Author:
Baker, Samuel E

Title:
The consequences of early life disease exposure

digitisation, new data, and empirical analyses

The consequences of early life disease exposure

Digitisation, new data, and empirical analyses

Samuel Edward Baker

A dissertation submitted to the University of Bristol in accordance with the requirements for award of the degree of PhD in the Faculty of Social Sciences & Law, School of Economics.

Submission 03/2022 Word Count: 32,827

Abstract

This thesis explores whether early life exposure to disease has long-term consequences for one's health and economic outcomes, measured in older age. It builds upon the Developmental Origins of Health and Disease hypothesis (DOHaD), which suggests that adverse early life circumstances can contribute to long-lasting, irreversible effects on one's health and well-being in older age. Research investigating developmental origins requires both detailed data on early life circumstances and later-life outcomes, which can be challenging to undertake.

In this interdisciplinary thesis, we expanded the literature with new research, software, and data. Chapter 2 shows ArchiveOCR: a new custom-coded software solution to digitising historical tables and its first use case of digitisation of 40,000 tables of weekly regional disease notifications. Chapter 3 details WeightGIS, and how it was implemented in the construction of a time-invariant 1931-1971 district structure for England and Wales. Chapter 4 contains a large new historical database on 20th Century England and Wales using the methods within Chapter one and two.

The data was then utilised in two research chapters. Chapter 5 shows that UK Biobank participants who experienced increased exposure to scarlet fever had a higher risk of later-life heart disease and declined fluid intelligence. Finally, chapter 6 found UK Biobank participants with high genetic risk to asthma were at less risk of developing asthma in later-life if exposed to scarlet fever or pertussis in early life, but those with moderate risk had no associated decline in risk.

In memory of my ancestors

Acknowledgements

I express my gratitude and apparition to the extensive advice and guidance of my four supervisors of Professor Stephanie von Hinke, Dr Neil Davies, Professor Frank Windmeijer, and Professor Tim Frayling. I would not be in the position I am now without their support, understanding, patience, and encouragement. I thank them not only for making it possible to work towards a PhD but to undertake research that I felt mattered along the way.

I am immensely grateful for the opportunity to undertake my research from the funding I was awarded from the MRC doctoral training program of GW4 BIO-MED. This work was supported by GW4 BIOMED DTP awarded to the Universities of Bath, Bristol, Cardiff and Exeter from the Medical Research Council (MRC)/UKRI grant reference MR/N0137941/1. Their simple process of application, posting topics for individuals to apply for, and supportive demeanour from their staff are likely the only reason I have been able to undertake a PhD. I would also like to offer my thanks to NORFACE DIAL, grant reference 462-16-100, and the international research team I have had the pleasure of being a part of from Erasmus University Rotterdam and University of Zurich for all their support. This thesis includes maps based on historical material which is copyright of the Great Britain Historical GIS Project and the University of Portsmouth. They can be found at www.VisionofBritain.org.uk.

Finally, I wish to thank my family for their consistent love, support, and sacrifices throughout the years, despite failing to achieve results at times along the way. I would also wish to thank my grandparents for their inspirational lives that motivated me forward. I may not be able to achieve a Doctorate of Science or obtain heroism akin to that of serving my country during the Second World War. However, I hope my actions in the future can at least live up to my ancestor’s legacy that they left behind.

Author's declaration

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's Regulations and Code of Practice for Research Degree Programmes and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

SIGNED: DATE: 30/03/2022

Contents

Figures

Tables

Algorithms

Abbreviations

Abbreviations	Definitions
ICD	International Classification of Diseases
OCR	Optical Character Recognition
ASCII	American Standard Code for Information Interchange
DOHaD	Developmental Origins of Health and Disease
GIS	Geospatial Information System
BIO-HGIS	Biobank Historical Geospatial Information System
OOP	Object-Oriented Programming
JSON	JavaScript Object Notation
ISO	International Organization for Standardization
SQL	Structured Query Language
CSV	Comma-Separated Values
GBHDHHC	Great British Historical Database on Health and Health Care
Polio	Poliomyelitis
CNS	Central Nervous System
BMI	Body Mass Index
COVID-19	Coronavirus Disease 2019
S. Pyogenes	Streptococcus pyogenes
MMR	Measles Mumps and Rubella
BCG	Bacille Calmette-Guérin
API	Application Programming Interface
ACE	Adverse Childhood Experiences
MR	Mendelian Randomisation

PheWAS	Phenome-wide Association Study
PANDAS	Paediatric Autoimmune Neuropsychiatric Disorders Associated with Streptococcal Infections
AMI	Acute Myocardial Infarction
IHD	Ischemic Heart Disease
CVD	Cardiovascular Diseases
LPM	Linear Probability Modelling
OLS	Ordinary Least Squares
GWAS	Genome Wide Association Study
SNPs	Single Nucleotide Polymorphisms
IgE	Immunoglobulin E
IgG	Immunoglobulin G

1 Chapter 1: Introduction

Until the 1950s, it was still believed that a foetus was akin to a parasite, therefore protecting it from the actions of the mother[1]. After thalidomide was introduced as a treatment for morning sickness in the 1950s, which subsequently led to an epidemic of birth defects, researchers began to take in utero exposures seriously[1]. However, given obvious ethical concerns of human experimentation, researching the effects of exposures within utero or childhood was challenging. Crucially, one must also wait to observe a change to determine the effect[1], which delays warnings of potential harm for years or even decades.

Researchers have sought to work around these limitations with the use of historical records. David Barker, whose work spawned the foetal origins' hypothesis, used historical data extensively. One of Barkers' first papers found an association between infant mortality rates and later life heart diseases[2]. Barker utilised 212 local authorities' data on infant mortality and extracts from the international classification of diseases (ICD) 8 records from 1968 to 1978[2]. Barker argued that the infant mortality rate could be used as a proxy for the unobserved early life deprivation of nutrition, which subsequently led to an increased probability of later life ischemic heart disease[2].

Barkers' work specifically received some criticism, such as from Mervyn Susser, for being mostly correlational yet drawing conclusions without sufficient rigorous testing[3]. However, the principles of using historical data itself were supported by sound evidence from a natural shock of the 'Dutch Hunger Winter'. Dutch citizens had their rations of food-restricted to only 400-800 calories a day for nearly six months from 1944[4]. These restrictions led to women in multiple stages of pregnancy being affected[4].

Mothers who experienced the famine in the third trimester had children born with starkly decreased birth weight[5], but the children were relatively healthy later in life[1]. Conversely, those exposed in the first half of the pregnancy reported standard birth weights but had increased incidence of later life health issues such as heart disease[1]. As such, a child born of healthy birth weight but with negative exposures in utero could still experience consequences in adulthood[4].

A central strength of studies investigating the Dutch Hunger Winter came from having a famine that was extensively and reliably documented[6]. However, just as important were the records of military conscripts who provided their place and date of birth cohorts[6], and could be tested for health and well-being later in life. Without these well-kept historical statistics, studies investigating the effects of starvation in differing trimesters would have been difficult, if not impossible, because of the ethical concerns of experimentation on humans[4].

There is no shortage of historical statistics, as the idea of a census dates as far back as Babylon[7]. The UK has a wealth of historical records spanning back centuries. The Labour Gazette, the Annual, Quarterly, or even weekly Registrar-General's reports, and even census records all offer crucial insights of the UK across time. The UK also has detailed cohort studies for external data to be utilised with, especially during the 20th, such as the 1946 National survey on health and development[8], the 1958 National Child Development study[9], the 1970 British Birth cohort[10], and the UK Biobank[11,12]. Therefore, the UK is in theory an exceptional location to undertake research investigating for long-term consequences of early life exposures or circumstances.

Despite a wealth of historical records existing within the UK and efforts by both government, researchers, and public bodies to digitise these records, limited amounts are available. The National Archives estimated that only 10% of its documentation is available in any digital form[13] and will include simple scans and images which, whilst crucial for historical preservation, are not conducive to research.

Converting the information into data tables faces multiple issues. Manual digitisation can be extremely expensive due to labour costs[14]. Said costs often act as barriers for larger projects, limiting digitised material to easy but often aggregated statistics. Conversely, many projects ignore these risks but then fail to complete[15].

An alternative is to use technology that utilises optical character recognition (OCR). OCR takes pixel data and converts it to ASCII data a computer can use[16]. Many solutions, such as Google's Tesseract, are both open source and free to use. However, larger projects often require a level of computer science to get the most out of these solutions, which can be a significant barrier to entry.

Even when digitisation is successful, such as the extensive efforts of Vision of Britain by the University of Southampton, digitising the exact information does not make it usable. The modifiable area unit problem[17], or that locations can vary both in name and boundaries, frustrates comparing regions over time. The UK's regional levels change significantly over the 20th century, making most regional data suffer from the modifiable area unit problem[18]. Construction of a time-invariant location would therefore require extensive historical knowledge of each of the UK's regions to link them appropriately.

These barriers are not insurmountable, with many papers from David Barker[19,20,21], among many others[1], utilising parts of these records successfully to undergo their research. However, if there was a greater enthuses on data preservation for purpose, in addition to preservation, there would be a greater potential for a wider group of disciplines and professions to utilise these statistics for public good.

Within this interdisciplinary thesis, I expanded the literature investigating the Developmental Origins of Health and Disease (DOHaD) hypothesis with new research, software, and data. The new data constructed as part of this thesis represents some of the most detailed records of 20th century infectious diseases that are currently available to research. However, a diverse set of data sources have been digitised beyond this. This has allowed for research within this thesis to differentiate itself by being able to both expand with greater detail beyond existing research questions and explore completely new ones.

1.1 Contributions

As many of the current off-the-shelf options for digitisation are complicated to use, this thesis further seeks to reduce the barrier of entry of digitisation so that others can also undertake similar projects. Chapter 2 provides a significant contribution through the construction of a new piece of software called ArchiveOCR, which was specifically tailored to assist large scale digitisation of historical numerical tables. ArchiveOCR achieved a predictive accuracy of 99.5% from 3,058,921 characters from around 40,000 tables of weekly regional disease notifications for England and Wales from 1941 to 1973. The accuracy itself is not a contribution, OCR software frequently advertises an accuracy of 99%[22]. However, when applied to historical data, this advertised rate often declines. Digitisation of newspapers from 1803 to 1954 led to accuracy that varied from as a high as 98.02% to as low as 71%[22]. Here, ArchiveOCR managed to worked fairly consistently, even with older material.

The data constructed from ArchiveOCR had complex geospatial qualities which made standardisation and quality control complex. The data was reported at a district level, regional zones within the UK prior to 1974[23]. However, district definition is not static. Districts change names and boundaries within the 20th century, know as the modifiable area unit problem[24], which can limited comparability of regions across time. Chapter 3 details the software solution to these problems of WeightGIS, which can construct time-invariant locations using areal or sub-unit population weighting. The software is itself a significant contribution, simplifying standardisation of complex geo-spatial geographies and linking data to said geographies into just a few steps. However, WeightGIS has also been utilised to create time invariant districts between 1931-1974 within England and Wales. This allows any data within districts between 1931-1974 to be standardised to the 1951 census, allowing for compatibility within locations and removes the modifiable area unit problem.

Both software solutions have been used in tandem to construct a new database from historical administrative data and linked it to the UK Biobank. Chapter 4 details the construction of the Biobank Historical Geospatial Information System (BIO-HGIS), and how it is designed to assist research new research questions whilst utilising one of the largest cohort studies in the UK. The disease notifications represents one of the largest new contributions of new data, but this thesis has also processed a wide range of additional resources. Administrative data on population, unemployment of counts, and geo-locating each individual air raid dropped during the Blitz represent just a few of the additional data contributions construct during this thesis.

Whilst the digitisation and quality control of the historical data is itself a contribution, the ability to link to the UK Biobank strengths its importance. Briefly, the UK Biobank is a prospective cohort study of UK adults aged 40-69 at time of recruitment[12]. The UK Biobank contains extensive later life information on the health and well-being of its participants, most of whom have been genotyped. However, the UK Biobank itself has collected relatively little information on individuals' early life circumstances. BIO-HGIS's ability to link administrative data to the UK Biobank participants allows construction of early life environments and exposures for its participants. Therefore, BIO-HGISs other main contribution is that it allows one main weakness of the UK Biobank, of a lack detail on early life, to be partially mitigated.

The data within BIO-HGIS was used to investigate two research papers. Each paper focused on the exploring the potential for long-term consequences of early life exposure to childhood diseases, linking itself to the DOHaD research literature. Due to the use of brand-new data, each paper offers answers to question that have previously been difficult or impossible to investigate. Within Chapter 5, we explored if scarlet fever could lead to later life cardiovascular and cognitive health, in addition to educational attainment. Whilst we found limited evidence for long-term effects on cognitive health or educational attainment, we found a positive association between increased early life exposures to scarlet fever and later life ischemic heart disease. The crucial contribution of this paper is that it represents on the most well powered studies to date to show an association between streptococcus infections and increased risk of later life cardiovascular health.

Within Chapter 6, we examined whether exposure to disease in early life affects the development of asthma. Presently the literature is filled with hypothesis on why asthma rates rose so rapidly during the later half of the 20th century within the UK. Many point to change in the environment, but evidence is often inconclusive and ignores the highly genetic component of asthma. In this paper, we tested a hypothesis based on lab experimentation that found that other components of the immune system[25^] may attenuate the excessive immune response, produced by those with high genetic risk of asthma[26,27]. Many of these components remain heightened after infection, so we hypothesised that during an era of heightened childhood infections, that individuals would have reduced risk of asthma. To differentiate from the literature, we also constructed and controlled for genetic risk to asthma given its high heritability[28].

We found that increased exposure to childhood diseases of scarlet fever and pertussis were association with reduced risk of developing later life asthma, but conditional on genetic risk. That is, that those exposed to scarlet fever or pertussis had a reduced probability of developing asthma, compared to those with low exposures of these diseases and the same PGS. However, those exposed to higher rates of scarlet fever or pertussis still have a higher risk than those exposed to said diseases, but with a lower PGS. This studies association suggests that declining disease incidence within the 20th century may in part be relate to the rise in asthma, but also stresses the importance of considering gene environmental interactions within research.

1.2 Thesis Outline

This thesis comprises eight chapters, including the introduction and conclusion, which are organised as follows:

Chapter 1: Introduction: This chapter contains an introduction and overall contribution from the contents within the thesis and outlines were said content is placed within it.

Chapter 2: ArchiveOCR: Scalable and accessible digitisation for tables: Chapter 2 covers the construction of ArchiveOCR, and how, in principle, it works to digitise tabular historical records. Chapter 2 also covers how ArchiveOCR was used for its first project of digitising 40,000 tables of historical records, detailing its accuracy.

Chapter 3: WeightGIS: To Automate Standardisation of Regions: Chapter 3 covers why WeightGIS was constructed, discussing the issues surrounding standardising geospatial data. How WeightGIS works and what is required from the end user at each step is detailed. Chapter 3 concludes with its use case to create a time-invariant set of weights for England and Wales between 1931 and 1974.

Chapter 4: Data Resource Profile: Biobank Historical Geospatial Information System (BIO-HGIS): This chapter summarises the data constructed throughout the PhD and motivates why the following two software chapters were required. Chapter 4 also covers some background knowledge of the UK geospatial make up during 1931 to 1974 to help to understand in future chapters. This Chapter further motivates why this data is of such importance by covering each data source, why it is important, and how this data has been used within the thesis or will be used within future work.

Chapter 5: Early life exposure to scarlet fever is associated with ischemic heart disease later in life.: *Co-authored with Neil Davies, Frank Windmeijer, and Stephanie von Hinke*. Chapter 5 investigates if there are further long-term consequences of exposure to scarlet fever in childhood and later life cardiovascular diseases, fluid intelligence, and educational attainment.

Chapter 6: Can gene-environment interactions explain the rise in asthma incidence in the 20th century?: *Co-authored with Neil Davies, Frank Windmeijer, and Stephanie von Hinke*. Chapter 6 investigates a plausible biological mechanism to explore if a gene-environment interaction between the declining disease incidence and genetic predisposition to asthma explains part of the increase in asthma prevalence in the 20th century.

Chapter 7: Conclusion: This section will conclude the thesis and highlight future work.

2 Chapter 2: ArchiveOCR: Scalable and accessible digitisation for tables

2.1 Introduction

Digitisation is transferring media from its original physical form to a digital one using computational hardware and software[15]. Modern type scripts are easier to digitise as they are regular, so can usually utilise large generalisable training sets of multiple fonts and achieve low Character Error Rates (CERs)[29]. Older text has much larger variation, which often requires source specific training[29]. Therefore, whilst the problem of digitally transcribing modern text is practically a solved problem, most of the cultural corpus of humanity remains a challenge to digitise.

There are many state-of-the-art, commercial, or commonly used solutions to Optical character recognition (OCR), many of which have been used to digitised historical records to varying degrees of accuracy. A common commercial option is ABBYY FineReader. One of the largest known uses of ABBYY FineReader 11 was the digitisation of a repository of historical newspapers published in Finland between 1771-1929[30]. However, the resulting CER ranged between 8 to 13%[30], which is highly undesirable.

Over the years, many open source solutions now offer similar or better performance, such as calamari[31], OCRopyor its fork Kraken[32], and Tesseract[33]. A Calamari model achieved a 70% improvement over ABBYY FineReader, and a CERs of close to 1% from digitising various German literature spanning 1781 to 1892[29]. Tesseract 3.04 reported a 9.16% improvement over ABBYY FineReader 11, with further improvements within Tesseract 4.0[34]. The F1 score (average precision and recall) of Tesseract could also be substantially improved by stronger impact

Whilst a wealth of open-source options for OCR exists, 4 of the largest big 5 technology companies have also invested heavily in this field. Microsoft Azure, Amazon Web Services (AWS) Textract, Google Vision, and Facebooks Dectron2[35] are all examples of commercial options in this field. Some of these have already seen publications using them for historical documentation directly. For example, AWS Textract and Comprehend were used in digitising 22,436 card indexes on bank loans from 1932 to 1957 with a spot check on a random 100 cards resulting in a 3.65% error rate[36].

Facebooks Detectron2's could be used itself, but because its licence is open and more permissible, it has also been utilised as part of the OCR pipeline in other software, such as LayoutParser[37]. This full pipeline allows for complex page segmentation, and the use of pre-trained models to help extract a wide variation of attributes, such as images, tables, headings and more[37]. Crucially, this software still depends on training models, but the ability to create custom models, and the aim to facilitate and openly encourage sharing of them, makes it an appealing option. Sadly, this option did not exist in 2019 when this project was undertaken but will be a core point of comparison at the point of release.

The scale of material that requires digitisation has also led to some considerations of crowdsourcing some of the work. A project that achieved significant success using crowdsourcing was the Bentham papers Transcription Initiative, which resulted in 2.6 million words being transcribed over 4 years[38]. Critically, this implementation is the polar opposite of OCR, with 100% of the work being transcribed directly. Whilst it may seem potentially inefficient, it keeps digitisation open to anyone. There have, however, been community efforts on a more technique front, such as Hugging Face[39], which aims to bring machine learning enthusiasts together in the attempt to further generalising training methods.

Therefore, the proposed problem at present appears to be a lack of a middle ground. Many highly technical solutions exist, but they potentially require more knowledge than a potential 'citizen scientist' may have. Given the scale of documentation requiring digitisation, this is problematic. Whilst many of these advanced solutions are impressive, if they require source specific training, then the generalisability benefit is lost. However, ignoring all software benefits and just undertaking manual transcription, whilst it makes the project open, is a slow option that cannot scale. As of 2020, only ***8%*** of the National archives within the UK are available digitally[13], with considerably less, therefore, expected to have been digitised beyond pure images.

2.1.1 A middle ground solution

My proposed solution, ArchiveOCR, is a currently Python-based software solution focused on the digitisation of tables. It can scale from notebooks to super-computers, whilst also reducing barriers to entry through graphical user interfacing front ends. Whilst not a derivative of pre-existing digitisation software packages, it does extensively use the pre-existing open-computer vision library [40] and numpy[41]. However, ArchiveOCR specifically uses a custom written wrapper for CV2 of imageObjects, that makes CV2 work in an image based Object-Oriented Programming (OOP) approach; publicly available on GitHub.

ArchiveOCR is positioned as a middle ground solution due to how it works. ArchiveOCR has zero learning functionality, and instead is a purely guided identification via the user, with said parameters designed to mirror how a human interacts with source material. Instead of setting more transitional machine learning hyper-parameters, which guide models but are not part of the end result, ArchiveOCR uses hyper-parameters explicitly. These parameters are explicit in that they directly affect what happens on the page rather than merely guiding a model. For example, if the user states there are two columns of data on the page, ArchiveOCR will only attempt to find two columns of data. ArchiveOCR will never attempt to find anything it has not been told to, even if it exists. If ArchiveOCR cannot match a page to the hyper-parameters, it will simply declare the page does not conform and stop.

Whilst ArchiveOCR hyper-parameters start set at a project level, the hyper-parameters can be adjusted on a per page or even element, like a table, basis. This means that outliers can be handled separately, or within smaller groups, with their own set of hyper-parameters. They can also be copied across to similar projects and adjusted if the sources are similar. When used on a set of tables that is reported as a series, with the same structure but with new data, most hyper-parameters should only be required to be set once. If you have tens of thousands of similar pages, it then will undertake the same operation consistently, assuming tolerances set by the user have been set appropriately, without any further guidance.

The following sections detail how ArchiveOCR works through its main method of table identification, and any relevant hyper-parameters that are required for that process. Each process of ArchiveOCR has its hyper-parameters explained before an example model explanation, and, where appropriate, a rough pseudo code of said processes algorithm. It concludes with ArchiveOCR's first major use, the digitisation of approximately 40,000 tables of weekly disease notifications from the Registrar-General's Weekly Return for England and Wales (1941-73)[42].

2.2 Isolating tables from lines

Identification of tables requires knowing the horizontal and vertical lines of the table's bounds. Frequently, table detection uses standard algorithms such as Hough line transformations[43]. Here, for identification of a straight line, each pixel found via edge detection is incrementally compared and voted on if it meets the slope-intercept parametric equation of a line[43]. Whilst robust to noise, it can be complex to parameterize when the position and rotation of the image can differ across the sample. Here, we use a simple custom-made algorithm, that operates on the structure of the underlying pixel matrix instead.

2.2.1 Definition of an image

An image is a matrix (M) of pixels of height (y) X width (x) of rows and columns. As with most programming languages, python is base zero, and the cv2 library is top down, so the top left element of the matrix is p0,0 and exists up to lowermost right pixel of py,x as shown in Equation 2-1. The value of a given py,x element depends on the colour dimension of the image. For example, a binary image constructed of single Bits only allows each element to be zero or one, which leads to a monochrome image of black or white[44]. In binary or other forms of monochromatic 'grey scale' images, py,x is equal to an integer. However, most modern images use 'True colour', which by default is a vector of three 8-Bit integers, allowing 256 values ranging from 0 to 255, for red, green, and blue respectively[45]. In this case, the value of py,x is a vector of length three for the red, green and blue values. Some images may also contain a fourth 8-Bit integer, so that pixels can store additional information, such as transparency.

Equation 2-1: A matrix representation of Figure 2-1

$$\boldsymbol{M}^{y\times x} = \begin{pmatrix} \boldsymbol{p}_{00} & \cdots & \boldsymbol{p}_{0x} \\ \cdots & \cdots & \cdots \\ \boldsymbol{p}_{y0} & \cdots & \boldsymbol{p}_{yx} \end{pmatrix}$$

2.2.2 Hyper-parameters of line isolation

A line that is not perfectly horizontal or vertical will be composed of rows or columns of smaller perfectly horizontal or vertical lines, as shown in Figure 2-1. For the custom line isolation, all images are inverted, so black text becomes white, and are converted into binary images so that only white and black integers of zeros and ones represent the pixels. The inversion is used as by default, the CV2 library identifies white elements from a black background.

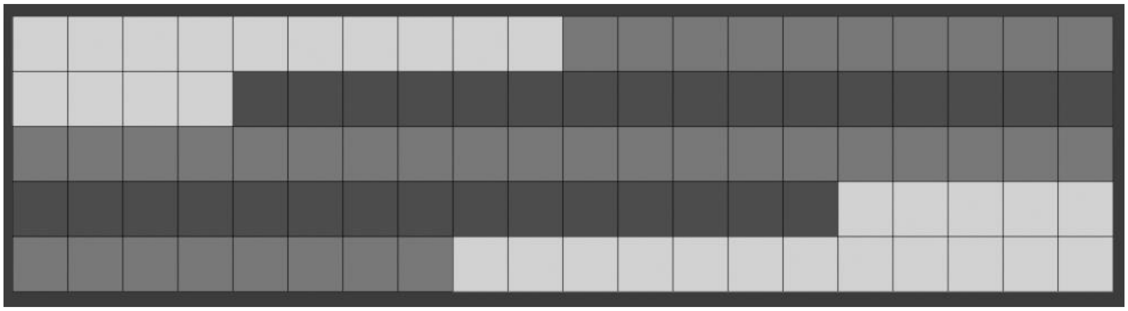


Figure 2-1: An example of a pixel representation of a horizontal line with a positive gradient. Here, the black line has been broken into individual smaller lines with grey scale shading for clarity.

The algorithm exploits one of two key rules to simplify line detection. For a horizontal line, there should be longer sequences of sequential white pixels within the rows of the matrix. Conversely, if the line is vertical, the longer sequential sequences of white pixels should be within the columns. ArchiveOCR will run the line identification twice, horizontally and vertically, but to identify what is a line and not another element requires a set of hyper-parameters. Users specify the minimum length a line segment can take for a given horizontal or vertical line. This can be visually determined by zooming into the image, and usually be specified by investigating the font size, or sizes of other elements on the page.

For example, suppose we specify an image of 5000 X 8000 pixels that is purely constructed of lines and text without noise. If the text font size is 120 pixels high, then setting the vertical line minimum length threshold above the font size guarantees only the lines will be identified. Clearly, noise, page curvature, and other defects can make line identification challenging. Curvature, in particular, may result in parts of a line being smaller than the text, so have to be removed. However, as long as enough of the line segments can be captured by this algorithm, then they can be grouped and reconstructed later.

Grouping represents the second line identification hyper-parameter. If the lines are badly damaged, only a small part of the line might be available. This could mean that, for example, a vertical line 10 pixels wide is only able to identify parts of the line in the first and last two columns. The user specifies how wide or tall, for vertical and horizontal lines respectively, a given line identification should be. Pixels found within that range will be grouped together, rather than assuming them to be separate parts of a different line.

2.2.3 Custom line isolation

Algorithm 2-1 shows an example process for extracting a vertical line from within an image. However, the only difference between the vertical and horizontal isolation methods is that vertical isolation searches for sequential sequences within columns, whereas the horizontal isolation searchers within rows.

The algorithm takes two input parameters. The first is the input image, which is constructed as a matrix (M) akin to equation 1, but with each pixel element (p) equal to either zero or one. The second parameter is the minimum length a line is allowed to take (L), which helps remove smaller elements that may not be a line. L is a hyper parameter set by the user. For vertical line isolation, each column is isolated as a vector (V).

Algorithm 2-1: Rough line detection (Example for vertical line)

Input: A Matrix (M) of size width (w) X height (h) where each $m^{w,h}$ is a 0 or a 1, and a minimum length target parameter (L).
Output: Matrix M

```
for i = 0 to w do
    create a vector (V) isolated as the $i^{th}$ column from M
    Create a vector of sub vectors from the sequential white pixel elements
    for SV in V do
        if the length of SV >= L
            Save the pixel coordinates from SV in terms of y, x
        end if
    end for
end for
Set all elements of M to 0, then set all saved y, x coordinates to be equal to 1
return M
```

The algorithm then groups sequential white pixel elements together into sub vectors (SV) which are separated via the black pixels. If a given SV's length is greater than the minimum length L, these pixel coordinates are stored. This process loops until all columns have been processed, then a matrix of dimension of M is returned, but all pixels other than the stored elements set to black.

For example, suppose we have a binary image matrix of 10 X 2 and the minimum length L of four. The example image contains only one pixel vertical line of length five in column one, but also some noise in the same column. The algorithm would isolate column zero as a vector of $p_{0,0}$ to $p_{9,0}$ with each p element equal to zero. The algorithm searches for white pixels, which would be equal to one, and finds none. Therefore, no p elements from column zero are saved.

The algorithm then isolates the first column from $p_{0,1}$ to $p_{9,1}$ and finds the first five elements are white, then three are black, then two are white. Which will result in two new vectors of lengths five and two. However, as the minimum length was set to four, only the first of the two vectors will be stored, which means when the algorithm finishes, only these five pixels will remain in white.

2.2.4 Table isolation through line identification

For each table, lines are constructed based on the rough estimates created in Algorithm 2-1. An example of this output on a table with some curvature damage is shown in Figure 2-2. For vertical lines, shown in Figure 2-2, the line grouping searches each row within the image for white pixels. For the list of pixels found, they are then grouped based on their x position with a spacing parameter, said spacing parameter being a hyper-parameter provided by the end user. The row then counts how many groups exist within this row, and proceeds to the next.

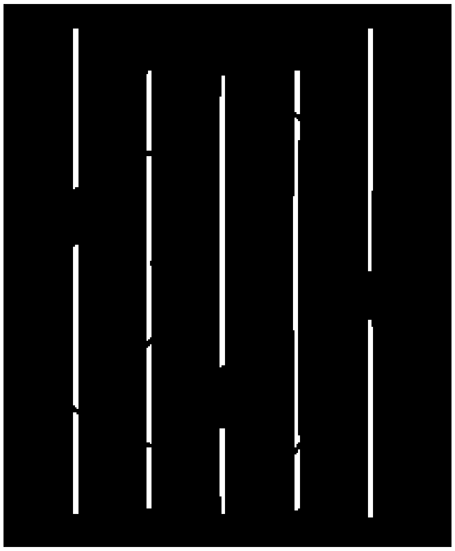


Figure 2-2: An image showing the output of Algorithm 1 on a damaged page

Once all rows have been evaluated, it looks for the most common row count of pixel groups and keeps those rows for line identification. Each group can then be transposed with a line of best fit plotted through each group. In more complex setups, it is possible to plot lines of best fit *between* line groups, rather than through the whole set of line groups. This allows for curvature that would otherwise, with linear approximation across line grounds, result in table elements being cut off. Each table is then saved to disk for the next process. If the number of columns meets the user target, it is given a success tag, which can assist larger scale operations from crashing by conditioning the follow-up process on previous success.

2.3 Isolating table contents

Tables are traditionally made up of columns and rows. Isolation of rows can be undertaken by the same methodology as finding tables, just by looking for continuous white space between words or characters. However, the input image is likely to not be perfectly rectangular, and some level of rotation is probable; this may interfere with row identification. To avoid this, ArchiveOCR applies a 2D perspective transformation by extracting the first and last position of each bounding line using the perspective transformation within CV2.

2.3.1 Hyper-parameters of Table isolation

ArchiveOCR is built primarily for series table extraction, so adds considerable additional hyper-parameters to try to ensure tables conform to a set standard. Therefore, by default, ArchiveOCR expects each table that is isolated to have a hyper-parameter of column and row counts, specified by the end user. These can be relaxed, which for less consistent tables is required, but ArchiveOCR cannot explicitly check the table is correct without a known specification.

Users must provide a pixel font size hyper-parameter, which can be calculated by measuring (in pixels) the width and height of characters in the source material. The height and width of the font are both equally important. The height is used to help split rows, within a given column, that might not be possible through traditional white / black space delimiting. If we know the font has a height of around 30 pixels, but we find a row that is 92 pixels in height, then we know something is wrong with this row. Similarly, when splitting characters, if we know characters are around 40 pixels wide, but find a character that is 82 pixels wide, then the system can identify that something is incorrect. Finally, the user provides a tolerance for these widths and heights, which can allow characters slightly too large to not be split. If an element requires splitting, then the number of new splits is calculated as the floor of the element dimension divided by the font size dimension.

2.3.2 Rough row isolation

Rows are roughly isolated through white space delimiting, with a tolerance of black pixels to be overwritten set by the user shown in Figure 2-3. Whilst perspective transformations will reduce errors in identification, issues may still persist from obstructions, damage, erosion, and poor lighting. Rows that have been damaged or defaced may be large, as they were not delimited. By calculating, for each column of data, each rows height, an average row height can be constructed to assist delimiting damaged rows.

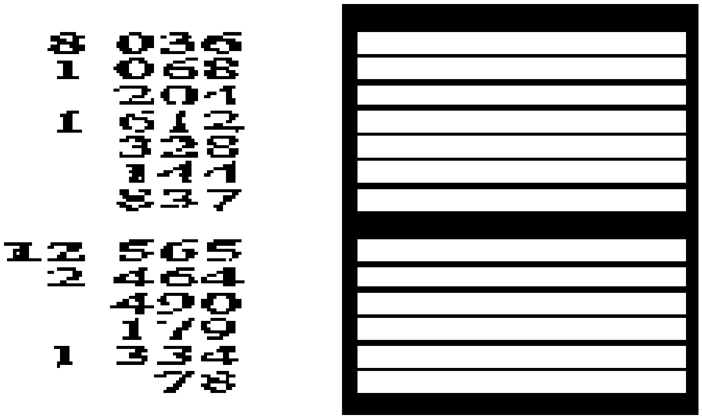


Figure 2-3: Rough row isolation via white space delimiting. Left-hand side shows an example column, with the right-hand side being the mask constructed from white space delimiting.

For the rows that are larger than the average, they can be split by ⌊row height / average row height⌋ row height new rows. By flooring the result, only rows that are significantly larger are changed, and small variations remain unaltered. However, rows may still yet not be in the correct order, due to missing rows for example, so Algorithm 2-2 was derived to sort rows into their positions in the table, as described next.

2.3.3 Row Sorting

The rows that where roughly isolated are ordered vertically within columns but may be incorrect. For example, if the first row is missing in any of the columns, the second row will appear as the first. A simple solution to solve this would be to construct each rows centroid, by isolating the first and last pixels vertically from the row and averaging them. For example, if the row existed between (4, 4) and (24, 36), we know the row is 20 pixels tall and 32 pixels wide, with its (y, x) centroid being at (14, 20).

However, even after perspective transformation, page warping may mean that items closer to a page's seam start lower on a page and finish sooner. Furthermore, if a row's height varies because the information spans multiple lines, then comparing the distance between the centroids of the rows will lead to row elements being incorrectly rejected. This issue is avoided by using relative distance, constructed as the centroid of the distance between the row elements. Using the right-hand side of Figure 2-3, this represents constructing the centroid from the black horizontal lines between the rough isolated rows. Whilst the relative distance between rows will still differ based on page warping, it will differ significantly less than the rows absolute positions.

Algorithm 2-2 shows the process of sorting rows, which continues until all rows have been accepted or until more than half the columns no longer have entries; with the remaining entries assumed to be noise. Algorithm 2-2 takes in the rough rows that where isolated for each column (CR), a bounding parameter (B), and the average height of the rows (AHR). The bounding parameter determines the maximum distance in terms of y a row can be from a row in another column and still be considered part of the current row.

Algorithm 2-2: Row sorting

Input: A List of Lists (CR), where each sub list represents a column's (C) rough row (R) extraction using white space delimiting, a bound (B), and average height (ARH).
Output: A List of Columns, which each column containing the rows in that column

```
i = 0
While length of empty C is less than number of C / 2
    Create the present row (PR) by extracting $R^i$ for each C in CR
    for j = 0 to length of PR do
        Compare Relative distance (RD) for $R^{i,j}$ in PR to all other R in PR
    end for
    if Each $R^i$'s RD is within B distance of at least one other $R^i$'s RD
        Append PR to out list, remove each $R^i$ from CR
    Elif All $R^i$'s are not within B distance to another $R^i$
        All elements are random, remove each $R^i$ from CR
    Elif Some $R^i$'s are not within B distance of another $R^i$
        for f for index positions of failed $R^i$ do
            for subsequent row (RS) in $C^f$ do
                if RS RD is within B distance of at least one other non-failed $R^i$
                    Replace $R^i$ at index position f within RS, remove rows in $C^f$ before
                    RS
                end if
            end for
            if No RS found for $R^i$ at index position f
                Row is missing, add a row of dimensions ARH to $C^f$
            end if
        end for
        Append fixed PR to out list, remove each $R^i$ from CR
    end if
    i += 1
```

Algorithm 2-2 iterates through the rough column-row data set and isolates the present row (PR) from each row (R) at the current index position i (R^i) in each column. If the relative distance (RD) between R^i and R^{i+1} is within B distance of at least one other R^i in PR, then we assume it is in the correct position. If all rows in PR have at least one link, the row is assumed to be correct and accepted. Conversely, if no rows are within B distance of at least one other row in PR, then it is impossible to know which row is correct, so PR is deleted.

