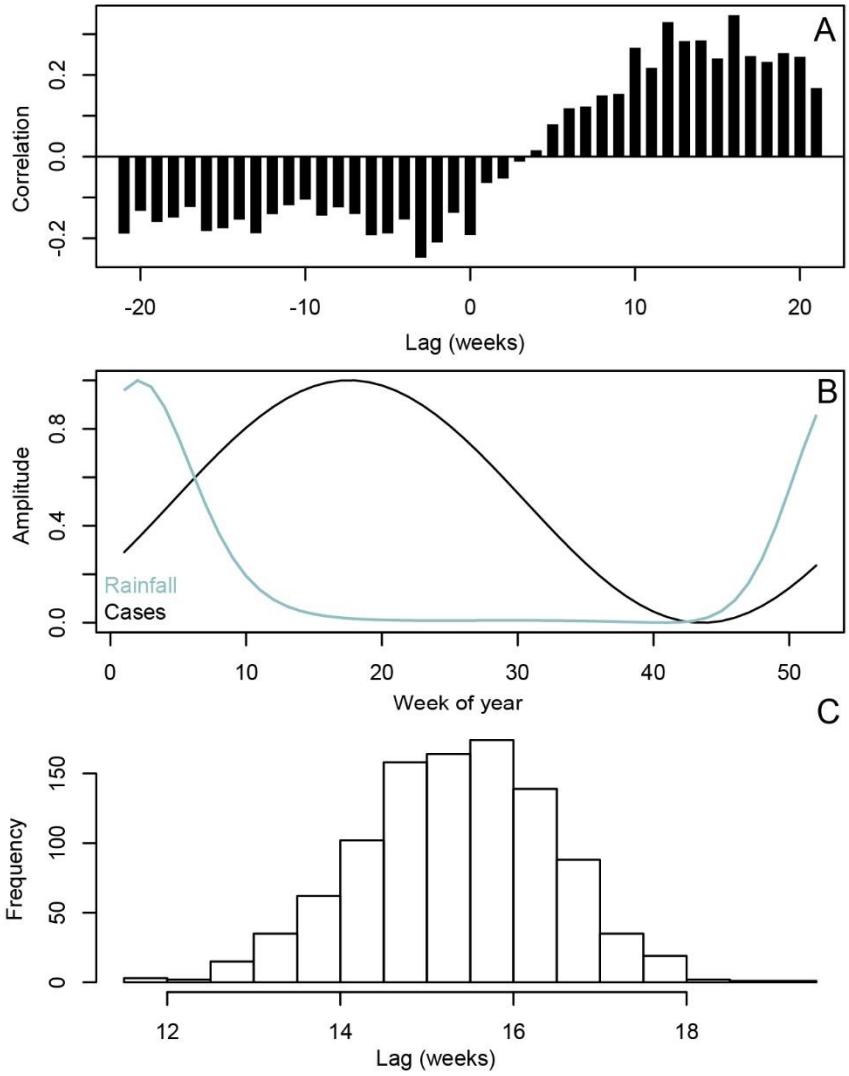


Figure 4.3 A. Cross-correlation of detrended cases and rainfall, B. Best-fit seasonal amplitude for cases (black line) and rainfall (blue line), C. Histogram of the calculated seasonal lags generated from 1000 realizations of the multivariate normal distribution parameterized by model covariates.

4.3.3 Predictive model results

Although model covariates for lagged rainfall anomalies were not found to be significantly different from zero assuming a log-linear relationship between rainfall and cases according to the Wald test [p = 0.178], marginal significance was found assuming a log-quadratic relationship [p=0.0524]. After investigating

the log-quadratic model [equation 4.6] further, observed that the lag of two weeks was highly significant [Table 4.1].

Table 4.1 Summary of estimates from log-quadratic model with all lags included.

Coefficient	Value	Std. err	P-value
Intercept	0.023	0.027	0.402
1-week lag rainfall anomaly	0.004	0.006	0.445
2-week lag rainfall anomaly	0.008	0.006	0.170
3-week lag rainfall anomaly	0.004	0.005	0.497
4-week lag rainfall anomaly	-0.002	0.005	0.727
1-week lag rainfall anomaly^2	0.0002	0.0005	0.622
2-week lag rainfall anomaly^2	-0.002	0.0006	0.006
3-week lag rainfall anomaly^2	-0.0003	0.0005	0.472
4-week lag rainfall anomaly^2	0.001	0.0005	0.144

We therefore re-ran the model including only the 2 week-lagged linear and quadratic coefficients, which resulted in a significantly improved fit of the model to the data compared to the null model, as assessed by the likelihood ratio test (scaled deviance =11.46, df=2, p = 0.003, Table 4.2). The negative coefficient of the quadratic effect indicates a concave effect, with low and high anomaly values being protective compared to medium (close to zero) anomaly values. The effect of rainfall anomaly on incidence rate predictions is visualized in Figure 4.4.

Table 4.2 Summary 2-week lagged quadratic rainfall anomaly model.

Coefficient	**Value**	**Std. err**	**P-value**
Intercept	0.039	0.025	0.123
2-week lag rainfall anomaly	0.007	0.005	0.165
2-week lag rainfall anomaly^2	-0.001	0.0005	0.005

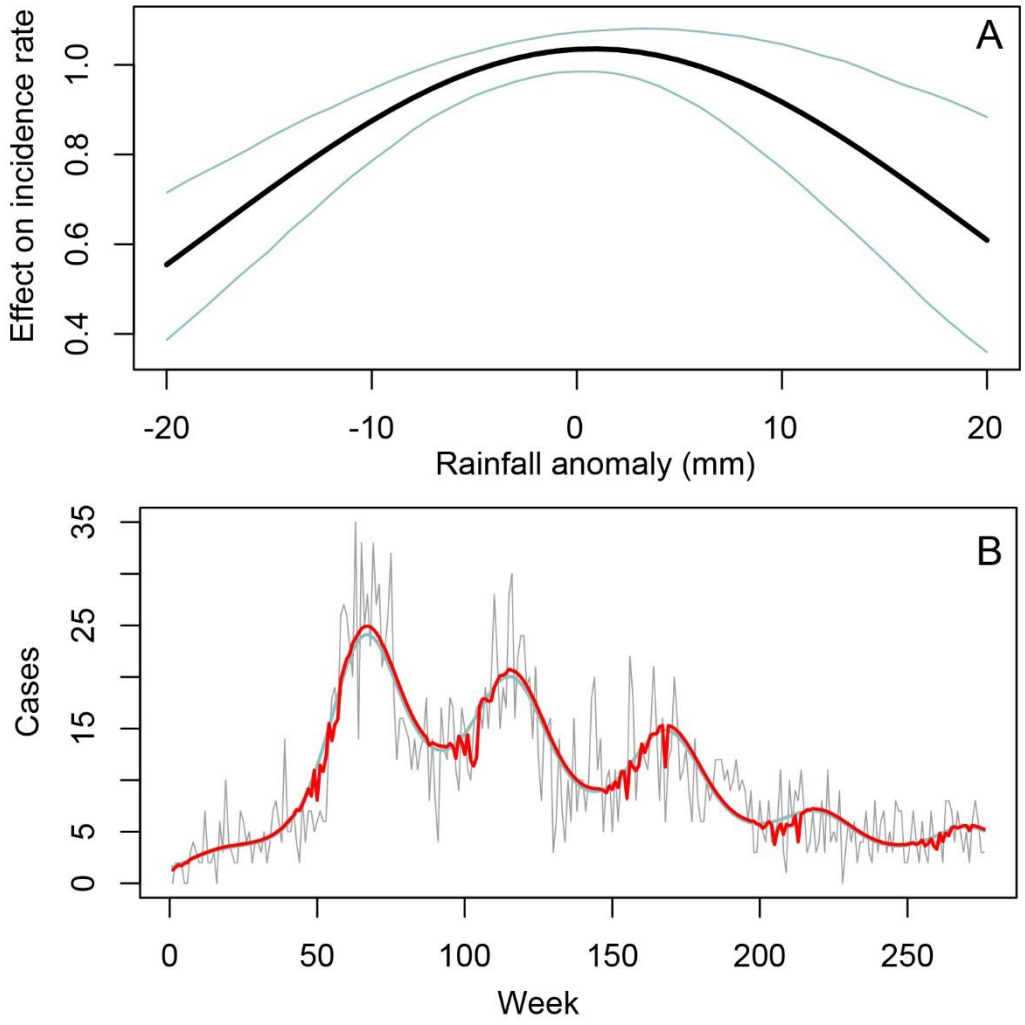


Figure 4.4 A. Predicted effect of 2-week lagged rainfall anomaly on case incidence, B. Model predictions with (red) and without (blue) rainfall anomaly included, and total cases in light grey.

4.4 Discussion

The pathway between shedding of *S.* Typhi and subsequent ingestion by an exposed individual is poorly understood. The primary reservoir of *S.* Typhi is humans, however it must survive in the environment long enough to permit transmission to the next human host. Therefore, rainfall may act as a mediator in this process. In this study, we aimed to explore the relationship between rainfall and typhoid in Blantyre, Malawi.

Both rainfall and typhoid cases exhibit seasonal patterns in Blantyre, and we found that the peak in rainfall precedes the peak in cases by approximately 15 weeks. Given that the incubation period of typhoid fever is typically between 1 and 4 weeks [21,118], this does not suggest that rainfall is a primary driver of typhoid incidence without an intermediate step, which is yet to be identified. We still aimed to explore the ability of rainfall anomalies to predict typhoid cases at biologically plausible lags of 1-4 weeks. We found a significant association between quadratic rainfall anomaly and typhoid cases, with the negative coefficient indicating that a larger or smaller rainfall anomaly than expected is protective.

It is biologically plausible that weeks with either more or less rainfall than expected are protective against typhoid transmission in different ways. If rainfall is a mechanism that disseminates *S.* Typhi and thus facilitates exposure to susceptible humans, for example through flooding of pit latrines, or runoff of sewage into rivers, it is plausible in this context that a dry period could be protective. Conversely, more rainfall than expected may have a cleansing effect on the environment, in that any pathogens present may exist in low, non-infectious concentrations. This protective effect of heavy rainfall has been reported for other enteric diseases, for example diarrheal disease following wet

periods in Ecuador [69]. A high infectious dose of *S.* Typhi is required to cause infection, so after a rainfall event, even if the environment isn't completely clean of the bacteria, individuals may be at a lower risk of developing typhoid fever [29,118].

Rainfall anomalies, however, are distinct from the potential effects of total overall rainfall, which was not correlated with typhoid cases within biologically plausible lags, and was therefore not included in a predictive model. For example, an "expected" amount of rain during any given week of the rainy season would not be counted as an anomaly in this model, only if there was more rain than expected according to the seasonal predictions. The mechanistic differences between total rainfall and rainfall anomaly are not well defined with regards to transmission of typhoid, thus it is plausible that the consistent and predictable seasonal cycles of rainfall are acting as a consistent mediator despite overall precipitation amounts, with anomalous weather events leading to protection through deviation from this consistent pattern. In contrast to our findings, the recent global burden estimates for typhoid found the proportion of the population living in the monsoon belt was a significant predictor of incidence, indicating these extreme events may put individuals at higher risk of typhoid fever [4]. However, these data are based on large-scale global models, and do not include the granular weekly predictions of our current model.

There are some limitations to this study. First, the time series of typhoid cases reflects the date of blood culture diagnosis of a patient, however the time at which an individual is infected precedes this by the incubation period, and to a lesser extent by individual variations in treatment seeking. We know that both of these factors are likely heterogenous: a large range of incubation periods have been found in challenge studies [118], additionally the geographic span of Blantyre (approximately 20 km), may indicate differential propensities to seek

care based on distance to the hospital. Therefore, our time series may not be representative of the date of infection. Further, we do not know what proportion of *S.* Typhi transmission is accounted for by short-cycle transmission (independent of the environment) and this may dampen the signal from rainfall/environmental interactions. It is important to highlight the dual approaches for seasonality exploration used in this paper. When two processes are seasonal, a significant correlation is almost always a given between them. When incorporating weather events as predictive processes, constraining lagged effects by known biological processes is critical for interpretation.

Overall, this study describes an extended lag between the seasonal patterns of typhoid fever and rainfall in Blantyre, Malawi. Although the study does not provide evidence of a directly causal linkage between total rainfall and typhoid fever, we do find evidence that rainfall anomalies (either more or less than expected) are protective. Improved data can help strengthen these observations, including prioritizing the detection of typhoid cases closer to their time of exposure, through active surveillance, and optimized environmental sampling and detection to understand the distribution of *S.* Typhi in the environment and over time. Further work to explore these relationships in other locations, and better understand the ecological niches of *S.* Typhi, will help advance our understanding of the link between weather patterns and typhoid transmission.

5 Intestinal perforations associated with a high mortality and frequent complications during an epidemic of multidrug-resistant typhoid fever in Blantyre, Malawi

Under review at Clinical Infectious Diseases

Franziska Olgemoeller[1,2], Jonathan J. Waluza[3], Dalitso Zeka[3], **Jillian S. Gauld[5,6]**, Peter J. Diggle[6], Jonathan M. Read[6], Thomas Edwards[2], Chisomo L. Msefula[1,7], Angeziwa Chirambo[1,7,8], Melita Gordon [1,8], Emma Thomson[3], Tiya Chilunjika[4], Robert S. Heyderman[9], Eric Borgstein[4], Nicholas A. Feasey[1,2]

1. Malawi-Liverpool Wellcome Trust Clinical Research Programme, Blantyre, Malawi
2. Department of Clinical Sciences, Liverpool School of Tropical Medicine, United Kingdom
3. Surgical department, College of Medicine, University of Malawi, Blantyre, Malawi
4. Surgical department, Ministry of Health, Queen Elizabeth Central Hospital, Blantyre, Malawi
5. Institute for Disease Modelling, Bellevue, Washington, United States of America
6. Centre for Health Informatics, Computing, and, Statistics, Lancaster University, United Kingdom
7. Pathology Department, College of Medicine, University of Malawi, Blantyre, Malawi
8. Institute of Infection and Global Health, University of Liverpool, United Kingdom
9. University College London, UK

Abstract

Typhoid fever remains a major source of morbidity and mortality in low-income settings. Its most feared complication is intestinal perforation. However, due to the paucity of diagnostic facilities in typhoid-endemic settings, including microbiology, histopathology and radiology, the aetiology of intestinal perforation is frequently assumed, but rarely confirmed. This poses a challenge for accurately estimating burden of disease.

We recruited a prospective cohort of patients with confirmed intestinal perforation in 2016 and performed enhanced microbiological investigations (blood and tissue culture, plus tissue Polymerase Chain Reaction (PCR) for *Salmonella* Typhi [*S.* Typhi]). In addition, we used a Poisson generalized linear model to estimate excess perforations attributed to the typhoid epidemic, using temporal trends in *S.* Typhi bloodstream infection and perforated abdominal viscus at Queen Elizabeth Central Hospital (QECH) from 2008-2017.

We recruited 23 patients with intraoperative findings consistent with intestinal perforation. 50% (11/22) of the patients recruited were culture- or PCR-positive for *S.* Typhi. Case fatality rate from typhoid-associated intestinal perforation was substantial at 18% (2/11). Our statistical model estimates that culture-confirmed cases of typhoid fever lead to an excess of 0.046 perforations per clinical typhoid fever case [95% CI: 0.03-0.06]. We therefore estimate that typhoid fever accounts for 43% of all bowel perforation during the period of enhanced surveillance.

The morbidity and mortality associated with typhoid abdominal perforations are high. By placing clinical outcome data from a cohort in the

context of longitudinal surgical registers and bacteremia data, we describe a valuable approach to adjusting estimates of the burden of typhoid fever.

5.1 Introduction

Typhoid remains a major public health problem in many low- and lower-middle-income countries (LMIC), with 10.9 to 17.8 million cases estimated to occur each year [3,4]. Whilst most cases present with non-focal sepsis [120], typhoid can be complicated by intestinal perforation [121]. Surgical complications of typhoid fever are well described and typically occur in the third or fourth week after onset of fever and typically arise from necrosis of Peyer´s patches in the terminal ileum [122]. Estimates of the case fatality rates of typhoid perforation remain high at 15.4% globally and a case fatality rate estimate of 20% for sub-Saharan Africa [123], with important regional differences ranging widely between 5% and 80% [124].

In cases of perforated abdominal viscus presenting in typhoid-endemic settings, the aetiological agent is often assumed to be *Salmonella* Typhi (*S.* Typhi), however this is rarely confirmed because there are few diagnostic microbiology facilities in LMIC [125]. Furthermore, publicly available datasets describing longitudinal trends in abdominal perforations in LMIC are rare [121,126–128]). Consequently, data describing "surgical typhoid" are not currently incorporated into global burden of disease (GBD) estimates of typhoid [129]. In the absence of these data, global burden estimates will underestimate the true morbidity and mortality of typhoid.

Routine, quality assured diagnostic blood culture facilities have been available at Queen Elizabeth Central Hospital (QECH), Blantyre, Malawi, since 1998. Until 2010, *S.* Typhi was an uncommon cause of bloodstream infection (BSI). Since 2011, however, there has been a substantial increase in the number

of culture-confirmed cases of typhoid at QECH, increasing from an average of 14 cases per year between 1998 and 2010, to 843 cases in 2013 [46]. Although QECH does not have the capacity to routinely identify the aetiological agent responsible for perforated abdominal viscus, the Department of Surgery has systematically recorded the occurrence of macroscopic perforations identified at laparotomy since 2008.

To identify the microbial cause, and to describe morbidity and mortality of perforated abdominal viscus associated with typhoid fever in this setting, we recruited a prospective cohort of patients undergoing laparotomy for suspected intestinal perforation at QECH. Further, we placed these cases in the context of longitudinal BSI and perforation surveillance data.

5.2 Materials and Methods

We prospectively recruited an observational cohort of patients presenting with perforated abdominal viscus to the QECH, the largest hospital in Malawi, which serves the city and district of Blantyre and acts as a referral hospital to 13 districts in the Southern Region of Malawi. Patients undergoing laparotomy for suspected typhoid perforation or with intraoperative findings deemed by the operating surgeon to be consistent with possible typhoid perforation between February 2016 and February 2017 were eligible for inclusion. Blood cultures were taken either on admission or in theatre and intraoperatively debrided tissue (debridement of perforated bowel edges, resected bowel, pus, lymph nodes) was retained for culture and DNA extraction. In critically ill patients unable to give consent at presentation, consent was sought postoperatively.

Microbiological samples were tested at the diagnostic microbiology laboratory of the Malawi- Liverpool-Wellcome Trust Clinical Research programme (MLW). Blood samples were incubated in an aerobic BacT/Alert

bottle (bioMérieux, Marcy l´Étoile, France) on an automated system and suspected Salmonellae were identified by biochemistry and antisera processed as previously described [88].

Tissue from intraoperative debridement was enriched in 9 ml of buffered peptone water and cultured overnight at 37 ° C in air. On Day 2, 2 mls of this broth was subcultured in sodium biselenite and again cultured overnight at 37 ° C in air. On Day 3, a 10µl loop was taken from the top of the broth and inoculated onto Xylose Lysine Deoxycholate (XLD) agar plates and cultured overnight at 37 ° C in air. Suspected *Salmonella* colonies were identified by biochemistry using API 20E tests and serotyped according to the White-Kauffmann-Le Minor scheme by the following antisera: polyvalent O and H, O4, O9, Hd, Hg, Hi, Hm and Vi antisera (Pro-Lab Diagnostics). A further 2 mls were taken from the top of the selenite broth and stored at -20°C for DNA extraction.