However, if only some rows fail it is possible to reconstruct PR which is undertaken as a two-step process. First, Algorithm 2 assumes that the failed R^i has failed due to noise existing before the actual row. This means that there would be more rows present in this column than entries to fill. Each failed R^i's column is iterated through searching for an R that meets the distance condition to any non-failed R within PR. If found, R's before this matching R^{i+n} in this column are deleted and R^{i+n} replaces R^i in PR. If no match is found, then the second step is assumed, that the row is missing. Here a row is prepended into the current column of size with the height derived from the average height of the rows or AHR.

2.3.4 Character isolating

Once we have isolated a row, we can use the same method used to isolate the rough rows to isolate the characters within a row. This time, instead of isolating the horizontal distance between elements, to isolate rows, the vertical white space between elements in the row is isolated and used to separate the individual characters. As before, in the case of characters that cannot be split purely on white space, users provide a pixel size of the font to be isolated. Then, characters that are not isolated through white space delimiting, determined by the width of the isolated characters, are then split based on $\lfloor \text{Character Width / pixel font size} \rfloor$.

2.4 Optical Character Recognition (OCR) through pattern matching

ArchiveOCR does not use a deep learning mechanism, so the user needs to provide a list of samples to compare against for pattern matching. Pattern matching compares two images' pixels, normally a binary image, where a match is simply where the ones or zeros are the same for a given pixel between each image. A success is defined by an acceptance threshold, representing a percentage number of pixels that must match. Pattern matching can be highly accurate, and in modern text and when using characters of the same font, can lead to accuracy close to 100%[16].

2.4.1 Hyper-parameters of OCR

At its core, the OCR within ArchiveOCR is simply an implementation of a pattern match algorithm, as described above, that is within the cv2 python library. Whilst ArchiveOCR has a 'training' mode, it does not involve any machine learning, and is merely a way of making the software export characters rather than identify then. Therefore, a 'hyper-parameter' of sorts, is that the user must manually sort the letters that are exported into folder bins. So, all the capital A's go into a folder with the same character, and so on.

The user then provides two thresholds, a sufficiency threshold, and an identification threshold. The sufficiency threshold governs at which point a training character can be replicated by an existing character. This threshold is a float between 0 and 1, representing a percentage. If the user sets the sufficiency threshold to be high, say 0.9, then a training character must be 90% similar to another sample in order to be removed. Reducing the sufficiency threshold leads to more characters matching within the training characters, resulting in fewer characters being used in the predictive stage.

The predictive threshold is the second hyper-parameter required by the end user, which governs how many pixels a training character must match a detected character for it to be classified as that character. If the predictive threshold is high, say 0.9, then little deviation is permitted for a match to a given character. Whilst this prevents type

1 error of predicting a character is actually a different character, setting the threshold too high on varied texts will increase type 2 error, where a character cannot predict itself. In high-quality text, said risk is less prevalent, as character variation is low.

The user needs to manage these two thresholds carefully. Whilst setting the sufficiency threshold lower may help reduce needless characters being kept, it may have consequences for the accuracy of the actual digitisation. Characters that are removed may be predictive of unseen characters not in the training sample in ways that characters that have been kept cannot. Therefore, initially, some source specific experimentation on these thresholds will be required to set them so that type 1 and 2 errors, as well as additional training runs, are minimised.

Finally, whilst optional, the user can specify a hyper-parameter of the content of the column data. For example, if the data is purely numerical, then there is little point trying to match to alphabetical characters. This reduces error and speeds the process up, so is a recommended hyper-parameter where the structure of the table is known.

2.4.2 Constructing the training set

After setting ArchiveOCR to a training model on an example set of pages, it isolated the characters shown in Table 2-1. The user can then try different sufficient thresholds to determine if they have done enough training for the characters to predict unseen characters. With this example set, we can see, using character 0, that whilst we start with 170 characters, that only 36 of them are unique when requiring samples to be at least 85% unique. This further reduces to 12, when requiring characters to only be 75% unique. This means that the predictive rate for these characters on the training sample leads to a 78.67% and 92.94% predictive rate, respectively.

How high a predictive rate you should aim for depends on the text. The higher the variation the text holds, the fewer characters are likely to match. For example, using the character 8, when requiring characters to be 85% similar or more to be removed, the model still keeps 103 of the 183 characters we observe. Whilst reducing the sufficiency threshold increases the predictive rate, it is still only 76.5%. This suggests that there is extensive variation within 8s in our sample, and we may yet need more observations of 8's to predict it. ArchiveOCR will export a table like Table 2-1, but for a single predictive rate, each time the user asks to re-compile the samples. By comparing the output of these tables at different sufficiency thresholds, and looking at the predictive rate, it is designed to suggest if more training is required.

Table 2-1: Based on a training sample of characters, and two acceptance rates of 85% and 75%, how many characters were kept as 'sufficient' and what is the predictive rate of those characters.

Character	Total Characters provided	85% Kept Characters	85% Predictive Rate	75% Kept Characters	75% Predictive Rate
0	170	36	78.67	12	92.94
1	403	38	90.57	15	96.28
2	256	40	84.38	10	96.09
3	246	46	81.38	15	93.90
4	181	16	91.16	5	97.24
5	201	46	77.11	17	91.54
6	183	48	73.77	13	92.90
7	163	12	92.64	7	95.71
8	183	103	43.72	43	76.50
9	181	45	75.14	17	90.61

2.5 Example use case of digitisation of weekly notifiable diseases

ArchiveOCR was built originally to assist digitisation of weekly reports of notifiable diseases from the Registrar-General's Weekly Return (1941-73)[34]. Each page has two tables side by side, shown in Figure 2-4, documenting the number of disease notifications that were present in each district; regional zones in the UK prior to 1974[23].

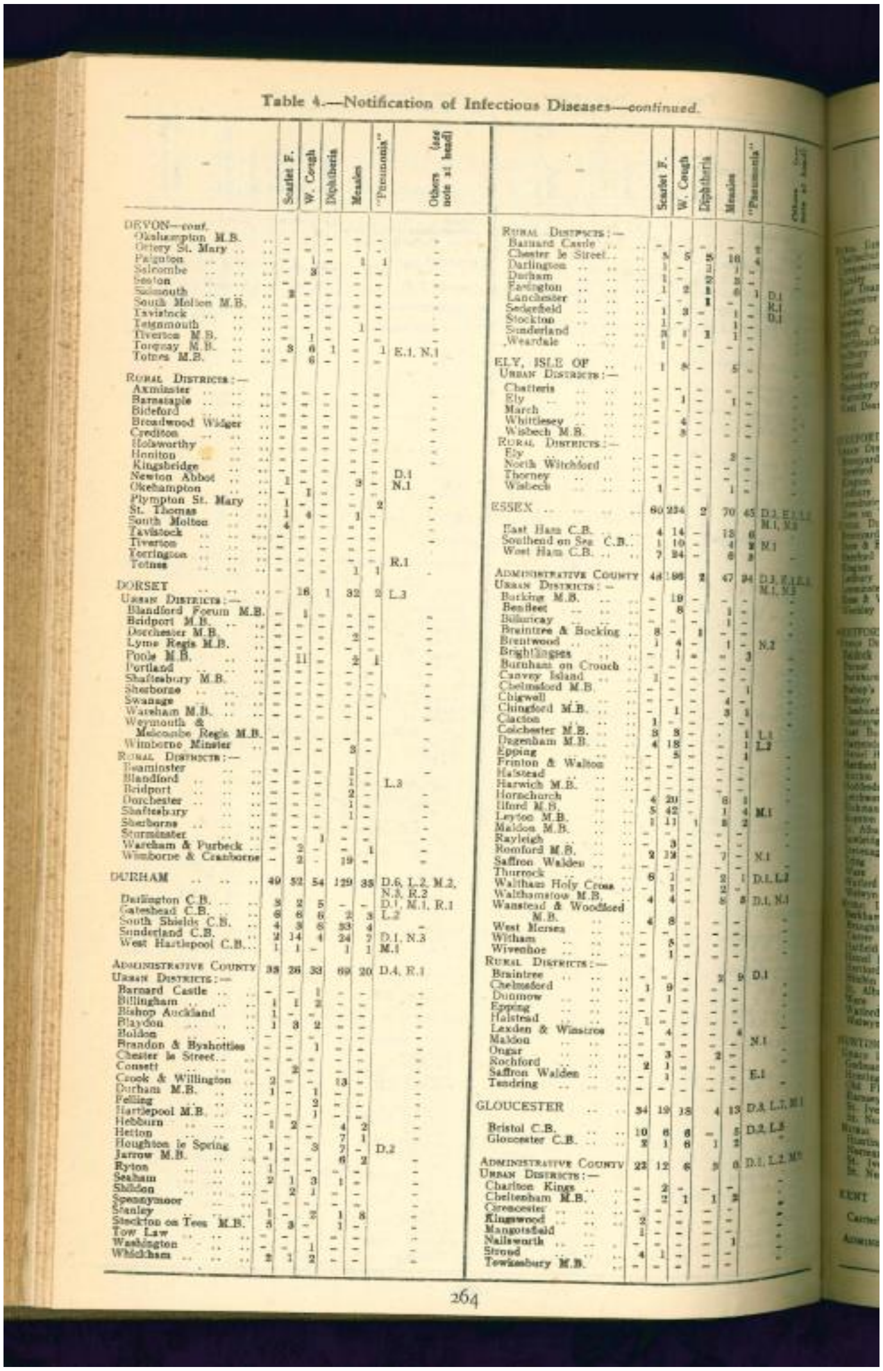


Table 4.—Notification of Infectious Diseases—*continued.*

—	Scarlet F.	W. Cough	Diphtheria	Measles	"Pneumonia"	Others (see note at head)
DEVON—*cont.*						
Okehampton M.B.	–	–	–	–	–	–
Ottery St. Mary	–	–	–	–	–	–
Paignton	–	1	–	1	1	–
Salcombe	–	3	–	–	–	–
Seaton	–	–	–	–	–	–
Sidmouth	2	–	–	–	–	–
South Molton M.B.	–	–	–	–	–	–
Tavistock	–	–	–	–	–	–
Teignmouth	–	–	–	1	–	–
Tiverton M.B.	–	1	–	–	–	–
Torquay M.B.	3	6	1	–	1	E.1, N.1
Totnes M.B.	–	6	–	–	–	–
RURAL DISTRICTS:—						
Axminster	–	–	–	–	–	–
Barnstaple	–	–	–	–	–	–
Bideford	–	–	–	–	–	–
Broadwood Widger	–	–	–	–	–	–
Crediton	–	–	–	–	–	–
Holsworthy	–	–	–	–	–	–
Honiton	–	–	–	–	–	–
Kingsbridge	–	–	–	–	–	–
Newton Abbot	1	–	–	3	–	D.1
Okehampton	–	1	–	–	–	N.1
Plympton St. Mary	1	–	–	–	2	–
St. Thomas	1	4	–	1	–	–
South Molton	4	–	–	–	–	–
Tavistock	–	–	–	–	–	–
Tiverton	–	–	–	–	–	
Torrington	–	–	–	–	–	
Totnes	–	–	–	1	1	R.1
DORSET	–	18	1	32	2	L.3
URBAN DISTRICTS:—						
Blandford Forum M.B.	–	1	–	–	–	
Bridport M.B.	–	–	–	–	–	–
Dorchester M.B.	–	–	–	2	–	–
Lyme Regis M.B.	–	–	–	–	–	–
Poole M.B.	–	11	–	2	1	–
Portland	–	–	–	–	–	–
Shaftesbury M.B.	–	–	–	–	–	–
Sherborne	–	–	–	–	–	–
Swanage	–	–	–	–	–	–
Wareham M.B.	–	–	–	–	–	–
Weymouth & Melcombe Regis M.B.	–	–	–	–	–	–
Wimborne Minster	–	–	–	3	–	–
RURAL DISTRICTS:—						
Beaminster	–	–	–	1	–	–
Blandford	–	–	–	1	–	L.3
Bridport	–	–	–	2	–	–
Dorchester	–	–	–	1	–	–
Shaftesbury	–	–	–	1	–	–
Sherborne	–	–	–	–	–	–
Sturminster	–	–	1	–	–	–
Wareham & Purbeck	–	2	–	–	1	–
Wimborne & Cranborne	–	2	–	19	–	–
DURHAM	49	52	54	129	35	D.6, L.2, M.2, N.3, R.2
Darlington C.B.	3	2	5	–	–	D.1, M.1, R.1
Gateshead C.B.	6	6	6	2	3	L.2
South Shields C.B.	4	3	6	33	4	–
Sunderland C.B.	2	14	4	24	7	D.1, N.3
West Hartlepool C.B.	1	1	–	1	1	M.1
ADMINISTRATIVE COUNTY	38	26	33	69	20	D.4, R.1
URBAN DISTRICTS:—						
Barnard Castle	–	–	1	–	–	–
Billingham	1	1	2	–	–	–
Bishop Auckland	1	–	–	–	–	–
Blaydon	1	3	2	–	–	–
Boldon	–	–	–	–	–	–
Brandon & Byshottles	–	–	1	–	–	–
Chester le Street	–	–	–	–	–	–
Consett	–	2	–	–	–	–
Crook & Willington	2	–	–	13	–	–
Durham M.B.	1	–	1	–	–	–
Felling	–	–	2	–	–	–
Hartlepool M.B.	–	–	1	–	–	–
Hebburn	1	2	–	4	2	–
Hetton	–	–	–	7	1	–
Houghton le Spring	1	–	3	7	–	D.2
Jarrow M.B.	–	–	–	6	2	–
Ryton	1	–	–	–	–	–
Seaham	2	1	3	1	–	–
Shildon	–	2	1	–	–	–
Spennymoor	–	–	–	–	–	–
Stanley	1	–	2	1	8	–
Stockton on Tees M.B.	5	3	–	1	–	–
Tow Law	–	–	–	–	–	–
Washington	–	–	1	–	–	–
Whickham	2	1	2	–	–	–

—	Scarlet F.	W. Cough	Diphtheria	Measles	"Pneumonia"	Others (see note at head)
RURAL DISTRICTS:—						
Barnard Castle	–	–	–	–	2	–
Chester le Street	5	5	5	16	4	–
Darlington	1	–	1	1	–	–
Durham	1	–	2	3	–	–
Easington	1	2	1	6	1	D.1
Lanchester	–	–	1	–	–	R.1
Sedgefield	1	3	–	1	–	D.1
Stockton	1	–	–	1	–	–
Sunderland	3	1	1	1	–	–
Weardale	1	–	–	–	–	–
ELY, ISLE OF	1	8	–	5	–	–
URBAN DISTRICTS:—						
Chatteris	–	–	–	–	–	–
Ely	–	1	–	1	–	–
March	–	–	–	–	–	–
Whittlesey	–	4	–	–	–	–
Wisbech M.B.	–	3	–	–	–	–
RURAL DISTRICTS:—						
Ely	–	–	–	2	–	–
North Witchford	–	–	–	–	–	–
Thorney	–	–	–	–	–	–
Wisbech	1	–	–	1	–	–
ESSEX	60	234	2	70	45	D.3, E.1, L.1, M.1, N.3
East Ham C.B.	4	14	–	15	6	
Southend on Sea C.B.	1	10	–	4	2	N.1
West Ham C.B.	7	24	–	6	3	–
ADMINISTRATIVE COUNTY	48	186	2	47	34	D.3, E.1, L.1, M.1, N.3
URBAN DISTRICTS:—						
Barking M.B.	–	19	–	–	–	–
Benfleet	–	8	–	1	–	–
Billericay	–	–	–	1	–	–
Braintree & Bocking	8	–	1	–	–	–
Brentwood	1	4	–	1	–	N.2
Brightlingsea	–	1	–	–	3	–
Burnham on Crouch	–	–	–	–	–	–
Canvey Island	1	–	–	–	–	–
Chelmsford M.B.	–	–	–	–	1	–
Chigwell	–	–	–	4	–	–
Chingford M.B.	–	1	–	3	1	–
Clacton	1	–	–	–	–	–
Colchester M.B.	3	3	–	–	1	L.1
Dagenham M.B.	4	18	–	–	1	L.2
Epping	–	5	–	–	1	–
Frinton & Walton	–	–	–	–	–	–
Halstead	–	–	–	–	–	–
Harwich M.B.	–	–	–	–	–	–
Hornchurch	4	20	–	6	1	–
Ilford M.B.	5	42	–	1	4	M.1
Leyton M.B.	1	11	1	3	2	–
Maldon M.B.	–	–	–	–	–	–
Rayleigh	–	3	–	–	–	–
Romford M.B.	2	12	–	7	–	N.1
Saffron Walden	–	–	–	–	–	–
Thurrock	6	1	–	2	1	D.1, L.2
Waltham Holy Cross	–	1	–	2	–	
Walthamstow M.B.	4	4	–	5	5	D.1, N.1
Wanstead & Woodford M.B.	4	8	–	–	–	–
West Mersea	–	–	–	–	–	–
Witham	–	5	–	–	–	–
Wivenhoe	–	1	–	–	–	–
RURAL DISTRICTS:—						
Braintree	–	–	–	2	9	D.1
Chelmsford	1	9	–	–	–	–
Dunmow	–	1	–	–	–	–
Epping	–	–	–	–	–	–
Halstead	1	–	–	–	–	–
Lexden & Winstree	–	4	–	–	4	–
Maldon	–	–	–	–	–	N.1
Ongar	–	3	–	2	–	–
Rochford	2	1	–	–	–	–
Saffron Walden	–	1	–	–	–	E.1
Tendring	–	–	–	–	–	–
GLOUCESTER	34	19	18	4	13	D.3, L.1, M.1
Bristol C.B.	10	6	6	–	5	D.2, L.3
Gloucester C.B.	2	1	6	1	2	–
ADMINISTRATIVE COUNTY	22	12	6	3	6	D.1, L.2, M.1
URBAN DISTRICTS:—						
Charlton Kings	–	2	–	–	–	–
Cheltenham M.B.	–	2	1	1	2	–
Cirencester	–	–	–	–	–	–
Kingswood	2	–	–	–	–	–
Mangotsfield	1	–	–	–	–	–
Nailsworth	–	–	–	–	1	–
Stroud	4	1	–	–	–	–
Tewkesbury M.B.	–	–	–	–	–	–

264

Figure 2-4: An example page from the Registrar-General's Weekly Return[34]

Weekly returns between February 1941 and December 1973 were digitised, leading to approximately 1670 weeks, 20,000 pages, and 40,000 tables requiring digitisation. ArchiveOCR can, for a series data set, allow users to provide row and column names to accelerate the process. There were only 12 major changes to row names, so these were digitised once separately and then loaded for the relevant pages. However, at present, a column labelled 'Other' could not be digitised due to it having multiple nested rows within potential entries; work since has been undertaken to address this limitation.

2.5.1 Training OCR

The only columns processed were numeric, meaning only characters 0-9 required training. Training was done from a single week of each year, representing 33 manual sorting training runs. A sufficiency acceptance rate of 85% led to 909 sufficient characters being isolated across the ten digits outlined in Table 2. The predictive acceptance rate was set to 75%.

Table 2-2: Number of sufficient characters used for this process.

Character	Count
0	41
1	89
2	221
3	144
4	50
5	85
6	101
7	34
8	75
9	69

2.5.2 Validation of output

There were 15,098,150 characters for identification, although only 3,058,921 of these were not dashes and processed; dashes were isolated based on height and set to zero. Each page was then manually validated using the inline totals of the counties that the districts are within. After analysing two years of data, it became clear automation could be applied, and a script was built to calculate where the errors were and highlight them to accelerate the process. Between the years of 1943 and 1973, this script identified there were 13,927 errors (0.45%) that required correction shown in Figure 2-5.

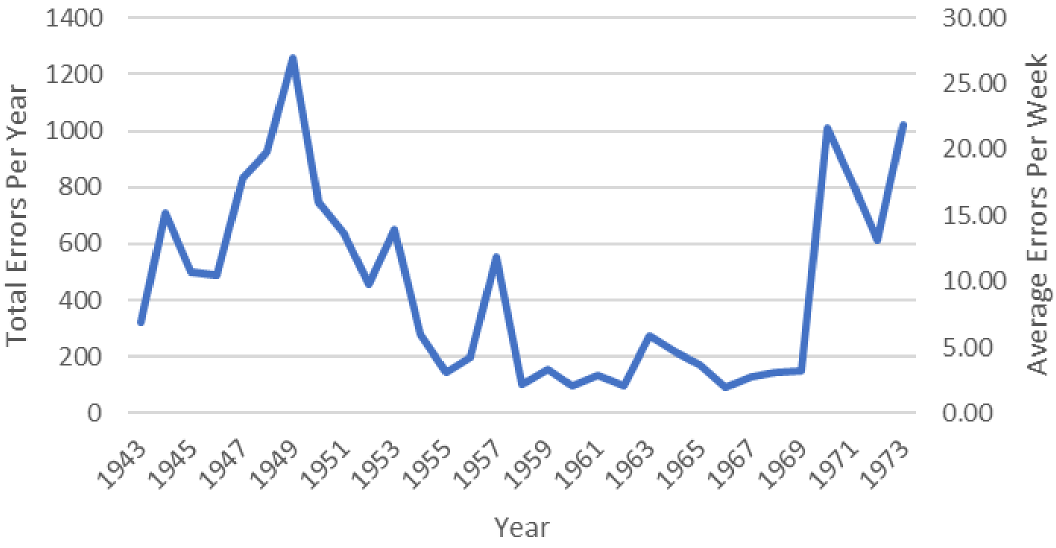


Figure 2-5: Number errors within each year. The total corrections per year are on the left-hand horizontal axis and average per week on the right-hand horizon

To assist future revisions, a simple script and GUI was built to iterate through these errors, where each error could be assigned one of eight reasons why the error had occurred. Figure 2-6 shows the result of this process, where the error clause has been broken into two categories. Misprinted values N = 1658, (0.054%), printing location errors (N = 119, 0.004%), no character being present (N = 374, 0.012%), the character being eroded (N = 1579, 0.052%), and the character being obstructed (N = 1257, 0.041%) are errors deemed not explicitly the fault of ArchiveOCR.

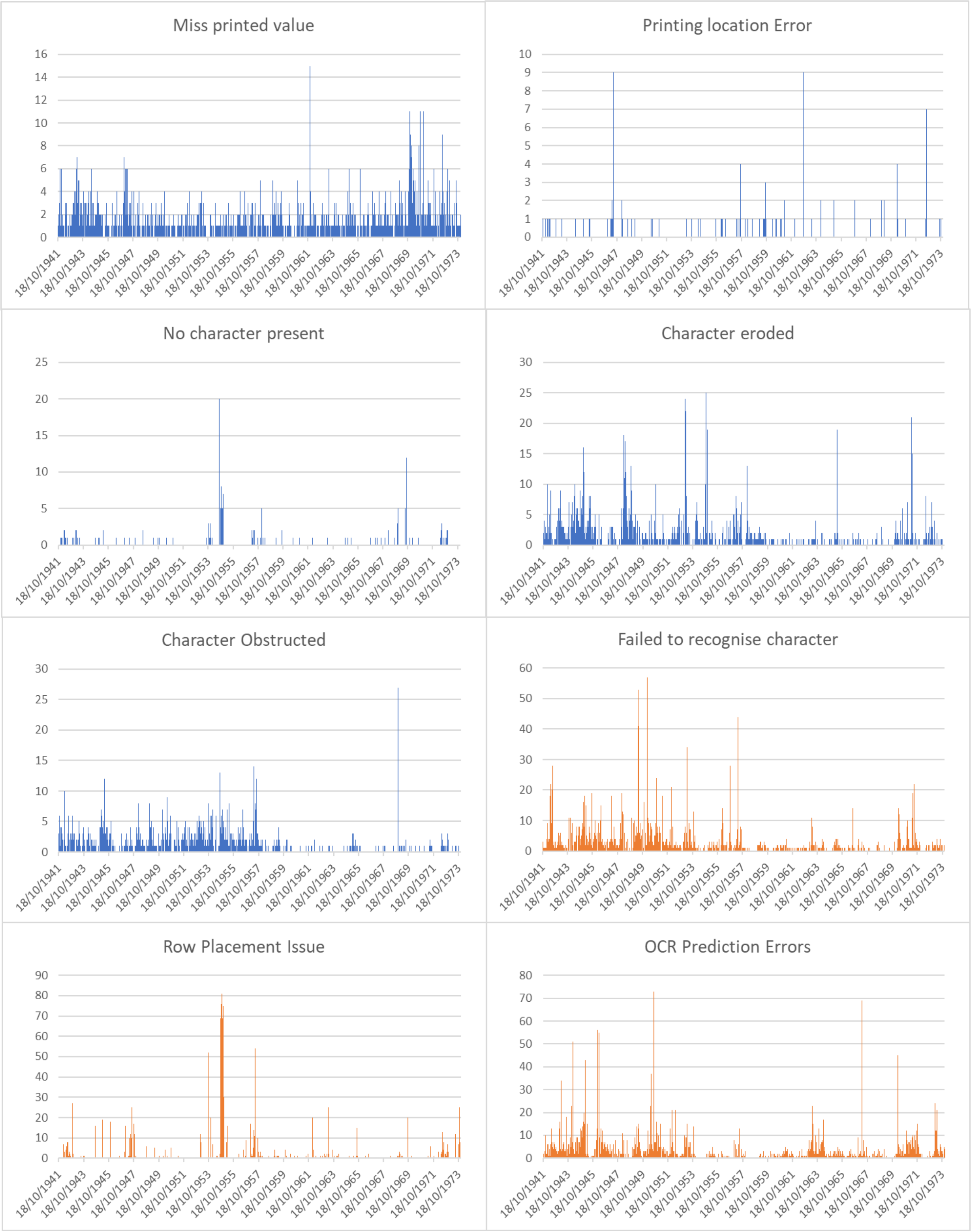


Figure 2-6: Error by types by week across 33 years of weekly reports

Conversely, failing to recognise a character representing type one error (N = 2778, 0.091%), row placement issues from Algorithm 2-2 (N = 3790, 0.124%), and prediction errors representing type 2 error (N = 3851, 0.126%) are deemed a fault of the setup and the system itself. Whilst the script to find the errors could only find regions of the page, this review process could identify individual rows, with the total rows with errors being 15,419. Whilst this system regrettably did not note how many characters required correction, given on average there is 1 character per row, for the 3,058,921 characters isolated this means the predictive accuracy was 99.5%.

Accuracy is a complex metric, as OCR accuracy is both the accuracy of isolation of elements, and the correct prediction of the text within the elements[34]. Most professional software reports an accuracy of 99%[22], which is the required standard from many governments for OCR software[46], but this is on clean ink-jet text that is likely already available in a digital form. Digitisation of newspapers from 1803 to 1954 led to accuracy that varied from as a high as 98.02% to as low as 71%[22]. Therefore, given the historical nature of the documentation, we are confident ArchiveOCR represents a contributed to this field.

2.6 Conclusion

ArchiveOCR offers an easy-to-use alternative to many OCR packages and software at present and has been tested on a large dataset. It has continued to be used for further projects and avoids overly complicated methods of operation, focusing on structural components of pixels within tables, columns, or rows for identification. Using ArchiveOCR allowed for 15,098,150 characters to be processed and cleaned within the span of two months, mostly by a single individual with little financial cost, something which would be challenging using presently available means.

3 Chapter 3: WeightGIS: Automate Standardisation of Regions

3.1 Introduction

The aggregation of individual-level data into larger districts can change the underlying mean and standard deviations of the data, which can change associations to phenotypic outcomes[47]. Bias resulting from changes of areal data, because of scaling or zoning, is known as the Modifiable Area Unit Problem[48,49]. This bias is a form of selection bias which can significantly change the value and direction of regression analysis if not account for[50]

To demonstrate this, let us use an example. In this case, the reported data is at the district level containing 10,000 individuals, a lower order unit of 1km grid squares exists, as well as households (and their location) where the data originates from. Figure 3-1 shows this example visually, with the circles representing a household with an individual infected with a notifiable disease. The district is locally also unofficially divided in terms of north-south areas, with the north area shaded in Figure 1. If this represents weekly data, then this district reported 20 cases of a notifiable disease in this week.

The scaling issue here is that, by reporting the aggregate of these disease cases, the observed associations this disease may have when aggregated may not be representative of the underling truth. For example, an outbreak of 20 cases might not be considerable for an area with 10,000 individuals, as this only results in 0.2 cases per 100 individuals. However, if only 500 individuals live across the North, then given 17 of these cases occurred in the north, this instead results in 3.4 cases per 100 individuals. When we aggregated, we lose the ability to utilise this underlying variation and potentially miss important associations.

Aggregation like this can also inflate correlation values when comparing between districts[49]. This is because as you aggregated smaller zones into large ones, the overall variation in the data is reduced, making areas more artificially more comparable. This can cause difficulty in interpreting analysis, as the homogenous distribution assumption of the variable across the larger geospatial units becomes far harder to hold[51]. The larger the aggregation, the more bias is likely introduced as a result to any subsequent analysis[48].

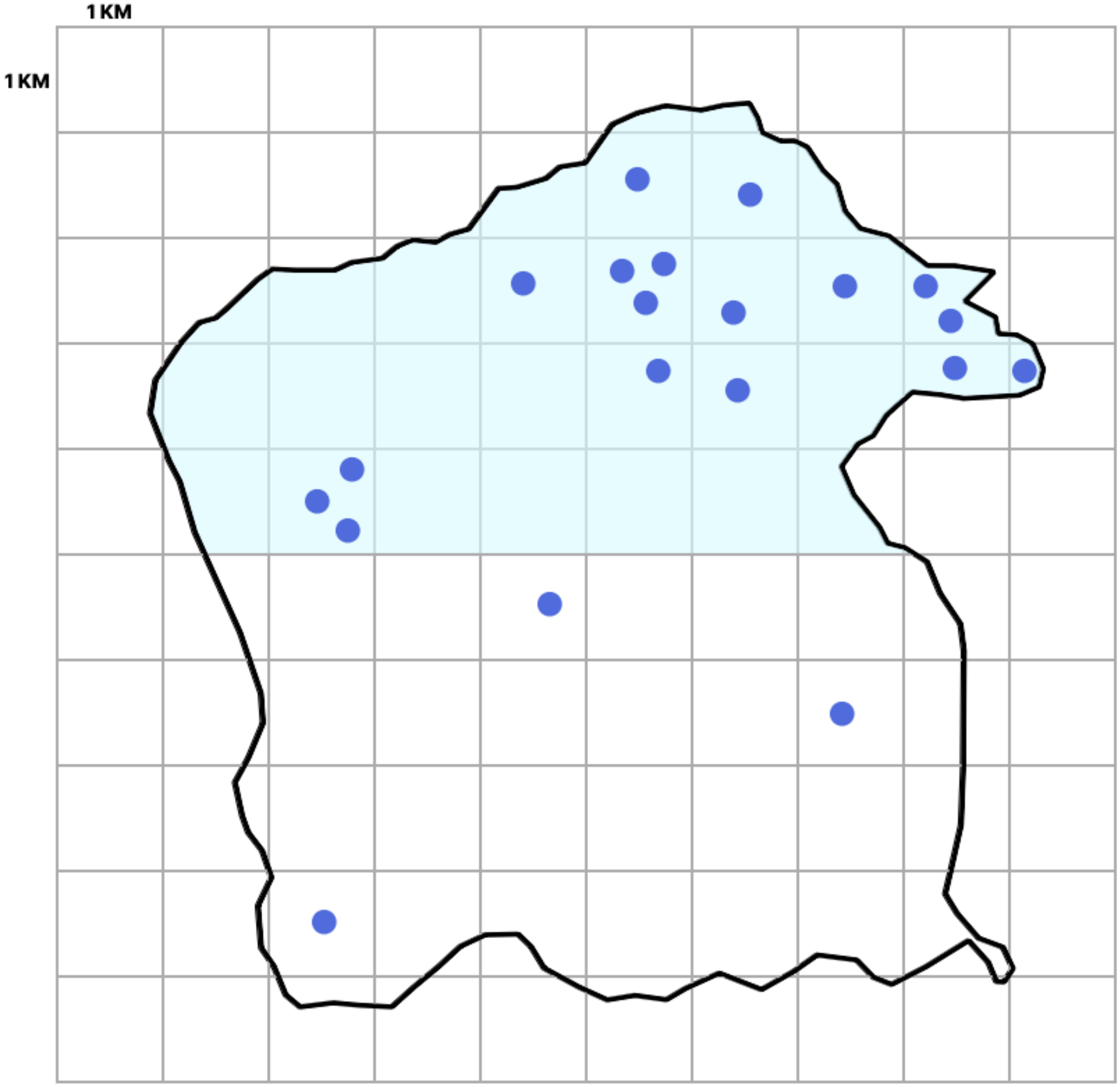


Figure 3-1: An example district (outlined in black), with 1km grid squares (grey outlines), households (circles), and a division of the district in a north (shaded) and area.

The next issue that can result from the Modifiable Areal Unit Problem is that of zoning. Suppose in this example that the government is in the process of re-drawing the electoral boundaries and decides to split the current district on the north-south division. However, other parties complain that this would result in ruling power maximizing their voting share, and instead propose it be divided east and west, so that both new districts remain politically diverse. Regardless of the choice, the decision is purely arbitrary, yet may still have significant consequences for areal data. Despite reporting the exact same week, the counts and subsequent means and standard deviations of the rates of these east-west and north-south splits of our example weeks disease notifications would be completely different, depending on how the district is partitioned[49]. This could change the outcome of any analysis using this data, despite the reasoning for this change being an arbitrary decision on where to draw a boundary.

If these boundaries change frequently over time, this limits any potential for comparison of population characteristics or dynamics in a time series setting when using areal data[52]. This is a widespread problem[53,48,52,54], for example, between 1841 and 1972 there were 4247 changes to the districts in Britain, and over 20,000 at a smaller geographic level of parishes between 1876 and 1972[48].

There have been various solutions to limit the bias caused from Modifiable Areal Unit Problem. One of the original solutions was aggregation into larger static geospatial units that remain constant over time[51]. Whilst simple to apply, aggregation like this has significant consequences when interpreting analysis. Similar to the concerns of the aggregation of individual to regional data, the means and standard deviations can be manipulated on aggregation which may change the result. You also in this instance are implicitly giving up useful information that exists, which is highly undesirable.

Despite these problems, aggregation is still commonly applied, as alternative methods can be complex. For example, a recent paper undertook a Genome Wide Association Study (GWAS) investigating the association between regional Infant Mortality Rates

(IMR) as a proxy for early life environment[55] and natural selection. The hypothesis being that certain genetic variants may not survive in populations that have a harsh early life environment but could if one's early life environment was better. However, the authors aggregated district level IMR (approximately 1472 districts) to 62 counties [55], despite the extensive variation that exists in IMR at a district level[56]. Given the level of aggregation this results in, serious questions should be asked on *what* the inference of said results mean.

Whilst more challenging, other techniques found within a GIS background mostly involve weighting. One of the simplest forms of weighting is areal interpolation, where the difference between areas in geographic units (such as meters squared) is used to represent the weight[52]. Multiplying the population base parameter by said weight, scales it to the original shape of the location, allowing it to remain consistent over time[52]. This too, has significant problems, the most obvious being that of the assumption of homogenous populations across geospatial space[48]. If 60% of the population live in the 5% of the area that was transferred, then the weights will not account for this.

More complex solutions have utilised additional source data to construct the weights. Satellite data of density of infrastructure[57], night-time light[58], road map data[59], and sub-unit populations[52,48], are all examples of methods used to address the homogenous population problem of areal weighting. However, in principle, all methods of weighting at present require an estimation of the source population that moved to another district[52]. Although, many of the more advanced techniques require data, such as from satellites, that cannot be utilised in a historical context which is of particular concern in this instance.

3.1.1 Pipeline Justification

A review of data constructed after digitising district level data from the registrar generals' weekly disease notifications[42] in England and Wales from 1941 to 1973 led to significant concerns. Considerable changes to both the number and shape of districts occur within this period, which would make the study of population and regional characteristics over time extremely challenging without introducing bias. The UK especially has established long-running geospatial disparities. The north-south divide, based on evidence of migration pattern to London from 1851 to 1911, was already established in the 19th century[60]. However, even in adjoined locations, it is common to find wages and living conditions that differ and remain entrenched [61]. Therefore, these circumstances would warrant weighting to limit any potential selection bias that is a result of zoning issues that occur over the course of the 32-year period.

Given our data was historical, we had limited options to use for weighting. Whilst aggregation would be simple, any interpretation of the results would be questionable, as it would reduce the variation in the data set from around 1472 locations to only 62. Previous literature suggested that for the UK, parishes would be sufficiently small that they could act as a sub-unit population weighting parameter for larger areas such as districts[48]. There are approximately 17,000 parishes, relative to the 1472 districts as of 1951 in England and Wales, making weighting highly specific to settlement level populations. Weighting via sub-units is not new but can involve a considerable amount of laboriously manual labour. Given the scale of the task ahead, I designed a new pipeline for undertaking these procedures.

Within this paper, I first present a python package, WeightGIS, that is designed to act as a pipeline to standardise regional boundaries by using sub-units. It has been generalised so should be of use to other researchers undertaking similar problems in other countries, which given the frequency of the problem[52], should be useful. I then present the sub-unit population weights constructed via WeightGIS and compare said weights to standard areal weighting that is also commonly used. Weights allow for any external data that can be linked to districts to be standardised between the years of 1931-1974 to the 1951 census by default. Finally, I explain how, in future work, these weights will be validated, and with this validation, any underlying issues corrected. The future aim being to construct a time-invariant view of the UK from the first census undertaken by a centralised office of the General Register Office in 1851[48], to the recent 2021 census.

3.2 Construction of weights

WeightGIS uses shapefiles, a format for boundary data, to compare changes relative to a base year by investigating overlapping geometry and sub geometry. For instance, for WeightGIS, shapefiles' polygons represent the regions or locations for WeightGIS to compare. There are multiple python-based packages to load shapefiles presently. WeightGIS uses an altered version of pyshp[62], named ShapeObject[63], which loads geometry as Shapely[64] compatible geometry. The Shapely[64] library contains crucial methods for the calculation of overlap between polygons. Therefore, adjusting a pre-existing loader from pyshp[62] to construct shapefiles as objects of Shapely[64] geometry simplifies the WeightGIS process. Although developed for WeightGIS, ShapeObject[63] is available separately.

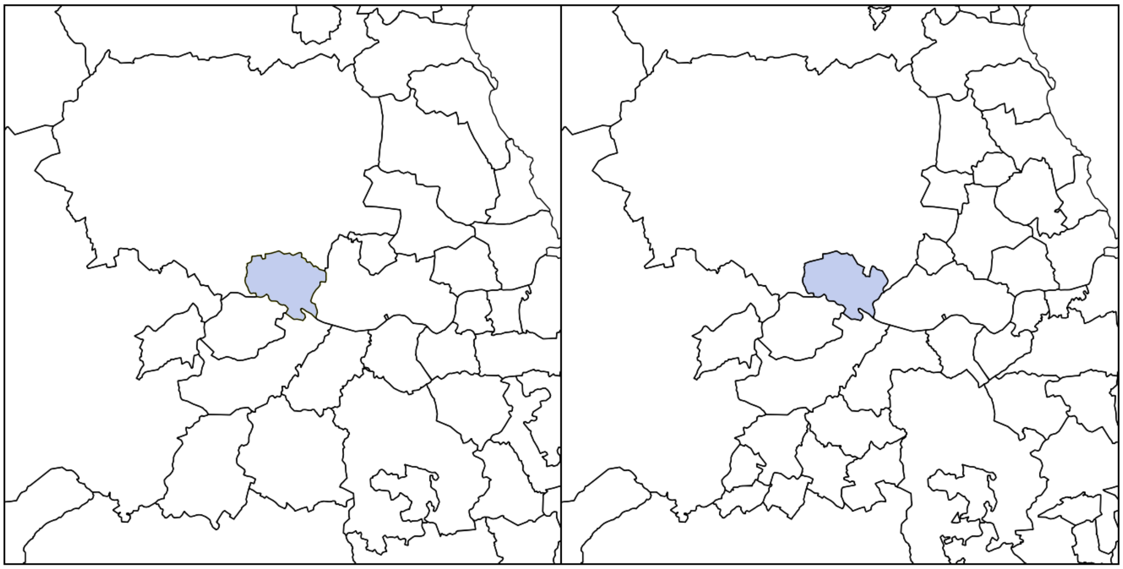


Figure 3-2: Changes to Newburn UD, highlighted polygon, in 1931 (Left) compared to 1951 (Right)

Here, we utilise data from the Great British Historical GIS project[65] to explain the process of WeightGIS through an example. Specifically, as shown in Figure 3, the example utilises changes involving Newburn urban district (UD) in 1931 and 1951. This example will also use the parish structure and population as of 1921 for sub-unit weighting.

3.2.1 Construction of base weights

The first step is to create the base weights. These weights represent total changes between geographical regions and act as the rough weights that may need to be adjusted. Whilst comparisons between shapefiles can determine if and to what extent changes occurred, they cannot infer when they occurred. In this stage, the base weights are the combined effect of all changes between shapefiles. After constructing the base weights, WeightGIS then has methods to help assign when they occur and to unpick combined changes.

Returning to the example, between 1931 and 1951, Newburn UD both ceded and gained territory. Newcastle Upon Tyne County Bough (CB) absorbed part of Newburn's eastern border. However, Newburn also gained part of the Castle Ward rural district (RD) to the north. Figure 3-3 details these changes. As we are using the base year of 1951, the only districts that overlap with the 1951 shape of Newburn UD are Castle Ward RD and Newburn UD. As Newcastle Upon Tyne CB absorbed territory before the base year shape, this change had already occurred by 1951. Therefore, weights before 1951 will not include Newcastle Upon Tyne CB.

WeightGIS can construct both area weights and sub-unit weights. The area differences between the shapefiles are the weight for area weighting. Despite potentially being less accurate, area weighting is useful if using highly detailed shapefiles that do not have a lower level available to use as a sub-unit weight. Area weighting works by isolating the percentage of the base shape that overlaps another shape, in another shapefile, at the same level. Referring to Figure 3-3, the 1951 shape of Newburn UD includes the polkadotted area, but not the line dashed area, with the opposite for the 1931 shape of Newburn UD. The area overlaps for the 1951 shape for Newburn UD are, as of 1931, 93.9% for Newburn UD and 0.4% for Castle Ward RD. Therefore, with area weighting, these values represent the weights.

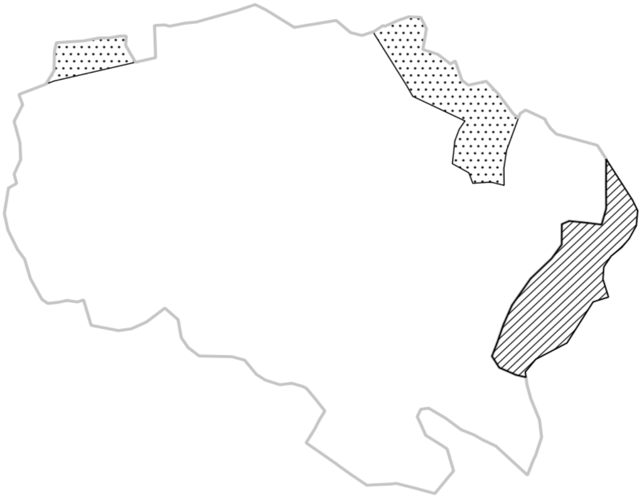


Figure 3-3: The grey outline shows Newburn UD borders as of 1951. The territory Newburn UD gained from Castle Ward RD shown with polka dots, were as territory ceded to Newcastle Upon Tyne CB line dashes.

Subunit weighting is more complicated. WeightGIS will divide the comparison shape into smaller areas using the subunit shapefile. Doing so creates a comparison shape but one made of subunits, with Figure 3-4 showing both the 1931 shapes of Newburn UD on the left and Castle Ward RD on the right after subdivision. WeightGIS then isolates the parts of this comparison shape of subunits that overlap the base shape, highlighted in grey in Figure 3-4. The weight is the total population in the highlighted area divided by the population total of the complete set of subunits.

Using the change between Newburn UD and Newcastle Upon Tyne CB as an example, the 1931 district shapefile comprises the seven parishes shown in Figure 3-4. Table 3-1 shows the seven parish populations before and after area weighting. The change

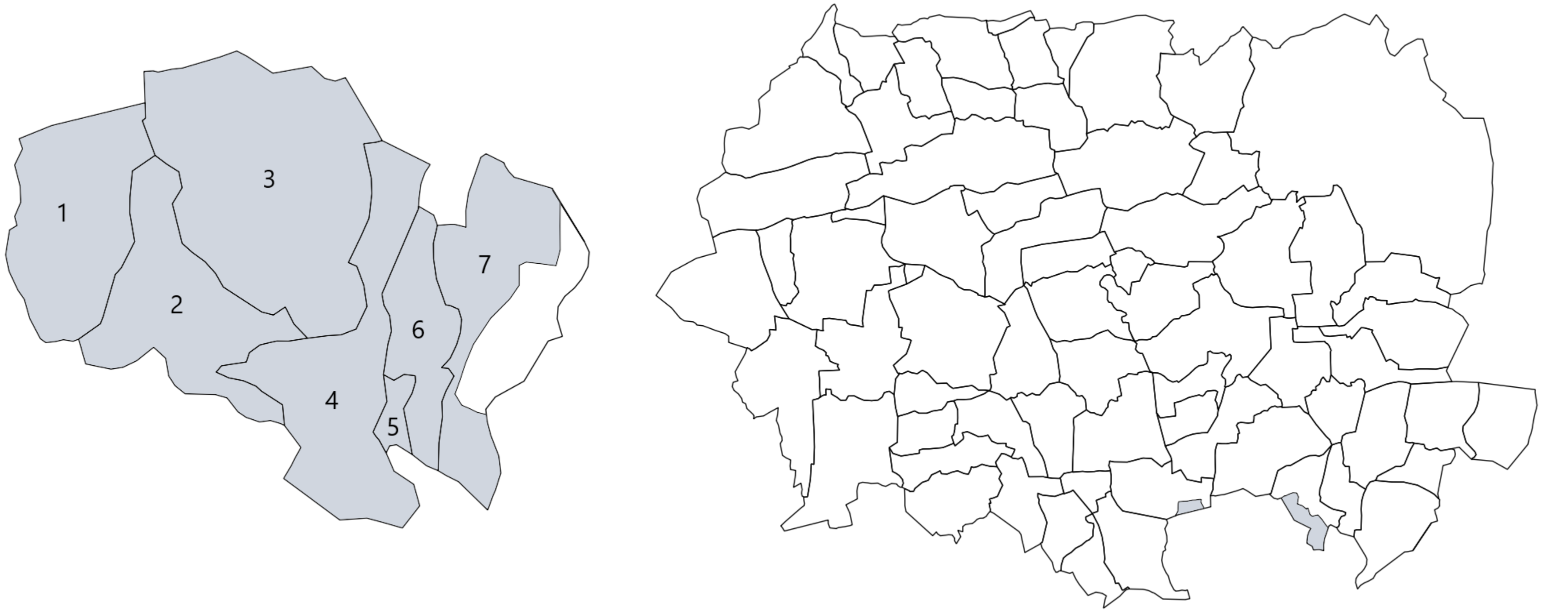


Figure 3-4: Areas of sub-units that are part of 1951 Newburn UD shown in grey

between Newburn UD and Newcastle Upon Tyne CB affects only a single parish of East Denton CP (7 in Figure 3-4).

Table 3-1: Population of parishes contained within Newburn UD, with the first entry corresponding to the parish that was partially absorbed by Newcastle Upon Tyne CB in 1951.

ID	Parish Name	1931 Population	1951 Population
1	Thockley CP	2640	2640
2	Newburn CP	4523	4523
3	WallBottle CP	3080	3080
4	Newburn Hall CP	4164	4164
5	Sugley CP	1054	1054
6	West Denton CP	504	504
7	East Denton CP	2865	1869
-	**Total Population**	**18829**	**17833**

Newburn UD in 1951, therefore, contains six whole parishes and 64.73% of the population of East Denton CP, which leads to population totals for Newburn as of 1931 and 1951 of 18829 and 17833, respectively. The weight is then the difference in subunit populations totals between the base year and the alternate shapefiles' year. With the base year of 1951, the weight is 17833 / 18829 or 94.71%. Applying the same process to the change with Castle Ward RD leads to a weight of 0.24%.

3.2.2 Algorithmic generalisation of base weight construction

Algorithm 3-1 shows the generalised iteration process explained in section 3.2.1. For a base shapefile of length (BL), determined from the number of polygons within the shapefile, WeightGIS isolates each ith polygon as the base polygon (BP). WeightGIS will then compare BP to each alternative polygon (AP) in each alternative shapefiles (AS). If BP overlaps part of AP, then WeightGIS will calculate a weight. If no change occurs, the area weight is 100%, and WeightGIS will also set the population weight to equal 100%. Otherwise, WeightGIS uses the sub-unit shapefile (SUS) AP into sub-unit polygons containing a weighting parameter (WP).

Algorithm 3-1: Construction of base weights

Input: A list of shapefiles (LS), the name of the shapefile to be set as the base, and a sub-unit shapefile (SUS) with a weight parameter (WP) within its attribute table.
Output: A json database of weights for location in the base shapefile, for each reference time.

```
Isolate base shapefile (BS) from other shapefiles of length (BL)
for i=0 to BL do
    Isolate the i$^{th}$ base polygon (BP) from BS
    for alternate shapefile (AS) in LS that is not BS do
        Isolate each alternate polygon (AP) as a list (APL) in AS that overlaps BP
        for AP in APL do
            Calculate area overlap between BP and AP
            if area overlap != 100%
                Divide AP and into subunits from SUS, calculate total $\sum_{i=0}^{AP} WP$
                Isolate within subunits shaped as AP by BP, calculate total $\sum_{i=0}^{BP} WP$
                Calculate the population weight from WP totals as $\frac{\sum_{i=0}^{BP} WP}{\sum_{i=0}^{AP} WP} * 100$
            Elif area overlap == 100%
                population weight == 100%
            end if
        end for
    end for
end for
```

The subunits that overlap AP are first made fit, with the fitted sub-unit area used to scale the WP. Once all sub-units that make up AP have been isolated, the total of WP from these fitted subunits is calculated. The BP is then overlaid on the AP shape made up of subunits, with any areas not included removed. The weight is then the summation of WP in BP divided by the summation of WP in AP. This process iterates until all BP weights have been constructed from comparisons to AP in all alternate shapefiles. Then, WeightGIS stores the calculated percentages weights as python floats in a JSON database. Python floats are equivalent to C double, which gives 14 floating points of precision.

3.2.3 JSON datastructure

Figure 3-5 shows an example JSON database for a base shapefile with a single entry, with the JSON structure used by WeightGIS as follows. First, the database states the name of BP, constructed from one or multiple entries within the BS attribute table. Second, the date for the weights of a given AS, which is derived from the names of the shapefiles themselves. The third level includes the names of AP that overlap the base shape within that AS's date level. Finally, each third-level overlap stores both the area and population weights calculated. In the base year, there will only be a single level

3.3 Quality control weights

The base weights both lack standardised names and dates. Specifically, the changes do not show exactly when the change occurred, only that a change occurred between the years of the shapefiles. If the difference in time between shapefiles is large, this leads to many years being inappropriately weighted. Also, a lack of standardised names can lead locations to appear to change drastically despite the location only changing the name. The weights, therefore, require quality control procedures.

3.3.1 Quality controlling names

WeightGIS can create a standardised set of names, here referred to as a place reference, for each location in the base year. This look-up database standardises all names to those used in the base year. Locational name changes represent one of the most common sources of place variation over time[66], so this is a required step to use the weights.

```json
{
  "10108949__NEWBURNUD": {
    "1931": {
      "10108937__CASTLE WARDRD": {
        "Area": 0.3510946943497809,
        "Population": 0.23678662780411672
      },
      "10108949__NEWBURNUD": {
        "Area": 93.86227723032636,
        "Population": 94.71453886174326
      }
    },
    "1951": {
      "10108949__NEWBURNUD": {
        "Area": 100.0,
        "Population": 100.0
      }
    }
  }
}
```

Figure 3-5: JSON data structure of a weight entry

To standardise names, WeightGIS constructs unique IDs or utilises pre-existing IDs that link to the location of the geometry. The unique location IDs can infer locations of different names as identical if they overlap. WeightGIS also allows for a multi-level structure of names. Multi-level names can be corrected simultaneously but can assist with frequent within-level name duplication.

Table 3-2: An example of a multi-level place reference for districts and counties.

GID	District	District Alternate 1	County	County Alternate 1
10002217	AXMINSTERUD		DEVON	DEVONSHIRE
10002229	BUDLEIGH SALTERTONUD	BUDLEIGH SALTERTON MB	DEVON	DEVONSHIRE
10002230	LYNTONUD	LYNTONMB	DEVON	DEVONSHIRE
10025369	IVYBRIDGEUD		DEVON	DEVONSHIRE
10026260	TIVERTONMB		DEVON	DEVONSHIRE

However, WeightGIS requires strict nesting of lower-level units. Therefore, if a lower level suffers from ambiguity, the end-user must manually assign the low-level to higher-level relation. Using the UK as of 1951 as an example, users may use a County-District structure of names, as shown in Table 3-2. WeightGIS will then standardise any alternative names back to the first instance. For example, WeightGIS would rename Lynton MB to Lynton UD.

3.3.2 Quality controlling dates

Shapefiles reflect the changes that occurred between their dates. However, it does not detail when these changes occurred. By investigating the base weights that are not equal to 100%, WeightGIS will produce a changelog detailing expected changes. Table 3-3 shows an example changelog with the changes the end-user added. WeightGIS compares each shapefile and adds an expected change if the weights are not 100%. The maximum expected change is the number of alternative shapefiles. It is then up to the user to locate when these changes occurred and fill in the dates. Removing any non-changing location saves the end-user some time.

Table 3-3: An example change log.

GID	Name	Expected Changes	Changes1	Changes2
10173322	STROUD RD	1	01/04/1935	01/04/1936
10108913	NEWCASTLE UPON TYNE CB	1	01/04/1935	-
10108925	AMBLE UD	0	-	-
10108937	CASTLE WARD RD	1	01/04/1935	-
10108949	NEWBURN UD	1	01/04/1935	-

To assign the dates, WeightGIS searches for the number of changes that occurred per observed change. If a single date exists between two shapefiles, WeightGIS assigns the date from the changelog to the observed change. Using the Newburn UD example of before, the change that is observed by 1951 occurred on 01/04/1935. Subsequently, the date of the weighted change is, therefore, 01/04/1935, rather than 1951. Crucially, if multiple changes occur, WeightGIS can only utilise the last change, as this change is the only observed change between the shapefiles. So, for Stroud RD, all the changes will have been set to have occurred on 01/04/1936. Further shapefiles can add additional changes if sufficient information allows for their construction.

3.4 Quality control external data

WeightGIS has methods to assist standardisation of external data into one that is compatible with the weights. Most quality control methods are generalisable, but the user must undertake some steps themselves. WeightGIS requires individual files for each point in time, with the file names as ISO 8601 or yyyy-mm-dd.

3.4.1 Standardisation of names

Standardisation of names uses the place reference constructed initially from shapefiles. Each name in the external data is cross-referenced against the names in the place reference. If a match is found, but it is not a reference name, then WeightGIS reassigns the name as the reference. Crucially, external files may contain names not used within the shapefiles, with a common reason being spelling mistakes. Users can submit a custom spell check sheet to correct all instances to a set reference name. For genuine alternative names, the user must update the place reference itself.

3.4.2 Solve Ambiguity

Sometimes, records may break down regions into ambiguous locations. For example, records may split regions into North and South sub-regions. Unless these regions have a distinct polygon in the base shapefile, then merge issues may arise. Setting all instances of sub-regions to the same reference name prevents the data from being lost. However, this leads to ambiguities, which WeightGIS detects and merges during this quality control step. In addition, the merge warnings may reveal improperly standardised names which, unless corrected, WeightGIS will aggregate.

3.4.3 Relational databases

As WeightGIS uses JSON to store the weights, it also requires the external data to be in the same format before it can weight the external data. JSON loads as dictionaries which increases both the speed of assignment and merging of external datasets. However, unlike SQL, the plain text storage of JSON can assist with debugging for those less familiar with database structures. Once the individual dates have been quality controlled, WeightGIS will restructure the CSV data of date-place-attribute-values to a JSON database format per location of attribute-date-value. This also allows for multiple databases to be quickly joined by location, which can reduce the number of weighting processes that are required.

3.5 Weighting External Data

When both the names and dates of the weights have been quality controlled, they can then be applied to external data. First, WeightGIS constructs a master database from all the quality-controlled external datasets. Then, this master database undergoes weighting using the process shown in Algorithm 3-2. WeightGIS isolates, for each weight location, the dates of any changes. If there is only a single date, no changes occur, and the unweighted data is the weighted version.

Algorithm 3-2: Construct weighted database

Input: Master database from quality controlled external data (MD), and the quality controlled JSON Weights database (WD).
Output: A weighted JSON database

```
For i=0 to length of WD do
    Extract dates of changes (DC) for WD location i
    if The length of DC == 1
        No changes, assign unweighted data to output database for all dates
    Elif The length of DC > 1
        For weight group (WG) in DC do
            Set the start and end date for this WG
            For Place weight (PW) in WG do
                For each attribute-date-value set in MD for $WD^i$
                    If start data <= date < end date
                        Weighted value = (value * (weight / 100))
                    end if
                end for
            end for
            Sum the weighted value for each attribute-date
            Assign weighted value to output
        end for
    end if
end for
```

If more than a single date exists, then values require weighting. WeightGIS sets the start and end dates between weight groups (WG). If this WG is the last group, then the end date of the whole dataset becomes the last date instead. The calculated weight for each attribute-date-value weight is (value * (weight / 100)) because WeightGIS stores the weights as percentages. Across each date, WeightGIS sums the individually weighted values from each weight location to construct the weighted value. This process iterates through each place in WD, then stores the weighted values in a JSON database.

3.6 Example use case by weighting England and Wales between 1931-1974

WeightGIS was designed to assist in construction of a time-invariant district data set from the registrar generals' weekly disease notifications1 in England and Wales from 1941 to 1973. To apply WeightGIS to this data, we used district shapefiles from the Great British Historical GIS project[65] for the years 1931, 1951, 1961, and 1971 in addition to the parish shapefile of 1921. For the weight parameter, we linked the parish population from Vision of Britain[67]. We selected 1951 to construct as our base weight reference year, as it is the centre of the distribution of the shapefiles date range.

Table 3-4: The number of changes that occurred for each district taking 1951 as the base year (N = 1472)

Changes	Count
0	385
1	716
2	300
3	51
4	14
5	4
6	2

We constructed the changelog from Vision of Britain[67], with most districts changed at least once over the 40 years of study, as shown in Table 3. Of the 1472 districts in the base year of 1951, to which we standardised to, WeightGIS standardised 1395. The 77 districts which remained had greater than one change occurring between the shapefiles. To fix these remaining 77, we used information from Vision of Britain[67] that showed parishes associated with each change, which we used to reconstruct additional shapefiles. After rerunning WeightGIS, this allowed standardisation of all 1472 districts. Figure 3-6 shows the districts that exist in the standardised dataset, with those highlighted in colour requiring a border redrawing by hand, to handle more than a single change between census years.

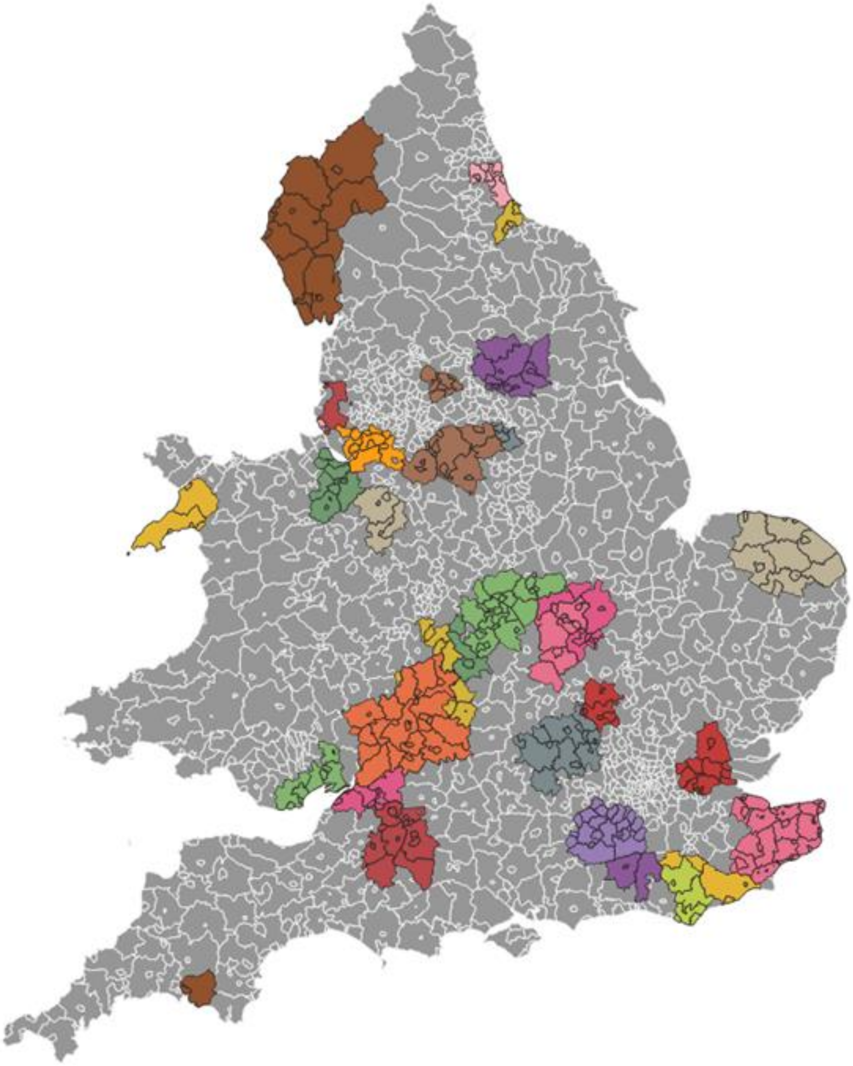


Figure 3-6: Areas in England and Wales that required at least one additional shapefile to be drawn between census shapefiles from additional data as shown in colour.

3.6.1 Evaluating differences between weight types

We calculated both area and sub-unit population weights using WeightGIS. The area weights are the difference in km squared between each district that changed. Whereas for population weighting, the shapefiles were subdivided into parishes, and the weight was calculated from the change in parish population. Of the changes that occurred, 73% had no difference between area and subunit population weighting, as shown in Figure 3-7.

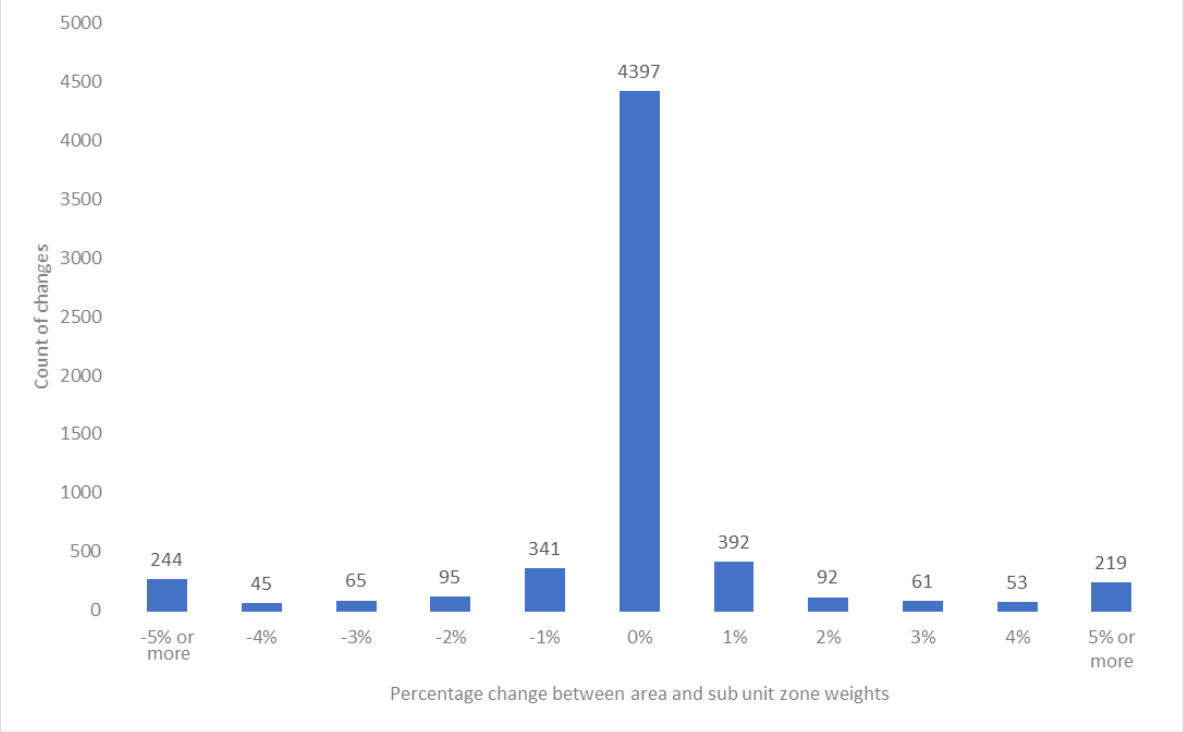


Figure 3-7: The percentage difference in weighting parameter between the area weight estimate, and the sub-unit population weight estimate.

Most of the changes that occurred are documented as reassigning parishes, or parts thereof, to another district. There was, therefore, bound to be similarity, although it was considerably higher than initially expected. Despite this, area weighting would still have led to under or over estimations. The area weights tended to be overestimated when involving changes with rural districts, but only 219 by over 5%, and an underestimated for urban districts, although only 244 by over 5%.

Rural areas are larger and made up of significantly more parishes than other district types and are best suited to show the strength of WeightGIS. Whilst rural districts have less population than many urban areas and with much lower density, the total population can still be considerable. A 10% area change can therefore easily be an overestimate if the specific parishes in question were of lower population density. Whilst the inverse is also possible, if the larger parishes were absorbed, as the changes for rural districts are also large, it was less likely to underestimate than overestimate.

In comparison, Urban districts comprise fewer parishes but at a higher density. If one of four parishes is part of a change, despite it potentially only making up 25% of the area, it is highly possible to have over 25% of the population. As with rural districts, the inverse is also true, and if one of the smaller parishes were involved in the change, then area weighting would overestimate.

3.6.2 Next steps: Validation

Whilst the differences between area and population weighting are interesting, this comparison does not validate how well the underlying weighting procedure has worked. For validation to work, it must be possible to compare the weighted value to the actual value that existed for the weighted shape in another time period. For example, if the district of Leeds has grown by 30% by 1971 compared to its 1951 size, it is required that sub-unit data exists sufficiently small enough to recreate the 1951 version of Leeds as of 1971. This task is challenging, as the only subunit below districts is parishes, and outside of basic population counts and demographics, no information was recorded[48].

Therefore, in a historical context, validation requires an assumption. As the only data available consistently at each census is parish population, the population is the only variable that can be weighed and then validated reliably. All other variables will have to hold the assumption that they are proportioned relative to the population of these parishes, as we cannot validate how weighting works at an individual phenotypic level, as no such data is known to exist. Work has currently started on this validation. The aim initially is to compare how weighting has performed by location by comparing the 1951 weights to the 1931, 1961, and 1971 census reports on parish population. This work is still ongoing, as further data is required to be acquired and digitised in order for this validation to be completed.

3.6.3 Conclusion

The modifiable area unit problem can frustrate the creation of panel data as locations change names or boundaries. WeightGIS offers simple, mostly automated methods to create weights, and also contains a suite of quality control measures to help clean external data before weighting. As not all areas will be suitable for subunit weighting, WeightGIS still allows for both area weighting to increase its potential use. Whilst WeightGIS still requires manual efforts, such as getting the dates of location changes, it offers a more automated solution that is currently known to exist. Therefore, whilst still containing manual efforts, often undertaking these efforts increases the understanding the underlying geography of the area. If the individual is to become the data manager of said geospatial files and weights, this in invaluable. However, as it remains intensive work which will likely precluded individuals not focused a given country or time period within it. WeightGIS offers a pipeline of resources, so regardless of weight type, WeightGIS should be of use to those working with time-varying locations and data. We hope that the validation procedure being undertaken will provide clarity of its importance and use in the future.

4 Chapter 4: Data Resource Profile: The Biobank Historical Geospatial Information System (BIO-HGIS)

The UK Biobank is a prospective cohort study of UK adults aged 40-69 at time of recruitment[12]. It contains extensive later life information on the health and well-being of its participants, most of whom have been genotyped. Many external sources of data have already been linked to the UK Biobank, such as the Hospital Episode Statistics[68], allowing for the UK Biobanks' potential to grow over time. However, the UK Biobank itself has collected relatively little information on individuals' early life circumstances.

A solution to this is to reconstruct individual exposures by using administrative level data. Whilst there is limited data on early life circumstances within the UK Biobank, crucially, the UK Biobank did collect the participants' place of birth as coordinates[69]. By utilising the birth coordinates, therefore, individuals within the UK Biobank can be mapped to regional areas of the UK during the 20th century. Using these regional Statistics, link via location of birth, to construct exposures allows the furthering of the UK Biobank's use case for even more research.

However, at present, there is a lack of extensively detailed administrative data from the 20th century at lower enough levels of geographic density which could be utilised to reconstruct early life exposures. Even when said data exists, due to its complexity such as due to issues of geographic areas changing over time[48], it can be challenging to use. This is heightened by the fact that research focused on the UK Biobank tends to be health related, with such researchers not necessarily having a Geographic Information System (GIS) background.

Within this chapter, I present the accumulation of interdisciplinary work and methods to create a new database, called the Biobank Historical Geographic Information System (BIO-HGIS). At the start of this research project, the scope was to just link current administrative data from the past that exists, or undertake small digitisation projects, to the UK Biobank; hence the databases current name. Since then, the project has grown drastically, with large new detailed datasets being digitised. A simple data return to the UK Biobank, therefore, would limit the benefits of the current research that has been undertaken.

With so much new data, many other cohort studies with cohort data after 1931 such as the 1946 National Survey on Health and Development[8], the 1958 National Child Development study[9], or the 1970 British Birth Cohort[10], also stand to benefit from this new data. Furthermore, for many researchers less focused on health outcomes directly, detailed geospatial data is of extensive benefit.

Depositing the data in a data repository like the UK Data service also risks the data not sufficiently being used, despite its potential impact. Therefore, a decision was made to turn a simple data return into a new database project focused on the digitisation, protection, and provision of historical statistics at geospatial administrative locations. This involves the construction of a new website, which will allow for the vast amounts of data to be visually explored for more general members of the public or media. It will also allow the data to be downloaded.