DNA extraction was performed from tissue Selenite supernatants using the QIAamp® Fast DNA Stool Mini Kit (Qiagen, Hilden, Germany), pathogen detection protocol. Elution was done using 30 µl elution buffer instead of 200 µl. Multiplex real-time polymerase chain reaction (PCR) tests were performed in a CFX96 thermal cycler (Bio-Rad, CA, US) using the Quantifast Pathogen PCR + IC Kit® (Qiagen, Hilden , Germany), targeting the pan-*Salmonella* invasion A gene, the *S.* Typhi fimbriae gene [130], and the kit´s internal control. The pan-*Salmonella*, *S.* Typhi and internal control probes were labelled with FAM, Texas Red and VIC, respectively. A 5 min Taq activation step at 95°C was followed by 40 cycles of annealing/extension (30 sec, 60°C) and denaturation (15 sec, 95°C). PCR signals were analyzed using the CFX Manager 3.1. Software with default threshold settings. Valid PCRs required the cycle threshold signal of the internal control to range from 29-31. A cycle threshold

$<$ 40 was considered positive in the presence of a typical exponential amplification curve. Detection of *S.* Typhi required both pan-*Salmonella* and typhoid-specific signal to be positive.

Demographic and clinical data, intraoperative findings and outcomes were captured using OpenDataKit (https://opendatakit.org) at time of recruitment and at time of discharge or death. Data analysis for quantitative data was performed using STATA/SE14.1 version.

Retrospective data summarizing monthly counts of surgically reported intestinal perforations from January 2008 to May 2017 were collected by the department of surgery at QECH. In brief, all patients taken to theatre are recorded in a log book, which is transcribed into an electronic database. Cases clearly not attributable to typhoid, i.e., appendicitis, trauma and perforated peptic ulcer, were excluded. Monthly counts of patients presenting to QECH with typhoid fever diagnosed through routine blood culture surveillance were available for the same time period [88]. We generated a generalized linear model with Poisson error distribution and an identity link to estimate excess perforations attributed to typhoid fever (Supplementary Material 5.1). We used the fitted values of a smoothed seasonal model of monthly typhoid cases through the study period as the predictor variable. Results from this model were then used to estimate the proportion of intestinal perforations attributed to the typhoid epidemic. This analysis was implemented using R, version 3.5.1 [89].

The study was approved by the Malawi College of Medicine Research and Ethics committee (COMREC P.08/14/1617).

5.3 Results

5.3.1 Patients

Between March 2016 and February 2017, 24 patients undergoing laparotomy were recruited. No eligible patient declined to participate. One patient had an intraoperative finding of a gallbladder perforation and was not included in the subsequent analysis. The median age of patients was 15 years (range 6-46 years) and 18 patients (78%) were male. Fever was recorded for 20 participants. 19 of 20 (95%) reported fever prior to admission, which began a median duration of two weeks prior to admission (range 2-30 days).

All patients had a history of abdominal pain (median duration seven days, range 2-30 days). Vomiting, constipation or diarrhea were reported by 43%, 43% and 35%, respectively. Three patients (14%) reported both constipation and diarrhea. Three patients (14%) presented with symptoms suggestive of gastrointestinal bleed. Two patients (9%) presented with reduced conscious level. On examination, most patients had a tender abdomen and frank peritoneal signs denoted as generalized abdominal guarding were present in 80% (Table 1).

Abdominal and/or chest radiograph was performed for 22 patients before undergoing surgery (in 18 patients both investigations were done) and 13 of 22 (59%) were reported as having evidence of free gas under the diaphragm. The median time between admission and operating theatre was one day (range 0-32 days).

5.3.2 Antibiotic treatment

All patients were treated with ceftriaxone from admission for a median of nine days (range 3-48 days) and metronidazole (median nine days, range 3-69 days),

whilst 14 patients received an additional course of ciprofloxacin (median 10 days, range 4-28 days).

Table 5.1 Demographics and clinical features of cohort.

Characteristic	**Value**
Demographic	
Age, median years (range)	15 (6 - 46)
Male, n (%)	18/23 (78%)
Clinical symptoms or signs	
Fever prior to admission, n (%)	19/20 (95%)
Duration of fever prior to admission, median days (IQR)	14 (14 - 21)
Abdominal pain, n (%)	23/23 (100%)
Duration of abdominal pain, median days (IQR)	7 (4 - 14)
Vomiting, n (%)	10/21 (48%)
Constipation, n (%)	10/23 (43%)
Diarrhea, n (%)	8/22 (36%)
Symptoms of gastrointestinal bleed, n (%)	3/20 (15%)
Jaundice, n (%)	1/20 (5%)
Abdominal tenderness, n (%)	22/22 (100%)
Generalized abdominal guarding, n (%)	16/20 (80%)
Reduced level of consciousness, n (%)	2/22 (9%)

5.3.3 Intraoperative findings and surgical treatment

Small bowel perforations with a single pin hole were identified in 16 patients, whilst five patients had multiple perforations and two patients had no visible perforation. Intestinal perforations were all located in the ileum (summarized in

Figure 5.1). In one case, the ileum was found to be inflamed without a visible perforation, there was a pelvic fluid collection and fibrinous deposits in all quadrants. This patient underwent an abdominal washout. One patient presented with a frozen abdomen with no visible intestinal injury and underwent primary adhesiolysis.

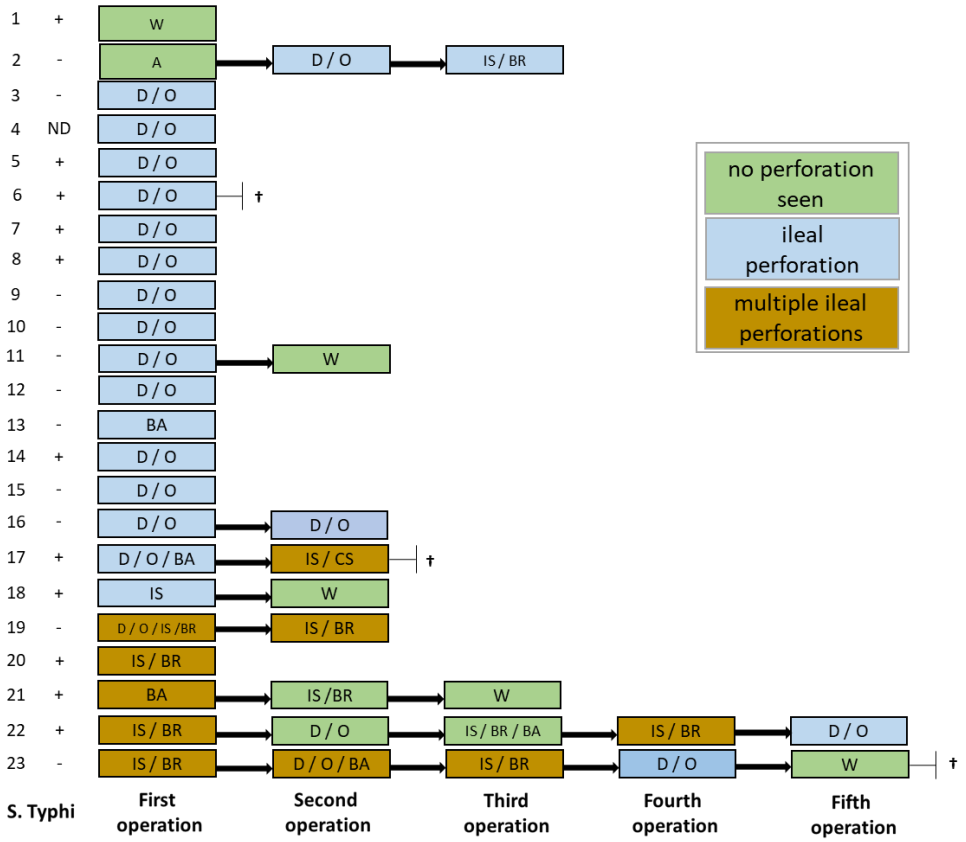


Figure 5.1 Confirmation of *S.* Typhi, relating to intraoperative findings, procedures and postoperative deaths. A: adhesiolysis; BA: bowel resection and anastomosis; CS: colostomy; D/O: debridement/oversew; IS: ileostomy; IS/BR: ileostomy with bowel resection; W: washout; *S.* Typhi: + confirmed by blood culture and/or tissue PCR, - not confirmed; †: patient died.

Primary debridement and oversewing of the perforation was performed in 14 of 16 (87.5%) patients with single ileal perforations. In one of these patients there was the coincidental finding of a tumor mass, prompting the fashioning of an ileocolonic anastomosis in addition to the perforation repair. One patient had a primary ileostomy done, and one patient underwent bowel resection and end to end anastomosis. The five patients with multiple perforations underwent primary ileostomy and bowel resection in four cases, with an additional separate debridement and oversew in one case. One patient underwent bowel resection and end to end anastomosis.

Nine patients (39%) required re-laparotomy two to 12 days after the initial operation (median four days). Secondary perforations—all located in the ileum—were seen in five patients. Three of these had more than one secondary perforation. Bowel anastomotic leaks were seen in seven (77%) of the nine re-laparotomies. In one case, there was an isolated pus collection. Four patients underwent a third operation. Two patients underwent a total of five operations due to recurrent ileal perforations, anastomotic breakdowns, fluid collections and adhesions.

5.3.4 Microbiological and molecular confirmation of *S.* Typhi

Blood cultures were taken from 14 patients on admission or on the hospital wards and *S.* Typhi was isolated from four, with two yielding contaminants and eight no bacterial growth. Eleven patients had intraoperative blood cultures taken. Intraoperative tissue samples were taken from 19 patients. *S.* Typhi was not isolated from any of the intraoperative blood or tissue samples, however other Enterobacteriaceae were identified in 16 tissue samples.

Twenty-two intraoperative tissue samples from 19 patients were analyzed by multiplex PCR. *S.* Typhi DNA was detected in 10 tested samples from nine

of 19 patients (47%). An additional three tissue samples were positive for the pan-*Salmonella* invasin A gene. Overall, 11 of 22 patients (50%) had a diagnosis of typhoid fever made by either blood culture, tissue PCR or both tests.

5.3.5 Mortality and postoperative complications

Three of the 23 patients (13%) died. A 17-year-old male died from sepsis two days post initial laparotomy. A 43-year-old male, who additionally had disseminated malignancy, died post second laparotomy. These two patients had confirmed typhoid infection, representing a case fatality rate of 18% in patients with confirmed typhoid. A 17-year-old male had multiple recurrences of perforations and died after his fifth laparotomy, six weeks after initial admission to the hospital. This patient had a negative admission blood culture, and no intraoperative tissue was submitted in this case.

There were four cases of post-operative pneumonia and a further three of severe sepsis. Seven patients required admission to the intensive care unit for respiratory support. Four patients had a Bogota bag fashioned for abdominal closure either after initial or after secondary surgeries. Twelve patients developed wound infection, 10 of which developed wound dehiscence. Nine patients developed malnutrition despite nutritional support and 7 were discharged on nutritional supplements. The median duration of hospital stay was 21 days (range 4-74 days).

5.3.6 Correlation of *S.* Typhi bloodstream infections and the intestinal perforation register in QECH

Monthly counts of typhoid fever and intestinal perforations at QECH from January 2008 to December 2017 are shown in Figure 5.2A. Results from the generalized linear model indicate that monthly case counts of *S.* Typhi are predictive of monthly intestinal perforations ($p<0.001$, Table 5.2). The intercept estimate of 1.5 indicates that 1.5 [95% CI: 1.16 - 1.85] intestinal perforations

occur each month, independent of typhoid cases. The model estimates that for every culture confirmed case of typhoid, 0.046 [95% CI: 0.033 - 0.058] perforations occur; approximately 1 perforation for every 20 culture confirmed cases of typhoid fever presenting to QECH. Predicted intestinal perforations and their attributed causes are shown in Figure 5.2B. The proportion of surgical perforations predicted by typhoid fever cases is heterogeneous over time. During the recruitment period of the cohort, March 2016 to February 2017, the model independently estimates that 43% of intestinal perforations were due to typhoid fever.

Table 5.2 Intercept and coefficient estimates from the generalized linear model, predicting intestinal perforations from monthly typhoid cases over the study period.

Variable	**Estimate**	**Standard error**	**P-value**
Intercept	1.505	0.17474	<0.001
Smoothed monthly typhoid cases	0.046	0.00649	<0.001

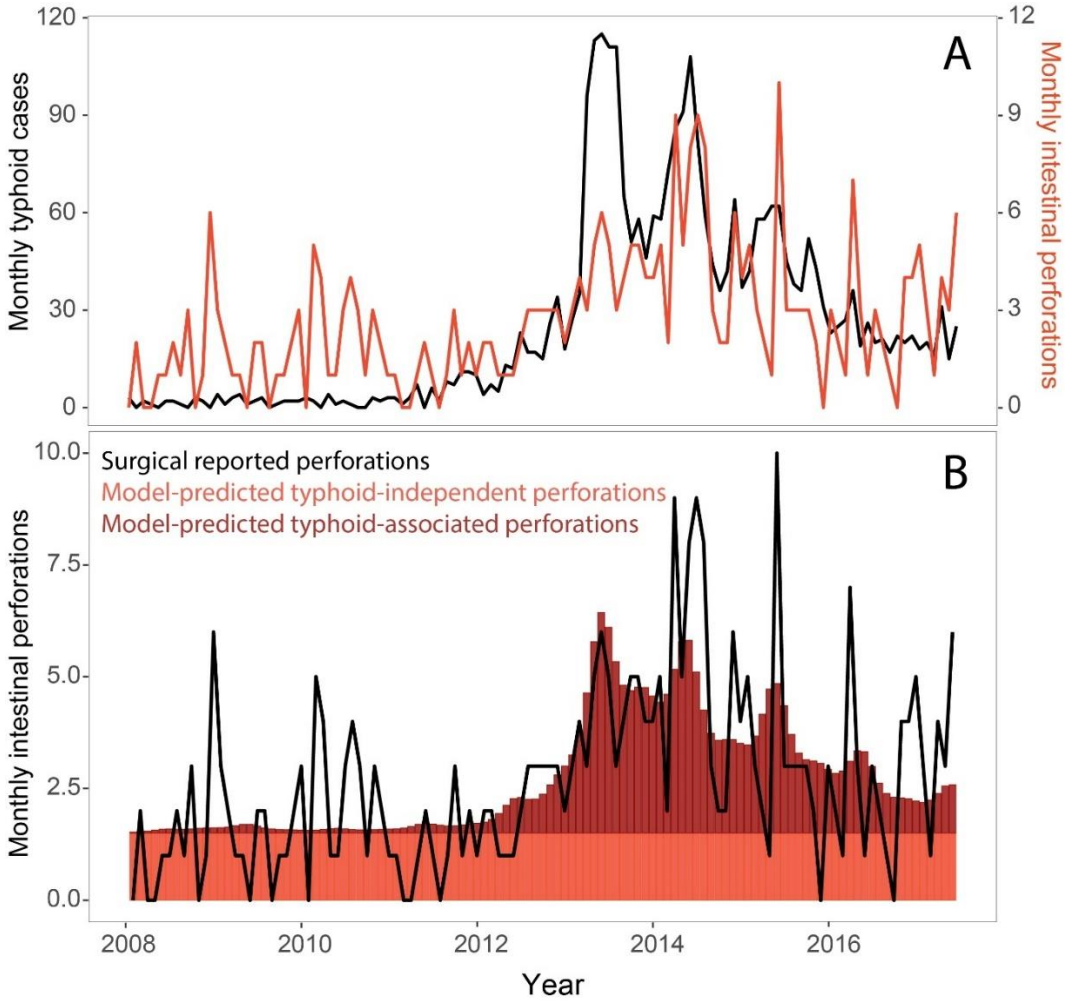


Figure 5.2 A. Monthly counts of intestinal perforations and typhoid cases between January 2008 and June 2015. B. Model predicted surgical perforations, colored by whether the predicted perforation is typhoid independent or typhoid-associated, along with monthly reported surgical perforations.

5.4 Discussion

Accurate estimates of disease burden are critical to prioritize public health interventions; however, this is difficult for typhoid fever, which requires advanced diagnostics. This is particularly true of "surgical" typhoid as surgical teams do not routinely have the capacity to send blood or tissue for culture in LMICs. Furthermore, longitudinal data from surgical teams based in LMICs are

rarely systematically recorded. Even when blood or tissue culture are performed, prior antibiotic therapy and limited sensitivity frequently compromise the sensitivity of culture-based assays. As a consequence, data describing the most feared complication of typhoid fever are not represented in GBD estimates, which will lead to important underestimation of the burden of morbidity and mortality associated with typhoid fever.

In this study we have attempted to confirm infection with *S.* Typhi in cases of perforated abdominal viscus by culturing both peripheral blood and intraoperative samples, and using PCR on tissue. None of the tissue samples analyzed in our study were culture-positive for *S.* Typhi by conventional microbiology. There are several possible reasons for this. *S.* Typhi might simply have been outcompeted in the media by other enteric pathogens, unlike in blood which is a normally sterile site. Alternatively, as perforation is a late complication, it is possible that patients had taken antibiotics prior to presentation, rendering the samples culture-negative. It has also been hypothesized that typhoid intestinal perforation may be the result of an exaggerated host response at the Peyer´s patches—the predilected site of typhoid intestinal perforation—resulting in microvascular changes, rather than a direct result of high bacterial burden [131]. We identified *S.* Typhi DNA by multiplex PCR in nearly half of the tested tissue samples. These results highlight a potential role for PCR in diagnosing surgical typhoid.

Correlation of the longitudinal surveillance of *S.* Typhi BSI and the register of intestinal perforations at QECH showed convincing evidence that the recent surge in intestinal perforation cases coincided with the typhoid epidemic in Southern Malawi. The model estimated that, although there is a baseline monthly rate of non-typhoid attributed intestinal perforation, for every typhoid case 0.046 perforations occur. Results from the generalized linear model were

consistent with observed data from the cohort; while 50% of the 22 patients recruited to the cohort were culture- or PCR-positive for *S.* Typhi, for the same period the model predicts that 43% of intestinal perforations were due to typhoid fever.

These results highlight the potential contribution of non-microbiological methods to understand the aetiology of intestinal perforations. The long-term surveillance capacity for both surgical perforations and routine blood cultures, an unusual resource in this setting, has enabled this exploration. Further, the microbiological testing of surgical cases contributed an independent validation of this methodology, and indicates moderate agreement.

The mortality in our cohort was substantial, with three deaths among 23 patients and two in 11 patients with confirmed infection with *S.* Typhi. Given our modeled estimates that approximately one in 20 cases of culture-confirmed typhoid will predict a case of intestinal perforation, and that our observed case fatality rate was 18% (consistent with other series from sub-Saharan Africa [123]), we estimate that capturing mortality due to typhoid intestinal perforation will increase the case fatality estimates of typhoid fever by 1.0% in our setting. Recent case fatality estimates for typhoid fever were 0.95%, but do not factor in intestinal perforation [4]. If our findings are replicated at other sites, the inclusion of this data into GBD estimates may double mortality burden estimates for typhoid fever. The postoperative morbidity was also substantial. Nine of 23 patients required one or more repeated laparotomies during their illness, and several more had wound infections, pneumonia or malnutrition. These data are also lost to GBD estimates.

Some limitations exist. This was a single center study, however as QECH is the only government hospital with surgical facilities in Blantyre, our data are likely to be representative of the city, although we will not have captured out

of hospital deaths from perforation or patients seeking private care. We did not record antibiotic use prior to admission, and therefore cannot estimate the contribution to missed microbiological diagnosis. We did not have access to histopathology or tuberculosis culture. We did not perform a systematic long-term follow-up after hospital discharge and may, therefore, have underestimated morbidity and mortality.

We reveal an expected, but hitherto undescribed burden of surgical typhoid in Blantyre, and report the associated high morbidity and mortality in the context of a general African epidemic. The systematic capture of these data may lead us to double estimates of mortality attributable to typhoid. Further data from studies of severe and complicated typhoid fever are critical to inform GBD estimates as they will support the case for widespread roll-out of typhoid conjugate vaccination.

Funding

This work was supported by the Bill and Melinda Gates Foundation (Grant number OPP1128444).

Acknowledgements

We would like to thank all patients who participated in this study. We do not have any conflict of interest to declare.

6 Discussion

6.1 Chapter overviews

This thesis explored the spatial, genomic, temporal characteristics of typhoid fever in Blantyre, Malawi, and contributed to insights regarding the mechanisms of transmission in this location for more effective control and targeted surveillance. The morbidity, carriage, and genomic epidemiology of typhoid (MCET) project was designed, and commenced recruitment, prior to the start of my PhD research. As part of my thesis, I have worked with the data collected through this study to answer pre-specified objectives, including analysis of the case-control and cohort datasets. In addition, I have expanded beyond the initial objectives of the MCET study to ask new questions of existing datasets, leading to the inclusion of the time series and spatial-genomic analyses.

Chapter 2 analyzed data from a detailed case-control study of typhoid fever in children, where risk factors and demographic characteristics were surveyed, in an attempt to shed light on possible transmission pathways across the city. Using a variable selection process, which reduced the large risk factor survey containing 97 variables to 14, a number of significant risk factors were found, reflecting the complexity of transmission of typhoid in this setting. Among water-related factors, river water used for cooking and cleaning was identified as a risk factor. No sources of drinking water were identified as risk factors. Distinguishing drinking from household water as possible risk factors is uncommon in case control-studies for typhoid (see review of risk factors in Chapter 2.1), and the finding of a significant non-drinking water exposure as a risk factor highlights the importance of distinguishing these in future studies. The study also found risk-factors related to social interactions, including attendance at school or daycare, indicating that common exposures in these

settings may play a role in transmission. Spatial correlation of the residuals from this model was tested for, as geolocations for the households of cases and controls were available, but no significant correlation was found. However, spatial matching occurred within residential wards, which would essentially correct for any spatial correlation between these residential wards on a larger scale (residential wards span approximately 5 kilometers).

Next, Chapter 3 utilized a cohort recruited from Queen Elizabeth Central Hospital (QECH), within which the spatial case-control dataset was nested. The aim was to understand the spatial distribution of incidence, and to assess spatial patterns of genomic data, gathered from whole genome sequencing of isolates. The spatial incidence mapping found incidence was heterogenous across the city. By employing a geostatistical modelling framework, a number of areas of unexpectedly high or low incidence were additionally identified. Although available covariates were incorporated, the remaining unexplained spatial heterogeneity indicates one or more processes are acting on the population that were not captured by the current study. A map in the supplementary material of this chapter (Figure S3.2.3) was included that can be studied further, for example to generate hypotheses about the unmeasured processes that may be acting in these locations. The mapped cohort is likely only a subset of the overall cases occurring across Blantyre, an observation bolstered by the lack of predictive power of the covariate for distance to hospital (an interpretation of this may be that the most severely ill will travel to care regardless of distance). Future work to capture less severe cases across the city through enhanced surveillance and diagnostics would be useful to provide a more accurate estimate of the overall incidence rate.

Chapter 3 additionally included an analysis of whole genome sequences (WGS) from a subset of the cohort. The finding of a correlation between spatial

distance and genetic relatedness even within a small geographic region, offers a promising approach to improving our understanding of the epidemiology of typhoid fever. It demonstrated that small-scale spatial resolution is possible using *Salmonella enterica* serovar Typhi WGS data, despite the clonal nature of these organisms.

Further, it was found that hydrological river catchment was able to predict some of the observed patterns of genetic relatedness across the city. This indicated that genetic relatedness of the isolated *S.* Typhi is greater within individuals living in the same river catchment than those living in different catchments, and suggests transmission may occur on these scales. Accounting for the ecological context of typhoid fever transmission, including hydrology, has been proposed in Fiji [117]. These findings further highlight the need to study the environmental context of typhoid endemic areas.

Chapter 4 explored the relationship between rainfall and typhoid fever in a time series analysis. It was found that the peak of typhoid fever incidence occurs approximately 15 weeks after the peak of rainfall, a lag that does not present a biologically plausible link, in the context of known incubation periods and survival of *S.* Typhi in the environment. However, the predictive ability of rainfall anomalies was further explored, and it found that a significant log-quadratic relationship exists between rainfall and case anomalies. The coefficient estimates indicated that either more or less rainfall than expected given the time of year, is protective. This suggests potential 'washing' effects of rainfall on the environment during extreme rain events, as well as, during times of less than expected rain, a lack of flushing of *S.* Typhi from open defecated feces or pit latrines into the exposure pathway.

Chapter 5 is methodologically linked to the time series analyses of Chapter 4, and focused on intestinal perforation, one of the most serious complications of typhoid fever. In the absence of routine microbiological confirmation, the attribution to typhoid fever is frequently assumed, but rarely confirmed. The contribution to this chapter was the development of a modelling framework to help determine the aetiology of intestinal perforations. A model was fitted to the seasonal and long-term trends of the case series, which then was used to predict the time series of surgically reported perforations. It was found that, although not all surgical perforations are predicted by typhoid fever case counts, a large proportion of them are. The modeled rates were consistent with what was observed through a small surgical cohort, where microbiological testing was done. This offers a useful framework for understanding intestinal perforation rates without direct culture of surgical tissues, which is not routinely performed in these settings.

6.2 Implications for typhoid fever transmission

This thesis provides evidence that typhoid transmission may be facilitated, at least in part, by exposure through domestic non-drinking river water. Chapters 2-4 suggest this through independent analyses and datasets: first, as an identified risk factor in the case-control study, second, through the ability of river catchment to predict genomic patterns, and finally, by proposing a mechanism of contamination into the rivers, from the finding of protective effects of extreme rain events.

However, given other findings of small-scale spatial correlation, alternative risk factors, and the extreme lag between typhoid and rainfall seasonality, transmission is likely very complex in this setting. Although rivers were suggested as an environmental reservoir, both through the case-control and

spatial-genomic analyses, the case-control study identified a number of other risk factors relating to common social exposures including daycare and school attendance. Further, the spatial-genomic analysis revealed small-scale spatial correlation that may also be predicted by social exposures or small communities; however, this resolution of exposure data was not collected for this cohort. Finally, although rainfall anomalies were predictive of case anomalies, the 15-week lag between the seasonal components of cases and rainfall does not indicate that rainfall is the primary driver of typhoid incidence, given our current knowledge of biological mechanisms of survival and persistence.

6.3 Novel contribution of the work

Chapter 2 is one of the few case-control studies of typhoid fever conducted in Africa to-date, and is unique outside of an outbreak-control context. Additionally, it was the first of these studies to identify non-drinking water usage as a risk factor [132–134]. It is often the case that the typhoid research community place a singular focus on drinking water as the primary exposure source. This tends to be justified through reference to historic data from the United States relating the decline in typhoid to chlorination of drinking water [135], without considering all the other societal improvements that would have been occurring at the same time, for example in sanitation. The singular focus on drinking water was never likely to be realistic given what was already known about typhoid transmission, and this study provides critical evidence that the research and public health community should cast the net wider than simply drinking water when considering long cycle transmission routes of *S.* Typhi.

Chapter 3 is the first city-level investigation to integrate WGS and geostatistical models for typhoid fever. So far, publications investigating genetic patterns on the city-scale have not utilized WGS of *S.* Typhi [64], and those

that have utilized WGS have been primarily global or regional studies [9,49]. The findings of a significant correlation between genetic relatedness and physical proximity within a city is important, in that analysis of *S.* Typhi sequences may be able to reflect small-scale transmission patterns. The ability to predict these genetic markers with spatial covariates is additionally unique for typhoid fever research, and this work may encourage the use of WGS for exploration of transmission insights within this research community. These findings are consistent with the view that WGS can enhance genetic analyses beyond what is possible through SNP-typing, and supports the continuation of sequencing of *S.* Typhi isolates when studying typhoid fever at a city-level scale. There is an opportunity to expand beyond descriptive studies and utilize WGS alongside on-the-ground epidemiology and control.

The fourth chapter highlighted an aspect of analysis of weather and disease patterns that is often overlooked or ignored, that is, the essential cross-correlation of two seasonal patterns. Expanding analyses beyond cross-correlation by identifying extreme events is less frequently conducted, and has not been previously explored for typhoid fever [50,69,70]. A publication that incorporates both types of analyses can offer a useful framework for these exploratory studies and may encourage more caution when interpreting models comparing seasonal patterns of weather and disease.

Finally, the ability to attribute intestinal perforation to typhoid fever has been thus far dependent on enhanced microbiological testing. The final chapter of this thesis proposes a modelling framework to determine the aetiology of intestinal perforations independent of this process. To-date, mortality due to typhoid intestinal perforation is not reliably included in case-fatality estimates [6], or entered into global estimates of mortality [4]. This study may enable the future admission of intestinal perforations into mortality estimates, expanding

the underlying mortality estimates of typhoid fever and advocacy as a global health problem.

6.4 Limitations and challenges of the data and approaches utilized

Geo-locating individuals in their household may be one spatial aspect of exposure, but individuals move between school, work, and other locations throughout a day. Some studies attempt to overcome this by focusing on young children who may not move as far from the home [51]. However, children are not the only individuals at risk, and additionally may attend school or daycare in other regions of the city. Geo-locating places of work, school, or food markets may be a useful way of exploring differing exposure locations and how they may compare with household location. Although drinking water sources of individuals were geo-located, individuals lived within close proximity of their drinking water source, and therefore the patterns of spatial correlation were very similar to that of household location. Instead of geo-locating water source locations, mapping networks of water sources (linking the water sources among individuals) and categorizing the type of source, may provide more insight. This may include the location of access to rivers, for those that use them.

Case-control studies are useful for identifying risk factors, but inherently limited in their ability to determine the source of contamination of the risk factor, when the risk factor is not the source itself. For example, ice cream could be contaminated by the water used to make it, or by an infectious individual through food handling. This case-control study additionally only focused on children under the age of 9, and therefore these exposures may not represent all potential pathways of transmission in this setting, if exposures differ among cases, older children, and adults.

Despite the finding of spatial patterns of genetic relatedness in the spatio-genomic analysis, there were some limitations. Multidimensional scaling of the SNP matrix reduced this complex relational dataset to two representative axes. Because of this reduction in dimensional space, some of the more subtle aspects of genetic relatedness are likely being missed. Further, the majority of analyses were conducted on the second of the principal coordinates, due to the first principal coordinate being dominated by a small number of highly related individuals. Although they did not appear to be related in space or time, further work to understand the reason behind the distinct genetic patterns of these individuals would be valuable, as they could represent infections from a chronic carrier, or distinct transmission route.

Throughout this thesis, only hospitalized cases who have sought care at QECH were studied. Sub-clinical and mild infections of typhoid are known to exist [118], and identifying more of these cases would not only benefit this research by increasing the study size, but provide data that is more representative of all typhoid infections. Lower-dose infections are known to result in more sub-clinical illness [118], so distinct risk factors may exist for individuals who are not seeking care at a hospital. As an example, if drinking water exposures are associated with a lower infectious dose, these may not be as easily identified through a hospital-based case-control study.

We additionally did not combine spatio-temporal analyses in the incidence mapping or genomic analyses. This decision was based on both the small number of geo-located cases for these analyses, as well as the research questions that were proposed, which aimed to provide insights into transmission during the study period. However, future dynamic modelling work incorporating the time and spatial observations of the datasets may be useful. A model could explore whether linking these cases by hypothesized transmission routes (common river

catchment or close proximity) within time periods that reflect the biological processes are able to recreate the spatio-temporal trends observed. This could also incorporate variation in force of infection over time due to seasonality or epidemic processes such as immunity after the sharp increase of cases in 2011.

6.5 Future work

Typhoid conjugate vaccines are a promising new tool for the control of typhoid fever, however it is yet to be determined whether the clinical protection observed in challenge models will translate to a reduction in shedding of the disease and subsequent herd protection [15]. Therefore, to achieve elimination of typhoid fever as a public health problem, a major unresolved research gap is how to rapidly assess transmission routes for the planning of water and sanitation intervention methods.

Although microbiological sampling of the environment, including potential exposure pathways, would ideally elucidate intervention points, at the start of this project these methods were not reliable for *S.* Typhi. Culture of environmental samples has historically not yielded high sensitivity for the detection of S. Typhi, even when placed in the sewage discharge of known shedders [37]. Additionally, identifying DNA through PCR is subject to specificity issues, given the large array of pathogens likely present in a single environmental sample [41]. Further, identifying specifically what components of the environment are important for sampling is challenging without any prior hypotheses. Since this study was conducted, improvements in culture media and PCR primer identification have been made, making environmental sampling a technically viable method for future work.

This thesis identified potential transmission routes. Microbiological confirmation of these routes through environmental sampling would greatly

strengthen the current evidence base and, critically, provide data that local policy makers could not ignore. For example, when the results of the case control study were presented to the Blantyre district health officer, investigators were advised to come back when they had confirmation of their findings [personal communication N. Feasey]. To address the hypothesis of exposure to *S.* Typhi through rivers, confirming the presence of *S.* Typhi in river water would be a first step. Ideally this would include whole genome sequencing, so strains found in acute cases presenting to the hospital, sub-clinical community cases, or chronic carriers could be compared with sequences found in environmental samples. The optimal methods for environmental sampling *Salmonella* Typhi in this or any setting are yet to be determined, so a proposed a pilot study based on the findings of this thesis was developed to help address this in river systems in Blantyre. Specific components of the pilot included a cross-section of measurements throughout a day to explore the importance of diurnal variation in rivers (an observed phenomenon in sewage systems [136]), sampling over a period representing a cross-section of the variation in rainfall, and comparing multiple points along a river to assess whether detection rates increase as sampling moves downstream (accumulation of material), or decrease (die-off of the bacteria over time). Because of the continued geo-location of cases, maps of cumulative downstream case counts can be generated (Figure 6.1). These will help prioritize sampling junctions for the initial pilots, which aim to prioritize areas with the highest numbers of cases for an increased chance of detection, given sensitivity challenges.

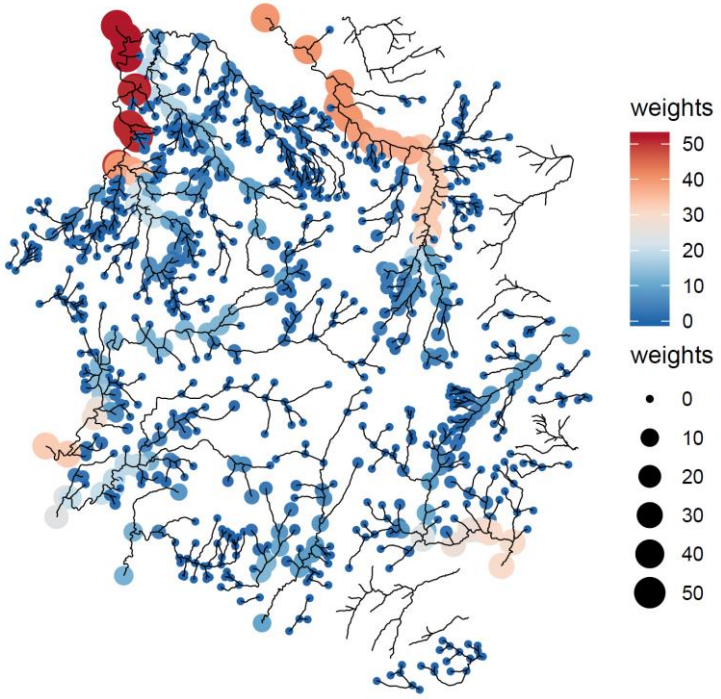


Figure 6.1 Cumulative downstream case-counts ('weights') for each river, as a proposed method of sampling prioritization.

In order to better disentangle contaminating sources of risk factors, future work could include linking environmental sampling with a case-control study, where the food and drinking water of cases and controls at home, school, and work were sampled to understand which exposures are the most important for transmission. If WGS was available, this additionally provides an opportunity to link these samples with the sequenced river or other environmental samples from a range of households.

However, understanding the link between shedding and exposure is still poorly understood, and will likely limit the interpretation of the above findings. Linking in sociological research to map out cooking, cleaning and water usage practices, agricultural research to understand how and where produce is irrigated or washed, as well as infrastructure and ecological work to understand sewage, pit latrine runoff, and river catchments are very much needed. Such

work will be important to describe the structural barriers to engaging with optimal WASH practice, and enable individuals to protect themselves from unsafe environmental exposures. While this may be a slow process, it will be valuable to have a single, in-depth study site where the complexities and societal drivers of typhoid transmission can be explored and described, in order to more efficiently make public health decisions interventions in new locations.

At a minimum, continuing the current hospital-based surveillance at QECH would be valuable. Each year, more observations add to the evidence base for observations such as the protective effects of rainfall anomalies, and the attribution of intestinal perforations. This surveillance can additionally provide a monitoring system in the case of introduction of new strains, such as XDR typhoid [11]. It would additionally be valuable to continue to geo-locate the homes of these cases. Currently, the surveillance in Blantyre is expanding: active surveillance has been initiated through clinics in two neighborhoods within Blantyre, in order to observe cases that may not typically present to QECH. Repeating the analyses contained in this thesis using data from these new sites is a necessary next step to understand whether milder or sub-clinical cases exhibit different epidemiological characteristics, and what role these cases play in typhoid transmission.

Finally, although potential routes of transmission were identified in this thesis, their relative importance to the overall transmission of typhoid fever was not precisely defined. It would be useful to incorporate these hypothesized transmission frameworks (households, schools, river catchments) into a mathematical model, and test whether the relative importance of these transmission routes can be resolved using the current data available. Although more cases may be needed to fit this model with enough confidence to

distinguish these transmission routes, exploring the possibility of this approach would be a useful exercise.

6.6 Conclusions

In conclusion, this thesis outlined multiple methodological approaches to better understand typhoid fever transmission in Blantyre, Malawi. As part of the MCET study, some of these analyses were pre-specified and designed before the commencement of my research. However, this thesis additionally harnessed routinely collected blood culture surveillance data from Blantyre, as well as spatial data, to expand on the initial aims of the project and provide hypothesis-generating results for future studies. Transmission of the disease remains difficult to disentangle in endemic settings. However, this work has illustrated the potential roles of spatial, genomic, and time series data, and highlighted the importance of better understanding the ecological context of typhoid fever. A need still exists for rapid assessment of transmission in endemic settings, however the increased ease of geo-location of cases, falling costs of sequencing, and advances in environmental sampling will certainly aid in this process in the future. This work further highlights the value of interdisciplinary collaboration between statistical modelers, epidemiologists, clinicians, bioinformaticians and WASH experts. These combinations allowed for unique and innovative data collection, methodological approaches atypical to the field, and interpretations that consider the ecological context as well as the epidemiological.

References

1. Mogasale V, Maskery B, Ochiai RL, et al. Burden of typhoid fever in low-income and middle-income countries: A systematic, literature-based update with risk-factor adjustment. Lancet Glob Heal **2014**; 2:e570–e580.