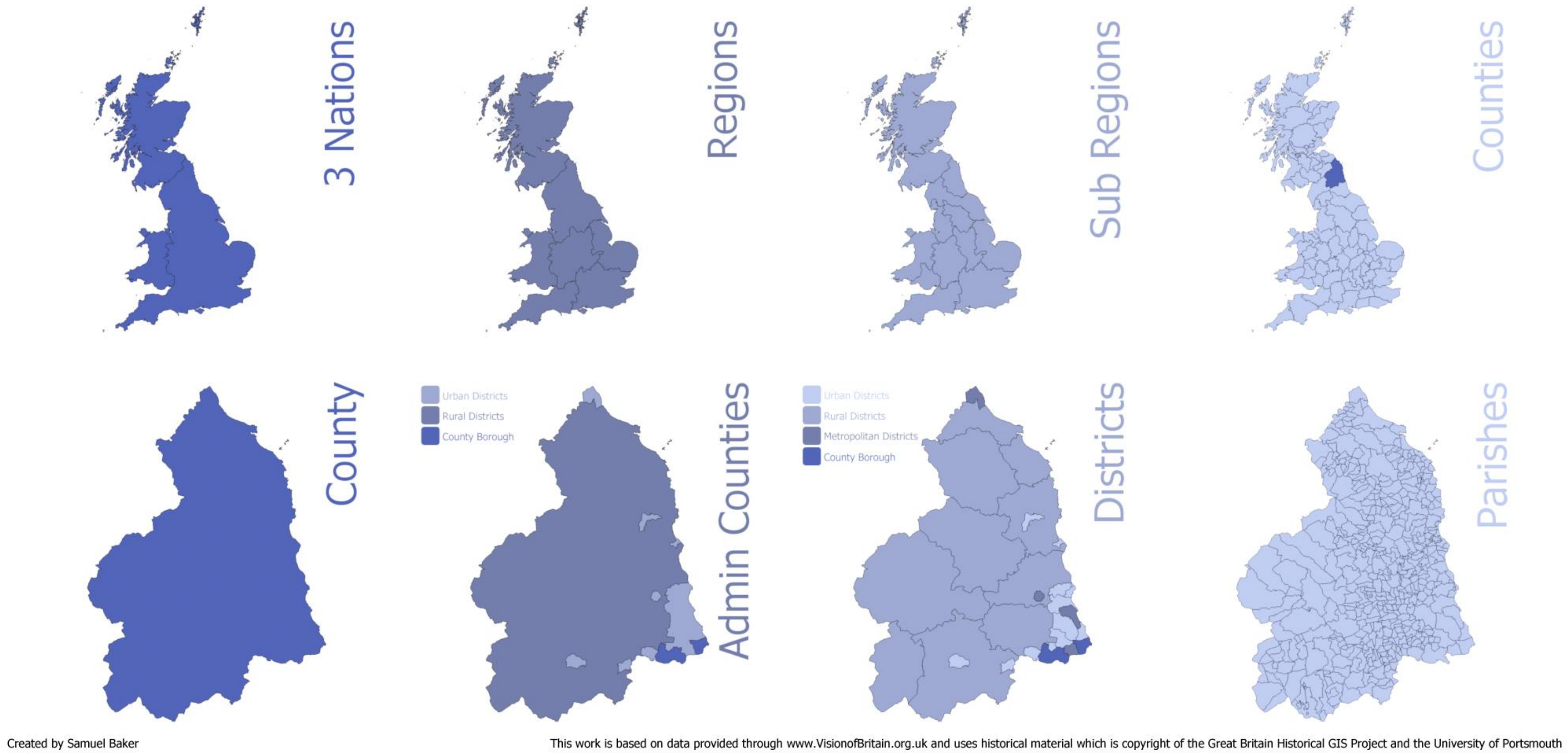
At present, the construction of this new website is still underway. However, once finalised, the aim is to release the data within this paper in stages. Once a chapter involving the data has been accepted for publication, the aim will be to ensure that links within that paper link to the source data on the website. This should aid replication studies, but also ensure that there is no lag between interest in the data source and the ability to use it. Over time, it should also expose different audiences to more and more data, as links will take people to a larger database than at the point of publication. It is hoped that papers within this thesis will start being published over the course of 2023, and with that, so too will the data start be being made available.

4.1 Geographical levels within the UK

Before considering the data within BIO-HGIS, an understanding of UK geography is required in understanding the structure of BIO-HGIS, and the level of detail within the data sources. This next section briefly summarises how the UK was structured from 1931 to 1974. The United Kingdom (UK) had multiple regional levels from 1931 to 1974. First, the UK comprises the four countries of England, Wales, Scotland, and Northern Ireland. England and Wales have the following sub-divisions, from the most aggregated to lowest level of detail: regions, subregions, counties, administrative counties, districts, and parishes; shown in Figure 4-1.

Regions and sub-regions represent broad areas, such as the North or Northwest, that represent large historical regions that subdivide the nations within the UK. The sub-regions have a similar structure and use as the current government office regions. Counties are like regions but are more detailed, with 64 counties as of 1951. Many of these county's stem from their original classification in the Middle Ages Domesday record from 1086[70]. However, others have changed since, with Yorkshire becoming divided and areas of the North aggregated.

Administrative counties divide counties further into rural and urban areas in addition to the county boroughs. Administrative counties are mostly a division for statistical reporting as opposed to a historical measure. County Boroughs are a form of a district, which could act independently of county controls for matters such as sanitation and health care[71], hence their separate inclusion within administrative counties. As for districts, there were within England and Wales four core types, with some additional London specific classifiers, shown in Table 4-1. Broadly, higher-order districts represent increasing levels of density of development, with there being approximately 1870 districts in Great Britain and 1472 districts in England and Wales.



Created by Samuel Baker

This work is based on data provided through www.VisionofBritain.org.uk and uses historical material which is copyright of the Great Britain Historical GIS Project and the University of Portsmouth

Figure 4-1: A visual break down of regions within Great Britain during 1931-1971 using shapefiles from Vision of Britain. The second row of images represents the highlighted county in the upper right of the first row.[44]

Table 4-1: Explanation of what each district constitutes from Vision of britain[71].

District Break Down	Code	Description
Rural District	RD	Contains market-towns of differing size
Urban District	UD	Containing small towns
Municipal Borough	MB	Towns which do not possess the more dignified title of city
County Borough	CB	Towns able to free themselves from county control
London specific	Various	Various London types exist

Parish level data is significantly denser, allowing for extensive variation, with nearly 14,000 parishes within Great Britain. The parishes are strongly associated with the church[72], with the church's original aim that every settlement, no matter its size, should have a church with a priest[72]. However, very little information was recorded for parishes outside of census years.

Most of the data in the BIO-HGIS is at a district level and focused on England and Wales. The focus on England and Wales is in part due to the devolved nature within the UK[73]. Devolution meant records could be recorded separately for different nations, especially Northern Ireland and Scotland, which would lead to additional costs and difficulty in digitisation. Conversely, England and Wales's records were reported together from 1931 to 1974, so easy to collect.

To allow linkage to the UK Biobank, we used all geographic layers in Figure 4-1 as shapefiles[67] to geo-locate everyone within the UK Biobank. To do so, we use the easting and northing coordinates of birth provided by the UK biobank, and then placed each coordinate within a given geographic layer. This means that each individual in the UK Biobank can link to any data from any of the regional layers within Figure 4-1. We aim to provide this linkage file to the UK Biobank, so anyone can link regional data, from BIO-HGIS or otherwise, to Biobank participants based on their birth coordinate.

4.2 Data Waves

BIO-HGIS has been constructed over many years, with each wave of digitisation focused on specific outcomes and continues to grow. Each individual data collection wave is summarised in Table 4-2. The following sections detail the content of each wave further, provide a context for its collection, and its present or future intended use.

4.3 Wave 1: Supporting data for the Great British Historical Database on Health and Health Care

The first wave focused on finalising data from the Great British Historical Database on Health and Health Care (GBHDHHC)[74]. The GBHDHHC contains the majority of Table 17 within the Registrar General's Statistical Review of England and Wales. Table 17 contained the estimated population, births by legitimacy and sex, deaths by sex, and various measures of infant mortality for each district annually. However, whilst the efforts undertaken by GBHDHHC were extensive, spanning most years from 1930 to 1974, it lacked five years of data from 1958 to 1962.

Whilst district population may not be a crucial outcome variable itself, it is vital for converting the counts that many historical records report to rates. Therefore, wave one focused on digitising the remaining five years of data that were missing with help of ABBYY FINEREADER 14. This process, despite its automation, still led to considerable manual time costs. Given larger datasets were to be processed, this experience inspired the creation of a custom software for digitisation of ArchiveOCR, as shown in Chapter 2.

Table 4-2: Data collection waves for the BIO-HGIS

Wave	Data Source	Variables	Dates	Time Dimension	Spatial dimension	Description
1	Registrar General's Statistical Review of England and Wales and the Great British Historical Database on Health and Health Care (GBHDHHC)	Estimated population, Births by legitimacy and sex, Deaths by sex, infant mortality	1930-1974	Annual	Districts	Mostly Quality controlling the Great British Historical Database on Health and Health Care (GBHDHHC), but also digitised the five years of data that was missing

2	Register Generals Weekly Return	Acute Meningitis, Acute Polio Non Paralytic, Acute Polio Paralytic, Diphtheria, Dysentery, Food Poisoning, Infective Jaundice, Measles, Pneumonia, Scarlet Fever, Tuberculosis Meninges and CNS, Tuberculosis Respiratory, Pertussis	1941-1974	Weekly	Districts	Notifiable diseases reported weekly for each district within England and Wales. Here we digitised the 40,000 tables these weekly reports between 1941-1974. Not all Disease remain notifiable across the whole date range.
3	War, State, and Society	Air Raid count, Deaths from Air raids, Injured from Air Raids, Expected Causality from Air Raids	1939-1945	Daily	Districts	Each of the 32,000 air raids were transcribed by the War, State, and Society research group. Here, we geo-locate each of these locations into a district.

4	Vision of Britain, NOMIS, CASWEB	Numerous	1931, 1951, 1961, 1971	Annual	Districts	The data collected through the census frequently changes, but this makes comparisons difficult. Here we standardised several of the variables, such as employment by age groups, and linked all of the possible variables for standardisation at a later date
5	Registrar General's Statistical Review of England and Wales	Mortality by age and sex in roughly 5-year bins	1947-1972	Monthly	England and Wales	Mortality data by rough age groupings by sex and month of death

5	Registrar General’s Statistical Review of England and Wales	Scarlatina and rheumatic deaths	1848-1901	Annual	England and Wales	19th century mortality data relating to Streptococcus pyogenes
5	Registrar General’s Statistical Review of England and Wales	Population for England and Wales, Scotland, Ireland, and Northern Ireland	1861-2018	Annual	National	Individual population counts, opposed to UK totals that are more readily available.
6	Hansard: Direct Grant Schools Vol 738	Grant value, Pupil Count, Pupil Teacher Ratio	1966	Annual	Parishes	The amount of money granted by the state for each individual grammar school, with number of pupils and that ratio to teachers

7	Labour Gazette	Unemployment of Males, Females, and under 18s	1945-1971	Monthly	Subset of Districts	Unemployment in the largest towns, cities, and urban areas were recorded monthly in the Labour Gazette.
7	Labour Gazette	Unemployment of Males, Females, and under 18s by Sector	1947-1971	Monthly	Great Britain, UK	Unemployment by Census groupings of professions, such as Mining, with some select individual professions also reported within each category, such as Coal Mining

Upon completing this process, it was attempted to merge the data into the UK Biobank. The extent of the modifiable area unit problem[18] had been underestimated, and the process of manual merging was time-consuming. As many districts were abolished or changed drastically in terms of boundary, a considerable number of locations could only be roughly merged, leading to drastic changes in variables within certain districts over time. This experience led to some useable data but showed a better method was required. This challenge would later lead to the creation of WeightGIS, as shown in Chapter 3, to handle the output of Wave 2 and to quality control all data waves, including Wave 1.

4.3.1 'Beyond' Barker: Infant mortality at Birth and Ischemic Heart Disease in Older Age

After the data was quality controlled by WeightGIS, we attempted a replication of David Barkers 1986 paper[2]. *This work was undertaken jointly with Stephanie von Hinke, Hans van Kippersluis, and Pietro Biroli.* Barker showed that regional infant mortality was associated with an increased risk of later-life heart disease but only used 212 local authorities' data on infant mortality as of 1921[2]. Here we worked on replicating the principle of Bakers' work but using the time-varying accounts for infant mortality across 1472 districts between 1934 and 1971[56]. We linked these data sources to the UK Biobank and constructed an early life environment with the infant mortality in the district and year of birth for each participant (N = 378,873).

The first crucial difference to Barker's paper was that, by utilising the UK Biobank, we could use a regional level construction of early life environments but *individually* measured outcomes of ischemic heart disease. Barker's paper, in comparison, had a regional measure for both the early life environment and outcome. We further used the genotyped participants within the UK Biobank to construct a polygenic score for each participant to investigate the potential of the gene-environment interplay between early life environment and ischemic heart disease. In doing so, we sought to investigate if genetic susceptibility can aggregate adverse early life circumstances. We then undertook further sensitivity analyses with a sibling sub-sample (N=33,069).

We found several key findings. Our first was from a direct replication of Barker, using a regression of a binary measure of ischemic heart disease on infant mortality rates in the year and district of birth. We found a strong association between the infant mortality rates, proxying for adverse early life circumstances, and later life ischemic heart disease. Even adding both the polygenic score for ischemic heart disease and allowing for gene-environmental interplay did not notably change the result. A one standard deviation increase in the infant mortality rate in the year and district of birth increased the probability of ischemic heart disease by 1.1 percentage points, but was stronger for those with a high polygenic score.

We then utilised the district fixed effects, which showed that over half of this association between ischemic heart disease and early life environment was capturing time-invariant differences between districts. This means that infant mortality rates are likely to capture far more than just individual level nutritional deficiencies that Barker suggested. Whilst our results were robust but attenuated for district fixed effects, the same was not true for family fixed effects, suggesting that infant mortality rates do not capture the within family variation. As we found a non-negligible gene-environment interaction, we also found evidence that even those at high genetic risk of later life diseases, such as ischemic heart diseases, can have this risk mitigated with interventions to their environments. Improving early life circumstances could reduce the variation in later life ischemic heart disease stemming from genetic risk.

4.4 Wave 2: Weekly notifiable diseases from 1941 to 1973

To survive childhood was historically difficult but doing so unscathed was considerably harder. The poliovirus could leave its victim paralysed or otherwise disfigured[75], streptococcus permanently damages the heart or limbs through rheumatism[76], or measles reduce the immune system antibody repertoire leaving individuals vulnerable to subsequent infection and mortality.[77]. Whilst many of these diseases are no longer prevalent, scarring from early exposures remains a risk even within the 21st century. Those infected with malaria continue to have a persistent risk of haemoglobinuria, jaundice and anemia[78,79], and those with pneumonia have an increased risk of chronic lung disease[80].

Infections can also lead to declines in mental health, well-being, or cognition through the impairment of brain functionality[81,82]. Early life infections during critical periods of brain development are associated with increased aggressive behaviour and subsequent violent criminal behaviour[82]. Multiple psychiatric disorders, from obsessive-compulsive disorders to depression, have also been similarly associated[83], although the evidence is still often poorly established[83]. One of the well-established pathways is because of inflammatory responses in the brain[84], with neuro-inflammation linked to multiple forms of paediatric autoimmune neuropsychiatric disorders and Sydenham chorea[85]. Longer-term consequences of difficult to recognise symptoms have also been described, such as declines in later life cognitive health[86].

Those affected in utero by the 1918 influenza pandemic reported declined socio-economic outcomes, such as educational attainment, lower income, and lower socio-economic status[87]. Reductions in mental health in childhood, which could continue into midlife for males, has also been further associated with exposure in utero to the 1918 influenza pandemic[88]. Each additional infection-related hospitalisation in Finland was associated with lower log earnings, fewer years in employment, and a higher likelihood of requiring social welfare[89]. Whilst infections are crucial, simply being in poor health, be it from birth weight, nutrition, or otherwise, have similar outcomes[90,91].

Vaccines[92,93,77], penicillin[94,95], and general improvements to living conditions[96] reduced childhood mortality throughout the 20th century. These innovations led to the leading cause of death in western nations to change from contagious diseases to the non-transmissible[94], such as cardiovascular disease. Despite this success, many infections once thought to be mostly banished have begun to return. Scarlet fever[97], pertussis[98], and measles[99] have all risen to a level that would have been unprecedented a decade ago.

Whilst still far from their peaks within the 19th century or before, this rise is concerning. Many of these diseases still have unknown quantities, and the research body continues to discover the increasing burden these diseases placed on survivors. For example, streptococcus pyogenes, the bacteria behind scarlet fever, may in some cases cause inflammation of the brain, leading to severe behavioural regression[85]. However, the return of these diseases occurred relatively recently, so any renewed investigation into the potential consequences of early life infections of more historical childhood diseases must utilise data from the past.

Limited data are available historically for diseases. However, research methods have adapted to meet this challenge through the utilisation of exogenous historical shocks in an environment[1]. For example, one of the most comprehensive examples of this methodology was for those affected in utero by the 1918 influenza pandemic, which resulted in reductions in their later life socio-economic outcomes[87]. Within the UK, extracts from the Registrar General's Weekly Return were used to construct indices of the Asian flu outbreak in 1957[100]. These indices were then utilised in combination with the National Childhood Development study and found that mean test scores, at both ages 7 and 11, were reduced[100].

Exogenous shocks alone cannot answer every question, but as most of the extensive historical records on disease notifications have not been digitised, there is little alternative at present. To alleviate this, we digitised the Registrar General's Weekly Return for notifiable diseases across the 1472 districts of England and Wales[42] between 1941 and 1973. If a notifiable disease was found by a general practitioner, it was required that these instances be reported to the central state. The total of these in each district is reported each week in the Registrar General's Weekly Return.

Digitisation of the main body of the table resulted in data from 13 notifiable diseases. We started in 1941 as the tables within the Registrar General's Weekly Return changed drastically before 1941, which would have added considerable time costs to the project. Furthermore, both measles and pertussis, which are reported consistently from 1941 to 1973, were also only notifiable from the week starting from the 4th November 1939.

Table 4-3: Diseases by dates of availability within BIO-HGIS

Disease	Start Date	End Date
Acute Meningitis	01/01/1970	31/12/1973
Acute Polio Non-Paralytic	08/01/1955	21/09/1968
Acute Polio Paralytic	08/01/1955	21/09/1968
Diphtheria	01/01/1941	08/01/1955
Dysentery	01/01/1970	31/12/1973
Food Poisoning	01/01/1970	31/12/1973
Infective Jaundice	01/01/1970	31/12/1973
Measles	01/01/1941	31/12/1973
Pertussis	01/01/1941	31/12/1973
Pneumonia	01/01/1941	08/01/1955
Scarlet Fever	01/01/1941	31/12/1973
Tuberculosis Meninges and CNS	01/01/1970	31/12/1973
Tuberculosis Respiratory	08/01/1955	31/12/1973

We only digitised the main body of the table, which was purely numeric, rather than the alphanumeric overflow column for less notifiable diseases. For example, diphtheria used to belong to the main body of the table because of its prevalence up to 1955, but thereafter it was simply placed in the overflow. As such, information for diphtheria at present is limited to 1941 to 1955. Other diseases become notifiable after the start of the sample, such as tuberculosis, which means there is no information for tuberculosis before 1955. The full list of diseases by dates available is shown in Table 4-3. A future wave will seek to digitise the additional years of data and the less common notifiable diseases.

Figure 4-2 shows the digitised weekly totals from 4 of the 13 notifiable diseases in their current form. Unlike the current annual totals available from the Office of National Statistics, this offers 52 times more detail. To our knowledge, this data source is currently the most detailed version of notifications in the UK, and the world, for the 20th century.

However, the true strength of the weekly notifications is that each week has 1472 data points representing the districts allowing for both within year and geospatial variance. Targeted efforts to digitise parts of the weekly records have been undertaken before[101], but to our knowledge, our version is the most complete version to date. The aggregation of the notifications of pertussis, measles, scarlet fever, and pneumonia within 1472 districts in England and Wales from September 1945, converted into rates per 100,000 population, is shown in Figure 4-6.

Given the extensive possibilities for this data source and 18 months of work to construct it, many research papers involving the diseases have been started or planned. The remaining parts of this section cover these papers and, if sufficiently developed, will also link to a relevant chapter within the thesis. These sections cover a brief background of each disease. However, only diseases with at least five years of data have been investigated, so any disease that was added after 1970 has been discounted. In some cases, papers utilised multiple disease notifications, but otherwise, the following sections follow the same order as Table 4-3.

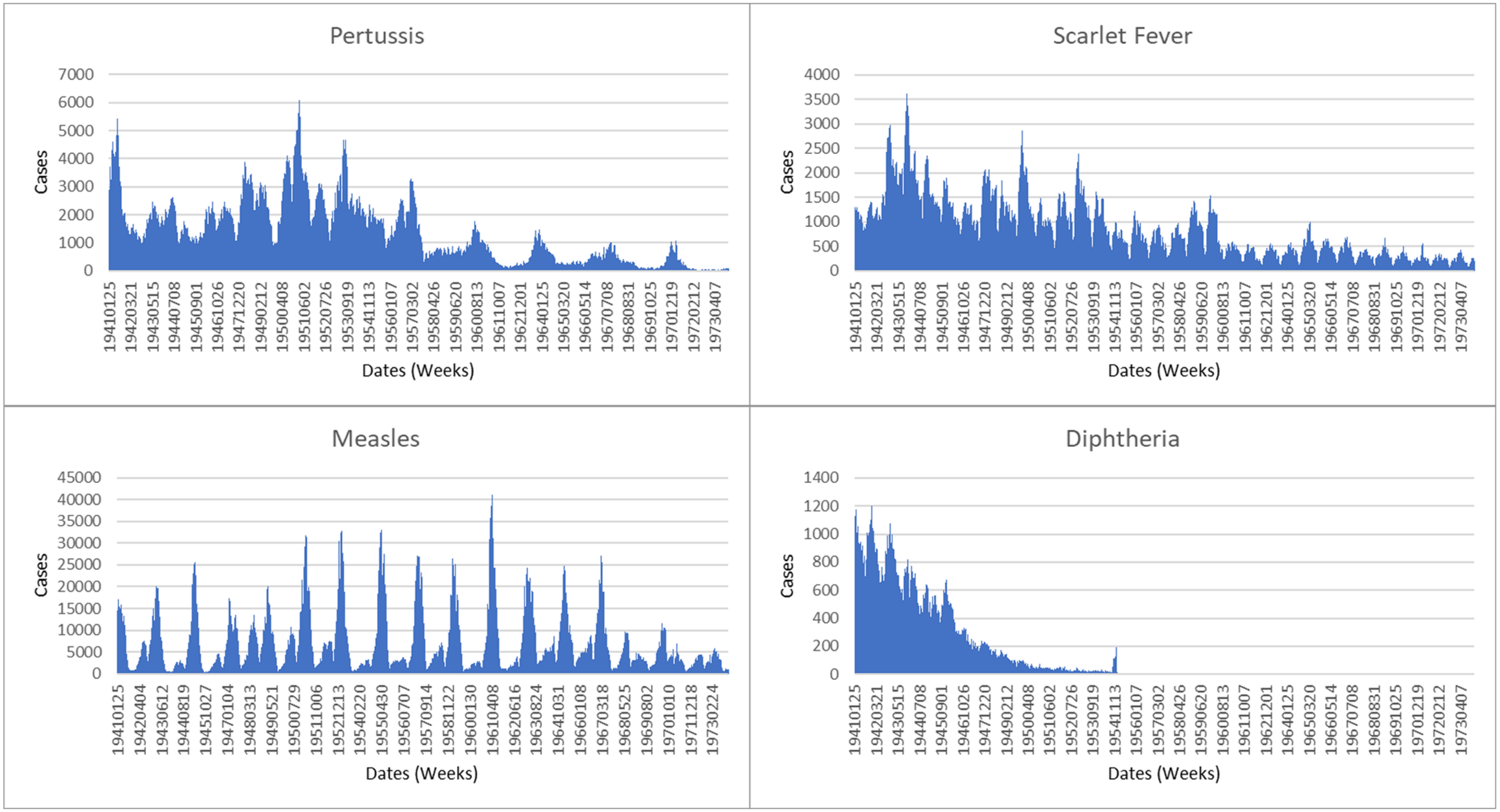


Figure 4-2: The number of cases of each of the 13 notifiable diseases per week

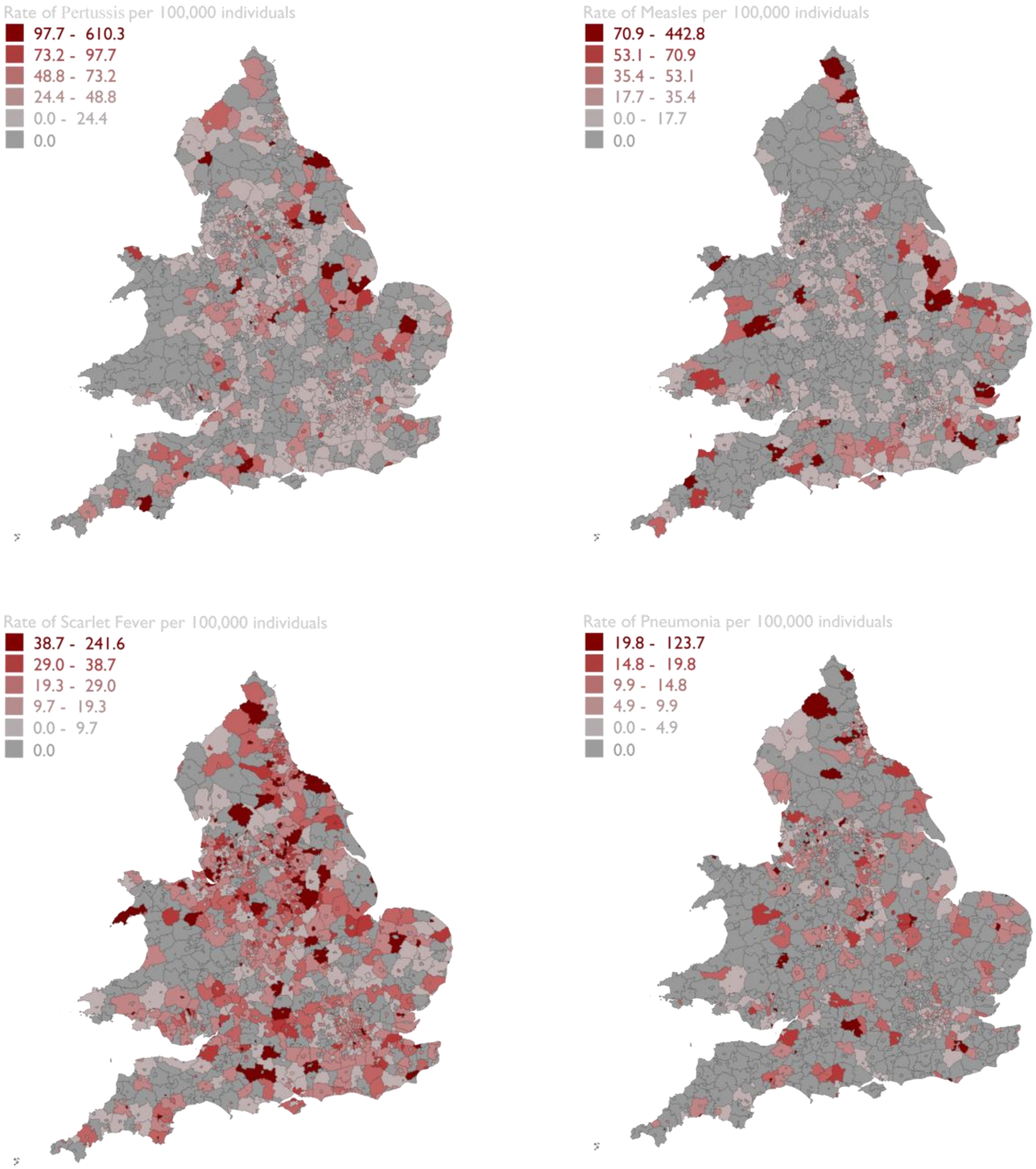


Figure 4-3: The aggregation of the notifications of pertussis, measles, scarlet fever, and pneumonia within 1472 districts in England and Wales from September 1945, converted into rates per 100,000 population

4.4.1 Acute Poliomyelitis

Poliomyelitis (polio) led to deformation and death for thousands of years[102] despite 90-95% of those infected being completely asymptomatic[103,102]. Another 4-8% would experience abortive poliomyelitis, where polio did not enter the central nervous system (CNS) and was a mild diseases[103,102]. Non-paralytic polio was a worse version of abortive poliomyelitis and less common, which further induced fever and pains in the neck and led to muscle weakness[103,75], but was rare. The most serious, yet most commonly known, outcome of paralytic polio, could result to irreversible limb paralysis from the disease entering the CNS[75], occurred in less than 1% of patients. Polio was considered endemic up to the 19th century, with paralytic outbreaks only surging from the unsanitary and cramped conditions of the era of industrialisation[102]. The polio vaccine introduced in 1950 reduced paralytic polio cases in the United States from 58,000 to 5600 within a year and is considered one of science's definitive success stories[102].

However, non-paralytic cases are thought to have been significantly under-reported[75]. Despite an expected ratio of 1:10:50 in paralytic, non-paralytic, and abortive cases respectively[75], the UK reported 13,491 paralytic and only 9097 non-paralytic cases during 1956-1968[42]. Abortive poliomyelitis, given its mild nature, is not a notifiable disease in the UK, so cases of abortive polio are unknown. Studies show conflicting evidence as to the prevalence and consequences of non-paralytic polio[104,105,75]. The main concern is if non-paralytic polio, considered fully recoverable, can itself lead to prolonged muscle damage even if less than full paralysis[75].

To estimate if non-paralytic polio cases are under-reported, we intend to use the actual observed cases from BIO-HGIS to investigate if exposure to non-paralytic polio was associated with increased later life muscle disorders, BMI, white matter scarring in the brain, and reduced bone density using the UK Biobank. We seek to investigate muscle disorders across the body. Muscle disorders affecting the lower limbs are the most known and widely established for polio[106]. However, instances of carpal tunnel syndrome in polio survivors have shown it is not exclusive to the lower limbs^[107]^. We seek to investigate scarring on the brain, as 92% of polio survivors suffered white matter scarring to only 1-2% in controls, which can lead to long-term fatigue[75]. Finally, we seek to investigate BMI and bone density as polio both increases the risk of osteoporosis[108], and as potential muscle pain leads to a loss in mobility, and higher BMI[109].

We will then further seek to utilise both the non-paralytic and paralytic cases, which, given their severity, should not suffer from under-reporting, as sensitivity analysis. The mortality associated with paralytic means it is unlikely that the UK Biobank will have sampled paralytic polio survivors. Therefore, given the known ratio of paralytic to non-paralytic cases, if combining the two notifications increases the predictive power, this estimates a degree of under-reporting within non-paralytic polio.

4.4.2 Diphtheria

Diphtheria has existed since ancient times[110] and was a frequent cause of death, with 5-10% of those affected not surviving[111].In more severe cases which required hospital treatment, the mortality rate could reach as high as 50% in the 19th century[112]. Whilst sometimes tracheotomy was attempted to save the patient, the operation itself was dangerous, with an 80-90% mortality rate in the 19th century[112].

The introduction of the vaccine in the 1940s led to cases of diphtheria dramatically declining[110]. However, the UK tended to have higher hesitancy and backlash against vaccination, which started with the first vaccine against small pox[113]. This backlash was higher than many other western nations, with lower levels of education and apathy being central issues[113]. Only 28% of those who left school before 14 were vaccinated against smallpox, compared to 61% of those leaving school after 15[113].

The decline in cases of diphtheria after the introduction of the vaccine led to diphtheria becoming a less known disease and apathy, with newer parents no longer concerned about the risks of diphtheria[113]. The social class structure continued to play a significant factor, with only 60% of those leaving education before 14 immunised, compared to over 85% of those leaving after 15[113]. Said apathy was also spatially clustered, with only 7 of the 26 areas needing improvements to vaccine uptake outside the North or the Midlands.[113].

Targeted measures such as vaccinations at school improved the diphtheria vaccination uptake[113], with targeted measures having a successful history in tackling vaccine apathy. The hookworm eradication campaign in 1910-1915 both reduced hook worms' instance and prevalence in the American south[114]. Similarly, targeted intervention and vaccination against tuberculosis in schools in Norway had a similar effect[115]. In both instances, the reduction in cases also led to individuals obtaining increased educational attainment[114,115]. These gains were higher for those born in areas that originally suffered higher rates of the disease[114,115].

In this paper, we seek to investigate the changes in weekly diphtheria rates by location and district type. In doing so, we may better understand how hesitancy existed and was eased in the past, which may assist with handling similar issues currently experienced around COVID-19. We hypothesise that areas with greater autonomy and higher initial cases would experience greater percentage reductions than their rural or urban centrally run counterparts. We can then further investigate if, similar to other known examples in the literature, individuals who experienced greater percentage reductions in cases attained higher levels of educational attainment.

4.4.3 Scarlet Fever

Records from the 17th and 18th centuries showed that scarlet fever cases were often benign[116]. However, scarlet fever transitioned from a benign illness to one of the most common causes of early life mortality within just a few years[116]. The scarlet fever case-fatality rate rose to 15% in 1834 and often exceeded 30% by the mid-19th century[116].

Scarlet fever is caused by streptococcus pyogenes (S. pyogenes)[117], and is a possible outcome of untreated prior infection of S. pyogenes, such as strep pharyngitis[118]. Strep pharyngitis represents 20-40% of all pharyngitis cases[119] and one of the most common causes for visiting a general practitioner, even at present[120]. In an era before penicillin, there was not an effective treatment, which meant each case of pharyngitis posed a risk of developing scarlet fever.

A significant link to later life mortality was rheumatic heart disease[121], where the heart valves are permanently damaged[76]. Rheumatic heart disease is a consequence of acute rheumatic fever, traditionally starting two weeks after an untreated infection of S. pyogenes[76] and avoided if treatment occurs within nine days[118]. The risk of developing rheumatic heart disease after an untreated infection was high. A case study in Austria found 35% had developed rheumatic heart disease a year after infection, which after ten years increased to 61% ten[76].

The case mortality rate for scarlet fever dramatically declined to close to 1% at the end of the 19th century within just a few years, at a rate not dissimilar to its original rise[116]. The theory behind the sharp rise and fall of scarlet fever cases is that older strains were more virulent but less transmissible. The industrialisation in the 19th century, and subsequent increases in density, allowed the higher virulent strains of scarlet fever to spread rapidly[116]. Over time, reduced overcrowding, improved sanitation, medical standards, and rising herd immunity reduced the case mortality rate[116,96]. As scarlet fever can be prevented by treating preceding infections[118], the introduction of penicillin further reduced the risk of developing scarlet fever later in the 20th century.

The latent risk of heart disease from early life infections of streptococcus declined with cases across the 20th century[121]. However, scarlet fevers' decline slowly reversed after the 1980s[122], despite further improvements to health and socio-economic circumstances across the 20th and 21st century. Scarlet fever cases have since risen further across Asia[123], Europe[124], and the UK[97] in the 21st century. Studies have since found multiple additional links to latent effects from S. pyogenes infections[125,126,127,85]. With cases rising and potential latent conditions still under-investigated, this recent surge could lead to considerable consequences for this generation.

In Chapter 5, we investigate if there were broader risks to heart disease, later-life cognition, and educational attainment outside the most established link to rheumatic heart disease[76,119] from exposure to scarlet fever using the UK Biobank. We constructed scarlet fever exposure using rates of scarlet fever in an individual's place of birth from data within BIO-HGIS. We found that increased exposure to scarlet fever across childhood was associated with increased risk of declines in later latent cognition and increased risk of heart disease.

4.4.4 Pertussis

Pertussis is highly infectious, with an expected mortality rate of around 10% before the introduction of the vaccine in the 1940s[128]. The introduction of the whole-cell pertussis vaccine led to a 157-fold reduction of pertussis cases between 1940 and 1973 within the United States[129]. However, pertussis cases, unlike vaccinations of measles, did not change their pattern or frequency of infection cycle[129]. Immunity from the pertussis vaccine wanes from as early as four years old[130,131,129]. Waning immunity ensures there are always susceptible individuals, but far less than before the vaccine[129]. Importantly, for vaccinated individuals who ultimately end up infected, the illness is significantly less severe than in unvaccinated children[129].

Pertussis cases have seen a recent resurgence, but the true nature of this is complex[129]. Pertussis can be challenging to diagnose[132]. Only 5-25% of pertussis cases were estimated to have been reported in England and Wales as of the mid-20th century[133]. There is not a known recent estimate for pertussis reporting in the UK. However, it is estimated that only 3-12% of pertussis cases in the United States were reported as of the end of the 20th century[134]. A similar study investigating Germany as of 2015 still estimated that only 54 to 61% of cases were being identified. Therefore, despite a resurgence in cases, a significant part of this resurgence has stemmed from improvements in reporting pertussis from new tools to assist diagnosis[129].

As pertussis and many other childhood infections declined over the 20th century, asthma cases grew rapidly. Asthma increased from just 4% of 9-12-year-olds in 1964 to nearly 30% in 2004[135]. Asthma instance fell to 18.6% in 2014[135], which was correlated with the rise in cases of diseases that declined in the 20th century rising again, such as scarlet fever[97,124], pertussis[136] and measles[137]. However, the literature has conflicting evidence for the impact of early life infections on later life asthma[138,139,140,141,142,143,144,145,140,146,147,148,149,150], which in part comes from the extensive changes to early life environments across the 20th century.

In Chapter 6, we aim to quantify the potential impacts of early life infections on later life instances of asthma using rates of scarlet fever, pertussis, and measles from BIO-HGIS. We constructed exposures for individuals from the UK Biobanks based on the participants' year, month, and place of birth. We found little evidence of any associations between exposures to scarlet fever, pertussis, or measles up to the age of 10 on later life asthma instance. However, this principal hypothesis, and the literature at present, almost completely ignores genetic liability despite asthma's high heritability[28].

Based on a plausible biological mechanisms[151], we hypothesised that diseases may have moderated asthma incidence for the genetically susceptible from a gene-environment interaction. This biological mechanism would mean that only those who were genetically predisposed to asthma could potentially have a protective association from infection, which may in part explain the conflicting evidence in the literature. We found evidence for both scarlet fever and pertussis, that increased exposure up the age of ten had a protective association against asthma, but conditional to those of heightened genetic susceptibility. However, we continued to find little evidence of a protective association to later life asthma from exposure to measles.

4.4.5 Measles

Measles is one of the most infectious diseases, with over 90% of children infected when measles was endemic[152]. However, unlike many diseases, measles mortality was often through secondary infections[153], with the most common secondary infection being pneumonia[154], which in part is because of measles immunosuppressive qualities[77]. Immunosuppression inhibits the host's immune system [155], making them susceptible to follow-up infections[77]. As measles immunosuppression specifically reduced hosts' antibody diversity, with a loss of nearly 20-50%[77], hosts were highly vulnerable after infections.

The measles vaccine reduced childhood mortality, especially in impoverished nations, ranging from 30 to 90 percent[92,93,77]. The impact of the MMR vaccine introduced later in 1988 has been estimated to have saved over 20 million lives worldwide from the prevention of measles alone between 2000-2015[156]. However, given most deaths from measles were because of secondary infections[157], this could well be a considerable underestimate.

Measles historic decline has since been reversed, with measles cases now 300% larger than in 2018, infecting over seven million children and causing 100,000 preventable deaths directly annually[158]. This has occurred even in western nations such as the UK[99], which has no longer been able to maintain its measles free status. Despite extensive efforts attempting to quantify the impact of immunosuppression on secondary infections[77,155], limited access to disease data has limited advances in this field of research.

In this paper, we seek to attempt to estimate how measles outbreaks led to subsequent increases in pneumonia, scarlet fever, pertussis, and respiratory tuberculosis using data collected within BIO-HGIS. We aim to construct 30 months of prior measles rates from the weekly reports and investigate the association between increases in measles in the prior 30 months and increases to the disease phenotypes.

4.4.6 Tuberculosis

Multiple forms of tuberculosis have affected humans. At the start of the 19th century, one in four deaths were from respiratory tuberculosis[159]. Bovine tuberculosis originated from cows due to milk consumption and was also a deadly killer before pasteurisation[160]. A post-mortem of 1420 children under 12 in 1880 Britain found that 30% had died from bovine tuberculosis[160]. The Bacille Calmette-Guérin (BCG) vaccine would eventually be created from attenuation of bovine tuberculosis[161], first introduced in 1921, and is one of the most administered vaccines in humans history[161].

It was hoped that respiratory tuberculosis (tuberculosis hereafter) could be banished to history[162], but these hopes were quashed in the 1980s and 1990s with resurgent rates of tuberculosis[162]. Apathy towards the vaccine and higher migration were key for the resurgence in tuberculosis cases[162]. Cases would rise year-on-year in the UK between 1980 and 2012 until it declined once more[162].

Unfortunately, cases in the developing world remain elevated, with World Health Organisation targets to eradicate the disease consistently missed[163]. There are currently nearly 9 million new cases of tuberculosis each year, with half expected to develop a pulmonary dysfunction as a result[164]. The damage caused by tuberculosis is highly heterogeneous, with increasing evidence that genetic susceptibility to tuberculosis may be the cause[164].

We seek to construct exposures to tuberculosis up to the age of ten using the exposure data from the BIO-HGIS and link it to UK Biobank participants using their place of birth. Here, we seek to explore two core questions. First, we will attempt to quantify the damage caused by tuberculosis to lung functionality. We will utilise the forced expiratory volume and forced vital capacity as our dependent phenotypes for measures of lung functionality. Then we will investigate if lung functionality damage from exposure is greater for those with higher polygenic susceptibility to tuberculosis and chronic obstructive pulmonary disease.

4.4.7 Pneumonia

Pneumonia was branded 'captain of the men of death' by William Osler[165,166], one of the most cited clinicians and argued to be one of the fathers of modern medicine[167]. In the late 19th to early 20th century, its case mortality rate ranged as high as 30-40%[165]. Streptococcus pneumoniae is a common cause of pneumonia which, despite its high burden, is not usually highly contagious[168]. Whilst those living in cramped conditions, prisons or shelters are associated with increased risk, schools and general workplaces are not[168]. However, in those who are hospitalised, most would die within the first day or one week[168].

Multiple medical advances have reduced the burden. Early treatment with penicillin or serum reduced the mortality drastically, with now nearly 90% surviving if treated promptly[168]. Later advances of conjugate vaccines in the 1980s[169] have helped reduce the case frequency of pneumonia in children by about 90%, relative to the early 20th century[168]. Despite the advances, mortality remains between 5-10%[168], with pneumonia accounting for the greatest burden of all childhood morbidity[80], representing one in every five deaths of children worldwide[168].

For those who survive infection with pneumonia in early life, there is a considerable risk of permanent lung damage[80]. In studies looking at the long-term consequences of pneumonia, both restrictive and obstructive lung function were found to be elevated in those affected with pneumonia[80], in addition to reductions in educational attainment[100]. In one such study, records of babies from those born in Derbyshire were used to contact individuals[80,170]. The 13 men who had survived pneumonia within the first two years of life had significantly lowered forced expiratory volume and forced vital capacity[80,170].

We seek to replicate the principle of this study but on a wider scale. Here we seek to construct exposures to pneumonia up to the age of five from rates within BIO-HGIS from participants within the UK Biobank. We then seek to investigate the association of increased exposure on participants' forced expiratory volume, and forced vital capacity, collected as part of the spirometry analysis by the UK Biobank. Doing so will utilise the largest prospective cohort study to date to further investigate an instance of prolonged effects of early life infections from pneumonia on lung functionality.

4.5 Wave 3: The Blitz

In 2015 Europe celebrated that it had achieved 70 years of continued peacetime. Said peace, however, ignores the breakup of Yugoslavia in 1992, and 140,000 lives lost in the process[171]. It also ignores that peace within Europe has not reciprocated elsewhere, with the frequency of war not declining itself[172]. Those who survive conflict can face significant life altering ailments, some of which are visually apparent such as post-traumatic stress disorders[173]. However, the potential consequences are far greater, and may even have intergenerational effects, with evidence that preconception parental trauma resulting in methylation within Holocaust survivors could also be found within their children[174].

Investigating prolonged consequences for survivors of conflict is complicated, as you require data on exposures to conflict and a cohort study that samples them later in life. Most cohort studies in the UK are post-war, with one of the earliest being the 1946 National Survey on Health and Development[8]. However, the UK Biobank sampled individuals aged 40-69 from 2006 to 2010[12], which provides one of the largest cohort studies that includes those born before or during the Second World War within the UK. However, as limited information exists in early life for those within the UK Biobank, exposure to wartime conditions is not a measured statistic.

Within Wave 3, we sought to remedy this by merging information from War, State and Society[175] to the UK Biobank. The War, State and Society digitised the individual reports of each of just over 32,000 reports of the Blitz across the United Kingdom. The database contains the settlement or rough area of each air raid, and a casualty report, with the intensity reported in 12 hourly windows[175]. Given the generalisation of the UK Biobank birth coordinate to a 1 km grid coordinate[69], not where they were at that specific moment, a proximity exposure based on distance to epicentre is not possible. This is furthered by the fact that the epicentre is just a settlement name or locations name, not exactly where it hit.

Therefore, to reasonably link the data to the UK Biobank, we used these settlement names and locations provided by War, State and Society[175] to geo-locate them as easting and northing coordinate using the Google Maps API. The API returns multiple results for location. We used the first result, which is at the lowest level of detail, as the coordinate of the centroid of that air raid. Each point was then mapped to a district, so that each district from 1939 to 1945 had daily measures of Blitz intensity, both morning and night-time, for each of the Blitz variables. This worked well except for London, which was aggregated as 'London' in the original reports, so sadly is just reported as a single area, rather than having individual boroughs.

4.5.1 War intensity on stillbirth and spontaneous miscarriage

As the Second World War loomed, psychiatrists made dire predictions about the consequences of mental trauma from air raids[176]. These reports were incorrect, with only a few cases of 'bomb neuroses', mental illness resulting from experiencing air raids, per week[176]. However, there are concerns that less serious cases of mental illness went unreported as most medics were pre-occupied with physical ailments[176]. Coronary symptoms, superstition, fatalism, and even miscarriages increased[176,177], as did senility amongst the elderly[176].

Stress during pregnancy poses risks to stillbirth, potentially by as much as 42%[178]. A pregnant mother losing their parent during pregnancy was associated with small consequences such as lower APGAR scores and birth weight[179]. However, evidence of long-term consequences of maternal stress is more varied with both no association to long-term harm[179] and increased uptake of ADHD, depression, and anxiety medication from family ruptures[180].

Those who experienced adverse childhood experiences (ACE) were nearly twice as likely to have experienced a stillbirth[181]. Wartime conditions are not covered under ACE, which focuses on physical, emotional, and sexual abuse within a household[182]. However, experiencing one of the largest human conflicts of the 20th century may have similar consequences. We intend to use the Blitz data to create an intensity measure for UK Biobank participants. Then, we seek to investigate if those who experience a higher intensity of the Blitz were more likely to have a stillbirth or spontaneous miscarriage later in life.

4.5.2 Blitz notifiable disease modelling

On the 1st of September 1939, the largest human population movement in British history began[183]. Over 96 hours, nearly 1.47 million children, pregnant women, and the disabled were relocated away from urban areas feared to be targeted by the Blitz to the rural countryside[183]. This process led to serious epidemiological concerns that it would be a super-spreading event, but in reality, the initial evacuation led to a below odds risk of disease[183].

Much of the current focus has been on whether the relocation of children led to increases in cases of childhood diseases for those who left. However, less attention has been placed around the role of how the Blitz directly affected the number of cases in areas targeted for destruction by the Luftwaffe. In many shelters, scabies and impetigo spread freely, although infectious diseases such as scarlet fever were lower than feared[176]. Despite this, how air raids contributed to increases in infectious disease has not been extensively modelled. We seek to model how air raids within a given week and district affected notifiable diseases, using the weekly disease notifications from BIO-HGIS.

4.6 Wave 4: Census standardisation

A census offers extremely in-depth accounts of society and is well suited to analyse within a census year. There have been multiple efforts to digitise census data. Vision of Britain attempts to show how the UK changed from 1801 to modern day and has digitised multiple extracts of census data such as the employed by industry counts, population counts, and provide shapefiles for census years of the UK[67]. Vision of Britain has been vital too much of our work, especially the shapefiles, with the data on industry counts being the only known digital source for industry at this era in the UK.

The NOMIS site contains extensive census records for more modern years, however, have recently started focusing on the digitisation of older records, and of relevance, are their efforts to digitise the 1961 census[184]. The efforts in this case cover the whole census but are working on a particular census in which there is a considerable amount of missing data. Therefore, whilst the number of topics and variables available is much higher than Vision of Britain for census data, some of it is hard to use due to said missing-ness. Whilst NOMIS contains a considerable amount of census data, it does not contain information for 1971. This instead was downloaded from the UK Data Service Census data service of CASWEB[185]. This data is by far the cleanest but has no historical data prior to 1971.

Whilst each census data is of use in of itself, difficulties arise when combining multiple census years[18], such as changes to locations or questions[17,18]. Whilst efforts have been made to standardise locations over time, as shown in Chapter 3, if variables themselves do not exist in each census wave, then census data will remain difficult to utilise in a time varying context. Further efforts have begun on standardising these variables, so that they exist in each census year. However, the missing-ness of the data in 1961 is proving to be a bottleneck, and so presently, efforts have been limited to amenities, densities, unemployment, and those of working age by age groups. Work on this wave is ongoing.

4.7 Wave 5: The Registrar General's Statistical Review of England and Wales

Within the Registrar General's Statistical Review of England and Wales there is a wealth of information, such as the annual population counts shown in Wave 1. Many other tables exist that could offer similar value, but they are large which adds costs to digitisation. Wave 5 was designed as a scoping wave, digitising smaller tables or parts of larger tables, to attempt to justify future work. Whilst less valuable that the other resources, this work still adds to the database and helps to further protect the past.

4.7.1 Wave 5A: Measures of Mortality

Currently, within BIO-HGIS, there are district-level reports on infant mortality, but only at an annual level. If infant mortality had a seasonal pattern, then the use of annual data could be problematic if a study seeks to assign infant mortality rates as a proxy for the early life environment.

Two known sources of infant mortality data exist outside the current annual data from the Annual Statistical Review. Infant mortality was reported weekly for a subset of districts, and monthly mortality by age groups for the whole of Great Britain. Whilst the weekly infant mortality rates offer a higher time dimension, they are only reported for 80~ districts, which limits the spatial dimension compared to the annual reports.

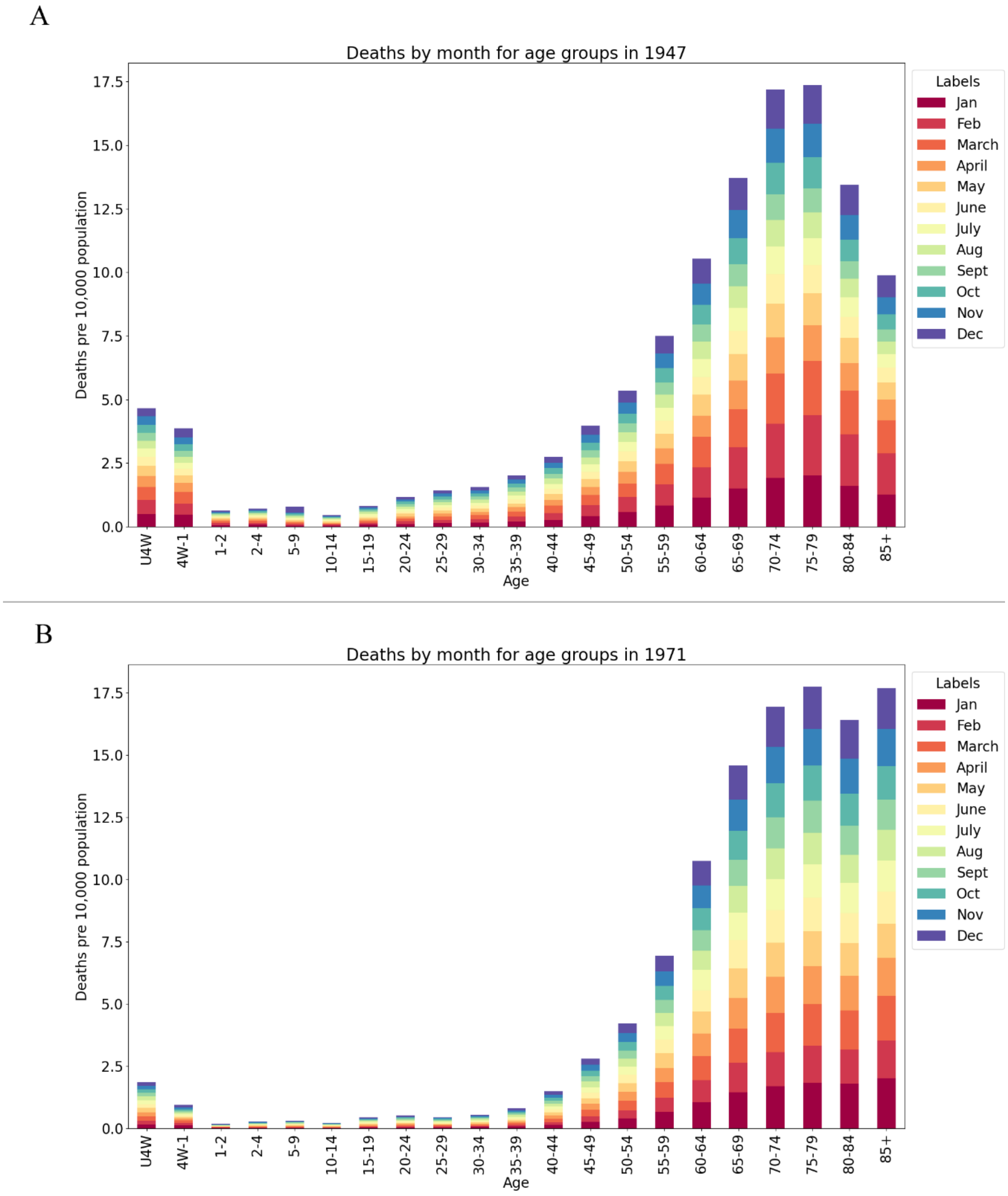


Figure 4-4: Mortality over time. Mortality over time. Panel A: Monthly mortality by age bins 1947; Panel B: Monthly mortality by age bins 1972;

Whilst the weekly reports are a valuable resource, we first digitised the monthly mortality for Great Britain to seek evidence of seasonality and if said seasonality also varied over time. This wave focused on digitising the monthly mortality for Great Britain from 1947 to 1972. Figure 4-4, Panels A and B, shows the monthly mortality by age bins in 1947 and 1972. This shows that infant mortality decreases over the 20th century, and that more children died in winter than in summer months.

However, as these are aggregate figures, it is still possible that individual districts experienced different seasonality. Towns experiencing dryer and warmer conditions have historically been linked to higher infant mortality than wetter areas due to summer diarrhoea, which occurred mostly in July to September[186]. Whilst beyond the scope of BIO-HGIS to date, there is still a justifiable reason to investigate infant mortality rates at a weekly level. A future Wave of BIO-HGIS will attempt to do this.

4.7.2 Wave 5B: Disease mortality distributions: Static or fluid?

Covid-19 has shown that age specific effects are common in infectious diseases, with mortality being starkly different in younger individuals than older cohorts[187]. Many diseases follow a U-shaped distribution of age mortality[188], but increasingly evidence is arising that J-shape and other mortality distributions exist for what are thought to be childhood diseases[187]. Measles, polio and tuberculosis are all thought as childhood infectious that can lead to mortality, yet this mortality is increasingly found in much later in childhood[187]. Both diseases and environments have changed significantly over the last century, but little focus has been placed on how these changes alters the distribution of deaths within diseases across the age spectrum[187].

We seek to further the literature in two ways. All international causes of deaths (ICD) by age for the 20th century to present at a national level have already been digitised or made available by the Office of National Statistics[189]. Within nation estimates also exist for administrative counties for an abridged list of the ICD. These records have not been digitised but will be the focus of a separate investigation and digitisation wave in the future.

Disease names and definitions change over time, which makes comparisons difficult when using ICD codes much before ICD 8. Our first contribution is therefore to assist standardization of older ICD codes to the more modern versions, such as ICD 10. In doing so, we should be able to investigate if the age distribution for diseases within the ICD changed significantly over the 20th century. We will then further investigate if any such changes were as a result of exogenous shocks to the environment.

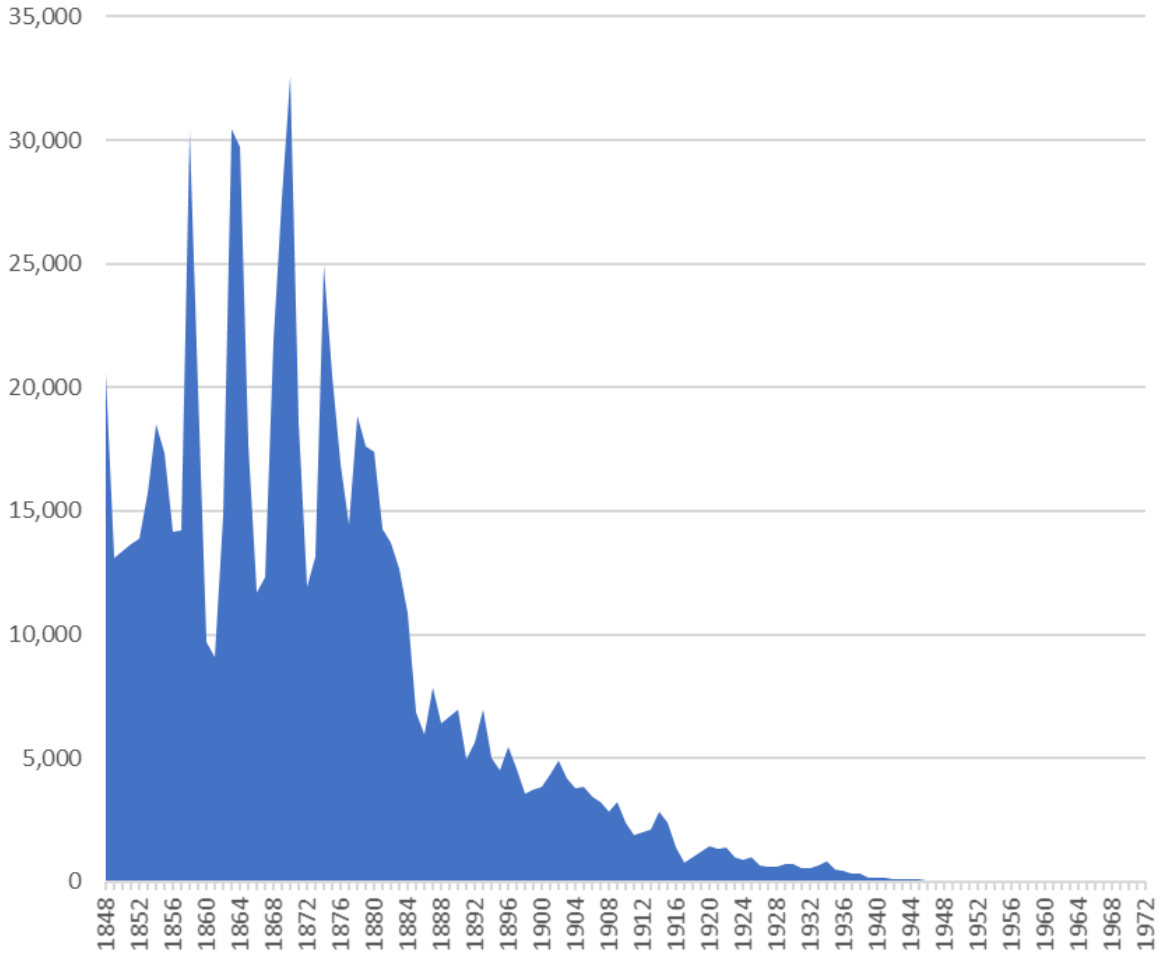


Figure 4-5: Scarlet Fever / Scarlatina Deaths 1848-1972 for England and Wales

Our second contribution comes from the inclusion of data from the 19th century. Many diseases have changed virulence over the 19th century, but this information remains mostly inaccessible. Most records of annual causes of death go back further in time than 1901 and can be constructed using currently known national records as far back as 1848. We seek to digitise the additional data on mortality per formation of the ICD and link them to ICD 10. Limited efforts have already been undertaken on digitisation, such as for scarlet fever, or as it was called in the 19th-century scarlatina, as shown in Figure 4-5. By adding additional years from the 19th century, we seek to expand our power to find any possible change to the diseases age-mortality distributions.

4.7.3 Wave 5C: Annual population counts for sub nations within the UK

If using the data Wave 5, and desiring the results as rates, then currently most of the available estimates are for the UK as a whole. This wave was designed to digitise the population for each individual nation so that rates could be constructed related to the national counts rather than the UK as a whole. Wave 5C digitised the population counts from 1861 to 1971, after which estimates for population counts for nations within the UK are more commonly available. This was used to convert the counts of age and month mortality into rates per 10,000 within Figure 4-4.

4.8 Wave 6: Grammar Schools

The UK has a history of educational class division, with vocational education traditionally being looked down upon as for the less able, less motivated, less employable, and ultimately designed for the lower social classes[190]. Whilst certain regions have moved away from the tracked tripartite system introduction in 1944, England continues to have resistance to fully comprehensive schools[191]. Grammar schools still have advocates, with selected schools increasing their intake in the 21st century[192], despite concerns about the effects of segregation on widening inequalities and reducing social cohesion[192].

Wave 6 focused on digitising locations of grammar schools in existence as of 1968 in England and Wales from the parliamentary papers[193]. Similar to the data from War, State, and Society, the grammar schools were only provided as names. Therefore, we used the same procedure as the Blitz's bombing points and geo-located the grammar schools names using the Google Maps API. Each school contains information as of 1968 for the amount of government funding they received, pupils in attendance, and the pupil-teacher ratio at said school. These schools are aggregated within each parish, where we also derive the total number of schools per parish. The location of these schools is shown in Figure 4-6.

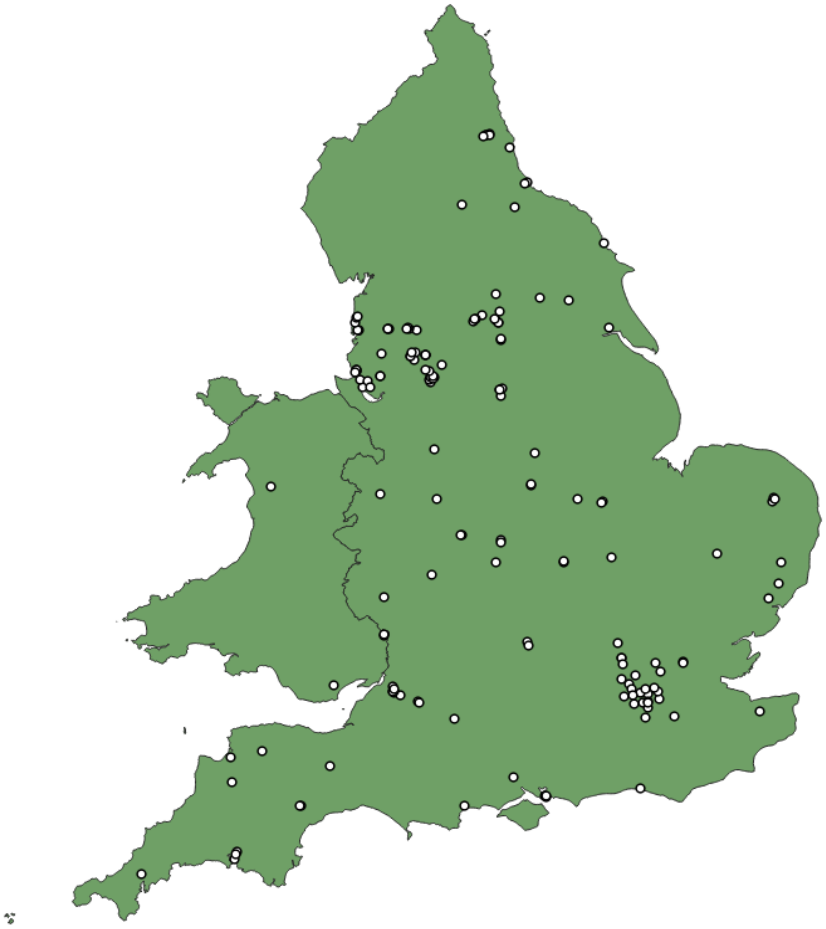


Figure 4-6: Locations of direct grant grammar schools in existence as of 1968 in England and Wales

4.8.1 Structural differences in BMI across the UK regions

Obesity rates have increased in the latter part of the 20th century, a trend that is now spreading to traditionally leaner middle income and lower income countries[194]. Higher education is commonly associated with lower obesity rates in the western world[195,196]. Causality has often been challenged, but others have utilised changes in school-leaving age as an exogenous shock for causal inference[197]. Other studies have used mendelian randomisation (MR) to show that higher educational attainment is associated with lower BMI[196,198], as well as higher type two diabetes[199,198] and coronary artery disease[198].

A common limitation of studies in the literature seeking to show that higher educational attainment is associated with reductions in BMI, is that they ignore the potential of area differences[200]. Here, we aim to investigate if individually reported educational attainment is associated with BMI after controlling for individuals polygenic scores for education, proximity to grammar schools, and regional level educational attainment from the 1961 census. We seek to utilise the change in mandatory years of education that affect those born after 1957 to further push for casual inference.

4.9 Wave 7: Unemployment

Defining deprivation is complex, with multiple historical measures existing, such as the Townsend deprivation index or the Jarman index[201]. The original Townsend deprivation index was a function of four variables of non-car ownership, non-home ownership, the log of unemployment, and the log of overcrowding[202]. Many of these variables are only available in census years making a time-varying measure difficult. However, simple regional unemployment is often highly correlated to such deprivation measures. Regional standardised unemployment rates had a 92.4% correlation with the Townsend deprivation index and 86.6 to the Jarman index in 1995[203].

We seek to utilise monthly unemployment data from the unemployment Labour Gazette[204] to construct a time varying measure of deprivation from 1931 to 1971. We then seek to compare this measure to both the original Townsend index of 1971, and constructed Townsend indices from new historical data, or imputation, for the census years of 1931, 1951, and 1961. In doing so, we seek to explore how simple unemployment rates may be used in lieu of the Townsend index for capturing deprivation. To validate our findings, we will investigate how our measures of deprivation predict health outcomes utilising the ICD 10 codes and the UK Biobank through a Phenome-wide association study (PheWAS).

4.10 Conclusion

BIO-HGIS has been designed to help remember the past, in the hope this data can help prevent similar mistakes or outcomes occurring in the future and designed to iteratively add new and meaningful data for research and public information. Whilst the BIO-HGIS itself currently only contains UK data, much of the methods sections of this thesis that have been used to construct it are generalisable to other countries' data. The papers that have been started or shown later within this thesis as working papers represent a fraction of the potential of the database. Most research projects have been focused exclusively on the UK Biobank, despite the huge potential for use in the many additional cohort studies that exist throughout the 20th century. Whilst the focus of the work behind the construction of the data remains focused on health, many disciplines may find extensive use for such data that is outside the remit of our knowledge.

BIO-HGIS has limitations even when applied to its targeted source data of the UK biobank, such as from the generalisation of the birth coordinates within the UK Biobank. However, said generalisation mainly limits the use of *distance-dependent* measures. For example, if a participant's actual birth coordinate is within 1 km of the main road, but the generalisation moves them further away, this could lead to measurement error for distance-dependent analysis. Regional data limits this measurement error to only being born within 1 km of a region's bounds. The use of regional data could also lead to limited variation, as all those born within a given location are assigned the same exposure. When investigating exposures over prolonged periods, individuals born a single month apart in the same location of birth will still share extensively the same exposure. Identifying variation, therefore, comes more from individuals who were born many years apart within the same place of birth. Fortunately, given the UK Biobank has close to half a million participants born over nearly 40 years[205] this is often possible.

Whilst this is significant work left to be undertaken. Much the data within BIO-HGIS has already been used to start or write a paper, as mentioned throughout the subsections of this paper. Therefore, whilst not publicly available yet, as these papers begin to be published, hopefully across 2023, the data will be made available for others. The hope is that, given the extensive work undertaken to quality control and document said data, that when it is made available, it should be easy to use. This should further research in the UK Biobank for years to come, assist other disciplines answer questions that cannot currently be answered, and protect our heritage from the passage of time.

5 Early life exposure to scarlet fever is associated with ischemic heart disease later in life.

5.1 Introduction

Streptococcus pyogenes (S. pyogenes) can cause rheumatic heart disease, which causes permanent life-threatening heart damage[76,119], yet can arise from an otherwise minor infection of untreated streptococcal pharyngitis[119]. Whilst rheumatic heart disease is one of the most known complications, elevations of S. pyogenes anti-bodies have been linked to other diseases, such as Crohn's disease[125], narcolepsy[126], Henoch-Schoenlein purural[206], psoriasis[127], Sydenham chorea[85], and paediatric autoimmune neuropsychiatric disorders associated with streptococcal infections[85] (PANDAS). As infections from S. pyogenes are heterogeneous, both in anti-body response[207,208] and duration[207], there may be further yet unknown long-term consequences of infections from S. pyogenes.

In this study, we investigated the relationship between early life exposure to S. pyogenes and later life outcomes using regional incidence of scarlet fever in childhood. To do so, we used regional (district) level weekly scarlet fever notifications from the Registrar-General's Weekly Returns[42] and linked these exposures to UK Biobank participants. Given infections from S. pyogenes have already been associated to heart disease through rheumatism[76,209,210], we sought to investigate if early life infections are related to other cardiovascular diseases later in life.

We further investigated the relationship between S. pyogenes exposure and later life cognitive and educational attainment. We do so specifically because S. pyogenes infections are linked to neuropsychiatric disorders through PANDAS because of neuroinflammation[85]. Prolonged neuroinflammation is associated with age-related cognitive impairment[211] and neuropsychiatric disorders that are highly detrimental to educational attainment[212]. More generally, other childhood diseases have been associated with lower later life cognition[86], and for those in worse health with lower educational attainment[87].

5.2 Methods

5.2.1 Population and exposure

We used data from participants within the UK Biobank, a prospective cohort study of 502,506 UK adults aged 40-69 from 2006 to 2010, who lived within 25 miles of an assessment centre[12]. We used the UK Biobank for its detailed information on later life health and specifically for older cohorts born in an era of more frequent S. pyogenes infections. However, as the UK Biobank includes limited information on early life conditions, we constructed proxy measures of exposure to S. pyogenes by using the notifications of scarlet fever from the Registrar General Weekly Reports for England and Wales[42].

To construct this proxy exposure to scarlet fever, we used the weekly incidence of scarlet fever reported in districts between 1946 and 1973. We linked the participants' location of birth to one of the 1472 districts across England and Wales as of the 1951 census, using a shapefile available from Vision of Britain[67]. We isolated cases of scarlet fever within the district of birth 52 weeks after the first full week of the participants' year and month of birth. We then used the annual population estimates from the Great British Historical Database on Health and Health Care[74] to convert the number of scarlet fever cases to incidence rates per 10,000 population. We iterated forward and created district incidence rates of scarlet fever up to the age of ten, as scarlet fever is most common between the ages of 2-10[213], peaking between the ages of 4-6[214]. We used the annual incidence rates of scarlet fever to construct average exposures of scarlet fever by the ages of one, five, and ten. The average exposures were constructed as the mean of the annual exposure incidence rates experienced up to that age.

5.2.2 Phenotype definition

We used ICD 9 and 10 codes from the diagnosis and death register to define heart disease phenotypes. We used definitions for acute myocardial infarction (AMI), ischemic heart disease (IHD), stroke, and cardiovascular diseases (CVD) as previously defined[215]. For ICD 10 codes, we constructed AMI from I21-22, IHD from I21-25, stroke from I6 and G45, and CVD from I0-99 and G45. All ICD9 and ICD10 codes are within supplement Table 5-S1. S. pyogenes is known to cause valve complications[76], so we defined additional outcomes, including non-rheumatic damage to heart valves using ICD codes I34-38, and rheumatic damage using ICD 10 I0.

For later-life cognition, we used participants' fluid intelligence, UK Biobank variable 20016. The score is the sum of the correct number of answers to 13 logic and reasoning multiple-choice questions within two minutes[216]. For educational attainment, we use the questionnaire on qualifications achieved, UK Biobank variable 6138, and transformed the variable into years of schooling following the literature[217]. We then standardise both fluid intelligence and educational attainment to have a mean of zero and standard deviation of 1.