2. Crump JA, Luby SP, Mintz ED. The global burden of typhoid fever. Bull World Health Organ **2004**; 82:346–353.

3. Antillón M, Warren JL, Crawford FW, et al. The burden of typhoid fever in low- and middle-income countries: A meta-regression approach. PLoS Negl Trop Dis **2017**; 11.

4. Stanaway JD, Reiner RC, Blacker BF, et al. The global burden of typhoid and paratyphoid fevers: a systematic analysis for the Global Burden of Disease Study 2017. Lancet Infect Dis **2019**; 19:369–381.

5. Antillon M, Saad NJ, Baker S, Pollard AJ, Pitzer VE. The Relationship between Blood Sample Volume and Diagnostic Sensitivity of Blood Culture for Typhoid and Paratyphoid Fever: A Systematic Review and Meta-Analysis. J Infect Dis **2018**; 218:S255–S267.

6. Pieters Z, Saad NJ, Antillón M, Pitzer VE, Bilcke J. Case fatality rate of enteric fever in endemic countries: A systematic review and meta-analysis. Clin Infect Dis **2018**;

7. Colquhoun J, Weetch RS. Resistance to chloramphenicol developing during treatment of typhoid fever. Lancet **1950**;

8. Paniker CKJ, Vimala KN. Transferable chloramphenicol resistance in salmonella typhi. Nature **1972**;

9. Wong VK, Baker S, Pickard D, et al. The emergence and intercontinental spread of a multidrug-resistant clade of typhoid agent Salmonella enterica serovar Typhi. Lancet **2016**; 387:S10. Available at: https://www.sciencedirect.com/science/article/pii/S0140673616003974.

10. Crump JA, Sjölund-Karlsson M, Gordon MA, Parry CM. Epidemiology, clinical presentation, laboratory diagnosis, antimicrobial resistance, and antimicrobial management of invasive Salmonella infections. Clin. Microbiol. Rev. 2015;

11. Klemm EJ, Shakoor S, Page AJ, et al. Emergence of an extensively drug-resistant Salmonella enterica serovar typhi clone harboring a

promiscuous plasmid encoding resistance to fluoroquinolones and third-generation cephalosporins. MBio **2018**; 9.

12. Levine MM, Simon R. The gathering storm: Is untreatable typhoid fever on the way? MBio. 2018;

13. Gröschel DH, Hornick RB. Who introduced typhoid vaccination: Almroth Write or Richard Pfeiffer? Rev Infect Dis **1981**; 3:1251–1254.

14. Levine MM, Ferreccio C, Abrego P, Martin OS, Ortiz E, Cryz S. Duration of efficacy of Ty21a, attenuated Salmonella typhi live oral vaccine. In: Vaccine. 1999.

15. Jin C, Gibani MM, Moore M, et al. Efficacy and immunogenicity of a Vi-tetanus toxoid conjugate vaccine in the prevention of typhoid fever using a controlled human infection model of Salmonella Typhi: a randomised controlled, phase 2b trial. Lancet **2017**; 390:2472–2480.

16. Burki T. Typhoid conjugate vaccine gets WHO prequalification. Lancet. Infect. Dis. 2018;

17. Shakya M, Colin-Jones R, Theiss-Nyland K, et al. Phase 3 Efficacy Analysis of a Typhoid Conjugate Vaccine Trial in Nepal. N Engl J Med **2019**; 381:2209–2218. Available at: https://doi.org/10.1056/NEJMoa1905047.

18. Levine MM, Ferreccio C, Black RE, Tacket CO, Germanier R, Committee CT. Progress in vaccines against typhoid fever. Clin Infect Dis **1989**; 11:S552–S567.

19. Shuval HI. Investigation of typhoid fever and cholera transmission by raw wastewater irrigation in Santiago, Chile. In: Water Science and Technology. 1993: 167–174.

20. Marineli F, Tsoucalas G, Karamanou M, Androutsos G. Mary Mallon (1869-1938) and the history of typhoid fever. Ann Gastroenterol **2013**;

21. Glynn JR, Bradley DJ. The relationship between infecting dose and severity of disease in reported outbreaks of salmonella infections. Epidemiol Infect **1992**; 109:371–388. Available at: http://www.journals.cambridge.org/abstract_S0950268800050366.

22. Walsh K. 3 Cases Of Life-Threatening Typhoid Fever Linked To Qdoba In Firestone. CBS News Denver. 2015; Available at: https://denver.cbslocal.com/2015/11/02/3-cases-of-life-threatening-typhoid-fever-linked-to-qdoba-in-firestone/.

23. Lynch MF, Blanton EM, Bulens S, et al. Typhoid fever in the United States, 1999-2006. JAMA - J Am Med Assoc **2009**; 302:859–865.

24. Karkey A, Jombart T, Walker AW, et al. The Ecological Dynamics of Fecal Contamination and Salmonella Typhi and Salmonella Paratyphi A in Municipal Kathmandu Drinking Water. PLoS Negl Trop Dis **2016**; 10.

25. Vezzulli L, Pruzzo C, Huq A, Colwell RR. Environmental reservoirs of Vibrio cholerae and their role in cholera. Environ. Microbiol. Rep. 2010;

26. Ames WR, Robins M. Age and Sex as Factors in the Development of the Typhoid Carrier State, and a Method for Estimating Carrier Prevalence. Am J Public Heal Nations Heal **1943**; 33:221–230. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1527221&tool=pmcentrez&rendertype=abstract.

27. Hornick RB, Greisman SE, Woodward TE, Dupont HL, Hawkins AT, Snyder MJ. Typhoid Fever: Pathogenesis and Immunologic Control. N Engl J Med **1970**; 283:739–746. Available at: http://www.ncbi.nlm.nih.gov/pubmed/4916913%5Cnhttp://www.ncbi.nlm.nih.gov/pubmed/4916916.

28. Sinnott CR, Teall AJ. Persistent gallbladder carriage of salmonella typhi. Lancet. 1987;

29. Waddington CS, Darton TC, Jones C, et al. An outpatient, ambulant-design, controlled human infection model using escalating doses of salmonella typhi challenge delivered in sodium bicarbonate solution. Clin Infect Dis **2014**; 58:1230–1240.

30. Cho JC, Kim SJ. Viable, but non-culturable, state of a green fluorescence protein-tagged environmental isolate of Salmonella typhi in groundwater and pond water. FEMS Microbiol Lett **1999**; 170:257–264.

31. Ercolani GL. Bacteriological quality assessment of fresh marketed lettuce and fennel . Bacteriological Quality Assessment of Fresh Marketed Lettuce and Fennel. Appl Environ Microbiol **1976**; 31:847–852.

32. Castro-Rosas J, Escartín EF. Survival and growth of Vibrio cholerae O1 Salmonella typhi, and Escherichia coli O157:H7 in alfalfa sprouts. J Food Sci **2000**; 65:162–165.

33. Lynch JM. Microbial Survival in the Environment, Bacteria and

Rickettsiae Important in Human and Animal Health . E. Mitscherlich , E. H. Marth . Springer Science & Business Media, 1985.

34. Moore B. The detection of enteric carriers in towns by means of sewage examination. J R Sanit Inst **1951**; 71:57–60.

35. Sears SD, Ferreccio C, Levine MM. Sensitivity of Moore sewer swabs for isolating Salmonella typhi. Appl Environ Microbiol **1986**; 51:425–426.

36. Sears ASD, Ferreccio C, Levine MM, et al. The Use of Moore Swabs for Isolation of Salmonella typhi from Irrigation Water in Santiago,Chile. J Infect Dis **1984**; 149:640–642.

37. Sears SD, Ferreccio C, Levine MM. Sensitivity of Moore sewer swabs for isolating Salmonella typhi. Appl Environ Microbiol **1986**; 51:425–426.

38. Murphy JL, Kahler AM, Nansubug I, et al. Environmental survey of drinking water sources in Kampala, Uganda, during a typhoid fever outbreak. Appl Environ Microbiol **2017**; 83.

39. Zeng B, Zhao G, Cao X, Yang Z, Wang C, Hou L. Formation and resuscitation of viable but nonculturable Salmonella typhi. Biomed Res Int **2013**; 2013.

40. Saha S, Tanmoy AM, Andrews JR, et al. Evaluating PCR-based detection of Salmonella typhi and paratyphi a in the environment as an enteric fever surveillance tool. Am J Trop Med Hyg **2019**;

41. Nair S, Patel V, Hickey T, et al. Real-time PCR assay for differentiation of typhoidal and nontyphoidal Salmonella. J Clin Microbiol **2019**; 57.

42. Republic of Malawi. 2008 Population and Housing Census. Popul. (English Ed. 2008; :35. Available at: http://www.malawihighcommission.co.uk/MWCensus08.pdf.

43. Habitat U. Malawi: Blantyre Urban Profile. **2011**;

44. National Statistical Office; Government of Malawi. Welfare Monitoring Survey. 2011. Available at: http://catalog.ihsn.org/index.php/catalog/2943/download/44499.

45. Maoulidi M. Water and sanitation needs assessment for Blantyre city, Malawi. **2012**;

46. Feasey NA, Gaskell K, Wong V, et al. Rapid Emergence of Multidrug Resistant, H58-Lineage Salmonella Typhi in Blantyre, Malawi. PLoS Negl Trop Dis **2015**; 9.

47. Pitzer VE, Feasey NA, Msefula C, et al. Mathematical modeling to assess the drivers of the recent emergence of typhoid fever in Blantyre, Malawi. Clin Infect Dis **2015**; 61:S251–S258.

48. Alexander FE, Cuzick J. Methods for the assessment of disease clusters. In: Geographical and Environmental Epidemiology: Methods for Small Area Studies. 2009.

49. Pham Thanh D, Thompson CN, Rabaa MA, et al. The Molecular and Spatial Epidemiology of Typhoid Fever in Rural Cambodia. PLoS Negl Trop Dis **2016**; 10.

50. Dewan AM, Corner R, Hashizume M, Ongee ET. Typhoid Fever and Its Association with Environmental Factors in the Dhaka Metropolitan Area of Bangladesh: A Spatial and Time-Series Approach. PLoS Negl Trop Dis **2013**; 7.

51. Akullian A, Ng'eno E, Matheson AI, et al. Environmental Transmission of Typhoid Fever in an Urban Slum. PLoS Negl Trop Dis **2015**; 9.

52. Giorgi E, Diggle PJ. PrevMap: An R package for prevalence mapping. J Stat Softw **2017**; 78.

53. Diggle PJ, Tawn JA, Moyeed RA. Model-based geostatistics. J R Stat Soc Ser C (Applied Stat **2002**; 47:299–350. Available at: http://doi.wiley.com/10.1111/1467-9876.00113.

54. Stoyan D. Matérn, B.: Spatial Variation. 2nd Ed., Springer-Verlag, Berlin, Heidelberg, New York, London, Paris, Tokyo 1986, 151 S., DM 33,-. Biometrical J **2007**; 30:594–594.

55. Zhang H. Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. J Am Stat Assoc **2004**; 99:250–261.

56. Ediriweera DS, Kasturiratne A, Pathmeswaran A, et al. Mapping the Risk of Snakebite in Sri Lanka - A National Survey with Geospatial Analysis. PLoS Negl Trop Dis **2016**;

57. Stresman GH, Giorgi E, Baidjoe A, et al. Impact of metric and sample size on determining malaria hotspot boundaries. Sci Rep **2017**; 7.

58. Gardy JL, Johnston JC, Ho Sui SJ, et al. Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. N Engl J Med **2011**; 364:730–739.

59. Gire SK, Goba A, Andersen KG, et al. Genomic surveillance elucidates

Ebola virus origin and transmission during the 2014 outbreak. Science (80-) **2014**;

60. Roumagnac P, Weill FX, Dolecek C, et al. Evolutionary history of Salmonella Typhi. Science (80-) **2006**;

61. Parkhill J, Dougan G, James KD, et al. Complete genome sequence of a multiple drug resistant Salmonella enterica serovar Typhi CT18. Nature **2001**;

62. Baker S, Holt KE, Clements ACA, et al. Combined high-resolution genotyping and geospatial analysis reveals modes of endemic urban typhoid fever transmission. Open Biol **2011**; 1.

63. Wong VK, Baker S, Connor TR, et al. An extended genotyping framework for Salmonella enterica serovar Typhi, the cause of human typhoid. Nat Commun **2016**;

64. Holt KE, Baker S, Dongol S, et al. High-throughput bacterial SNP typing identifies distinct clusters of Salmonella Typhi causing typhoid in Nepalese children. BMC Infect Dis **2010**; 10.

65. Jombart T, Pontier D, Dufour AB. Genetic markers in the playground of multivariate analysis. Heredity (Edinb). 2009; 102:330–341.

66. Jallow M, Teo YY, Small KS, et al. Genome-wide and fine-resolution association analysis of malaria in West Africa. Nat Genet **2009**; 41:657–665.

67. Paschou P, Ziv E, Burchard EG, et al. PCA-correlated SNPs for structure identification in worldwide human populations. PLoS Genet **2007**;

68. Teklehaimanot HD, Lipsitch M, Teklehaimanot A, Schwartz J. Weather-based prediction of Plasmodium falciparum malaria in epidemic-prone regions of Ethiopia I. Patterns of lagged weather effects reflect biological mechanisms. Malar J **2004**;

69. Carlton EJ, Eisenberg JNS, Goldstick J, Cevallos W, Trostle J, Levy K. Heavy rainfall events and diarrhea incidence: The role of social and environmental factors. Am J Epidemiol **2014**;

70. Saad NJ, Lynch VD, Antillón M, Yang C, Crump JA, Pitzer VE. Seasonal dynamics of typhoid and paratyphoid fever. Sci Rep **2018**; 8.

71. Deen J, von Seidlein L, Andersen F, Elle N, White NJ, Lubell Y. Community-acquired bacterial bloodstream infections in developing

countries in south and southeast Asia: A systematic review. Lancet Infect. Dis. 2012; 12:480–487.

72. Reddy EA, Shaw A V., Crump JA. Community-acquired bloodstream infections in Africa: a systematic review and meta-analysis. Lancet Infect. Dis. 2010; 10:417–432.

73. Marks F, von Kalckreuth V, Aaby P, et al. Incidence of invasive salmonella disease in sub-Saharan Africa: a multicentre population-based surveillance study. Lancet Glob Heal **2017**; 5:e310–e323.

74. Neil KP, Sodha S V., Lukwago L, et al. A large outbreak of typhoid fever associated with a high rate of intestinal perforation in Kasese district, Uganda, 2008-2009. Clin Infect Dis **2012**; 54:1091–1099.

75. Hendriksen RS, Leekitcharoenphon P, Lukjancenko O, et al. Genomic signature of multidrug-resistant salmonella enterica serovar Typhi isolates related to a massive outbreak in Zambia between 2010 and 2012. J Clin Microbiol **2015**; 53:262–272.

76. World Health Organization. Typhoid vaccines: WHO position paper, March 2018 – Recommendations. Vaccine. 2019; 37:214–216.

77. Luxemburger C, Duc CM, Lanh MN, et al. Risk factors for typhoid fever in the Mekong delta, southern Viet Nam: A case-control study. Trans R Soc Trop Med Hyg **2001**;

78. Vollaard AM. Risk Factors for Typhoid and Paratyphoid Fever in Jakarta, Indonesia. JAMA **2004**; 291:2607. Available at: http://jama.jamanetwork.com/article.aspx?doi=10.1001/jama.291.21.2607.

79. Tran HH, Bjune G, Nguyen BM, Rottingen JA, Grais RF, Guerin PJ. Risk factors associated with typhoid fever in Son La province, northern Vietnam. Trans R Soc Trop Med Hyg **2005**; 99:819–826.

80. Black RE, Cisneros L, Levine MM, Banfi A, Lobos H, Rodriguez H. Case-control study to identify risk factors for paediatric endemic typhoid fever in Santiago, Chile. Bull World Health Organ **1985**;

81. Luby SP, Faizan MK, Fisher-Hoch SP, et al. Risk factors for typhoid fever in an endemic setting, Karachi, Pakistan. Epidemiol Infect **1998**; 120:129–138.

82. Hussein Gasem M, Dolmans WMVWMV, Keuter MM, Djokomoeljanto RR. Poor food hygiene and housing as risk factors for typhoid fever in

Semarang, Indonesia. Trop Med Int Heal **2001**; 6:484–490.

83. Velema JP, Van Wijnen G, Bult P, Van Naerssen T, Jota S. Typhoid fever in Ujung Indonesia - High-risk groups and high-risk behaviours. Trop Med Int Heal **1997**; 2:1088–1094.

84. Ram PK, Naheed A, Brooks WA, et al. Risk factors for typhoid fever in a slum in Dhaka, Bangladesh. Epidemiol Infect **2007**; 135:458–465.

85. Prasad N, Jenkins AP, Naucukidi L, et al. Epidemiology and risk factors for typhoid fever in Central Division, Fiji, 2014–2017: A case-control study. PLoS Negl Trop Dis **2018**; 12.

86. Alba S, Bakker MI, Hatta M, et al. Risk factors of typhoid infection in the Indonesian archipelago. PLoS One **2016**; 11.

87. Karkey A, Thompson CN, Tran Vu Thieu N, et al. Differential Epidemiology of Salmonella Typhi and Paratyphi A in Kathmandu, Nepal: A Matched Case Control Investigation in a Highly Endemic Enteric Fever Setting. PLoS Negl Trop Dis **2013**;

88. Musicha P, Cornick JE, Bar-Zeev N, et al. Trends in antimicrobial resistance in bloodstream infection isolates at a large urban hospital in Malawi (1998–2016): a surveillance study. Lancet Infect Dis **2017**; 17:1042–1052.

89. R Core Team, Team R. R: A Language and Environment for Statistical Computing. R Found Stat Comput **2011**;

90. WaterAid. Low-income Customer Support Units. 2016. Available at: https://washmatters.wateraid.org/sites/g/files/jkxoof256/files/LICSU Malawi case study_1.pdf.

91. Lilje J, Kessely H, Mosler HJ. Factors determining water treatment behavior for the prevention of cholera in Chad. Am J Trop Med Hyg **2015**; 93:57–65.

92. Huber AC, Mosler HJ. Determining behavioral factors for interventions to increase safe water consumption: A cross-sectional field study in rural Ethiopia. Int J Environ Health Res **2013**; 23:96–107.

93. Inauen J, Mosler HJ. Mechanisms of behavioural maintenance: Long-term effects of theory-based interventions to promote safe water consumption. Psychol Heal **2016**; 31:166–183.

94. Wang Y, Moe CL, Null C, et al. Multipathway quantitative assessment of exposure to fecal contamination for young children in low-income

urban environments in Accra, Ghana: the Sanipath analytical approach. Am J Trop Med Hyg **2017**; 97:1009–1019.

95. Ferreccio C, Levine M, Astroza L, et al. The detection of chronic Salmonella typhi carriers: a practical method applied to food handlers. Rev Med Chil **1990**; 118:33–37.

96. McMichael C. Water, sanitation and hygiene (WASH) in schools in low-income countries: A review of evidence of impact. Int J Environ Res Public Health **2019**; 16.

97. Manda M. Water and sanitation in urban Malawi: Can the Millennium Development Goals be met? A study of informal settlements in three cities. 2009.

98. Chipeta L. The water crisis in blantyre city and its impact on women: The cases of mabyani and ntopwa, malawi. J Int Womens Stud **2009**; 10:17–33.