5.2.3 Statistical analysis

We estimated the association between regional incidence of scarlet fever and cardiovascular outcomes using linear probability modelling (LPM) linear regression for cognition and educational attainment. Due to declining rates of scarlet fever across the sample, we controlled for participants' year of birth. We defined the year to start in September of year t until August of year t+1, which ensures that each year is specific to the scarlet fever season[218]. Doing so also controls for years specific effects of our outcomes. We also controlled for seasonality of scarlet fever through the month of birth and gender because of potential sex differences of both our scarlet fever exposure[219] and our outcomes. We controlled for the district of birth to control for stronger associations for those in lower socio-economic position[220], and all time-invariant differences between districts. Finally, we estimated the relationships between the phenotypes of interest and average exposure to scarlet fever, with the above set of controls.

The sample data was formatted in python 3.7. Python 3.7 package weightGIS[221] was used to standardise districts to 1951, and to map locations of birth to a district. Statistical analyses were performed in Stata version 14 with the reghdfe package[222]. QGIS 3.10 was used to construct maps, with other figures made with python 3.7 and Blender 2.83 using the pyBlendFigures package[223], Stata 14, or Excel.

5.3 Results

The UK Biobank contains 502,506 adults. As the disease notifications were limited to England and Wales, we removed all participants born in Scotland (n = 39 488). We also excluded those born before June 1946 for two reasons. First, because of the introduction of penicillin in 1946[224], those born before penicillin faced a different disease environment in early life. Second, those born between 1939 and 1945 may have moved in early childhood because of the Second World Wars evacuation orders[183]. These issues could lead to measurement error, so we excluded participants before June 1946, leading to a loss of 122,177 participants. To ensure all participants had ten years of exposure data and because the disease data ended in 1973, we excluded those born after 1963 (N=41,339). As linkage to the disease notifications required a birth coordinate, we exclude participants who had not reported one (n=35,022). After removing six districts which only had a single participant within them, this led to an analysis sample of 241,679 individuals, observed within 1,452 districts.

The descriptive statistics for the outcomes of interest in the analysis sample are shown in Table 5-1. Some instances of heart disease are rare, such as rheumatic disease, but the analysis sample still observed hundreds of cases, even for these rarer definitions. The sample participants on average correctly answered 6.26 (2.10 standard deviations (SD)) out of the 13 total questions used to construct the measure of fluid intelligence.

The analysis samples fluid intelligence is slightly higher, with a lower standard deviation, than the full UK Biobank sample of 5.99 (2.14 SD). On average, the sample also had attained more education than the population given the sample were born between 1946 and 1963.

Table 5-1: Descriptive statistics of count, mean, and standard deviation. All heart related phenotypes come from the full sample of 241,679, with the count in the table representative of the number of cases. For fluid intelligence and educational attainment, there we only 99,189 and 207,242 observations respectively.

Variable	Unit	Mean	Standard Deviation
Age	Years	68.541	5.134
Sex	Male	0.445	0.497
Rheumatic Disease	With disease	0.003	0.059
Vascular Disease	With disease	0.008	0.091
Acute myocardial infarction	With disease	0.017	0.128
Ischemic heart disease	With disease	0.047	0.212
Stroke	With disease	0.018	0.132
Cardiovascular diseases	With disease	0.274	0.446
Fluid Intelligence	Questions correctly answered out of 13	6.259	2.098
Educational Attainment	Years of Schooling	16.555	3.989

5.3.1 Distribution of sample

Figure 5-1A shows the share of the population covered by our analysis sample from the UK Biobank of each district. We used the 1951 population estimates and then calculated the share of the number of UK Biobank participants born in that district. In most rural districts, our sample population represents about 0.05-0.5% of the total population. Whereas in more urban areas, we observed a higher share of between 0.5% to 1% of the total population. Figure 5-1B shows the distribution of the scarlet fever totals within 1951 per 100,000 population within each district. Within 1951, there is notable variation in the rate of scarlet fever notifications across the districts within England and Wales.

5.3.2 Exposures

Figure 5-2 shows the mean exposure for the average scarlet fever incidence rates at ages one, five, and ten for each birth cohort. Given the declining trend in cases, shown in supplementary Figure 5-S2, those born later experienced on average lower exposure. Variation between years and districts is higher for exposures by younger age, where individual year variations have a greater effect on the incidence rates than exposures averaged out by age five or ten.

5.3.3 LPM estimates of associations between district-level scarlet fever exposure and individuals' later-life cardiovascular health

Figure 5-3 presents the associations of an additional case of scarlet fever per 10,000 in a district per year by the age of one, five, and ten, with cardiovascular outcomes. We found little evidence that rates of scarlet fever associated with rheumatic diseases. We found weak evidence that participants exposed to higher levels of scarlet fever were less likely to subsequently develop vascular disease: for exposure at age one (-0.18, 95%CI: -0.72; 0.36), age five (-0.77, 95%CI: -2.21; 0.66) and age ten (-1.16, 95%CI: -3.38; 1.06).

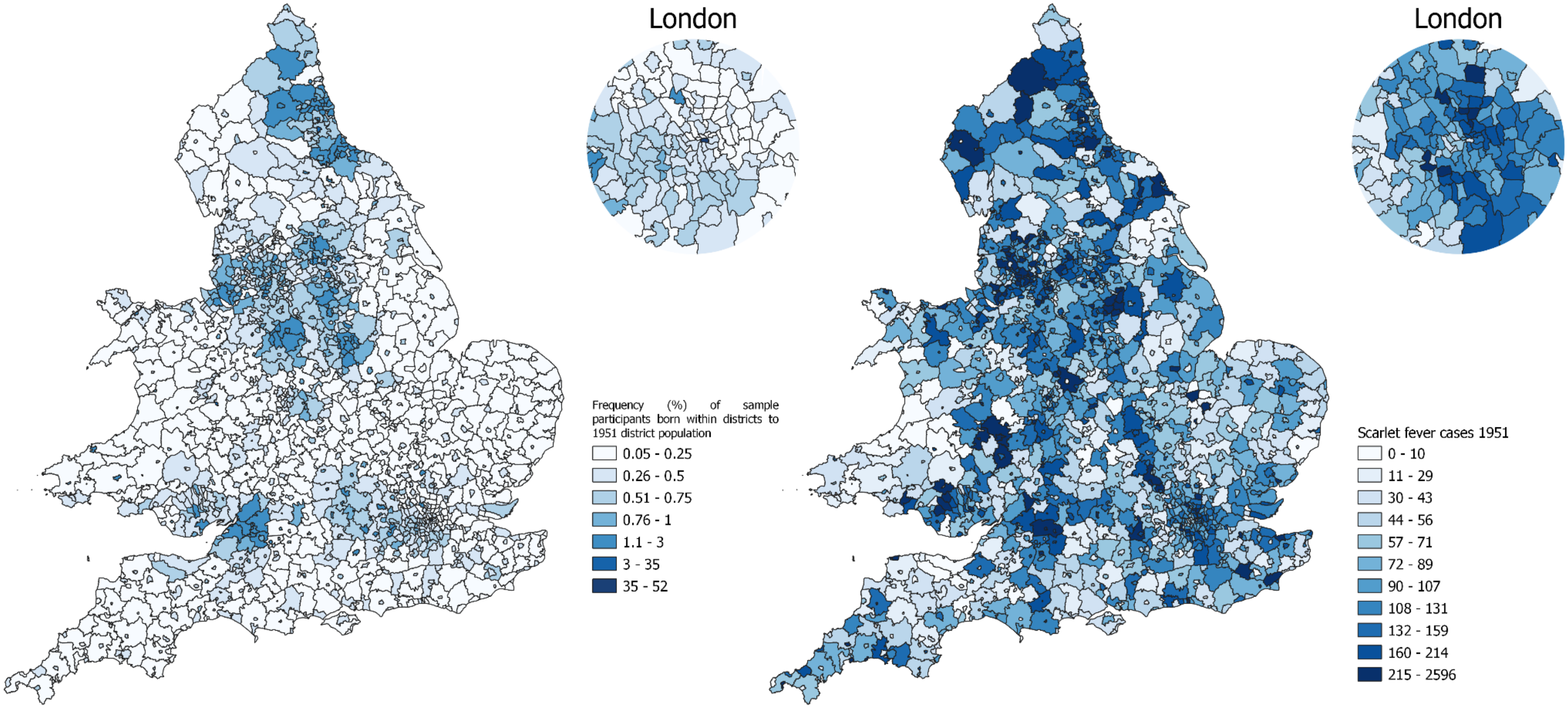


Figure 5-1: Left-hand side shows the share of the population covered by our analysis sample from the UK Biobank. The right-hand side shows the distribution of the total scarlet fever cases within 1951 per 100,000.

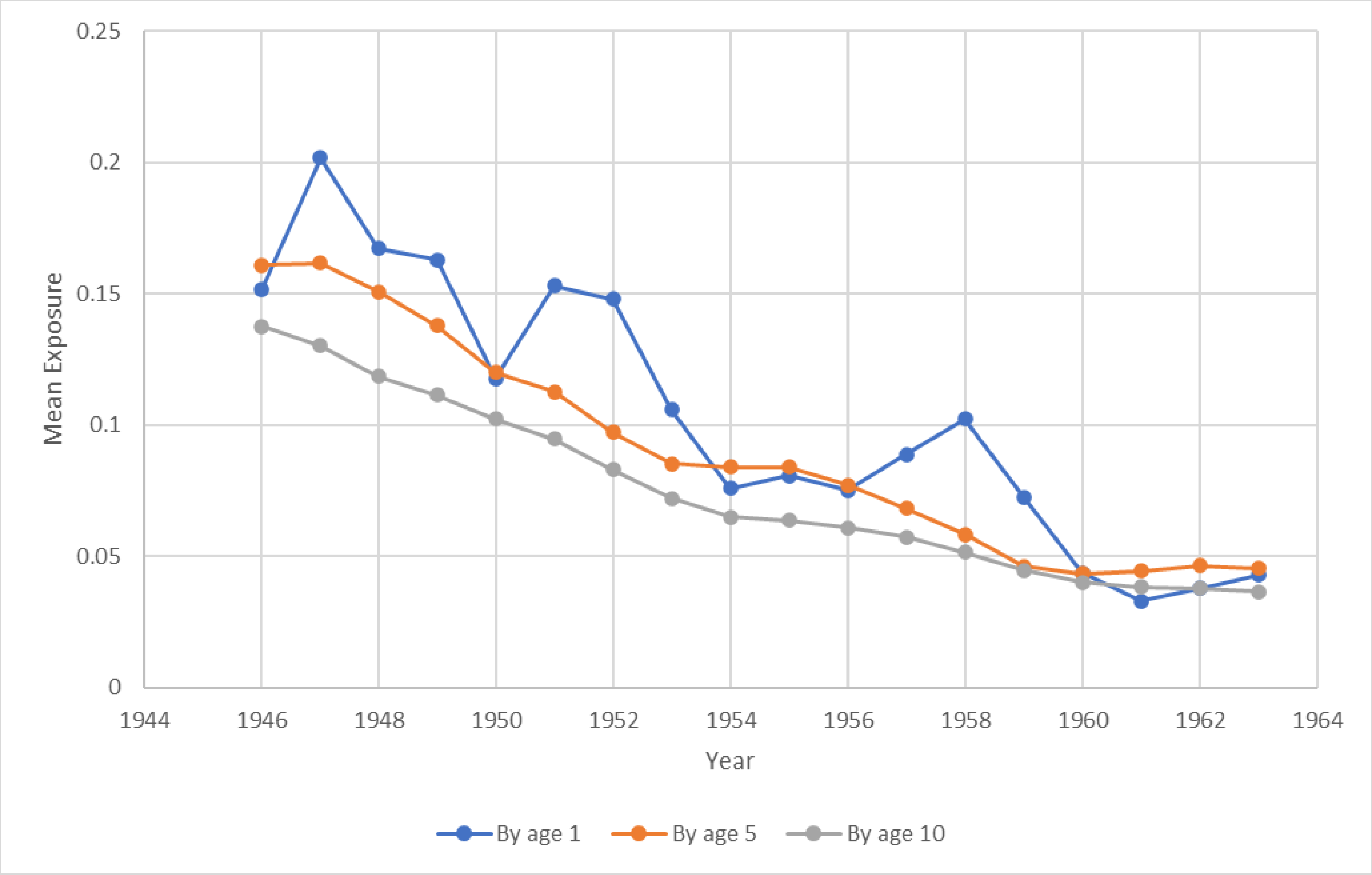


Figure 5-2: The mean exposure for participants, derived from their district of birth, per birth cohort for average scarlet fever incidence rates at the age one, five, and ten.

There was some evidence of a positive association between exposure to scarlet fever and AMI, IHD, and stroke. The strength of the association increased with age for AMI and IHD, but not with stroke. The association between exposure to scarlet fever and risk of AMI increased from age one (0.51, 95%CI: -0.31; 1.33), to age five (1.59, 95%CI: -0.23; 3.42), and age ten (3.24, 95%CI: 0.31; 6.17). Similarly, the association of scarlet fever and IHD increased with age: at age one (0.76; 95%CI: -0.53: 2.04), age five (2.62, 95%CI: -0.64; 5.87) and age ten (7.25, 95%CI 1.58; 12.92).

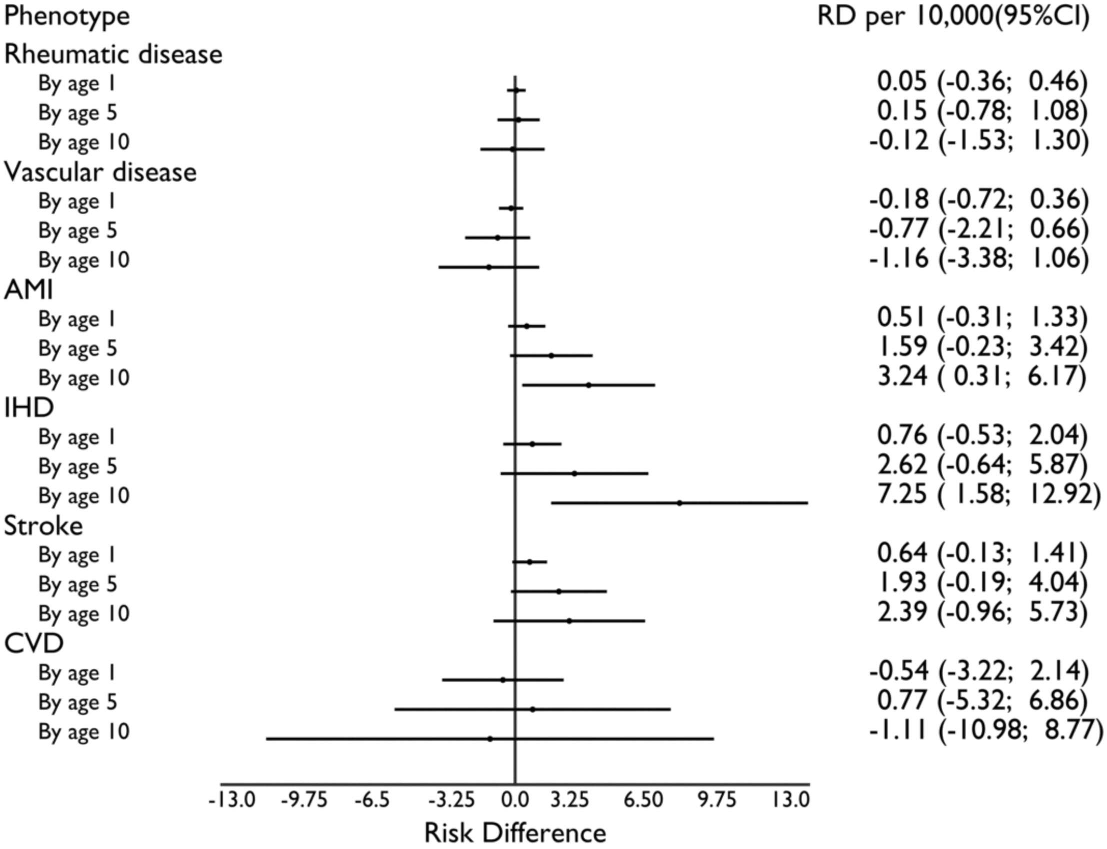


Figure 5-3: Risk difference associations and 95% CIs of an additional case of regional incidence of scarlet fever experienced on average by the age of 1, 5, and 10 per 10,000 individuals on cardiovascular outcomes. N = 241,679 for all LPM models.

A positive association was found for stroke by age one (0.64, 95%CI: -0.13; 1.41), age five (1.93, 95%CI: -0.19; 4.04), and age ten (2.39, 95%CI: -0.96; 5.73). However, the CIs for stroke widen at a greater pace than the increased strength of association, making the CIs at all ages overlap zero. Finally, we found little evidence of an association between early life exposure to scarlet fever and later-life CVD.

5.3.4 OLS estimates of associations between district-level scarlet fever exposure to individuals' cognition and educational attainment

Figure 5-4 presents the association of an additional annual regional case of scarlet fever per 10,000 in the district of birth with educational attainment and fluid intelligence, with 95% CIs, by the age of one, five, and ten. We found little evidence that exposure to scarlet fever in childhood associated with educational attainment. Scarlet fever was negatively associated with fluid intelligence: for exposure at age one (-0.13, 95%CI: -0.23; -0.04), age five (-0.16, 95%CI: -0.39; 0.08), and age ten (-0.17, 95%CI: -0.54; 0.19). However, the confidence intervals widen greatly at later ages of average exposure to scarlet fever and overlap zero.

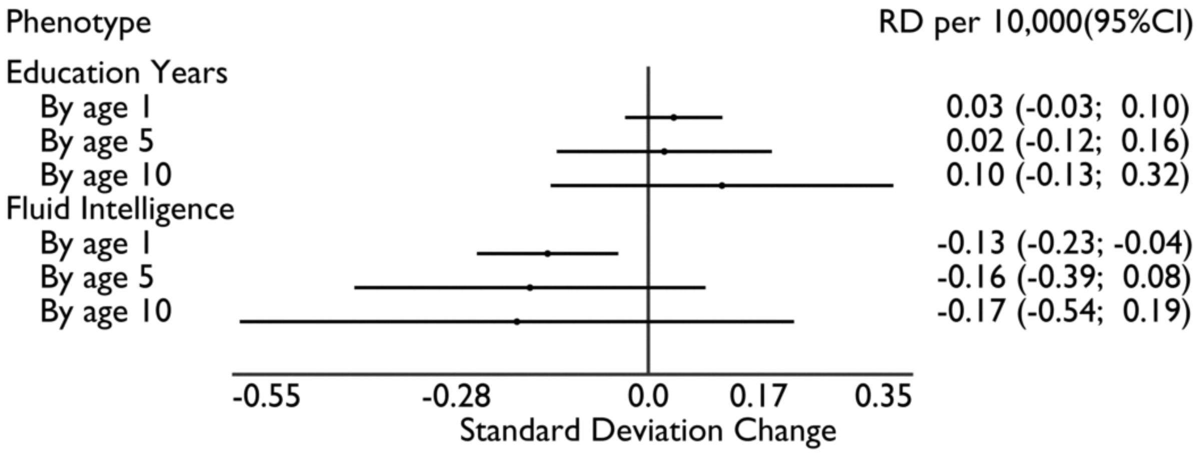


Figure 5-4: The standard deviation change associated with an additional case of scarlet fever per 10,000 individuals. N=99,189 for fluid intelligence; N=207,242 for educational attainment.

5.3.5 Multiple hypothesis testing

As we have investigated the impact of exposure to scarlet fever on multiple phenotypes, it is appropriate that we undertake multiple hypothesis testing, so we validate our results using the Bonferroni correction. As the exposures of interest (exposure at age 1, age 5 and age 10) are nested, we are testing 8 hypotheses. The pre and post corrected P values are shown in Table 5-2. Using this Bonferroni correction for multiple hypothesis testing, only our result for fluid intelligence at age 1, in addition to IHD at age 10, are found to remain significant.

Table 5-2: P values for exposure of scarlet fever by age 1, 5, and 10 each phenotype before and after Bonferroni correction. As exposures are nested, each exposure group is corrected for 8 hypotheses.

	By Age 1		**By Age 5**		**By Age 10**	
Phenotype	P	Bonferroni P	P	Bonferroni P	P	Bonferroni P
Rheumatic Disease	0.801	1.000	0.749	1.000	0.873	1.000
Vascular Disease	0.514	1.000	0.292	1.000	0.307	1.000
AMI	0.222	1.000	0.087	0.696	0.030	0.240
IHD	0.247	1.000	0.115	0.920	0.012	0.096
Stroke	0.101	0.808	0.074	0.592	0.162	1.000
CVD	0.692	1.000	0.804	1.000	0.826	1.000
Educational attainment	0.305	1.000	0.771	1.000	0.399	1.000
Fluid Intelligence	0.005	0.040	0.187	1.000	0.353	1.000

5.4 Discussion

In this paper, we investigate the potential for long-running consequences from exposure to scarlet fever in childhood to worse later life cardiovascular health and cognitive performance. We found weak evidence of an association for AMI and stroke, with stronger evidence robust to multi-hypothesis testing for IHD and fluid intelligence. Unlike previous annual national data from Public Health England[225], our study has three core benefits. First, it better captured individual exposures by specifying 52 weeks from the participants' year and month and birth. Second, the weekly analysis was recorded within areas of England and Wales, rather than the aggregate within England and Wales, which meant we could exploit extensive spatial variation across the districts. Finally, the weekly data allowed for controlling both within year variance in outcomes and the seasonality of the exposure and outcomes. To our knowledge, this is one of the first papers to investigate the association between early life exposure to scarlet fever and later-life heart disease, education, and intelligence within the context of England and Wales in the 20th century in a very large sample.

Whilst our new data is a strength, that lack of a strong source of existing data for scarlet fever prior to our investigation means we have few papers to paper our results to. A study that investigated 19th century Sweden and found a negative association between exposure to scarlet fever in early life and later life cardiovascular disease[226]. The only known prior study to have investigated educational attainment for this birth cohort was at age 11 on a sample of 43,820 Birmingham children [227]. Children infected with scarlet fever in the years prior to the eleven plus examinations had similar exam results compared to non-infected peers[227]. Our investigation into fluid intelligence is based on a potential rare complication of S. pyogenes of PANDAS, but results investigating PANDAS to fluid intelligence are not comparable to the effect of scarlet fever, and we know of no former studies that have reported comparable estimates of the associations of scarlet fever and fluid intelligence directly. However, a previous study investigating early life exposures and multiple childhood diseases at ages one and two[86] found negative associations to later life fluid intelligence, which is similar to our analysis for exposure to scarlet fever by age 1.

Many of our phenotypes result in null findings, especially after accounting for a Bonferroni correction to multi hypothesis testing. This could be argued as a strong negative result, as our data is some of strongest to data. Whilst possible, there are considerations as to why our result may yet be underestimated and therefore underpowered to find results. A key example of this is rheumatic disease. Rheumatic fever, the cause of rheumatic heart disease, is caused by S. pyogenes, and specifically requires a previous untreated infection of S. pyogenes[76], such as scarlet fever. Yet, our result would suggest that areas with higher infection rates of an S. pyogenes disease does not link to any increase in rheumatic fever, despite this being a known biological pathway.

Part of this is due to the low number of cases, as whilst rheumatic heart disease is still a leading cause of premature death, it declined across the western sharply since 1950[121]. Even for those unfortunate enough to be infected, as the UK Biobank sampled individuals aged 40-69, few individuals with rheumatic heart diseases as a result of early life exposures would likely have survived. Therefore, our result for rheumatic diseases is in part due to survivor bias, which is likely to significantly decrease our ability to find any associations.

Another bias that is likely to have attenuated our ability to detect effects is that of selection bias. The UK Biobank suffers from its participants being healthier and of higher socioeconomic status from the population sample[228,229]. Of particular concern is that individuals are also less likely to live in deprived areas[229]. Whilst this traditionally relates to participants current home address, the birth coordinates, as shown within Figure 5-1, also shows a strong bias for larger urban areas, with little rural participation. Therefore, we have reason to believe that our results suffer from considerable downward bias, as our participants were less likely exposed than an average member of the population.

This paper uses WeightGIS to standardise districts over time, to ensure greater comparability and ensure consistent within district characteristics. Whilst WeightGIS uses parish level population weights, subdividing England and Wales into nearly 17,000 locations, it is still a percentage-based weight and may result in variables with a degree of imprecision. This may result in measurement error, as individuals are assigned exposures that are higher or lower than they otherwise should be. Whilst this may bias our estimates if the measurement error is high, the areas that suffer from errors in

WeightGIS worst are rural areas. Given the sample is predominately living in urban areas, which do not change as drastically, any bias caused by weighting should be minimal.

A further source a measurement error is due to residential mobility, which the UK Biobank lacks detailed information on. As many of our larger findings exist for those exposed constantly up the age of ten, if individuals moved before the age of ten, this would result in measurement error. However, as we measured the exposure at a district level, this would only occur if individuals moved outside the district of birth. Estimates of residential mobility between 1938-1947 have been estimated to be 7% for individuals aged 0-19 per year[220], which suggests bias resulting from residential mobility may not be small. However, this particular estimate includes war years, such as the evacuation during blitz, so is believed to be a much higher estimate than for those born later in the sample. There is also the consideration that individual birth locations are generalised to 1km[69] grid points. Those born closer to district borders have a higher potential of measurement error from being assigned the wrong district of birth which may further result in underestimates of our outcomes of interest.

Therefore, as we believe that our estimations suffer from downward bias, the magnitude of the result for IHD should be of concern. Just under 4.7% of our analysis sample in the UK Biobank, which is healthy by far than the standard population, have IHD. Yet, our results indicate showed that being exposed to a sustained additional case of scarlet fever each year, on average, by the age of ten, increases the probability of being diagnosed with IHD by 7.25 percentage points, or the absolute risk difference. This increase in absolute risk difference suggests that consistently highly exposure to scarlet fever in early life may have stark impacts to later life health. The underlying selection bias in the sample result in this estimation not generalising to the population. However, given S. Pyogenes pre-established links to heart disease through rheumatic fever and the resulting rheumatic heart disease, the associations we have found within this paper are important to be further considered, especially given scarlet fevers return in recent years in the UK[230].

5.5 Supplementary Information

Table 5-3: ***(5-S1)*** *ICD 9 and 10 codes used for heart definitions*

Definition	ICD9 Codes	ICD 10 Codes
RD	3909-3929	I0
V	4240	I34-I38
IHD	4109-4149	I21-I25
AMI	4109,4129	I21-I22
STROKE	4309-4359	I6, G45
CVD	3909-4599	I, G45

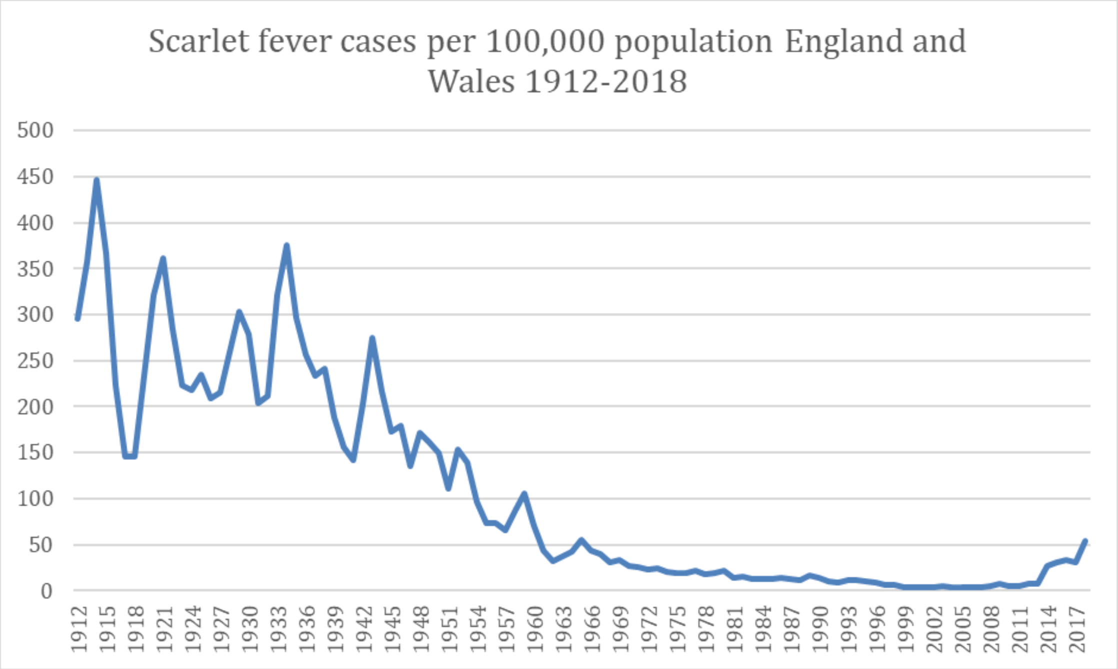


Figure 5-5: ***(5-S2)*** *Rate of scarlet fever cases in England and Wales per 100,000 from 1912-2018206,34*

6 Can gene-environment interactions from childhood diseases explain the rise in asthma incidence in the 20th century?

6.1 Introduction

Over 334 million people globally suffer from asthma,[231] with the prevalence growing from just 4% of 9-12-year old's in 1964, to 29.5% in 2004, then more recently back to 18.6% in 2014.[135] The cause for the rise in cases is still not fully understood. The environment changed substantially over the latter half of the 20th century and has been found to be associated with increased levels of asthma,[135232]. Reduced breastfeeding,[233] increased numbers of caesarean sections,[234] increased pre- and neonatal antibiotic treatments,[235],[236], reductions in exposure to livestock,[237] pets,[238] dust,[239] common childhood diseases[138], and unpasteurised milk,[240],[241] have all been associated with increased asthma incidence. Many of these findings have themselves been disputed,[242],[243],[244],[245],[148],[146] leading to a lack of definitive evidence as to the cause of the rise in asthma cases. However, none of these studies take underlying genetic liability to asthma into account.

Failing to control for genetic risk is importance is due to how those who are genetically at risk of asthma respond to allergens. Those at genetic risk of asthma are susceptible to produce an excessive immunoglobulin E (IgE) response to otherwise harmless environmental allergens[26],[27]. However, asthma is both polygenic and with strong environmental components, with the risk of developing asthma highly dependent on the interaction between these risk elements[246]. Therefore, studies that investigate the impact of a particular environment on the risk of developing asthma without considering individuals' genetic risk, fail to consider the potential difference in risk those with predetermined risk face.

With such a sharp rise in asthma rates, this precludes the rise in asthma cases being purely genetic, as changes in genetic makeup would take generations to occur[247]. However, if the genetic risk is moderated by certain beneficial environments, then, whilst the prevalence of the underlying genetics cannot change so quickly, the prevalence of the actual disease could. Many environments changed, but here we focused on the changes the underlying exposure to early life disease based on a paper investigating IgE inhibition[25].

A Lab study found that specific expressions of Immunoglobulin G (IgG) can inhibit excessive IgE Responses[25]. IgG levels are increased in response to an infection[248] and individuals can remain IgG positive for months or years[249]. Currently, one of the few affective treatments for asthma, immunotherapy, does much of the same thing, by using micro doses of the allergen to build up IgE regulation over time[250]. If diseases exposure can lead to increased persistence of IgG, and can inhibit an excessive IgE response, then we hypothesize that disease exposure in early life may have resulted in a protective environment.

The disease environment changed over the 20th century considerably. As over half of individuals develop asthma before the age of ten[251], the change to disease exposure early in life may be crucial to understanding the rise in asthma in the 20th century. Vaccines meant many previously common diseases, such as measles, tuberculosis, pertussis, diphtheria, mumps, rubella and polio declined[252],[253]. This saved millions of lives, with the MMR vaccine alone estimated to have prevented 20 million deaths just from measles between 2000-2015[156]. However, for those born before these introductions, the prevalence of common childhood disease would have been significantly higher than after. If the IgG levels from infection were able to help regulate the IgE response that asthmatics have, then for those with a high genetic risk, the disease environment may have had an unexpected protective effect. Conversely, for those without genetic risk, the change in the disease environment should conceivably have no effect.

Therefore, by incorporating individuals' genetic liability in the analysis, we could investigate how exposure differs not just among those with greater exposure, but also between those at greater vulnerability because of genetic risk. In this study, we investigated how the change across the 20th century of incidence of infectious diseases affected the risk of asthma. We focused on the three respiratory diseases of scarlet fever, pertussis, and measles, all of which have been associated with asthma[140],[146],[130],[148],[149]. This study has a key contribution, in that it may help explain conflicting evidence in the literature for the changing disease environment on asthma instance by including a genetic component[138],[139],[140],[141],[142],[143],[144],[145],[140],[146],[147],[148],[149].

6.2 Methods

6.2.1 Cohort population and data sources

The UK Biobank has been described elsewhere[11]. Briefly, the UK Biobank is a prospective cohort study of UK adults aged 40-69 at time of recruitment[12]. The UK Biobank contains extensive later life information on the health and well-being of its participants, whom crucially for our analysis, have also been genotyped. Unfortunately, the UK Biobank contains limited early life information. Therefore, we used data from the Registrar-General's Weekly Return[42], which provided notification totals for roughly 1472 districts, regional zones within the UK prior to 1974[23], to construct exposures for UK Biobank participants. Due to notifiable cases being reported as counts, we then merged in district population estimates from the Great British Historical Database on Health and Health Care[74] to allow construction of incidence rates.

6.2.2 Exposure Construction

Exposures were defined for each participant using their year, month, and location of birth. Each participant was assigned to a district based on their location of birth. We then defined the exposure in the first year as the rate of disease per 100 population in the 52 weeks from the first full week of the year and month of birth. We constructed exposures for scarlet fever, pertussis, and measles independently for each year up to the age of ten, to account for the ages they are most likely to be infected,[213],[214],[152],[254],[255].

We then used these annual incidence rates to construct average exposures for scarlet fever, pertussis, and measles by the ages of one, five, and ten. The average exposures were constructed as the mean of the annual exposure in the years up to that age.

6.2.3 Construction of asthma phenotype and PRS

We defined asthma cases from three sources. First, we used the hospital inpatient codes, stored both as ICD 9 and ICD 10. Specifically, we assigned anyone with ICD 9 493, ICD 10 J45 or ICD 10 J46 as asthmatic. We also used both self-reported and doctor diagnosed data fields from the UK Biobank variables of 3786 and 6152 to supplement anyone who lacked inpatient data. As our hypothesis is related to asthma specifically, we removed anyone with asthma alongside other atopic diseases of hay fever or dermatitis. We also ensured that controls do not have hay fever or dermatitis, in addition to not having asthma.

Participants genetic susceptibility to asthma was constructed as a polygenic score (PGS) using LDpred[256] with an R squared of around 1%. We used summary statistics of a prior GWAS (23,948 cases, 118,538 controls) using European ancestries and not including the UK Biobank on asthma[257]. We standardised our PGS to have a mean of zero and standard deviation of 1 for our analysis.

6.2.4 Statistical analysis

The association from early life exposures to scarlet fever, pertussis and measles on asthma was estimated using logistical regression and reported as odds ratios with 95% confidence intervals (CIs). We also clustered the standard errors by district of birth. For sensitivity analysis, we repeated our analysis, but only for those in the highest and lowest quartiles for exposure and genetic risk, to further investigate the robustness of our results.

Our covariates included were designed to capture the potential time varying and invariant characteristics that may be associated with either our outcome or exposure.

Both scarlet fever and pertussis cases gradually declined over our sample period[258], whereas measles cases remained constant given its immunisation introduction of around 1968. Asthma cases also began to rise significantly after the 1960s.[135] We controlled for these time trends' using participants' year and month of birth. Controlling for the year and month of birth of the participant also controls for the distinctive seasonal patterns of scarlet fever[218], pertussis[259], and measles[260].

Those in lower economic position tend to be at great risk of asthma[261]. However, asthma incidence tends to be lower in higher social economic areas, even amongst those of low income[262]. In addition, exposures are themselves often confounded through social economic position, as disadvantaged individuals had a higher likelihood of living in unsanitary or worse living conditions[263]. Whilst we do not know individuals specific social-economic position of birth, we control for regional differences by including fixed effects for the 1472 districts. In doing so, we controlled both for time invariant differences between districts and any confounding of social-economic position of our outcome or exposures.