99. Stanaway JD, Reiner RC, Blacker BF, et al. The global burden of typhoid and paratyphoid fevers: a systematic analysis for the Global Burden of Disease Study 2017. Lancet Infect Dis **2019**; 19:369–381. Available at: http://www.ncbi.nlm.nih.gov/pubmed/30792131. Accessed 20 March 2019.

100. Mercer LD, Safdar RM, Ahmed J, et al. Spatial model for risk prediction and sub-national prioritization to aid poliovirus eradication in Pakistan. BMC Med **2017**;

101. Lau MSY, Dalziel BD, Funk S, et al. Spatial and temporal dynamics of superspreading events in the 2014-2015 West Africa Ebola epidemic. Proc Natl Acad Sci U S A **2017**; 114:2337–2342.

102. Osei FB, Stein A, Nyadanu SD. Spatial and temporal heterogeneities of district-level typhoid morbidities in Ghana: A requisite insight for informed public health response. PLoS One **2018**; 13:e0208006. Available at: http://dx.plos.org/10.1371/journal.pone.0208006. Accessed 11 February 2019.

103. Mather AE, Vaughan TG, French NP. Molecular approaches to understanding transmission and source attribution in nontyphoidal salmonella and their application in Africa. Clin Infect Dis **2015**;

104. Ashton PM, Nair S, Peters TM, et al. Identification of Salmonella for public health surveillance using whole genome sequencing. PeerJ **2016**.

105. Gauld JS, Olgemoeller F, Nkhata R, et al. Domestic river water use and risk of typhoid fever: results from a case-control study in Blantyre, Malawi. Clin Infect Dis **2019**;

106. Harris RC. Informing development strategies for new tuberculosis vaccines: mathematical modelling and novel epidemiological tools. 2017. Available at: http://researchonline.lshtm.ac.uk/4648987/.

107. MacPherson P, Khundi M, Nliwasa M, et al. Disparities in access to diagnosis and care in Blantyre, Malawi, identified through enhanced tuberculosis surveillance and spatial analysis. BMC Med **2019**; 17.

108. Andrews JR, Barkume C, Yu AT, et al. Integrating Facility-Based Surveillance with Healthcare Utilization Surveys to Estimate Enteric Fever Incidence: Methods and Challenges. J Infect Dis **2018**;

110. U.S. Geological Survey. Shuttle Radar Topography Mission 1 Arc-Second Global. U.S. Geol. Surv. 2017; Available at: https://www.usgs.gov/.

111. Page AJ, Taylor B, Delaney AJ, et al. SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. Microb genomics **2016**; 2:e000056.

112. Nguyen LT, Schmidt HA, Von Haeseler A, Minh BQ. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol **2015**; 32:268–274.

113. Rambaut A, Lam TT, Max Carvalho L, Pybus OG. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). Virus Evol **2016**; 2:vew007.

114. Wailan AM, Coll F, Heinz E, et al. rPinecone: Define sub-lineages of a clonal expansion via a phylogenetic tree. Microb Genomics **2019**; 5:1–9.

115. Lo NC, Gupta R, Stanaway JD, et al. Comparison of Strategies and Incidence Thresholds for Vi Conjugate Vaccines Against Typhoid Fever: A Cost-effectiveness Modeling Study. J Infect Dis **2018**; 218:S232–S242.

116. Mogasale V, Mogasale V V., Ramani E, et al. Revisiting typhoid fever surveillance in low and middle income countries: Lessons from systematic literature review of population-based longitudinal studies. BMC Infect Dis **2016**;

117. Jenkins AP, Jupiter S, Mueller U, et al. Health at the Sub-catchment Scale: Typhoid and Its Environmental Determinants in Central Division,

Fiji. Ecohealth **2016**; 13:633–651.

118. Woodward WE. Volunteer studies of typhoid fever and vaccines. Trans R Soc Trop Med Hyg **1980**;

119. Morris P, Pawitan Y. In All Likelihood: Statistical Modelling and Inference Using Likelihood. Math Gaz **2002**; 86:375.

120. Parry CM, Hien TT, Dougan G, White NJ, Farrar JJ. Typhoid fever. N Engl J Med **2002**; 347:1770–1782.

121. Butler T, Knight J, Nath SK, Speelman P, Roy SK, Azad MAK. Typhoid fever complicated by intestinal perforation: A persisting fatal disease requiring surgical management. Rev Infect Dis **1985**; 7:244–256.

122. Bitar R, Tarpley J. Intestinal perforation in typhoid fever: A historical and state-of-the-art review. Rev Infect Dis **1985**; 7:257–271.

123. Mogasale V, Desai SN, Mogasale V V., Park JK, Leon Ochiai R, Wierzba TF. Case fatality rate and length of hospital stay among patients with typhoid intestinal perforation in developing countries: A systematic literature review. PLoS One **2014**; 9:1–11.

124. Contini S. Typhoid intestinal perforation in developing countries: Still unavoidable deaths? World J Gastroenterol **2017**; 23:1925–1931.

125. Obaro SK, Iroh Tam PY, Mintz ED. The unrecognized burden of typhoid fever. Expert Rev. Vaccines. 2017; 16:249–260.

126. Ameh EA. Typhoid ileal perforation in children: A scourge in developing countries. Ann Trop Paediatr **1999**; 19:267–272.

127. Uba AF, Chirdan LB, Ituen AM, Mohammed AM. Typhoid intestinal perforation in children: A continuing scourge in a developing country. Pediatr Surg Int **2007**; 23:33–39.

128. Bulage L, Masiira B, Ario AR, et al. Modifiable risk factors for typhoid intestinal perforations during a large outbreak of typhoid fever, Kampala Uganda, 2015. BMC Infect Dis **2017**; 17:1–7.

129. Qamar FN, Azmatullah A, Bhutta ZA. Challenges in measuring complications and death due to invasive Salmonella infections. Vaccine **2015**; 33:C16–C20.

130. Msefula CL, Olgemoeller F, Jambo N, et al. Ascertaining the burden of invasive Salmonella disease in hospitalised febrile children aged under four years in Blantyre, Malawi. PLoS Negl Trop Dis **2019**; 13:1–16.

131. Chanh NQ, Everest P, Khoa T, et al. A Clinical, Microbiological, and Pathological Study of Intestinal Perforation Associated with Typhoid Fever. Clin Infect Dis **2004**; 39:61–67.

132. Brainard J, D'hondt R, Ali E, et al. Typhoid fever outbreak in the Democratic Republic of Congo: Case control and ecological study. PLoS Negl Trop Dis **2018**;

133. Nyamusore J, Nahimana MR, Ngoc CT, et al. Risk factors for transmission of Salmonella Typhi in Mahama refugee camp, Rwanda: A matched case-control study. Pan Afr Med J **2018**; 29.

134. Kabwama SN, Bulage L, Nsubuga F, et al. A large and persistent outbreak of typhoid fever caused by consuming contaminated water and street-vended beverages: Kampala, Uganda, January - June 2015. BMC Public Health **2017**; 17.

135. US Centers for Disease Control and Prevention. Summary of Notifiable Diseases **1997**;

136. Salgado R, Marques R, Noronha JP, et al. Assessing the diurnal variability of pharmaceutical and personal care products in a full-scale activated sludge plant. Environ Pollut **2011**; 159:2359–2367.

Supplementary Material 2.1

Included for the purposes of completeness, but not written or contributed to by JSKG.

Samples were cultured for 24 hours at 37 ˚ C in air in buffered peptone water (BPW), then sub-cultured onto MacConkey agar for 24 hours. 2 mls of BPW were subsequently added into 8 ml of Selenite and incubated for 24 hours. DNA extraction was performed from Selenite supernatants using the QIAamp® Fast DNA Stool Mini Kit (Qiagen, Hilden, Germany) pathogen detection protocol. Multiplex PCRs were performed in a Biorad CFX96 thermal cycler using the Quantifast Pathogen PCR+IC Kit®, targeting the pan *Salmonella* invasion A gene, the *Salmonella* Typhi fimbriae gene, and the kit´s internal control. A 5 minute Taq activation step at 95°C was followed by 40 cycles of annealing/extension (30 sec, 60°C) and denaturation (15 sec, 95°C). PCR signals were analyzed using the CFX Manager 3.1. Software (Biorad). Valid PCRs required the cycle threshold signal to range from 29-31 for the internal control. A cycle threshold <40 was considered positive in the presence of a typical exponential amplification curve. Detection of *Salmonella* Typhi required both pan *Salmonella* and typhoid-specific signal to be positive.

Supplementary Material 2.2

Drafted by PJD and edited/ implemented by JSKG.

S2.2.1 Notation

N = population size; n = number of cases; m = number of controls, assumed to be sampled at random from non-cases.

Denote the following quantities for each person i in the population (suppressing the subscript i =1, ..., N temporarily):

P(sampled) = α, P(sampled|case) = 1, P(sampled|non-case) = $m/(N - n) = f$, P(case) = p, P(case|sampled) = p^*

S2.2.2 Derivation

Laws of probability now give:

α = P(case and sampled) + P(non-case and sampled)

= $p \times 1 + (1 - p) \times f$ [S2.2.1]

and

p = P(sampled and case) + P(not sampled and case)

= $\alpha \times p^* + (1 - \alpha) \times 0$ [S2.2.2]

Combining [S2.2.1] and [S2.2.2] gives

$p + (1 - p)f = p/p^*$,

hence

$p = p^* f/\{1 - p^*(1 - f)\}$ [S2.2.3]

S2.2.3 Application

Now extend the notation to $p(x_i,\ y_i)$: $i = 1,\ ...,\ N$ where y_i is the value of the exposure of interest for person i and x_i are the values of their other covariates. Similarly, write $p_i^*(x_i,\ y_i)$, noting that its value is hypothetical for a person i who is neither a case nor a control. Then, writing $x = (x_1,\ ...,\ x_N)$ and $y = (y_1,\ ...,\ y_N)$, equation S2.2.3 gives the hypothetical expected number of cases in the total population as:

$$\mu(x,y) = \sum_{i=1}^{N} p^*(x_i,y_i) f / [\{1 - p^*(x_i,y_i)(1-f)\}] = \sum_{i=1}^{N} q(x_i,y_i), \quad \text{[S2.2.4]}$$

If we now label $i = 1,\ ...,\ n$ as the cases and $i = n + 1,\ ...,\ n + m$ as the controls, then we can estimate [S2.2.4] by:

$$\hat{\mu}(x,y) = \sum_{i=1}^{n} q(x_i,y_i) + \frac{N-n}{m} \sum_{i=n+1}^{n+m} q(x_i,y_i), \quad \text{[S2.2.5]}$$

because the controls are a random sample of the non-cases.

Hence, the change in the expected number of cases if the actual sets of covariates x and exposures y change to hypothetical sets x' and y' is $\hat{\mu}\ (x\ ,\ y) - \hat{\mu}\ (x'\ ,\ y')$. For our application, we set $x' = x$ and $y' = 0$. Finally, note that as a reality check, we should get $\hat{\mu}\ (x,\ y) \approx n$. The attributable risk can be calculated as $\{\hat{\mu}\ (x,\ y) - \hat{\mu}\ (x'\ ,\ y'\ )\}/\hat{\mu}\ (x,\ y)$.

Supplementary Material 2.3:

Written by JSKG.

S2.3.1 Variable selection

We present the log likelihoods and p-values from the likelihood ratio tests for each iteration of the variable selection below.

Table S2.3.1 Results of the variable selection for each iteration.

Iteration	Variable	log likelihood	p-value
0	Residential ward	-310.84	-
1	Seeking care at QECH if child is severely ill	-287.5	8.28e-12
2	Age (years)	-273.8	1.73e-07
3	Number of drinking water sources used last three weeks	-263.6	5.96e-06
4	Stores drinking water in drum	-259.9	0.0066
5	Number of household members admitted to hospital for febrile illness in last four weeks	-255.6	0.0034
6	Distance to from household to primary water source (meters)	-251.7	0.016
7	Number of days water is stored	-248.72	0.029
8	Family grows crops	-246.1	0.023
9	Cooking and cleaning with river water in the previous three weeks	-243.9	0.034
10	Used stream or river water for drinking in the last three weeks	-240.2	0.0069
11	Child spends the day at school, preschool, nursery or any other daycare	-237.7	0.025

12	Cooking and cleaning using water from an open dug well in the previous three weeks	-235.5	0.033
13	Experienced water shortage in the house or surrounding area in the past two weeks	-231.8	0.031
14	Soap available to wash hands after the toilet in the previous three weeks	-229.7	0.040

S2.3.2 Spatial dependency

Because this study recorded GPS coordinates of participants' households, we were able to test for spatial correlation of the residuals from the fitted multiple logistic regression model. We extracted Pearson residuals, r, from the output of the fitted model, i.e. $r = (y - p)/\sqrt{\{p(1-p)\}}$ where, for each participant, y= 0/1 indicates control/case, respectively and p is the fitted probability that the participant is a case.

In order to test for spatial correlation, we randomly permuted the household locations of each individual. This was repeated 500 times, and a variogram of the residuals was calculated for each permutation up to 3 kilometers. This distance was chosen from the mean square-root of the area (2679m) of the residential wards, approximating an average length of each enumeration area, since spatial correlation on any larger scale was controlled for by matching on the residential ward. Next for each distance bin up to 5km, we calculate 90% tolerance envelope of the variogram as the interval from the 25th to the 475th of the 500 ordered values of the corresponding variogram ordinates. The calculated 95% tolerance envelope fully contains the whole of the empirical variogram from the final model (Figure S2.3.1), consistent with the absence of residual spatial correlation.

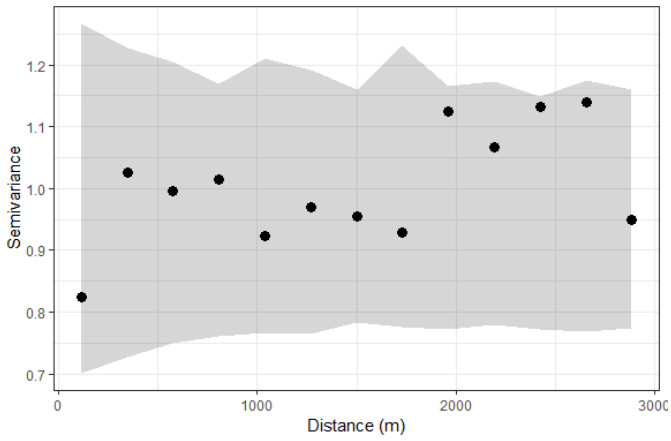


Figure S2.3.1 Variogram of Pearson residuals from the final multivariate model ($\cdot$), with the shaded area indicating the 95% tolerance envelope under the assumption of spatial independence.

We further define a test statistic to evaluate the variogram of the residuals from the final model against the null distribution generated by the randomly permutated household locations. This is generated for each permutation i, given in the equation S2.3.1 below:

$$Tstat_i = \sum_{j=1}^{K} n_{ij}(V_{ij} - \overline{V}_j)^2 \qquad [S2.3.1]$$

Where K is the number of variogram bins, V_{ij} is the calculated variogram ordinate in permutation i and bin j. $\overline{V}_j$ is the weighted average of the variogram ordinates in bin j for N permutations:

$$\overline{V}_j = \frac{\sum_{i=1}^{N} V_{ij} n_{ij}}{\sum_{i=1}^{N} n_{ij}} \qquad [S2.3.2]$$

We then compare the test statistic of the variogram from our final model, t, with the calculated values from the permutated locations (Figure S2.3.2). The p-value of the test is

$$p = \frac{1}{N}\sum_{h=1}^{N} I[T(h)>t] \qquad \text{[S2.3.3]}$$

Where $I\,[a > b] = 1$ if a > b and 0 otherwise. From this, we calculate the p-value to be 0.464, which is again consistent with the absence of residual spatial correlation.

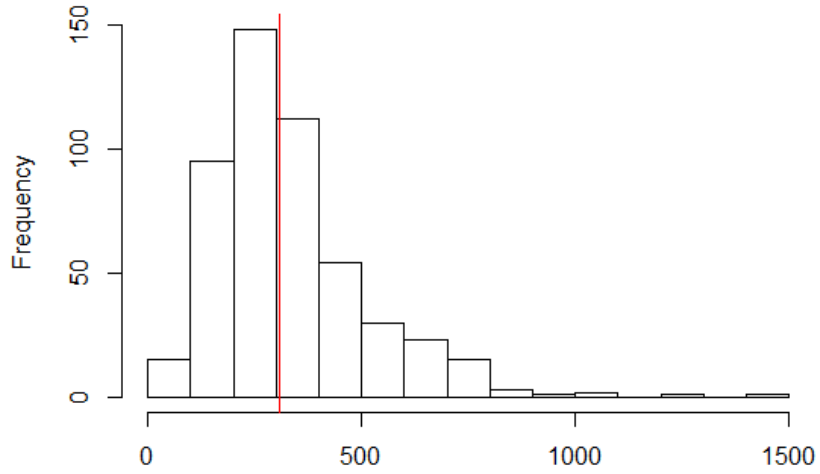


Figure S2.3.1 Histogram of the test statistics calculated from the 500 permutations, with the final model's calculated test statistic value marked by the red line.

Supplementary Material 3.1

Written by JSKG.

A digital elevation map (DEM) was downloaded from the United States Geological Survey (USGS). Two tiles spanning the Blantyre area were available, with data from the Shuttle Radar Thematic Mapper (SRTM) Version 3, recorded in 1 arc-second resolution (approximately 30 meters) [1].

All hydrological calculations used ArcGis Version 10.7 and ArcHydro tools 2.0. The DEM was reconditioned for consistency with a river map obtained from the Blantyre City Council. Flow direction was calculated, and estimated accumulation was visually compared to the known rivers, confirming agreement of the DEM with local maps (Figure S3.1.1). Pour-points were selected at the city limits, and along with the flow direction layer was used with ArcHydro's Watershed tool to estimate hydrological catchments for major rivers.

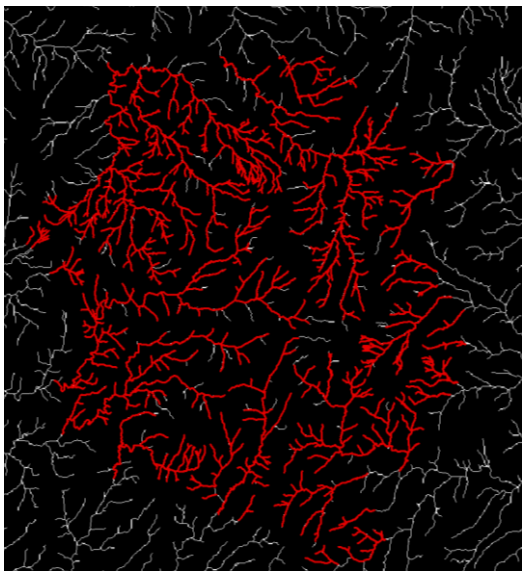


Figure S3.1.1 Map output from ArcMap showing estimated streams by flow accumulation (white), and known rivers (red).

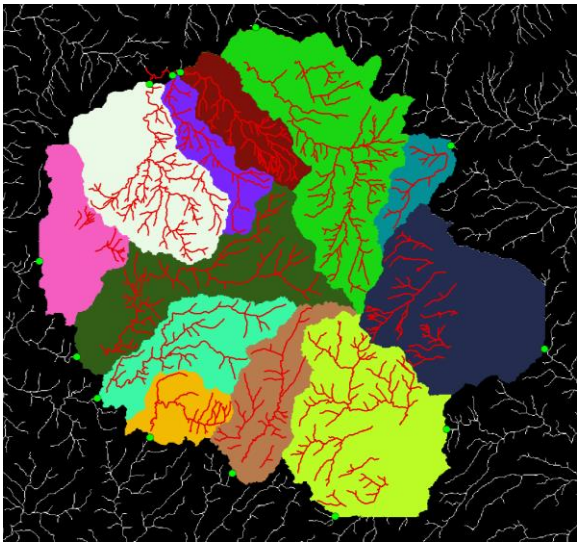


Figure S3.1.2 Pour points (green), and hydrological catchments displayed in multi-colored polygons.