The setup and cleaning of the data was undertaken in python 3.7, with the statistical analysis being undertaken in Stata 14. The project code, Stata logs of the analysis, and editable figures are available in this project's GitHub repository available at https://github.com/sbaker-dev/AsthmaDisease. The UK Biobank data cannot be made available but can be accessed to bona fide researchers by applying to the UK Biobank. Merged disease counts from the Registrar Generals Reports, converted as rates per 100 individuals, as well as district of birth identifiers, will be made available to other researchers through the UK Biobank.

6.3 Results

The UK Biobank contains 502,506 adults. However, participants not born in England or Wales were excluded (n = 39 488), as our exposure data was limited to England and Wales. As linkage to a district of birth is required to construct exposures, those lacking a birth coordinate were also removed from the study sample (n = 57 799). Both bacterial infections of scarlet fever and pertussis can be treated by penicillin, which was not widely available before 1946[224]. Those born before 1946 will have had a very different disease environment than those after, so our analysis excluded those born prior to 1946 (n = 122 177). This also excludes the period of the Second World War, where evacuations of children may have increased the measurement error of participants' location of birth[183]. To ensure all individuals had ten years of exposure, we excluded individuals born after 1963 (n = 50 468), as the digitised weekly returns end in 1973.

This left 232 556 participants within 1450 districts within the analysis sample. The analysis sample comprised of 128 446 (55.23%) females and 104 110 males (44.77%). Table 6-1 shows the analysis sample breakdown of for age, sex, and asthma status of the participants. Supplementary Figure 6-S1 shows the share of the population the analysis sample represents within each district to the district population estimates as of the 1951 census.

Table 6-1: Summary statistics of the analysis sample. Total sample size 232 556.

Variable	Unit	Mean	Standard deviation
Age	Years	68.72	± 5.11
Sex	Male (%)	0.45	± 0.5
Asthma Status	Asthmatic (%)	0.16	± 0.37

6.3.1 Exposures

Figures 6-1, 6-2, and 6-3 shows the mean rate and standard deviations for measles, scarlet fever, and pertussis respectively, with individual exposure in the UK Biobank assigned based on these regional measures. Figures 1-3 show that there is variation by district for each disease and that said variation changes over time, although for more common diseases such as measles, this variation is greater at. The mean rates for scarlet fever, pertussis, and measles by age do decline, although this is less the case for measles. Table 2 shows the mean and standard deviations of the exposures calculated for the UK Biobank participants from this regional data. As we average over a larger number of years, thus requiring data later in the sample, the averages do decline for scarlet fever and pertussis but less so for measles, since the measles vaccine was not widely introduced in the UK until 1968[258].

Table 6-2 shows the mean regional rate of scarlet fever, pertussis, and measles by age. The annual incidence for scarlet fever and pertussis declined over our sample periods, so as individuals age, they therefore experienced less exposure to these diseases. Whilst the exposure rate for measles does decline as participants age, since the measles vaccine was not widely introduced in the UK until 1968[237], it is far more consistent across the age range.

Table 6-2: The means and standard deviations, in parentheses, for the average exposures experienced by the ages of one, five, and ten for scarlet fever pertussis and measles.

Average Exposure by age	Scarlet Fever	Pertussis	Measles
One	0.11 (0.09)	0.22 (0.19)	0.96 (0.71)
Five	0.10 (0.06)	0.19 (0.14)	0.95 (0.26)
Ten	0.08 (0.05)	0.15 (0.11)	0.89 (0.23)

6.3.2 Early life exposures and their interplay with PGS of asthma on asthma incidence

The association of regional rates of childhood diseases and asthma incidence are shown in Figure 6-4 Panel A, and the interaction between a regional rates of childhood disease and the PGS of asthma on asthma incidence are shown in Figure 6-4, Panel B.

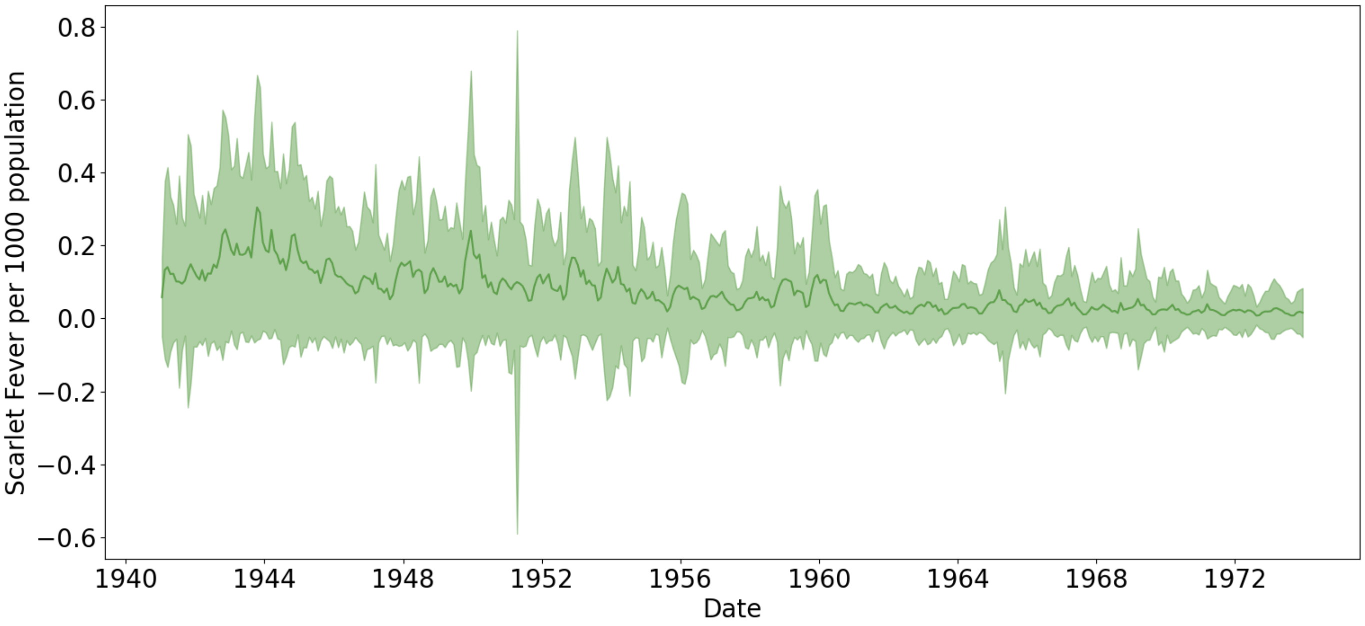


Figure 6-1: The means and standard deviations of scarlet fever rates per 1000 across 1472 districts

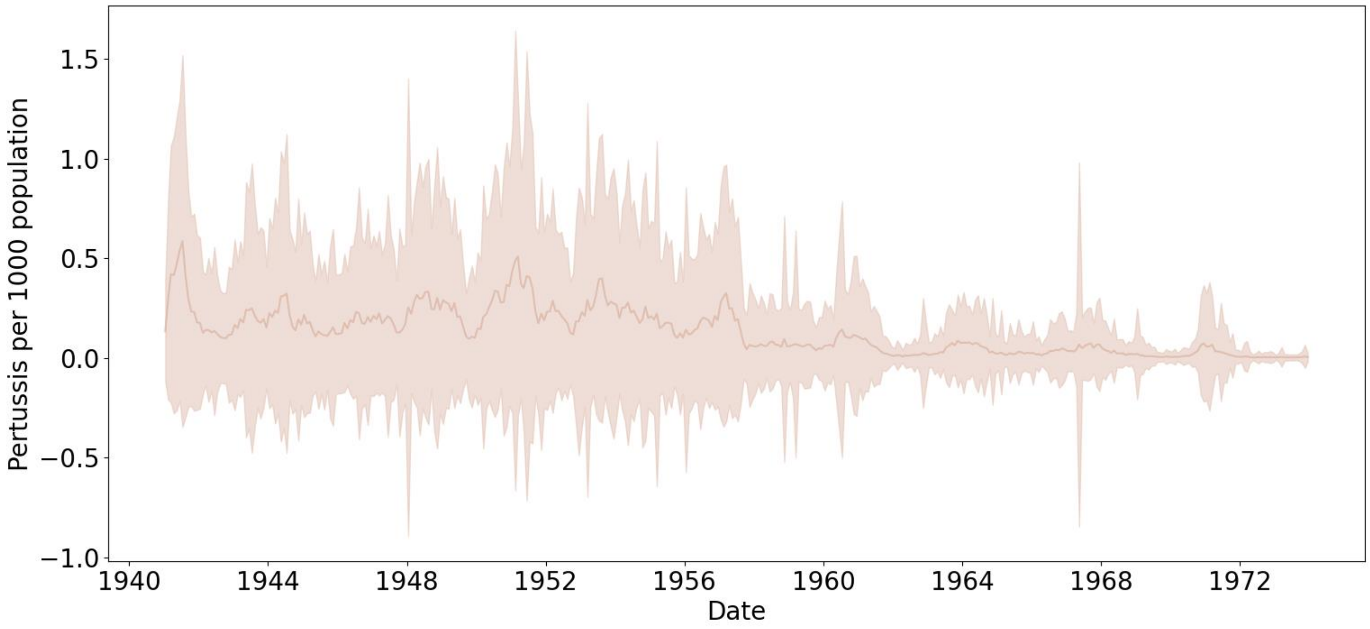


Figure 6-2: The means and standard deviations of pertussis rates per 1000 across 1472 districts

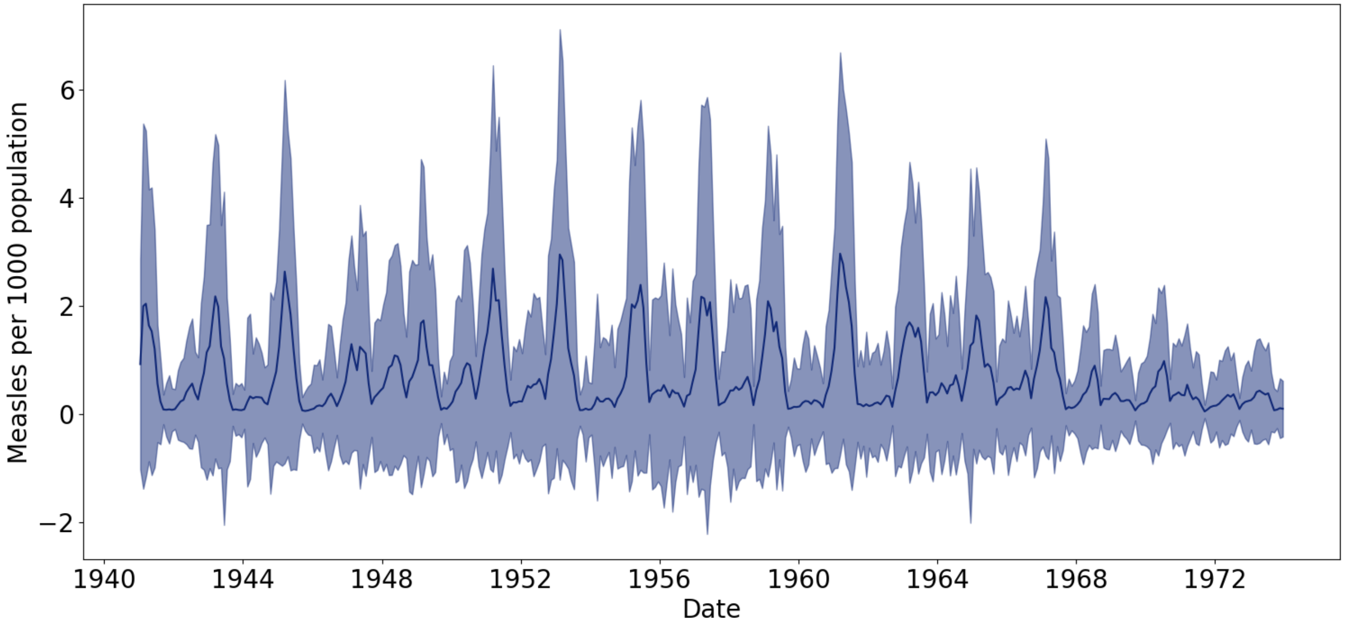


Figure 6-3: The means and standard deviations of measles rates per 1000 across 1472 districts

Finally, the association of a standard deviation increase of the asthma PGS on asthma incidence is shown in Figure 6-4, Panel C. The results are reported as odds ratios.

For the main associations between the childhood diseases and asthma shown in Figure 6-4, panel A, we found little evidence of associations between childhood diseases and asthma incidence. There was little evidence that increased rates of scarlet fever exposure at age one, five, and ten associated with incidence asthma (OR=1.21, 95%CI: 0.90; 1.61; 0.88, 95%CI: 0.50; 1.55; and 0.82, 95%CI: 0.37; 1.79 respectively). Similarly, there was little evidence that participants who experienced higher rates of pertussis had higher asthma incidence by age one, five or ten (OR=1.05, 95%CI: 0.92; 1.20; OR=1.05, 95%CI: 0.80; 1.38; OR=1.09, 95%CI: 0.74; 1.59). We found little evidence of any association between increased exposure to measles and asthma incidence.

We found evidence of a negative interaction between district level rates of scarlet fever and genetic liability for asthma (Figure 6-4, Panel B). We found evidence of an interaction between exposure to scarlet fever, and the genetic liability, on asthma incidence increased as from age one (OR=0.84, 95%CI: 0.74; 0.96), age five (OR=0.72, 95%CI: 0.60; 0.88), and age ten (OR=0.68, 95%CI: 0.54; 0.87). Similarly, we found evidence of a negative interaction between increased exposure to pertussis, and genetic liability on asthma incidence, although this was smaller than scarlet fever. Increased exposure to pertussis at age one, interacted with genetic liability, increased asthma incidence at age one (OR=0.95, 95%CI: 0.89; 1.01), age five (OR=0.92, 95%CI: 0.84; 1.00), and age ten (OR=0.87, 95%CI: 0.78; 0.97). We found limited evidence of an interaction of an exposure to measles and genetic liability, on asthma incidence.

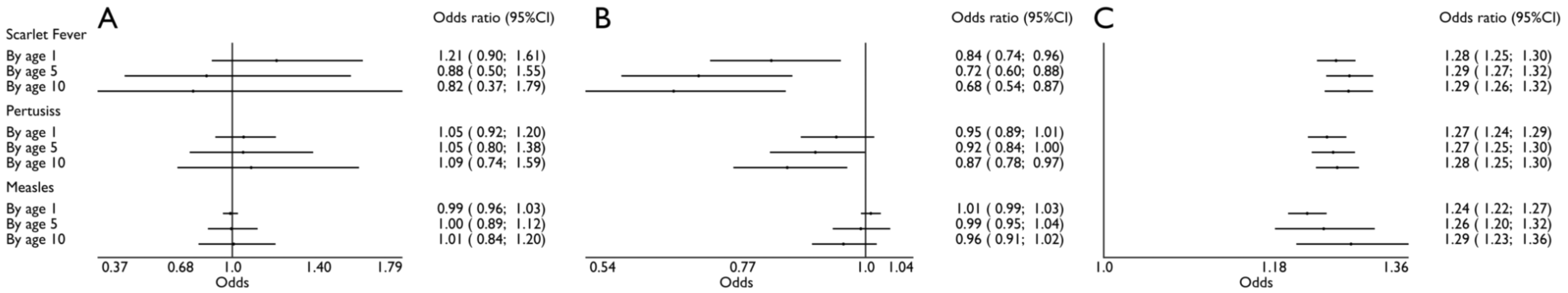


Figure 6-4: Associations of early life disease exposure rates per 100 individuals and their interplay with the PGS for asthma across childhood on asthma.

The protective associations shown in Figure 6-4, Panel B, must be made relative to the increased risk of asthma incidence participants face from having a higher PGS for asthma. Figure 6-4, Panel C, shows that an increased standard deviation of the PGS for asthma increased the risk of asthma by between 25-29%. Those exposed to scarlet fever or pertussis have a reduced probability of developing asthma, compared to those with low exposures of these diseases and the same PGS. However, those exposed to higher rates of scarlet fever or pertussis still have a higher risk than those exposed to said diseases, but with a lower PGS.

In a sensitivity we repeated the analysis comparing the upper and lower quartiles of genetic susceptibility (Supplementary Figure 6-S2 and 6-S3). For those in the highest genetic risk quartile, we found a similar but weaker relationship to Figure 6-4, Panel B for both scarlet fever and pertussis and still no association for measles. Conversely, for those in the lowest genetic ricks quartile, we found little evidence of any protective association from exposures to scarlet fever, pertussis, or measles.

As we have investigated the impact of three different exposures to asthma, it is appropriate that we undertake multiple hypothesis testing, which we do through the Bonferroni correction. As each exposure of interest (by age 1, 5 and 10) is nested, we are testing three different hypotheses. The pre and post corrected P values are shown in Table 6-3. Our results remain robust except the interaction effect between pertussis and the polygenic score for asthma by age 5, which is no longer significant.

6.4 Discussion

In this large prospective cohort study, we examined if increase exposure to childhood disease was associated with later life risk of asthma, and if this risk differed at different levels of genetic risk. There have been a few studies investigating the direct impact of the disease we investigate on asthma incidence. Previous studies suggested that scarlet fever was protective using national trend data[264], pertussis to exacerbate allergic airway inflammation[146], and measles having no associated protection[148]. Our analysis

Table 6-3: P values for exposure of scarlet fever, pertussis, and measles by age 1, 5, and 10, the polygenic score, and the interaction between the two, to asthma before and after Bonferroni correction. As exposures are nested, each exposure group is corrected for 3 hypotheses.

	By age 1		By age 5		By age 10	
Exposures	P	Bonferroni P	P	Bonferroni P	P	Bonferroni P
Scarlet fever						
Direct effect	0.202	0.606	0.659	1.000	0.615	1.000
Polygenic score	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
Interaction	0.009	0.028	0.001	0.003	0.002	0.006
Pertussis						
Direct effect	0.453	1.000	0.721	1.000	0.663	1.000
Polygenic score	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
Interaction	0.13	0.39	0.05	0.14	0.01	0.04
Measles						
Direct effect	0.608	1.000	0.938	1.000	0.946	1.000
Polygenic score	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
Interaction	0.304	0.913	0.746	1.000	0.189	0.567

found no evidence of a direct association between the disease environment and increased risk of asthma. As such, for a large proportion of the population, the declining rates of infectious diseases are unlikely to explain the rise in asthma incidence.

However, our paper is the first, to our knowledge, to investigate this question further by interacting early life exposure with genetic risk. For those with increased genetic risk of asthma, we found an association between increased exposure and reduced risk of developing asthma for both scarlet fever and pertussis. However, much of our results for pertussis are not robust to multi-hypothesis testing. However, for a those predisposed to be at risk of asthma, there is some evidence of an association that the changing disease environment may explain part of the rise in asthma prevalence over the latter 20th century.

Whilst our method of using district level data does result in less specific exposures than if measured individually, said individual level exposures may suffer from individual level confounding due to socioeconomic position. Instead, our regional exposure rates provide a summary measure of the environment within each district for participants exposed at a particular age. As such, this reduces socioeconomic confounding, with differences between districts controlled for by fixed effects. By controlling for the cohort year, we are also only comparing outcomes of participants with exposures with the outcomes of those born in the same local area but having been exposed to at a different rate. Hence, a strength of our modelling approach is that it limits the role of individual level confounding factors that can be challenging to address.

Many environments changed over the later 20th century at the same time such as diet, sanitation, pollution, and antibiotic use[148]. It is possible that the changing nature of diseases is also capturing other changing environments that we do not explicitly control for, which is a weakness of this study. Many of these, such as antibiotic treatment, are difficult if not impossible to control for at an individual level. Therefore, it is possible our diseases environment is confounded by these unobserved environments that may play equal or greater roles. To attempt to limit this bias, we have undertaken several steps. We controlled for individuals' year of birth, which captures general changes in both observed and unobserved environments that occur year-on-year. We further limited our study to only those born after introduction of antibiotics, to remove any potential of availability and exposure to them. As we lack a measure of individual-level antibiotic use in childhood within the UK Biobank, this is the best we can presently do with the data available.

The UK Biobank lacks detailed information on location of residence throughout the participants' lives, limiting this information to place of birth and residence at or after joining the study. Since individuals are assigned disease exposures that are derived from regional measures of disease incidence rates, we assumed that individuals have not moved by the age of ten. Any residential mobility up to the age of ten will have led to measurement error which may have attenuated our estimates. Before our sampling period, a historical estimate for annual residential mobility of 0–19-year-olds born between 1938-1947 was estimated to around 7%[220]. Therefore, with each additional year, the likelihood of measurement of individuals moving out of the district increases. However, this annual estimate is inclusive of those born during the war years, so is likely to not be representative of a post-war Britain.

UK Biobank participants tend to be healthier, leaner, smoke less, suffer from less disease, be older, and live in more socioeconomically advantaged areas which also results in selection bias[228]. This selection bias reduces the risk of infection of childhood diseases, due to the inverse correlation between wealth and disease exposure[263], so our results are likely to suffer from downward bias. They also have a higher proportion of asthmatics, with 16% of the sample having asthma compared to only 12% of the population[265]. This could in part be due to the higher density of individuals within the UK Biobank living in urban areas, which tend to have higher prevalence of asthma than rural areas[237]. It may also be, for those with milder symptoms, be that they were more likely to go to a doctor for a diagnosis.

This study used WeightGIS to create time invariant locations to ensure district level characteristics at a population level remained broadly constant. However, whilst weighting using 17,000 sub-locations in England and Wales should reduce weighting imprecision, a degree of measurement error will be caused using these methods. This may result in individuals being assigned higher or lower exposures than they would otherwise have experienced. As individuals in our sample are predominately born in urban areas, which proportionally change less, any measurement errors impact on our estimates as a result of the use of WeightGIS should be minimised.

Whilst our analysis is not perfect, it has replicated an interesting hypothesis in principle, that exposure to disease may be protective for certain subgroups. Should the lab results be further replicated, and IgG from infection be proven to be protective against asthma, then the policy implications are complex. The genetic component is fixed, and cannot be changed, therefore, the only thing that can be changed is the environment. However, we don't want to suggest that children should purposely get infected with unattenuated disease, as this will carry its own risk. Therefore, only if an attenuated bacterium could generate the regulatory response would it be implementable, as then it would be a treatment.

Given the polygenic nature of asthma, it is not possible to just screen for individual SNPs as predictors of asthma, as would be done for Mendelian traits like Huntington's[246]. Mass screening and storing genetic data has its risks and ethical concerns, see[266] for a review. As this is undesirable, the consideration for a policymaker comes to a trade-off between the potential benefits of treatment and costs. The associational results we presented suggest that there was no consequence for those who were not genetically liable to exposure. Universal treatment offers the maximum possible protection, but much of the financial cost for non-genetically liable individuals may be considered waste. It would be down to the policymaker to evaluate the benefits of maximising the reduction in asthma incidence against the potential financial cost of treatment.

If the cost is too high, then the alternative is to use family history and apply to those considered at risk. Asthma is highly heritable at around 82%.[28] Around 25% and 50% of the children with a single or both parents respectively having asthma, develop asthma themselves[251]. Therefore, determining whom should undergo treatment based on genetic susceptibility can be done through investigating parental and recent family history of asthma incidence. Whilst the coverage of this policy would be lower, so too will the cost, as only those most likely to benefit are the ones who are treated.

There remains the possibility that reduce exposure to childhood diseases may in part be responsible for the large increases in asthma incidence experienced in the latter half of the 20th century. Our analysis shows that for those who were at the greatest risk for asthma genetically, exposure to scarlet fever or pertussis was associated with reduced incidence of asthma. Crucially, these individuals are still at a greater level of risk, comparably, than those exposed to childhood diseases with lower genetic risk. However, our sensitivity analysis on the uppermost quartile of genetic risk found reduced results, which suggests that the protective association is not limited only to those of the highest risk. Therefore, whilst exposing children to infectious diseases will not lead to reductions in overall asthma incidence, it may in part explain this pattern for some of the most vulnerable. More research focus needs to be placed in exploring the potential for mechanisms that may exist, potentially exclusively, for those of the highest genetic susceptibility to diseases, such as asthma.

6.5 Supplementary materials

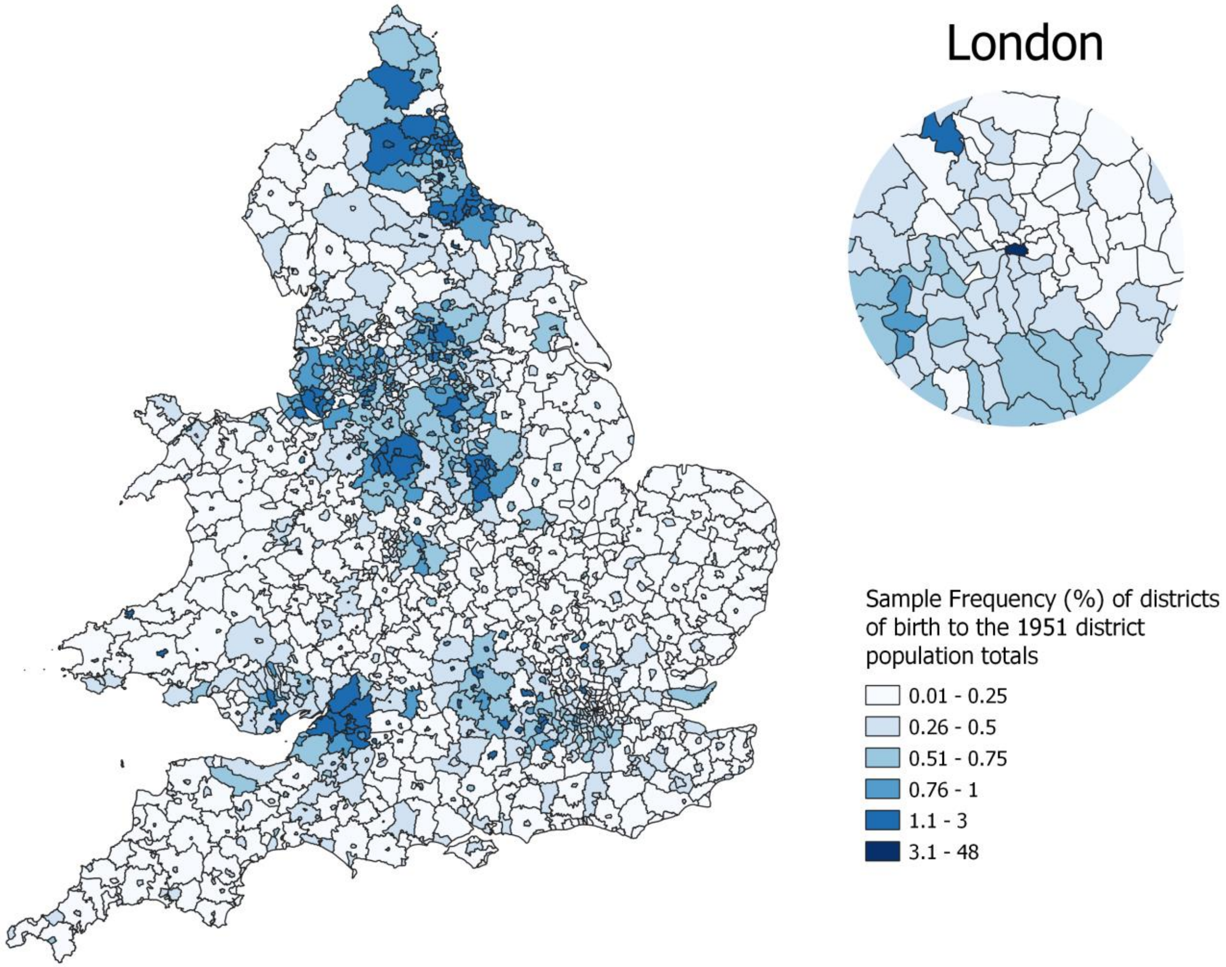

London
Sample Frequency (%) of districts of birth to the 1951 district population totals
0.01 - 0.25
0.26 - 0.5
0.51 - 0.75
0.76 - 1
1.1 - 3
3.1 - 48

Figure 6-5: (6-S1) The share of the population covered by our analysis sample from the UK Biobank relative to the population totals of each district within 1951

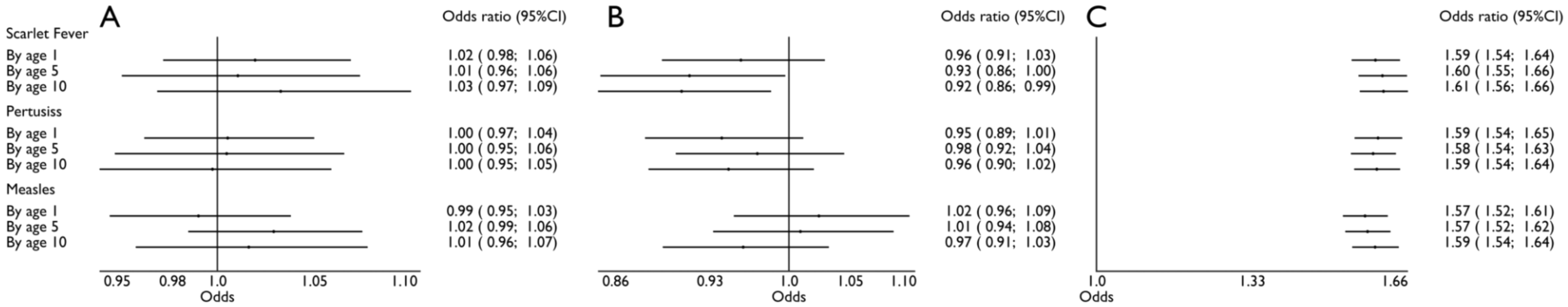


Figure 6-6: (6-S2) Associations of early life disease exposure rates per 100 individuals and their interplay with the PGS for asthma across childhood on asthma for those with a polygenic risk in the uppermost quartile.

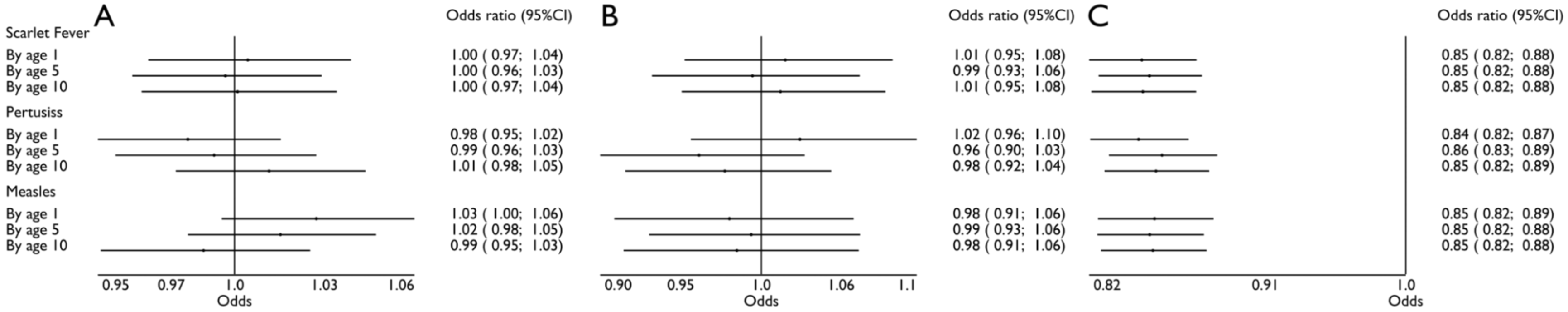


Figure 6-7: (6-S3) Associations of early life disease exposure rates per 100 individuals and their interplay with the PGS for asthma across childhood on asthma for those with a polygenic risk in the lowermost quartile.

7 Concluding remarks

The thesis has made multiple contributions to both the scientific literature. One its major contributions is the construction of new data. This allows for future contributions to the literature when this data is made public and has allowed this thesis to contain research that is distinctive from others in the literature. Availability is a key part of data quality, with collaboration and subsequent validation of research through data sharing crucial for rapid advancement in the literature[267]. However, data that fails to be usable, reliable, relevant, and readable will at best fail to facilitate such advancement[268], and at worst may even lead to erroneous findings. Whist the data will be made public, it will only be released when it meets all the criteria of good quality data.

Creation of said data required digitisation but using currently available methods would have been extremely time consuming and costly. The European Central Commission estimated that the cost of digitising Europe's cultural heritage would be upwards of 100 billion Euros, driven by the fact that costs of digitisation are not insignificant[269]. The money saved from using ArchiveOCR to process 40,000 tables of notifiable diseases is difficult to estimate, as the price per page varies based on the combination of scale, scope, and skills. However, an estimated cost can through establishment of a few assumptions. Any costs of scanning historical documentation are ignored, and it is assumed that it takes 15 minutes to fully digitisation and clean a page using commercially available OCR software. Using the minimum hourly wage of 2021 of 8.91 per hour, and if an individual works 8 hours a day, it would cost £89,100. It would also have taken three and a half years to process all 40,000 tables. Instead of spending the duration of my PhD constructing this resource, this thesis has created new tools to digitise historical records, digitised the diseases notifications and many other resources, in a fraction of the time and cost.

Without WeightGIS to standardise the data that has been produced, limited amounts could have been utilised for research. Most districts change over the 44 years of 1931 to 1974, limiting capacity to comparing within locations. Studies that have used small parts of the disease notifications have in the past simply had to drop any district that changed[101]. Dealing with boundary changes has proven historically to be significantly challenging[48], but with WeightGIS, any data from districts between 1931 and 1974 can be now standardised and used across time. Using GIS has also allowed for the data within BIO-HGIS to be linked to the UK Biobank. Linkage to the UK Biobank allows for studies using regionally measured exposures and environments, to be used to estimate individual level outcomes. This allows for a much broader reach and impact of research.

Streptococcus pyogenes has, within recent years, mutated into a new strain of M1T1, which is closer to its historical counterparts from the early parts of the 20th century. Should there be further mutations, there is a risk not just in a spike in childhood mortality but also increased latent consequences for survivors. This is in part due to the heterogeneity in immune responses, which allows at a population level for variation in outcomes from the exposure. Whilst an increasing number of long-term health outcomes have been established, many more have yet to be fully investigated. Within this thesis, we were unable to find robust associations between increased exposure to scarlet fever in childhood increased risk of heart disease outcomes. Given this established risk to rheumatic heart disease there is considerable concern that for some research outcomes, the underlying bias may still make the data less appealing. Despite this, we did still find some weak association with increased exposure to scarlet fever and reduced later life fluid intelligence. These associational findings are some of the first that we know of. If new strains of streptococcus pyogenes continue, and streptococcal disease become more prevalent, the consequences of infections such as scarlet fever may be broader than previously thought, but more research is required as where unable to prove anything causally here.

Whilst this thesis has not been able to causally prove why asthma cases have increased, it utilised a gene-environmental interaction to show the importance of both elements. Changes to the environment have frequently been associated to increase in asthma instances[135,232], but have failed to account for the genetic makeup of their sample. However, in order for a gene-environmental interaction study to be constructed, a cohort study with geneotyped individuals, and a sufficiently detailed list of exposures in early life is required. This does not currently exist, but by utilising BIO-HGIS we were able to modify the UK Biobank to allow for this study.

Within our study in Chapter 6 we found that despite diseases frequently being dismissed as a potential cause to rise in asthma[148], that when considering the innate genetic risk of participants, that the changing disease environment was associated to increased asthma risk. However, our results also contradict an older hypothesis, simply suggesting that child with higher exposures were less likely to have atopic disease[138]. Those at genetic risk were associated with less risk of later life asthma with exposure, but those with little genetic risk stood to gain little from exposure to childhood diseases. Studies must strive to consider that associations from exposures or early life circumstances may hide effects for genetic sub-groups of the population that may not be well represented, or over-represented, in that study sample.

In all, this thesis has striven to collect, digitised, standardise, innovate, and deliver data, software, methods, and research. Taken as a whole, this thesis represents a conclusion, but only to the prelude of the true project ahead. The methods developed within this thesis will be used to protect more of our past, and us it to try to inform those in the present of the potential future of their past actions if they remain unchanged. The skills learned to undertake research will be used to further seek to research ways of utilising our past to reduce the inequalities of health, place and birth. There is significant more work to be undertaken. However, when the data is finally made public, it is hoped that the research communities of multiple disciplines can explore our past to try to help further the research body as a whole.