1. USGS. SRTM Topography. SRTM Doc. **2009**;

Supplementary Material 3.2

Written by JG.

S3.2.1 Non-spatial Poisson log-linear model

A Poisson log-linear model was used to model incidence across the city, initially with the assumption of no spatial dependence. We utilized available covariates for each enumeration area (EA): distance to QECH, elevation and river catchment at the centroid of the EA, and average household size and population density per square km across the enumeration area. For each enumeration area, we have age-stratified data of the population sizes in age bins of <5, 5-14, and 15+ years of age, and therefore can explore incidence rates in each enumeration area (i) and age band (j), where $d_i'\beta$ represent enumeration area-specific predictors, α_j are age-band specific intercepts, and N_{ij} are age-band and enumeration area-specific offsets.

$$Y_{ij} \sim Poisson\left(\mu_{ij}\right) \qquad \text{[S3.2.1]}$$

$$\mu_{ij} = N_{ij}\, exp\left(\alpha_j + d_i'\beta\right)$$

Results from the model are shown in Table S3.2.1; estimated coefficients are relative to the 15+ age band. Average household size and age band were found to be significant predictors of incidence in the multivariate model. River catchment 6 was marginally predictive of an elevated incidence.

Table S3.2.1 Estimated parameters from non-spatial Poisson log-linear incidence model.

Parameter	**Estimate**	**Standard error**	**P-value**
Intercept	-3.18E+00	1.77E+00	7.30E-02
Distance to QECH	-3.97E-05	6.52E-05	5.43E-01
Elevation	-3.78E-04	1.33E-03	7.76E-01
Average household size	-1.05E+00	2.65E-01	7.06E-05
Density	-1.06E-05	7.28E-06	1.45E-01
Age 5-14	1.29E+00	1.32E-01	1.40E-22
Age <5	1.04E+00	1.68E-01	5.21E-10
Catchment 1	1.69E-01	2.43E-01	4.87E-01
Catchment 2	-1.25E-01	4.62E-01	7.86E-01
Catchment 3	8.69E-02	3.55E-01	8.07E-01
Catchment 4	2.41E-01	4.93E-01	6.24E-01
Catchment 5	-2.65E-01	3.98E-01	5.05E-01
Catchment 6	6.45E-01	3.42E-01	5.93E-02
Catchment 7	-1.00E-01	2.26E-01	6.58E-01
Catchment 8	-2.57E-01	7.23E-01	7.23E-01
Catchment 9	-1.50E-01	2.94E-01	6.10E-01
Catchment 10	4.78E-02	3.51E-01	8.92E-01

Covariates of average household size and age band were retained as the base model for further analyses. We explored the addition of any of the other four variables, but we found no significant ($p<0.05$) improvement in model fit with the addition of any of these variables (Table S3.2.2). Therefore, we use average

household size as predictor of incidence in each age band across the city for further analyses (Table S3.2.3).

Table S3.2.2 Summary of the contribution of added variables to the model, evaluated using the likelihood ratio test.

Model	**LL**	**P-value**
Base model	-596.66	-
Base model + elevation	-596.20	0.334
Base model + density	-595.68	0.161
Base model + hospital distance	-596.11	0.294
Base model + river catchment	-590.91	0.320

Table S3.2.3 Summary of final model coefficients.

Parameter	**Estimate**	**Standard error**	**P-value**
Intercept	-4.61	0.941	<0.001
Average household size	-0.90	0.132	<0.001
Age 5-14	1.29	0.168	<0.001
Age <5	1.04	0.222	<0.001

S3.2.2 Assessing spatial dependence

We explore whether spatial dependence of the residuals exists in the non-spatial model. We calculate the standardized Pearson residuals at the centroid of each enumeration area, i, by combining the expected counts $\widehat{\mu}_{ij}$ and observations y_{ij} for age band j:

$$r_i = \frac{\sum_{j=1}^{3} y_{ij} - \sum_{j=1}^{3} \hat{\mu}_{ij}}{\sqrt{\sum_{j=1}^{3} \hat{\mu}_{ij}}} \qquad \text{[S3.2.2]}$$

To test for spatial dependence, we randomly permutated the centroid locations s=500 times. We then constructed an empirical variogram for each permutation up to 10,000 meters, approximately half the linear dimensions of the study area. We calculated values for the empirical variogram as:

$$\hat{V}(u) = \frac{1}{2|K(j)|} \sum (r_h - r_k)^2 \qquad \text{[S3.2.3]}$$

Where, for each value of j, $|K(j)|$ is the number of pairs in distance bin j and the summation is over all pairs h and k corresponding to pairs of locations whose distance apart falls within distance bin j. We calculated a 95% tolerance envelope of the variogram as the interval from the 13th to the 487th of the 500 ordered values of the corresponding variogram ordinates for each distance bin. The lower limit of the tolerance envelope lies substantially above 1 at all plotted distances, suggesting over-dispersion relative to the Poisson distribution. Also, the 95% tolerance envelope does not contain all points in the empirical variogram (Figure S3.2.1), suggesting the presence of some residual spatial correlation.

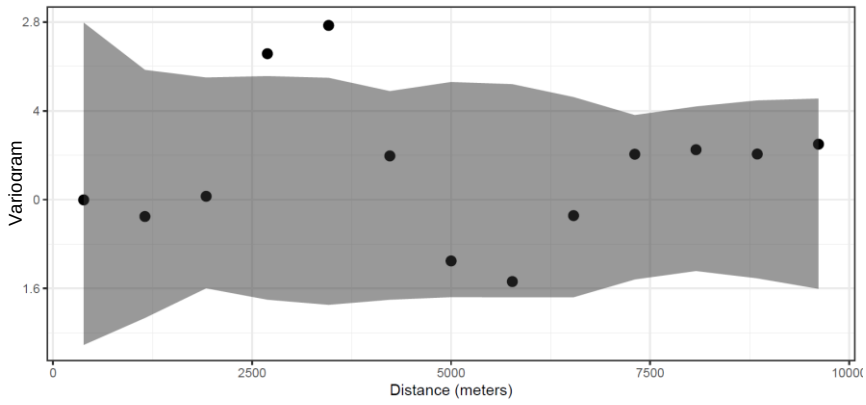


Figure S3.2.1 Empirical variogram of the residuals from the non-spatial generalized linear model, with the 95% tolerance envelope under the assumption of spatial randomness.

To test this formally, we define a test statistic to evaluate the variogram of the residuals from the final model against the null distribution generated by the randomly permutated centroid locations. This is generated for each permutation i, given in the equation below:

$$Tstat_i = \sum_{j=1}^{K} n_{ij}(V_{ij} - \overline{V}_j)^2 \qquad \text{[S3.2.4]}$$

Where K is the number of variogram bins, V_{ij} is the calculated variogram ordinate in permutation i and bin j. $\overline{V}_j$ is the weighted average of the variogram ordintes in bin j over N permutations:

$$\overline{V}_j = \frac{\sum_{i=1}^{N} V_{ij} n_{ij}}{\sum n_{ij}} \qquad \text{[S3.2.5]}$$

We then compare the test statistic of the variogram from our final model, t, with the calculated values from the permutated locations (Figure S3.2.2).

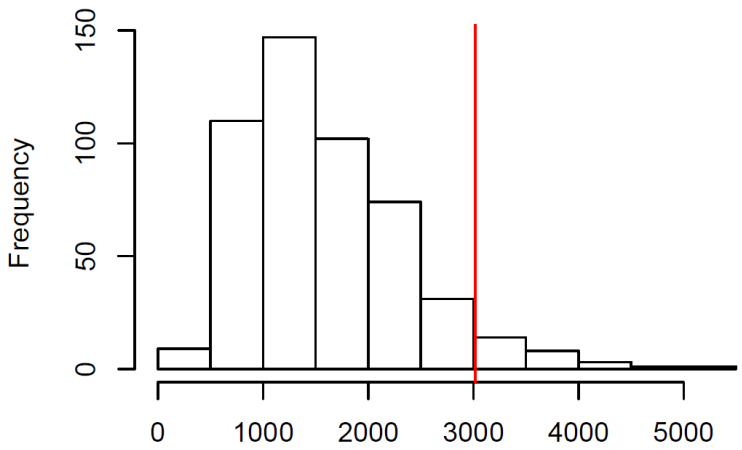


Figure S3.2.2 Histogram of the calculated test statistics from 500 permutations, with the empirical test statistic shown in red.

The p-value of the test is:

$$p = \frac{1}{N}\sum_{h=1}^{N} I\left[T(h) > t\right] \qquad \text{[S3.2.6]}$$

Where $I\left[a > b\right] = 1$ if a > b and 0 otherwise. From this, we calculate the p-value to be 0.054. Based on this statistic and the above visualization, there appears to be marginal evidence of spatial dependence in the data, meriting further analyses using an extended model that includes a spatial random effect.

S3.2.3 Geostatistical model

Next, we extend our model to allow for over-dispersion and spatial dependence:

$$Y_{ij} \sim Poisson\left(\mu_{ij}\right)$$

$$\mu_{ij} = N_{ij}\exp\left(\alpha_j + d_i'\beta + Z_i + S(x_i)\right) \qquad \text{[S3.2.7]}$$

Where $S(x)$ is a spatial random effect with a Matérn correlation function and kappa = 0.5 [54]. The model was fit using Monte-Carlo maximum likelihood (MCML), with 20,000 simulations, a burn-in of 10%, and a thinning parameter

of 10. Initial values for the regression coefficients were taken from the fitted parameters of the non-spatial model, while the spatial covariance parameters (σ^2, τ^2, φ, representing the variance of S(x), variance of Z_i, and the range of the spatial correlation, respectively) were estimated from a least-squares fit of the empirical variogram. MCML was repeated three more times, updating the initial values with estimates from the previous iteration. Diagnostics for the final iteration are shown for a randomly selected enumeration area in Table S3.2.4, with code used from PrevMap [52]. These diagnostics show little correlation between runs in thinned samples, visually apparent from the first and second columns, as well as a similar distribution of values in the first 900 and second 900 thinned samples, indicating stability in estimates over the iterations and convergence of the algorithm.

Estimated parameters for the non-spatial and spatial models are summarized in Table S3.2.5, and in the manuscript. Coefficient estimates appear similar to the nonspatial model estimates, though with reduced standard error. The covariance parameter estimates show that much of the variance in the model is captured in the nugget (τ^2) relative to the sill (σ^2), consistent with the over-dispersion observed in Figure S3.2.1, while the value phi indicates the practical range of spatial correlation (>5%) reaches approximately 1600 meters.

Table S3.2.4 MCMC diagnostic plots for the geostatistical model indicating convergence.

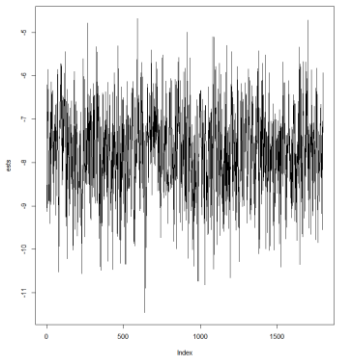
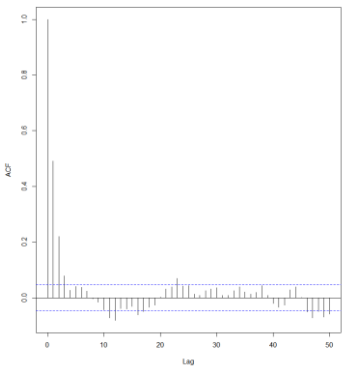
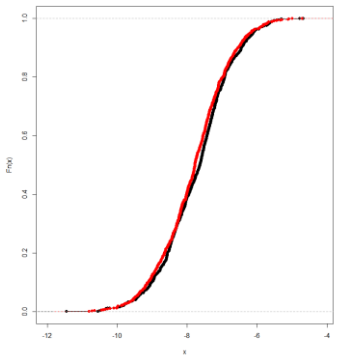
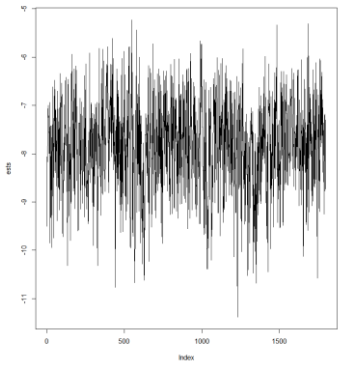
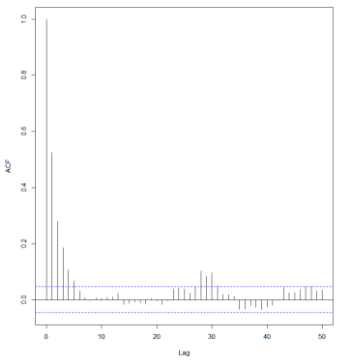
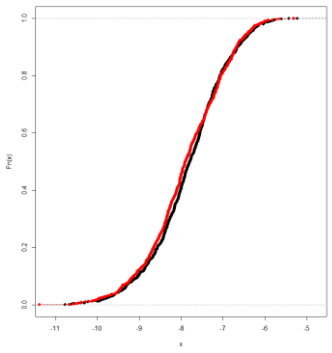
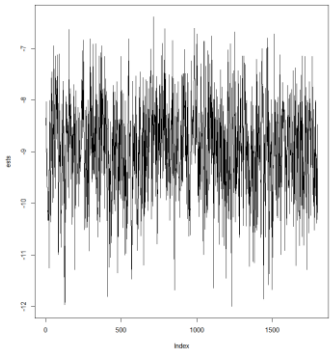
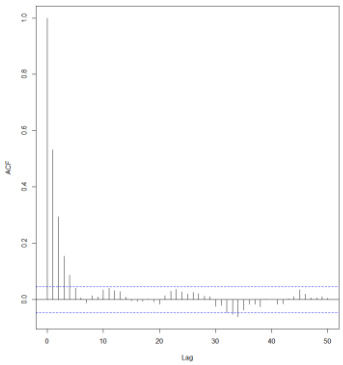
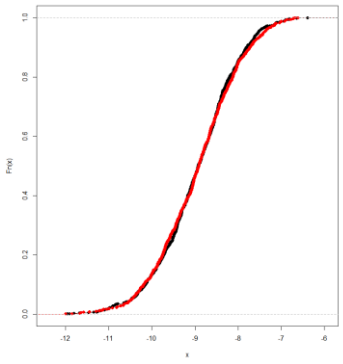
Age band	Trace of estimates by iteration	Autocorrelation of estimates between iterations	Cumulative distribution of estimates for first 900 and second 900 thinned samples
<5			
5-14			
15+			

Table S3.2.5 Parameter estimates for incidence model with and without spatial random effect.

	Nonspatial model			Spatial model		
Parameter	**Est.**	**Standard error**	**P-value**	**Est.**	**Standard error**	**P-value**
Intercept	-4.62	0.94	<0.001	-5.25	0.560	<0.001
Average household size	-0.90	0.22	<0.001	-0.829	0.129	<0.001
Age 5-14	1.29	0.13	<0.001	1.108	0.076	<0.001
Age <5	1.04	0.16	<0.001	1.168	0.075	<0.001
$\log(\sigma^2)$	-	-	-	-1.797	0.243	-
$\log(\varphi)$	-	-	-	6.269	0.331	-
$\log(\tau^2)$	-	-	-	-0.251	0.510	-

Predictions of incidence at each centroid are calculated in PrevMap, again with 20,000 simulations, a burn-in of 10%, and a thinning parameter of 10. We additionally calculate the rates attributed to the covariate, as well as the rates attributed to the spatial signal. These estimates can be separated into two components (colored in red and blue below, respectively):

$$Y_{ij} \sim Poisson\left(\mu_{ij}\right)$$

$$\mu_{ij} = \exp\left(\alpha_j + d_i'\beta\right) * \exp(S(x)) * N_{ij} \qquad [S3.2.8]$$

We want to estimate these across all age bins j and for individual enumeration areas i. Using the additive properties of Poisson rates, we can combine estimated μ_{ij} across all 3 age bands:

$$\mu_i = \Sigma(\mu_{i1} + \mu_{i2} + \mu_{i3}) \quad [S3.2.9]$$

$$\mu_i = \exp(S(x)) * [\exp(\alpha_{i1} + d_i'\beta) * N_{i1} + \exp(\alpha_{i2} + d_i'\beta) * N_{i2} + \exp(\alpha_{i3} + d_i'\beta) * N_{i3}]$$

We calculate the contribution of the estimated covariates to the incidence directly, using estimated coefficients as

$$C_i = \exp(\alpha_{i1} + d_i'\beta) * N_{i1} + \exp(\alpha_{i2} + d_i'\beta) * N_{i2} + \exp(\alpha_{i3} + d_i'\beta) * N_{i3} \quad [S3.2.10]$$

and the contribution of the spatial random effect as:

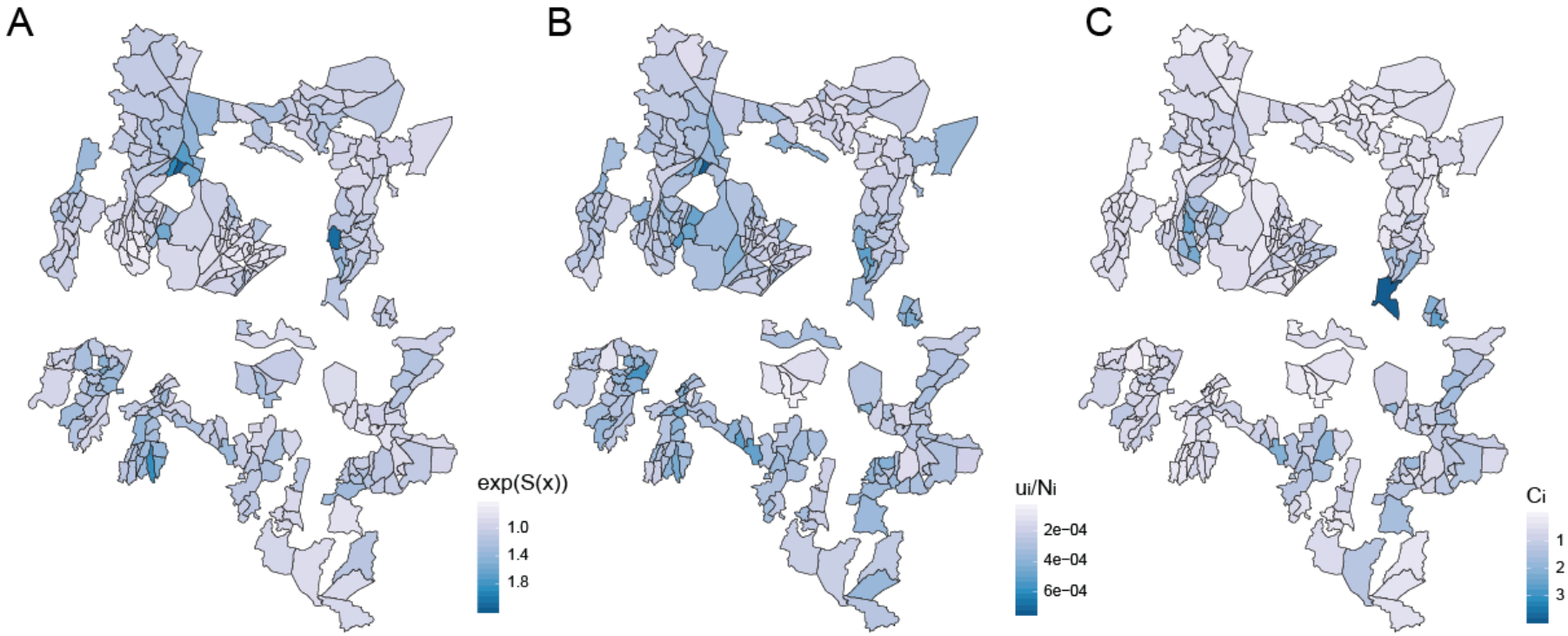
$$\exp(S(x)) = \frac{\mu_i}{C_i}. \quad [S3.2.11]$$

Each component is plotted in Figure S3.2.3. Some of the high incidence regions that appear in the model (Figure S3.2.3B) were attributed to the model covariate (Figure S3.2.3C), but others are not explained by measured covariates, and instead are captured by the spatial random effect (Figure S3.2.3A), indicating that there may be unmeasured processes contributing to these hot-spots.

1. Matérn, B.: Spatial Variation. 2nd Ed., Springer-Verlag, Berlin, Heidelberg, New York, London, Paris, Tokyo **1986**;

2. Giorgi E, Diggle PJ. PrevMap : An R Package for Prevalence Mapping. J Stat Softw **2017**;

Figure S3.2.3 Spatial (A) and covariate (C) attributed contributions to the model predicted incidence rate (B).



Supplementary Material 3.3

Included for the purposes of completeness, but not written or contributed to by JG.

S. Typhi from consenting participants were isolated, DNA extracted with the Qiagen Universal Biorobot® (Limburg, Netherlands) using Qiagen All-for-one® extraction kits, and subjected to whole genome sequencing on Illumina HiSeq2500 machines (Illumina, San Diego, CA, USA) generating 150 bp paired-end reads. For the pan-genome analysis, annotated assemblies were produced using the pipeline described in [1]. For each sample, sequence reads were used to create multiple assemblies using VelvetOptimiser v2.2.5 (Velvet Optimiser: For automatically optimising the primary parameter options for the Velvet de novo sequence assembler. Gladman, S & Seemann, T, Victorian Bioinformatics Consortium, 2008. http://bioinformatics.net.au/software.velvetoptimiser.shtml) and Velvet v1.2 [2]. An assembly improvement step was applied to the assembly with the best N50 and contigs were scaffolded using SSPACE [3] and sequence gaps filled using GapFiller [4]. Automated annotation was performed using PROKKA v1.5 [5] and a genus specific database from RefSeq [6].

All of the software developed by Pathogen Informatics at the WSI is freely available for download from GitHub (Pathogen Informatics, WSI, https://github.com/sanger-pathogens/vr-codebase; Bio-Assembly-Improvement: Improvement of genome assemblies by scaffolding and gapfilling, Pathogen Informatics, WSI, https://github.com/sanger-pathogens/assembly_improvement) under an open source license, GNU GPL 3. The improvement step of the pipeline is also available as a standalone Perl

module from CPAN (http://search.cpan.org/~ajpage/). The core- and pan-genome were analyzed using roary [7] for gene-based comparisons.

S3.3.1 Single nucleotide polymorphisms (SNPs)

Reads were mapped against the high-quality reference genome of *S.* Typhi 1036491 isolated in Blantyre, Malawi 2012 (GCA_001367555.3). All bases were filtered to remove those with uncertainty in the base call. The bcftools variant quality score was required to be greater than 50 (quality < 50) and mapping quality greater than 30 (map_quality < 30). If not all reads gave the same base call, the allele frequency, as calculated by bcftools, was required to be either 0 for bases called the same as the reference, or 1 for bases called as a SNP (af1 $<$ 0.95). The majority base call was required to be present in at least 75% of reads mapping at the base, (ratio < 0.75), and the minimum mapping depth required was 4 reads, at least two of which had to map to each strand (depth < 4, depth_strand < 2). Finally, strand_bias was required to be less than 0.001, map_bias less than 0.001 and tail_bias less than 0.001. If any of these filters were not met, the base was called as uncertain. An alignment was constructed by substituting the base call at each site (variant and non-variant) in the BCF file into the reference genome and any site called as uncertain was substituted with an N for each respective isolate.

S3.3.2 Phylogenetic analyses

A pairwise SNP distance matrix of this alignment was generated selecting only sites containing ACGT (no gaps or Ns) using snp_sites [8], resulting in 436 informative sites for the pairwise comparison, used for further geo-spatial modelling. Recombinant sites and mobile elements were removed following analysis of the mapping-based alignment with gubbins v2.3.4 [10] as well as phage characterization using PHASTER [11] and manually curating the output.

Informative sites were then extracted from this alignment using snp_sites [8]; only sites containing ACGT (no gaps or Ns) were used for the final analysis, resulting in 409 informative SNPs in the final alignment. For the phylogenetic analyses, the informative SNP alignment was used as input for iq-tree [12] for phylogenetic tree reconstruction under the general time-reversible (GTR) model, ascertainment (ASC) correction for a SNP-only alignment and under Gamma distribution (-m GTR+G+ASC), support was assessed using 1000 bootstrap replicates. The resulting tree was assessed for phylogenetic signal using tempest (v1.5.1) and the isolate collection days as recorded by QEH, however the root-to-tip correlation (0.07) indicated not enough temporal signal to allow a temporal analysis (supplement). The phylogenetic tree was reconstructed into a joint ancestral tree using pyjar, and rPinecone [13] was used to further group the isolates based on this tree, using 2 and 4 as relevant SNP cutoffs for minor and major clusters, respectively. Pairwise tip-to-tip distances were calculated using the adephylo package for R with the command distTips from the alignment before recalculation with pyjar.

1. Page AJ, De Silva N, Hunt M, et al. Robust high-throughput prokaryote de novo assembly and improvement pipeline for Illumina data. Microb Genomics **2016**;

2. Zerbino DR, Birney E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Res **2008**;

3. Boetzer M, Henkel C V., Jansen HJ, Butler D, Pirovano W. Scaffolding pre-assembled contigs using SSPACE. Bioinformatics **2011**;

4. Boetzer M, Pirovano W. Toward almost closed genomes with GapFiller. Genome Biol **2012**;

5. Seemann T. Prokka: Rapid prokaryotic genome annotation. Bioinformatics **2014**;

6. Pruitt KD, Tatusova T, Brown GR, Maglott DR. NCBI Reference Sequences (RefSeq): Current status, new features and genome annotation policy. Nucleic Acids Res **2012**;

7. Page AJ, Cummins CA, Hunt M, et al. Roary: Rapid large-scale prokaryote pan genome analysis. Bioinformatics **2015**;

8. Page AJ, Harris SR, Seemann T, et al. SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. Microb Genomics **2016**;

9. Wong VK, Baker S, Connor TR, et al. An extended genotyping framework for Salmonella enterica serovar Typhi, the cause of human typhoid. Nat Commun **2016**;

10. Croucher NJ, Page AJ, Connor TR, et al. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. Nucleic Acids Res **2015**;

11. Arndt D, Grant JR, Marcu A, et al. PHASTER: a better, faster version of the PHAST phage search tool. Nucleic Acids Res **2016**;

12. Nguyen LT, Schmidt HA, Von Haeseler A, Minh BQ. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol **2015**;

13. Wailan AM, Coll F, Heinz E, et al. rPinecone: Define sub-lineages of a clonal expansion via a phylogenetic tree. Microb Genomics **2019**;

Supplementary Material 3.4

Written by JG.

S3.4.1 Motivation for spatial analysis: Correlation

Though it is commonly assumed that epidemiologically-linked individuals tend to have genetically related isolates, due to differences in transmission patterns between diseases, it is less established that spatially-close individuals are genetically linked. Therefore prior to geostatistical modelling of genetic data, we explored the correlation between spatial and genetic distances in our dataset.

The SNP data is represented as a nxn matrix of genetic distances. Using the household location of the patients, we then generated spatial distances between all patients. Next, the correlation between physical distance and SNP distance for all combinations of isolates was calculated, resulting in a value of 0.071.

In order to test the significance of this value, we randomly permuted the location labels of the individuals included in the genetic distance matrix, and calculated the correlation between SNP distance and physical distance. This process was repeated 1000 times.

We then compared our empirical correlation statistic, t, with those generated from the randomized values, C(h), using the calculated p-value:

$$p = \frac{1}{N+1}\sum_{h=1}^{N} I[C(h)>t]+1 \qquad \text{[S3.4.1]}$$

Where N is the number of permutations, $I[a>b]=1$ if $a>b$ and 0 otherwise. The distribution of the permuted test statistics is shown in Figure S3.4.1, with the

empirical test statistic shown in red. The resulting p-value is <0.001, indicating that there is evidence of spatial-genetic correlation in our dataset.

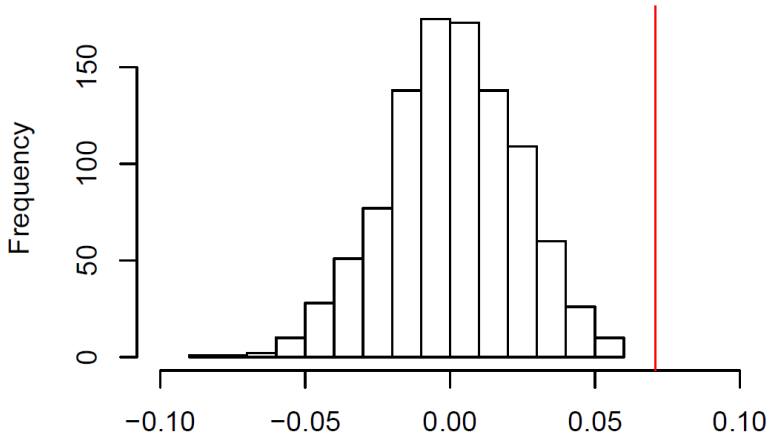


Figure S3.4.1 Histogram of calculated test statistics from 1000 permutations, with the empirical test statistic shown in red.

S3.4.2 Exploration of multidimensional scale: PC1

The variogram in Figure S3.4.2 does not suggest that PC 1 of the multidimensional scale has any spatial correlation up to 5 km.

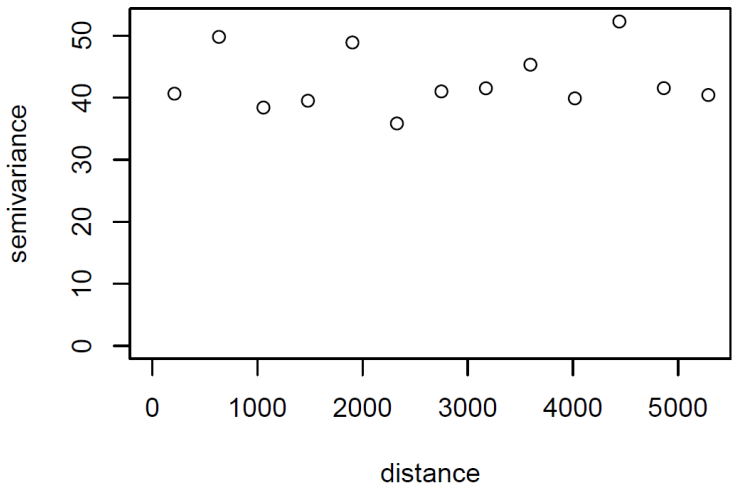


Figure S3.4.2 Empirical variogram of the genetic score for PC 1.

Regardless, there were 11 individuals with a genetic score of approximately -30 on PC 1 of our multidimensional scale of the pairwise SNP distance matrix. Available covariates to investigate these individuals were age, time of infection, and household location (Figure S3.4.3). No significant difference in average age exists between these individuals and the rest of the cohort (13.8 vs. 15.7, p=0.67).

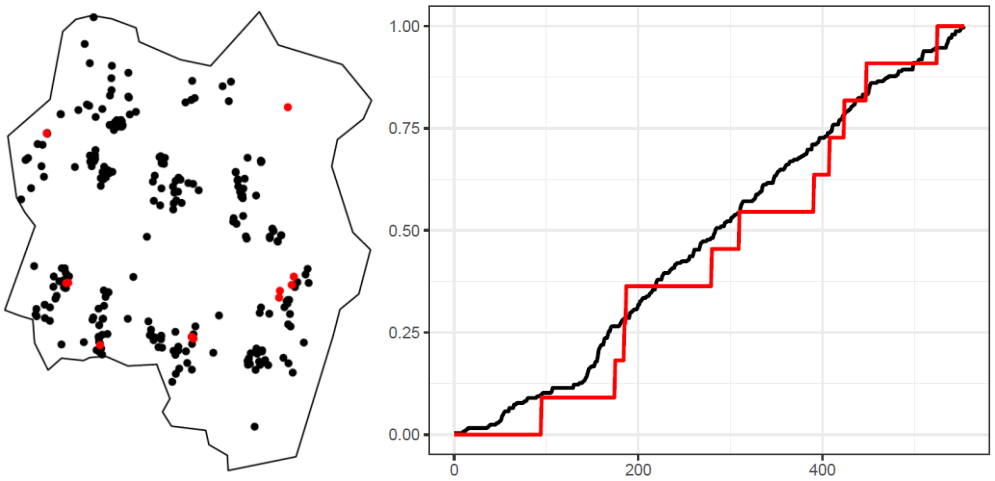


Figure S3.4.3 Spatial distribution of cases (left), and cumulative proportion of cases over the study period (right), with the investigated individuals highlighted in red.

To evaluate spatial clustering of these individuals versus the rest of the cohort, we generated K-functions across the study region up to 5000 meters (Figure S3.4.4), and used a statistical test for point process clustering [1]. The test statistic is evaluated as the difference between K-functions (Figure S3.4.4), divided by the standard error of these differences, across the study region:

$$T = \sum_{S_0}^{S} \frac{K_C(s) - K_X(s)}{SE(s)} \qquad [S3.4.2]$$

Randomly permutating the labels for K_C and K_X, and repeating 500 times to create a null distribution, we generate a p-value of 0.072 indicating weak evidence of spatial clustering compared to the rest of the cohort. Given the small number of individuals in the evaluated group, there is little evidence to contribute to further conclusions regarding these individuals.

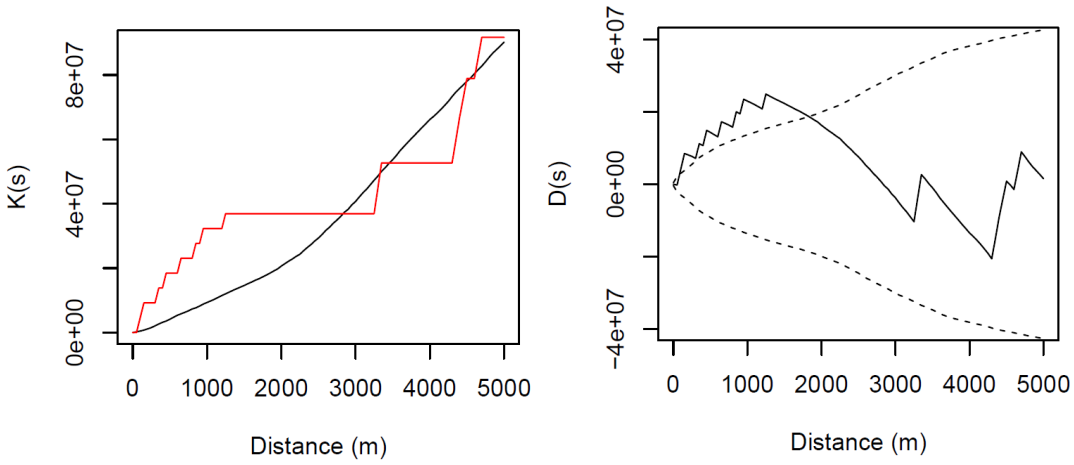


Figure S3.4.4 K(s) at evaluated distances for entire cohort (black) and evaluated individuals (red) (left), and the difference in K-function estimate for the evaluated individuals compared to the rest of the cohort with dashed lines indicating 2 +/- the standard error (right).

A similar approach was used to evaluate clustering over the study period, with the position in space (in two dimensions) replaced by one-dimensional position in time. The p-value for clustering over time was calculated as 0.5. Therefore, although this group shows distinct differences in genetic scores of PC1 in relation to the rest of the cohort, these individuals do not appear to be related in time or space, and do not show unique characteristics regarding age at infection.

S3.4.3 Exploration of multidimensional scale: PC2

The variogram of PC 2 shows visual evidence of spatial correlation (Figure 3.3C), therefore we conducted a statistical test to observe whether this pattern

is significant. To test for spatial dependence, we randomly permuted the labels for household locations of each isolate 500 times. We then constructed a variogram for each permutation up to 5000 meters, approximately 1/4 the range of the study area. We calculated 95% tolerance envelope of the variogram as the interval from the 13th to the 487th of the 500 ordered values of the corresponding variogram ordinates for each distance bin. The 95% tolerance envelope does not contain all points in the empirical variogram (Figure S3.4.5).

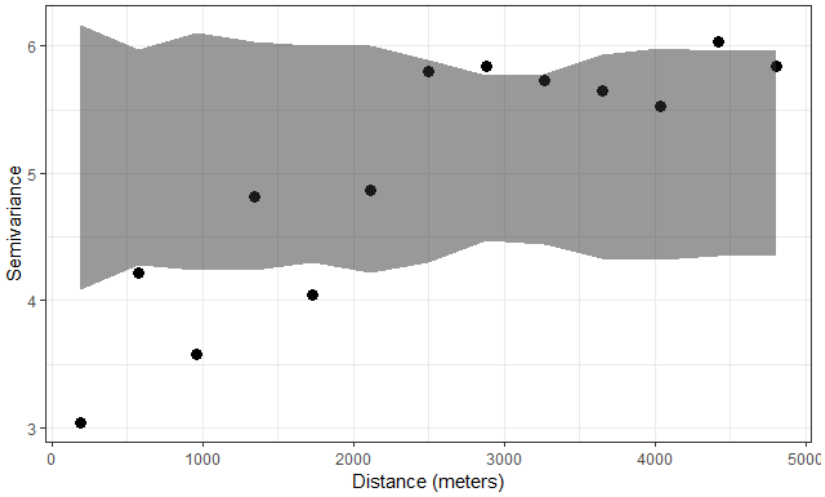


Figure S3.4.5 Empirical variogram values of PC 2 (points), with 95% tolerance envelope in shaded band.

We further define a test statistic to evaluate the variogram of the residuals from the final model against the null distribution generated by the randomly permutated household locations. This is generated for each permutation i, given in the equation below:

$$Tstat_i = \sum_{j=1}^{K} n_{ij}(V_{ij} - \overline{V}_j)^2 \qquad [S3.4.3]$$

Where K is the number of variogram bins, V_{ij} is the calculated variogram ordinate in permutation i and bin j. $\overline{V}_j$ is the weighted average of the variogram ordinates in bin j for N permutations:

$$\overline{V}_j = \frac{\sum_{i=1}^{N} V_{ij} n_{ij}}{\sum_{i=1}^{N} n_{ij}} \quad [S3.4.4]$$

We then compare the test statistic of the variogram from our final model, t, with the calculated values from the permutated locations (Figure S3.4.6). The p-value of the test is

$$p = \frac{1}{N} \sum_{h=1}^{N} I[T(h) > t] \quad [S3.4.5]$$

Where $I[a > b] = 1$ if a > b and 0 otherwise. From this, we calculate the p-value to be < 0.002 (none of the random statistics were greater than the empirical statistic), visualized in Figure S3.4.6. Based on this and the above visualization, there appears to be strong evidence of spatial dependence in PC 2 of the multidimensional scale.

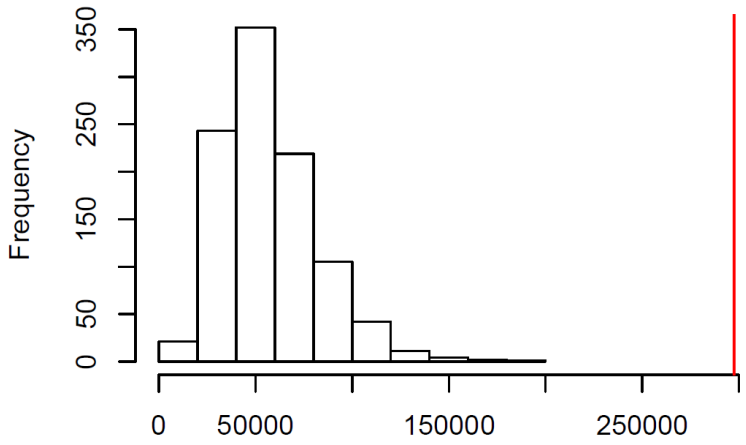


Figure S3.4.6 Histogram of randomly permuted test statistics, with the calculated value in red.

S3.4.4 Geostatistical modelling process

We first utilized an intercept-only linear model with a spatial random effect:

$$Y_i = \alpha + S(x_i) + Z_i \quad [S3.4.6]$$

S(x) is a spatial random effect with covariance parameters σ^2, φ, and τ^2, estimated from the data, with shape parameter of the Matérn function $\kappa = 1.5$ fixed, after evaluating the log likelihoods of the model at κ values at 0.5, 1, 1.5 and 2.

We can extend the above model to include river catchment:

$$Y_i = \alpha + \beta_{c(i)} + S(x_i) + Z_i \qquad \text{[S3.4.7]}$$

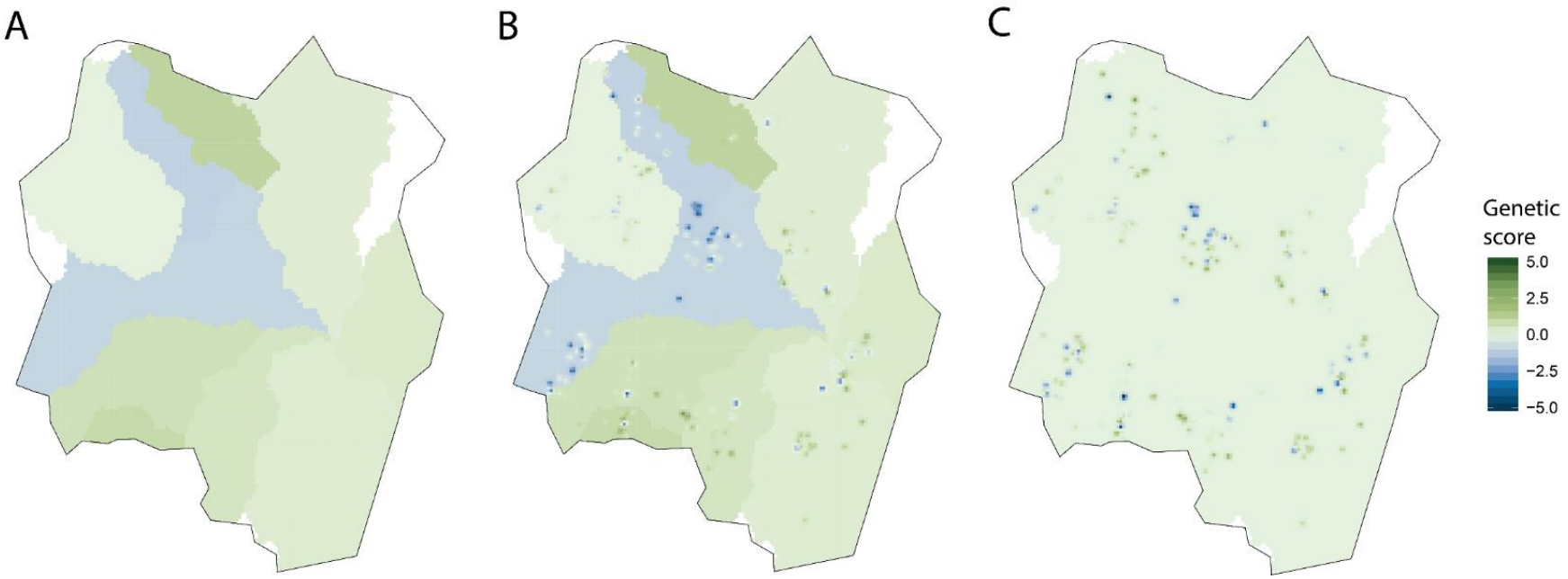
Where $c(i)$ is the catchment associated with location x_i for each location i, and $\beta_1 = 0$. Parameter estimates for both models S3.4.6 and S3.4.7 are summarized in Table S3.4.1. Catchment effects are relative to catchment 1.

Predicted genetic score across the city boundaries are shown in Figure S3.4.7 B, with the individual contributions from the covariate (Figure S3.4.7 A) and spatial random effects (Figure S3.4.7 C) shown.

Table S3.4.1 Covariance parameters and coefficient estimates from the geostatistical model with and without river catchment as a predictor.

	Intercept-only model			Intercept + river catchment		
Parameter	**Estimate**	**Standard error**	**P-value**	**Estimate**	**Standard error**	**P-value**
$sigma^2$	4.75	1.107	-	4.116	1.106	-
phi	50.49	1.175	-	40.496	1.119	-
tau^2	0.185	1.857	-	0.165	1.859	-
intercept	0.066	0.161	0.683	0.091	0.34	0.79
Catchment 2	-	-	-	-1.33	0.63	0.04
Catchment 3	-	-	-	1.21	0.92	0.19
Catchment 4	-	-	-	0.22	0.52	0.68
Catchment 5	-	-	-	0.40	0.81	0.62
Catchment 6	-	-	-	0.26	0.51	0.61
Catchment 7	-	-	-	0.99	0.75	0.19
Catchment 8	-	-	-	-1.18	0.48	0.01
Catchment 9	-	-	-	0.72	0.59	0.22
Catchment 10	-	-	-	0.55	0.57	0.33

Figure S3.4.7 Predictions from the intercept + river catchment model, A. Genetic score attributed to river catchment B. Total genetic score predictions across the city C. Estimated contribution of the spatial random effect



S3.4.5 Sensitivity analyses with household water source location

Individuals tended to live near their water source locations, with a median distance of 65 meters (IQR 26-112). Therefore, we do not expect using water source location instead of household location to change the predictive ability of the river catchment variable, which varies on a much larger spatial scale (>1km).

Regardless, we conducted a sensitivity analysis to compare results when using a geostatistical model using water source coordinates instead of household location. Although the small-scale spatial correlation changes, river catchment still significantly improves the fit of the model to the spatial-genomic patterns seen, although less significantly (LL -313.39 vs. -304.58, D= 17.623, p = 0.040). Coefficients for river catchments 2 and 8 remain distinct from the other catchments (Table S3.4.2).

Table S3.4.2 Covariance parameters and coefficient estimates from geostatistical model using GPS coordinates of water source instead of household.

Parameter	**Estimate**	**Standard error**	**P-value**
$sigma^2$	4.43	1.098	-
phi	22.46	1.330	-
tau^2	0.048	3.374	-
intercept	0.21	0.32	0.52
Catchment 2	-1.41	0.60	0.02
Catchment 3	0.55	1.00	0.58
Catchment 4	0.20	0.50	0.69
Catchment 5	-0.34	0.76	0.65
Catchment 6	0.32	0.50	0.53
Catchment 7	0.52	0.71	0.47
Catchment 8	-1.32	0.47	0.005
Catchment 9	0.21	0.55	0.71
Catchment 10	0.39	0.57	0.49

1. Diggle PJ, Tawn JA, Moyeed RA. Model-based geostatistics. J R Stat Soc Ser C (Applied Stat **2002**; 47:299–350. Available at: http://doi.wiley.com/10.1111/1467-9876.00113.

Supplementary Material 5.1

Written by JG.

The goal of this study is to estimate typhoid-attributed perforations, since surgical patients presenting with intestinal perforations are not routinely tested for *S.* Typhi. Monthly typhoid fever counts, and surgical perforations were collected between 2008 and 2017 (Figure S5.1.1).

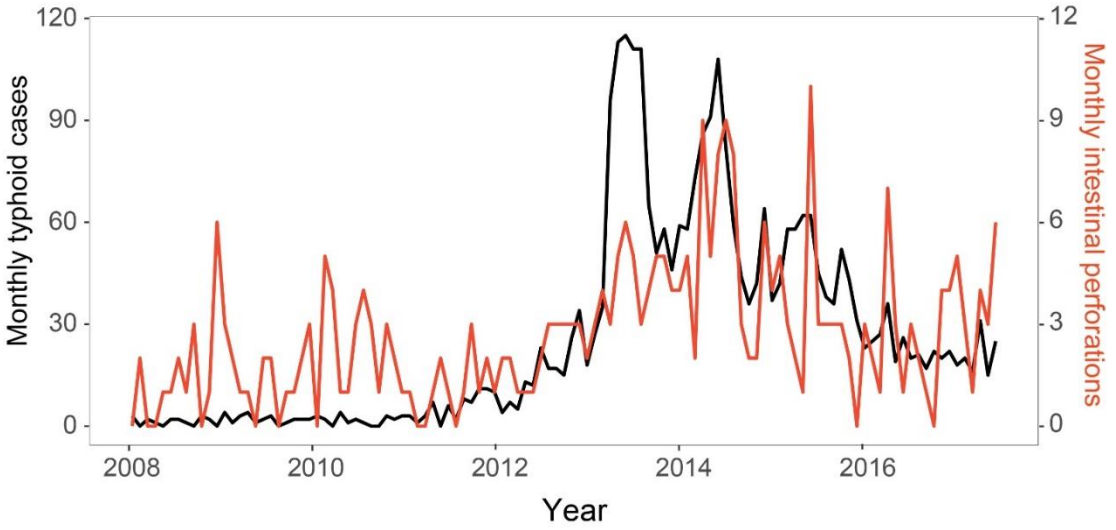


Figure S5.1.1 Monthly typhoid fever counts from QECH (black) and surgical perforations (red).

We first fit a model to the typhoid case counts over the study period. This is implemented using a Poisson log-linear generalized additive model using the mgcv package in R [89]. In addition to the non-linear time-trend that is apparent in Figure S5.1.1, a smoothed periodogram of the typhoid fever case counts shows peaks at 12 month and 6 month frequencies. We therefore specified the model as follows:

$Y_t \sim \text{Poisson}(\mu_t)$

$$\mu_t = \exp\left(\alpha + s(t) + \cos\frac{2\pi t}{12} + \sin\frac{2\pi t}{12} + \cos\frac{4\pi t}{12} + \sin\frac{4\pi t}{12}\right) \quad [S5.1.1]$$

μ_t=typhoid case counts at month t

In equation S5.1.1, t denotes numeric month, beginning January 2008, μ_t is the expected number of typhoid cases in month t, and the trend term $s(t)$ is a penalized regression spline with the default setting for the degree of smoothing, as implemented in the R package mgcv. The fit of the smoothed model to the expected monthly numbers of typhoid fever cases is shown in Figure S5.1.2.

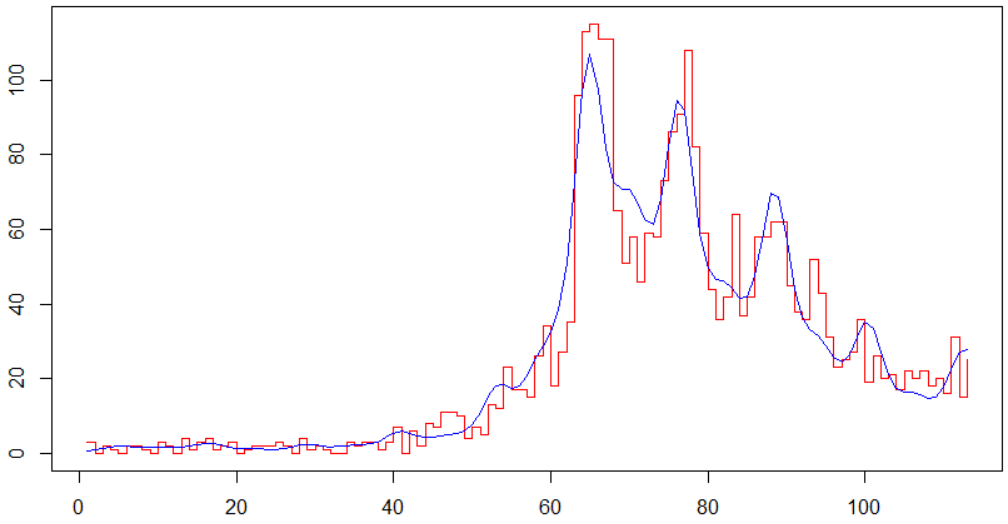


Figure S5.1.2 Fit of the model to estimated monthly typhoid fever case counts.

Next, we estimate excess perforations attributed to the typhoid epidemic. We use the fitted values, d_t, of the expected monthly numbers of typhoid cases [equation S5.1.1 and Figure S5.1.2, blue curve] as a predictor variable in a Poisson generalized linear model for the monthly numbers, Y_t, of intestinal perforations, with an identity link for interpretability of the covariate effects, hence

$$Y_t \sim \text{Poisson}\ (\alpha + \beta d_t) \quad [S5.1.2]$$

Results from the model [equation 8.2] are shown in Table S5.1.1. Monthly typhoid fever case counts are predictive of monthly intestinal perforations

($p<0.001$). The intercept estimate of 1.5 indicates that 1.5 intestinal perforations occur each month, independent of typhoid cases. The model also estimates that for every typhoid case, 0.046 perforations occur.

Table S5.1.1 Estimates from perforation model. Standard error is reported to 5 decimal places for comparison with the model incorporating uncertainty in the predictor.

Variable	**Estimate**	**Standard error**	**p**
Intercept	1.503	0.17386	<0.001
β	0.046	0.00641	<0.001

We now extend the above approach to include uncertainty in the fitted values d_t used in our estimation of perforations attributed to typhoid fever. We extract the covariance matrix of the regression parameter estimates from equation S5.1.1, generate 1000 realizations of the parameters from their multivariate Normal sampling distribution and use these to reconstruct the corresponding expected monthly numbers of typhoid cases (Figure S5.1.3).

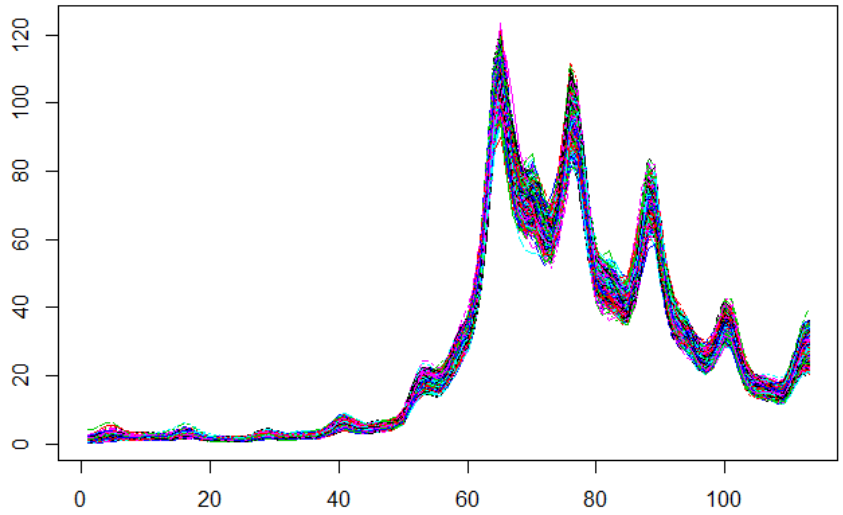


Figure S5.1.3 1000 realizations of the smoothed model from equation S5.1.1 estimating monthly typhoid fever case-counts, drawn from the multivariate Normal sampling distribution of the model parameter estimates.

We then re-estimate the parameters of model [S5.1.2] using each of these smoothed curves as inputs d_t. The resulting estimated values of β for each realization can be represented as $\widehat{\beta}_k$: k= 1, …,*1000*, with associated squared values of the reported standard errors as v_k: k = 1, …,*1000.*

We denote the sample mean of $\widehat{\beta}_k$ by $\overline{\beta}$, the sample variance of $\widehat{\beta}_k$ by s^2_{β}, and the sample mean of v_k by $\overline{v}$, and make use of the following theorem:

Let U and Y be any two random variables:

(a) $\mathrm{E}[\mathrm{Y}] = \mathrm{E}_{\mathrm{U}}\left[\mathrm{E}_{\mathrm{Y}}\left[\mathrm{Y} \mid \mathrm{U}\right]\right]$

(b) $\mathrm{Var}\{\mathrm{Y}\} = \mathrm{Var}_{\mathrm{U}}\left\{\mathrm{E}_{\mathrm{Y}}\left[\mathrm{Y} \mid \mathrm{U}\right]\right\} + \mathrm{E}_{\mathrm{U}}\left[\mathrm{Var}_{\mathrm{Y}}\left\{\mathrm{Y} \mid \mathrm{U}\right\}\right]$

Using (a), our point estimate of β is $\overline{\beta}$. Using (b), our estimate of the variance of this estimate is: $\mathrm{Var}\{\overline{\beta}\} \approx s^2_{\beta} + \bar{v}$

Figure S5.1.4 shows the distribution of the individual estimates $\widehat{\beta}_{\mathrm{k}}$ generated from the 1000 realizations. The distribution is tightly concentrated

around the original point estimate of 0.046 (Table S5.1.1). Consequently, the standard error of the point estimate $\bar{\beta}$ is only slightly larger than that of the original estimate (Table S5.1.2).

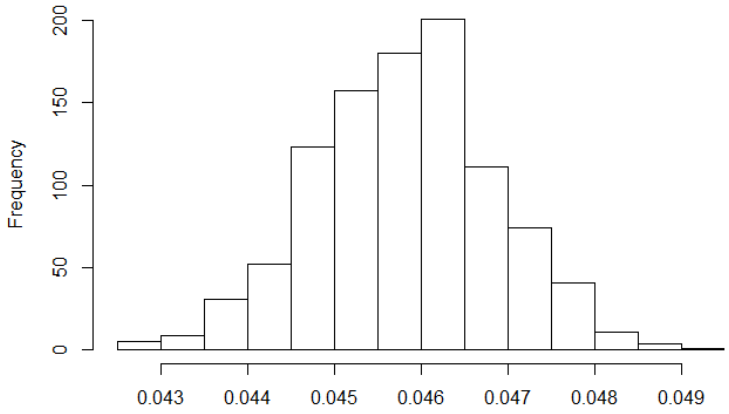


Figure S5.1.4 Histogram of estimates of $\hat{\beta}_k$ incorporating uncertainty of the smoothed predictor.

Table S5.1.2 Estimates from perforation model, incorporating uncertainty of the smoothed predictor.

Variable	**Estimate**	**Standard error**	**P-value**
Intercept	1.505	0.17474	<0.001
β	0.046	0.00649	<0.001

1. Team RC. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://wwwR-project.org/ **2013**;