7.1 Future Work

Whilst ArchiveOCR is still within an alpha stage and not yet available, it represents another significant contribution that will become part of future work. Considerable time constraints have left much to be desired, and to truly be accessible to all, it requires a much cleaner front end graphical user interface. Whilst the code base is simple to use, with only two main command calls, the number of arguments make it unwieldy for newer or less experienced users. Future work will focus not just on updating the code base to a 2022 standard, but also to focus on the accessibility aims of the project. In doing so, it is hoped that ArchiveOCR can be used as tool for others to protect their own past and heritage. Whilst WeightGIS has been publicly accessible for a while, it requires proper user documentation to ensure it can be used by others and that the findings within this thesis can be replicated.

BIO-HGIS requires extensive work before being made public, but work has started on construction of the front and back end of a web application for data investigation. However, extensive user documentation still is required for it be of much use for external researchers. Many data sources also require further digitisation or standardisation before they are complete. In the coming year it is hoped that all current digitisation projects can be finalised, so that the data can begin to be made public as the chapters go for publication. However, additional resources, such as weekly infant mortality, may also be produced depending on time and funding commitments.

However, this thesis has already produced more data than I could possibly utilise. Although, that is not to say that hypotheses have not been constructed for the wealth of data. Many of these hypotheses were used as justification for digitisation in the first place, many of which are shown within Chapter 3. The aim is to produce a paper for each of the major notifiable diseases, releasing the data from that disease after publication. Currently, scarlet fever, pertussis, measles, and pneumonia have all be utilised within research in this thesis. The next paper will seek investigate if exposure to non-paralytic polio, assumed to be self-limiting, still had the potential for later life declines in muscle functionality, white matter, or bone density in addition to increased BMI. Similar to Chapter 5, there is also a potential gene-environment interaction to exploit utilising the polygenic score for tuberculosis and exposure to tuberculosis from the disease notifications. However, unlike scarlet fever, pertussis, and measles the rate of tuberculosis is much lower, which will reduce the power to detect meaningful effects.

Whilst the Blitz data was constructed within 2020, I never managed to utilise it myself. Given current European events as of March 2022, further investigating the potential consequences of exposure air raids on increased risk of later life stillbirths and miscarriage has become unfortunately more relevant than ever. The other current focus is constructing an alternative to the Townsend index that can be used across the sample period of BIO-HGIS. Whilst district fixed effects assist in capturing time-invariant differences, the impact of deprivation is difficult to untangle. Digitisation of monthly unemployment data from 1931-1974 will allow for the construction of an index of deprivation. Given the Townsend index in 1971 was 92.4% correlated to unemployment[203] this will allow for a measure, even if imperfect, of capturing said inequalities. We then hope to use this index for a research paper, utilising it for a PheWAS, to show how inequalities from location of birth are associated with declines in later life health and well-being.

The proposals and chapters within this thesis represented a select few of an ever-increasing scope of research that could be undertaken. The hope going forward is that it will be increasingly possible to work collaboratively with pre-released data on research papers, to speed up the time to publication and eventual release of the data to public. For long term future work, there is a desire to set up a research group focused on protecting historical documentation. With a team of individuals, projects that are currently unfeasible for one individual to process, can be protected and hopefully used for research and public good for years and decades to come.

8 Bibliography

1. Almond D, Currie J. Killing me softly: The fetal origins hypothesis. *Journal of Economic Perspectives*. 2011;25(3):153-172. doi:10.1257/jep.25.3.153

2. Barker DJP, Osmond C. Infant Mortality, Childhood Nutrition, And Ischaemic Heart Disease In England And Wales. *The Lancet*. 1986;327(8489):1077-1081. doi:10.1016/S0140-6736(86)91340-1

3. Paneth N, Susser M. Early origin of coronary heart disease (the "Barker hypothesis"). *BMJ : British Medical Journal*. 1995;310(6977):411. doi:10.1136/BMJ.310.6977.411

4. Schulz LC. The Dutch Hunger Winter and the developmental origins of health and disease. *Proceedings of the National Academy of Sciences*. 2010;107(39):16757-16758. doi:10.1073/PNAS.1012911107

5. Stein Z, Susser M. The Dutch Famine, 1944–1945, and the Reproductive Process. I. Effects on Six Indices at Birth. *Pediatric Research*. 1975;9(2):70-76. doi:10.1203/00006450-197502000-00003

6. Stein Z, Susser M, Saenger G, Marolla F. Nutrition and Mental Performance. *Science*. 1972;178(4062):708-713. doi:10.1126/science.178.4062.708

7. Grajalez CG, Magnello E, Woods R, Champkin J. Great moments in statistics. *significance*. 2013;10(6):21-28. doi:10.1111/J.1740-9713.2013.00706.X

8. Wadsworth M, Kuh D, Richards M, Hardy R. Cohort Profile: The 1946 National Birth Cohort (MRC National Survey of Health and Development). *International Journal of Epidemiology*. 2006;35(1):49-54. doi:10.1093/IJE/DYI201

9. Power C, Elliott J. COHORT PROFILE: 1958 British birth cohort (National Child Development Study). *International Journal of Epidemiology*. 2006;35:34-41. doi:10.1093/ije/dyi183

10. Elliott J, Shepherd P. Cohort Profile: 1970 British Birth Cohort (BCS70). *International Journal of Epidemiology*. 2006;35(4):836-843. doi:10.1093/IJE/DYL174

11. Bycroft C, Freeman C, Petkova D, et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature 2018 562:7726*. 2018;562(7726):203-209. doi:10.1038/s41586-018-0579-z

12. Littlejohns TJ, Sudlow C, Allen NE, Collins R. UK Biobank: opportunities for cardiovascular research. *European Heart Journal*. 2019;40(14):1158-1166. doi:10.1093/eurheartj/ehx254

13. The National Archives. Digitised records - Freedom of Information. Published online 2020. Accessed March 11, 2020.

https://www.nationalarchives.gov.uk/about/freedom-of-information/information-requests/digitised-records-2/

14. Zhang AB, Gourley D. Planning and managing digitisation projects. In: *Creating Digital Collections*. Elsevier; 2009:7-17. doi:10.1016/b978-1-84334-396-7.50002-x

15. Zhang AB, Gourley D. Digitising material. In: *Creating Digital Collections*. Elsevier; 2009:55-72. doi:10.1016/b978-1-84334-396-7.50005-5

16. Mohammad F, Mohammad F, Anarase J, Shingote M, Ghanwat P. Optical Character Recognition Implementation Using Pattern Matching. *International Journal of Computer Science and Information Technologies*. 2014;5(2):2088-2090. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.661.1089

17. Goodchild MF, Anselin L, Deichmann U. A Framework for the Areal Interpolation of Socioeconomic Data. *Environment and Planning A: Economy and Space*. 1993;25(3):383-397. doi:10.1068/a250383

18. Martin D, Dorling D, Mitchell R. Weight In Infancy And Death From Ischaemic Heart Disease. *Area*. 2002;34(1):82-91. doi:10.1111/1475-4762.00059

19. Barker DJ, Osmond C. Death rates from stroke in England and Wales predicted from past maternal mortality. *British medical journal (Clinical research ed)*. 1987;295(6590):83-86. doi:10.1136/bmj.295.6590.83

20. Barker DJP, Osmond C, Law CM. The intrauterine and early postnatal origins of cardiovascular disease and chronic bronchitis. *Journal of Epidemiology and Community Health*. 1989;43(3):237. doi:10.1136/JECH.43.3.237

21. Barker DJP, Osmond C, Winter PD, Margetts B, Simmonds SJ. WEIGHT IN INFANCY AND DEATH FROM ISCHAEMIC HEART DISEASE. *The Lancet*. 1989;334(8663):577-580. doi:10.1016/S0140-6736(89)90710-1

22. Holley R. How Good Can It Get? *National Library of Astralia D-Lib Magazine*. 2009;15(3). http://www.dlib.org/dlib/march09/holley/03holley.html

23. Parliament: House of Commons. Local Government Act 1972 Chapter 70. Published online 1972. Accessed September 8, 2021. https://www.legislation.gov.uk/ukpga/1972/70

24. Netrdová P, Nosek V, Hurbánek P. Using Areal Interpolation to Deal with Differing Regional Structures in International Research. *ISPRS International Journal of Geo-Information 2020, Vol 9, Page 126*. 2020;9(2):126. doi:10.3390/IJGI9020126

25. James LK, Till SJ. Potential Mechanisms for IgG4 Inhibition of Immediate Hypersensitivity Reactions. 2016;16:1-7. doi:10.1007/s11882-016-0600-2

26. Bugajev V, Halova I, Draberova L, et al. Negative regulatory roles of ORMDL3 in the FcϵRI-triggered expression of proinflammatory mediators and chemotactic response in murine mast cells. *Cellular and Molecular Life Sciences*. 2016;73(6):1265-1285. doi:10.1007/s00018-015-2047-3

27. Froidure A, Mouthuy J, Durham SR, Chanez P, Sibille Y, Pilette C. Asthma phenotypes and IgE responses. *European Respiratory Journal*. 2016;47(1):304-319. doi:10.1183/13993003.01824-2014

28. Ullemar V, Magnusson PKE, Lundholm C, et al. Heritability and confirmation of genetic association studies for childhood asthma in twins. *Allergy: European Journal of Allergy and Clinical Immunology*. 2016;71(2):230-238. doi:10.1111/all.12783

29. Reul C, Springmann U, Wick C, Puppe F. State of the Art Optical Character Recognition of 19th Century Fraktur Scripts using Open Source Engines. Published online 2018. https://www.abbyy.com

30. Drobac S, Lindén K. Optical character recognition with neural networks and post-correction with finite state methods. *International Journal on Document Analysis and Recognition (IJDAR)*. 2020;23:279-295. doi:10.1007/s10032-020-00359-9

31. Wick C, Reul C, Puppe F. Calamari − A High-Performance Tensorflow-based Deep Learning Package for Optical Character Recognition. Published online 2018. https://github.com/Calamari-OCR

32. Breuel TM. The OCRopus Open Source OCR System. In: *Document Recognition and Retrieval Xv*. Proceedings of SPIE; 2008.

33. Thammarak K, Kongkla P, Sirisathitkul Y, Intakosum S. Comparative analysis of Tesseract and Google Cloud Vision for Thai vehicle registration certificate. *International Journal of Electrical and Computer Engineering (IJECE)*. 2022;12(2):1849-1858. doi:10.11591/IJECE.V12I2.PP1849-1858

34. Sporici D, Cușnir E, Boiangiu C-A. Improving the Accuracy of Tesseract 4.0 OCR Engine Using Convolution-Based Preprocessing. *Symmetry*. 2020;12(5). doi:10.3390/sym12050715

35. Wu Y, Kirillov A, Massa F, Lo W-Y, Girshick R. Detectron2. Published online 2019. Accessed October 15, 2022. https://github.com/facebookresearch/detectron2

36. Amujala S, Vossmeyer A, Das SR. Digitization and data frames for card index records. *Explorations in Economic History*. Published online July 2022:101469. doi:10.1016/J.EEH.2022.101469

37. Shen Z, Zhang R, Dell M, et al. LayoutParser: A Unified Toolkit for Deep Learning Based Document Image Analysis. Published online 2021. https://github.com/BobLd/DocumentLayoutAnalysis

38. Causer T, Terras M. Crowdsourcing Bentham: Beyond the Traditional Boundaries of Academic History. *https://doiorg/103366/ijhac20140119*. 2014;8(1):46-64. doi:10.3366/IJHAC.2014.0119

39. Face H. The AI community building the future. Published online 2022. Accessed October 15, 2022. https://huggingface.co/

40. Bradski G. The OpenCV Library. *Dr Dobb*. 2000;39(1). https://www.drdobbs.com/open-source/the-opencv-library/184404319

41. Harris CR, Millman KJ, Walt SJ van der, et al. Array programming with NumPy. *Nature 2020 585:7825*. 2020;585(7825):357-362. doi:10.1038/s41586-020-2649-2

42. Office GR. *Registrar-General's Weekly Return (1941-71)*. Her Majesty's Stationery Office; 1941.

43. Mukhopadhyay P, Chaudhuri BB. A survey of Hough Transform. *Pattern Recognition*. 2015;48(3):993-1010. doi:10.1016/J.PATCOG.2014.08.027

44. Zhang P. Data Communications in Distributed Control System. *Industrial Control Technology*. Published online January 2008:675-774. doi:10.1016/B978-081551571-5.50007-4

45. Raggo M, Hosmer C. Apple iOS Data Hiding. *Data Hiding*. Published online January 2013:107-131. doi:10.1016/B978-1-59-749743-5.00006-7

46. Booth J-MJ, Gelb J. *Optimizing OCR Accuracy on Older Documents : A Study of Scan Mode , File Enhancement, and Software Products*. Office of Innovation; New Technology U.S. Government Printing Office, Washington, DC; 2006. https://www.semanticscholar.org/paper/Optimizing-OCR-Accuracy-on-Older-Documents-{\%}3A-A-of-{\%}2C-Booth-Gelb/https://www.govinfo.gov/media/WhitePaper-OptimizingOCRAccuracy.pdf

47. Wang Y, Di Q. Modifiable areal unit problem and environmental factors of COVID-19 outbreak. *Science of The Total Environment*. 2020;740:139984. doi:10.1016/J.SCITOTENV.2020.139984

48. Gregory IN, Ell PS. Breaking the Boundaries: Geographical Approaches to Integrating 200 Years of the Census on JSTOR. *Journal of the Royal Statistical Society Series A (Statistics in Society)*. 2005;168(2):419-437. https://www.jstor.org/stable/3559970?seq=2{\#}metadata{_}info{_}tab{_}contents

49. Mennis J. Problems of Scale and Zoning. *Geographic Information Science & Technology Body of Knowledge*. 2019;2019(Q1). doi:10.22224/GISTBOK/2019.1.2

50. Ye X, Rogerson P. The Impacts of the Modifiable Areal Unit Problem (MAUP) on Omission Error. *Geographical Analysis*. 2022;54:32-57. doi:10.1111/gean.12269

51. Blake M, Bell M, Rees P. Creating a temporally consistent spatial framework for the analysis of inter- regional migration in Australia. *INTERNATIONAL JOURNAL OF POPULATION GEOGRAPHY*. 2000;6(1):155-174. https://onlinelibrary.wiley.com/doi/epdf/10.1002/{\%}28SICI{\%}291099-1220{\%}28200003/04{\%}296{\%}3A2{\%}3C155{\%}3A{\%}3AAID-IJPG180{\%}3E3.0.CO{\%}3B2-A?saml{_}referrer

52. Vrieling A, Melser C. Constructing boundary-consistent population time series for the municipalities of the Netherlands, 1988-2011. *Population Studies*. 2013;67(2):195-208. doi:10.1080/00324728.2012.754049

53. Syphard AD, Stewart SI, Mckeefry J, et al. Assessing housing growth when census boundaries change. *https://doiorg/101080/13658810802359877*. 2009;23(7):859-876. doi:10.1080/13658810802359877

54. Cromley RG, Ebenstein AY, Hanink DM. Estimating components of population change from census data for incongruent spatial/ temporal units and attributes. *http://dxdoiorg/101080/1449859620099635180*. 2010;54(2):89-99. doi:10.1080/14498596.2009.9635180

55. Wu Y, Furuya S, Wang Z, Nobles JE, Fletcher JM, Lu Q. GWAS on birth year infant mortality rates provides evidence of recent natural selection. *Proceedings of the National Academy of Sciences*. 2022;119(12). doi:10.1073/PNAS.2117312119

56. Baker S, Biroli P, Kippersluis H van, Hinke S von. Beyond Barker: Infant Mortality at Birth and Ischaemic Heart Disease in Older Age. Published online May 2022. doi:10.48550/arxiv.2205.06161

57. Langford M, Maguirem D, Unwin D. The areal interpolation problem: Estimating population using remote sensing in a GIS framework. In: Masser I, Blakemore M, eds. *Handling Geographical Information: Methodology and Potential Applications*. Longman; 1991:55-77.

58. Briggs DJ, Gulliver J, Fecht D, Vienneau DM. Dasymetric modelling of small-area population distribution using land cover and light emissions data. *Remote Sensing of Environment*. 2007;108(4):451-466. doi:10.1016/J.RSE.2006.11.020

59. Qiu F, Woller KL, Briggs R. Modeling urban population growth from remotely sensed imagery and TIGER GIS road data. *Photogrammetric Engineering and Remote Sensing*. 2003;69(9):1031-1042. doi:10.14358/PERS.69.9.1031

60. Schürer K, Day J. Social History Migration to London and the development of the north-south divide, 1851-1911. Published online 2019. doi:10.1080/03071022.2019.1545361

61. Lasker B. BUREAU OF LABOR STATISTICS THE BRITISH SYSTEM OF LABOR EXCHANGES. *US Department of Labor*. Published online 1916. http://fraser.stlouisfed.org/

62. Lawhead J. pyshp PyPI. Published online September 2011. Accessed March 20, 2020. https://pypi.org/project/pyshp/

63. Baker S. shapeObject PyPI. Published online 2020. Accessed October 29, 2021. https://pypi.org/project/shapeObject/

64. Gillies S. Shapely PyPI. Published online October 2007. Accessed March 20, 2020. https://pypi.org/project/Shapely/

65. Gregory I, Dorling D, Southall H. A century of inequality in England and Wales using standardized geographical units. *Area*. 2001;33(3):297-311. doi:10.1111/1475-4762.00033

66. Beall J. Geographical research and the problem of variant place names in digitized books and other full-text resources. *https://doiorg/101080/14649055201010766263*. 2013;34(2-3):74-82. doi:10.1080/14649055.2010.10766263

67. Project GBHG. *Great Britain Historical GIS*. University of Portsmouth; 2017.

68. Davis KAS, Bashford O, Jewell A, et al. Using data linkage to electronic patient records to assess the validity of selected mental health diagnoses in English Hospital Episode Statistics (HES). *PLoS ONE*. 2018;13(3). doi:10.1371/journal.pone.0195002

69. Biobank U. *UK Biobank Deriving the grid coordinates.*; 2012. https://biobank.ndph.ox.ac.uk/showcase/showcase/docs/UKgrid.pdf

70. Darby HC, Glasscock RE, Sheail J, Versey GR. The changing geographical distribution of wealth in England: 1086–1334–1525. *Journal of Historical Geography*. 1979;5(3):247-262. doi:10.1016/0305-7488(79)90071-9

71. Britain V of. Administrative Units Typology | Type definition: Local Government District. Published online 2017. Accessed October 14, 2021. https://www.visionofbritain.org.uk/types/type/LG{_}DIST

72. Francis LJ, Lankshear DW. The rural rectory: The impact of a resident Priest on local church life. *Journal of Rural Studies*. 1992;8(1):97-103. doi:10.1016/0743-0167(92)90033-3

73. Torrance D. Introduction to devolution in the UK. *House of Commons Library*. 2019;CBP 8599.

74. Ell P, Garrett EM, Galley C, Southall HR, Mooney G. Great Britain Historical Database : Health and Health Care Data : Mortality Statistics 1851-1973 [data collection]. *UK Data Service*. 2020;1(1). https://beta.ukdataservice.ac.uk/datacatalogue/studies/study?id=4570{\#}!/details

75. Bruno RL. Paralytic vs. "Nonparalytic" Polio. *American Journal of Physical Medicine & Rehabilitation*. 2000;79(1):4-12. doi:10.1097/00002060-200001000-00003

76. Sika-Paotonu D, Beaton A, Raghu A, Steer A, Carapetis JR. Rheumatic Fever and Rheumatic Heart Disease. In: *Streptococcus Pyogenes : Basic Biology to Clinical Manifestations*. Elsevier Inc.; 2016:357-362. doi:10.1016/B978-0-12-803678-5.00385-4

77. Mina MJ, Metcalf CJE, Swart RL de, Osterhaus ADME, Grenfell BT. Long-term measles-induced immunomodulation increases overall childhood infectious disease mortality. *Science*. 2015;348(6235):694-699. doi:10.1126/science.aaa3662

78. Singh H, Parakh A, Basu S, Rath B. Plasmodium vivax malaria: Is it actually benign? *Journal of Infection and Public Health*. 2011;4(2):91-95. doi:10.1016/J.JIPH.2011.03.002

79. Ajetunmobi WA, Orimadegun AE, Brown BJ, et al. Haemoglobinuria among children with severe malaria attending tertiary care in Ibadan, Nigeria. *Malaria Journal*. 2012;11(336). doi:10.1186/1475-2875-11-336

80. Grimwood K, Chang AB. Long-term effects of pneumonia in young children. *Pneumonia 2015 6:1*. 2015;6(1):101-114. doi:10.15172/PNEU.2015.6/671

81. Holding PA, Snow RW. Impact of Plasmodium falciparum Malaria on Performance and learning: Review of the Evidence. Published online 2001. https://www.ncbi.nlm.nih.gov/books/NBK2614/

82. Blomström Å, Kosidou K, Kristiansson M, Masterman T. Infection during childhood and the risk of violent criminal behavior in adulthood. *Brain, Behavior, and Immunity*. 2020;86:63-71. doi:10.1016/J.BBI.2019.02.026

83. Green MJ, Watkeys OJ, Whitten T, et al. Increased incidence of childhood mental disorders following exposure to early life infection. *Brain, Behavior, and Immunity*. 2021;97:376-382. doi:10.1016/j.bbi.2021.08.009

84. Dantzer R, O'Connor JC, Freund GG, Johnson RW, Kelley KW. From inflammation to sickness and depression: when the immune system subjugates the brain. *Nature Reviews Neuroscience 2007 9:1*. 2008;9(1):46-56. doi:10.1038/nrn2297

85. Chain JL, Alvarez K, Mascaro-Blanco A, et al. Autoantibody Biomarkers for Basal Ganglia Encephalitis in Sydenham Chorea and Pediatric Autoimmune Neuropsychiatric Disorder Associated With Streptococcal Infections. *Frontiers in Psychiatry*. 2020;11:564. doi:10.3389/fpsyt.2020.00564

86. Case A, Paxson C. Early life health and cognitive function in old age. In: *American Economic Review*. Vol 99. NIH Public Access; 2009:104-109. doi:10.1257/aer.99.2.104

87. Almond D. Is the 1918 influenza pandemic over? Long-term effects of in utero influenza exposure in the post-1940 U.S. population. *Journal of Political Economy*. 2006;114(4):672-712. doi:10.1086/507154

88. Turner AJ, Fichera E, Sutton M. The effects of in-utero exposure to influenza on mental health and mortality risk throughout the life-course. *Economics & Human Biology*. 2021;43:101059. doi:10.1016/J.EHB.2021.101059

89. Viinikainen J, Bryson A, Böckerman P, et al. Do childhood infections affect labour market outcomes in adulthood and, if so, how? *Economics & Human Biology*. 2020;37:100857. doi:10.1016/J.EHB.2020.100857

90. Case A, Fertig A, Paxson C. The lasting impact of childhood health and circumstance. *Journal of Health Economics*. 2005;24(2):365-389. doi:10.1016/j.jhealeco.2004.09.008

91. Case A, Paxson C. Causes and consequences of early-life health. *Demography*. 2010;47(Suppl 1):S65-S85. doi:10.1353/dem.2010.0007

92. Clemens JD, Stanton BF, Chakraborty J, et al. Measles Vaccination And Childhood Mortality In Rural Bangladesh. *American Journal of Epidemiology*. 1988;128(6):1330-1339. doi:10.1093/oxfordjournals.aje.a115086

93. Koenig MA, Khan MA, Wojtyniak B, et al. Impact of measles vaccination on childhood mortality in rural Bangladesh. *Bulletin of the World Health Organization*. 1990;68(4):441-447. https://pubmed.ncbi.nlm.nih.gov/2208557

94. Adedeji WA. The Treasure Called Antibiotics. *Annals of Ibadan postgraduate medicine*. 2016;14(2):56-57. https://pubmed.ncbi.nlm.nih.gov/28337088

95. Massell BF, Chute CG, Walker AM, Kurland GS. Penicillin and the Marked Decrease in Morbidity and Mortality from Rheumatic Fever in the United States. *http://dxdoiorg/101056/NEJM198802043180504*. 2010;318(5):280-286. doi:10.1056/NEJM198802043180504

96. McKinlay JB, McKinlay SM. The questionable contribution of medical measures to the decline of mortality in the United States in the twentieth century. *Milbank Memorial Fund Quarterly, Health and Society*. 1977;55(3):405-428. doi:10.2307/3349539

97. Lynskey NN, Jauneikaite E, Li HK, et al. Emergence of dominant toxigenic M1T1 Streptococcus pyogenes clone during increased scarlet fever activity in England: a population-based molecular epidemiological study. *The Lancet Infectious Diseases*. 2019;19(11):1209-1218. doi:10.1016/S1473-3099(19)30446-3

98. Choi YH, Campbell H, Amirthalingam G, Hoek AJ van, Miller E. Investigating the pertussis resurgence in England and Wales, and options for future control. *BMC Medicine*. 2016;14(1):121. doi:10.1186/s12916-016-0665-8

99. Paules CI, Marston HD, Fauci AS. Measles in 2019 — Going Backward. *The New England Journal of Medicine*. 2019;380(23):2185-2187. doi:10.1056/NEJMP1905099

100. Kelly E. The Scourge of Asian Flu: In utero Exposure to Pandemic Influenza and the Development of a Cohort of British Children. *Journal of Human Resources*. 2011;46(4):669-694. doi:10.1353/jhr.2011.0004

101. Munro AD, Smallman-Raynor M, Algar AC. Long-term changes in endemic threshold populations for pertussis in England and Wales: A spatiotemporal analysis of Lancashire and South Wales, 1940-69. *Social Science and Medicine*. Published online August 2020:113295. doi:10.1016/j.socscimed.2020.113295

102. Mehndiratta MM, Mehndiratta P, Pande R. Poliomyelitis: Historical Facts, Epidemiology, and Current Challenges in Eradication. *The Neurohospitalist*. 2014;4(4):223. doi:10.1177/1941874414533352

103. Vander Top E, Gentry-Nielsen M, Knoop FC. Poliomyelitis. In: *XPharm: The Comprehensive Pharmacology Reference*. Elsevier; 2007:1-4. doi:10.1016/B978-008055232-3.60926-2

104. Sabin AB. Paralytic Consequences of Poliomyelitis Infection in Different Parts of the World and in Different Population Groups. *American Journal of Public Health and the Nations Health*. 1951;41(10):1215-1230. doi:10.2105/AJPH.41.10.1215

105. Sigurdsson B, Sigurjónsson J, Sigurdsson JH, Thorkelsson J, Gudmundsson KR. A Disease Epidemic In Iceland Simulating Poliomyelitis. *American Journal of Epidemiology*. 1950;52(2):222-238. doi:10.1093/oxfordjournals.aje.a119421

106. Farbu E, Rekand T, Gilhus NE. Post-polio syndrome and total health status in a prospective hospital study. *European Journal of Neurology*. 2003;10(4):407-413. doi:10.1046/J.1468-1331.2003.00613.X

107. Mercan M, Bajrami A, Acır I, Yayla V. P349 Assessment of prevalence and risk factors for carpal tunnel syndrome in polio survivors. *Clinical Neurophysiology*. 2017;128(9):e291. doi:10.1016/J.CLINPH.2017.07.357

108. Grill B, Levangie PK, Cole M, Rosenberg D, Jensen L. Bone Mineral Density Among Individuals With Residual Lower Limb Weakness After Polio. *PM&R*. 2019;11(5):470-475. doi:10.1016/J.PMRJ.2018.08.387

109. Seo K-H, Lee JH, Lee S-Y, Lee JY, Lim J-Y. Prevalence and effect of obesity on mobility according to different criteria in polio survivors. *American Journal of Physical Medicine & Rehabilitation*. Published online 2020. doi:10.1097/phm.0000000000001556

110. Tiwari TSP. Diphtheria. *Hunter's Tropical Medicine and Emerging Infectious Disease: Ninth Edition*. Published online January 2013:402-406. doi:10.1016/B978-1-4160-4390-4.00037-0

111. Galazka AM, Robertson SE, Oblapenko GP. Resurgence of Diphtheria. *European Journal of Epidemiology*. 1995;11(1). https://www.jstor.org/stable/3582199

112. Opinel A, Gachelin G. French 19th century contributions to the development of treatments for diphtheria. *Journal of the Royal Society of Medicine*. 2011;104(4):173. doi:10.1258/JRSM.2010.10K069

113. Millward G. *Vaccinating Britain: Mass vaccination and the public since the Second World War*. (Cantor D, Waddington K, eds.). Manchester University Press; 2019. https://www.ncbi.nlm.nih.gov/books/NBK545997/

114. Bleakley H. Disease and Development: Evidence from Hookworm Eradication in the American South*. *Quaterly Journal of Economics*. 2007;122(1):73-117. doi:10.1162/qjec.121.1.73

115. Bütikofer A, Salvanes KG. Disease Control and Inequality Reduction: Evidence from a Tuberculosis Testing and Vaccination Campaign. *Review of Economic Studies*. 2020;87:2087-2125. doi:10.1093/restud/rdaa022

116. Katz AR, Morens DM. Severe Streptococcal Infections in Historical Perspective. *Clinical Infectious Diseases*. 1992;14(1):298-307. doi:10.1093/clinids/14.1.298

117. Ferretti J, Köhler W. History of streptococcal research. In: *Streptococcus Pyogenes: Basic Biology to Clinical Manifestations [Internet]*. University of Oklahoma Health Sciences Center; 2016. https://www.ncbi.nlm.nih.gov/books/NBK333430/?report=classic

118. Tanz RR. Sore Throat. In: *Nelson Pediatric Symptom-Based Diagnosis*. Elsevier Inc.; 2018:1-14.e2. doi:10.1016/B978-0-323-39956-2.00001-7

119. Wessels MR. *Pharyngitis and Scarlet Fever*. University of Oklahoma Health Sciences Center; 2016. http://www.ncbi.nlm.nih.gov/pubmed/26866221

120. Mantzourani E, Evans A, Cannings-John R, et al. Impact of a pilot NHS-funded sore throat test and treat service in community pharmacies on provision and quality of patient care. *BMJ Open Quality*. 2020;9(1):e000833. doi:10.1136/BMJOQ-2019-000833

121. Olivier C. Rheumatic fever—is it still a problem? *Journal of Antimicrobial Chemotherapy*. 2000;45(suppl_1):13-21. doi:10.1093/jac/45.suppl_1.13

122. Aziz RK, Kotb M. Rise and Persistence of Global M1T1 Clone of Streptococcus pyogenes. *Emerging Infectious Diseases*. 2008;14(10):1511-1517. doi:10.3201/eid1410.071660

123. Liu Y, Chan TC, Yap LW, et al. Resurgence of scarlet fever in China: a 13-year population-based surveillance study. *The Lancet Infectious Diseases*. 2018;18(8):903-912. doi:10.1016/S1473-3099(18)30231-7

124. Brockmann SO, Eichner L, Eichner M. Constantly high incidence of scarlet fever in Germany. 2018;18:499-500. doi:10.1016/S1473-3099(18)30210-X

125. Bermont A, Broide E, Matalon S, et al. New-onset of Crohn's disease is associated with antistreptolysin o positive titers. *Clinical and Experimental Gastroenterology*. 2020;13:187-191. doi:10.2147/CEG.S245770

126. Aran A, Lin L, Nevsimalova S, et al. Elevated anti-streptococcal antibodies in patients with recent narcolepsy onset. *Sleep*. 2009;32(8):979-983. doi:10.1093/sleep/32.8.979

127. Telfer NR, Chalmers RJG, Whale K, Colman G. The Role of Streptococcal Infection in the Initiation of Guttate Psoriasis. *Archives of Dermatology*. 1992;128(1):39-42. doi:10.1001/archderm.1992.01680110049004

128. Fanget N. *Pertussis: a tale of two vaccines*. Vol 1. Natureportfolio; 2020. https://media.nature.com/original/magazine-assets/d42859-020-00013-8/d42859-020-00013-8.pdf

129. Cherry JD. The History of Pertussis (Whooping Cough); 1906–2015: Facts, Myths, and Misconceptions. *Current Epidemiology Reports 2015 2:2*. 2015;2(2):120-130. doi:10.1007/S40471-015-0041-9

130. Rubin K, Glazer S. The potential role of subclinical Bordetella Pertussis colonization in the etiology of multiple sclerosis. *Immunobiology*. 2016;221(4):512-515. doi:10.1016/j.imbio.2015.12.008

131. Raeven RHM, Maas L van der, Pennings JLA, et al. Antibody Specificity Following a Recent Bordetella pertussis Infection in Adolescence Is Correlated With the Pertussis Vaccine Received in Childhood. *Frontiers in Immunology*. 2019;10(JUN):1364. doi:10.3389/FIMMU.2019.01364

132. Crowcroft NS, Johnson C, Chen C, et al. Under-reporting of pertussis in Ontario: A Canadian Immunization Research Network (CIRN) study using capture-recapture. *PLoS ONE*. 2018;13(5). doi:10.1371/JOURNAL.PONE.0195984

133. Clarkson JA, Clarkson PEMF, Institute R. The Efficiency of Measles and Pertussis Notification in England and Wales. *International Journal of Epidemiology International Epidemiological Association*. 1985;14(1). https://academic.oup.com/ije/article/14/1/153/694546

134. Deeks S, De Serres G, Boulianne N, Duval B, Rochette L, Déry P. Failure of Physicians to Consider the Diagnosis of Pertussis in Children. *Clinical Infectious Diseases*. 1999;28:840-846. https://academic.oup.com/cid/article/28/4/840/401614

135. Barnish MS, Tagiyeva N, Devereux G, Aucott L, Turner S. Diverging prevalences and different risk factors for childhood asthma and eczema: A cross-sectional study. *BMJ Open*. 2015;5(6). doi:10.1136/bmjopen-2015-008446

136. England PH. Laboratory confirmed cases of pertussis in England: annual report for 2019. *Health Protection Report*. 2020;14(8). https://www.gov.uk/government/publications/pertussis-laboratory-confirmed-cases-reported-in-england-2019

137. Roberts L. Why measles deaths are surging — and coronavirus could make it worse. *Nature*. 2020;580(7804):446-447. doi:10.1038/d41586-020-01011-6

138. Strachan DP. Hay fever, hygiene, and household size. *British Medical Journal*. 1989;299(6710):1259-1260. doi:10.1136/bmj.299.6710.1259

139. Okada H, Kuhn C, Feillet H, Bach JF. The 'hygiene hypothesis' for autoimmune and allergic diseases: An update. 2010;160:1-9. doi:10.1111/j.1365-2249.2010.04139.x

140. Rudwaleit M, Andermann B, Alten R, et al. Atopic disorders in ankylosing spondylitis and rheumatoid arthritis. *Annals of the Rheumatic Diseases*. 2002;61(11):968-974. doi:10.1136/ard.61.11.968

141. Yazdanbakhsh M, Kremsner PG, Van Ree R. Immunology: Allergy, parasites, and the hygiene hypothesis. 2002;296:490-494. doi:10.1126/science.296.5567.490

142. Flohr C, Tuyen LN, Lewis S, et al. Poor sanitation and helminth infection protect against skin sensitization in Vietnamese children: A cross-sectional study. *Journal of Allergy and Clinical Immunology*. 2006;118(6):1305-1311. doi:10.1016/j.jaci.2006.08.035

143. Cooper PJ, Chico ME, Vaca MG, et al. Effect of albendazole treatments on the prevalence of atopy in children living in communities endemic for geohelminth

parasites: a cluster-randomised trial. *Lancet*. 2006;367(9522):1598-1603. doi:10.1016/S0140-6736(06)68697-2

144. Lynch NR, Palenque M, Hagel I, Diprisco MC. Clinical improvement of asthma after anthelminthic treatment in a tropical situation. *American Journal of Respiratory and Critical Care Medicine*. 1997;156(1):50-54. doi:10.1164/ajrccm.156.1.9606081

145. Benn CS, Melbye M, Wohlfahrt J, Björkstén B, Aaby P. Cohort study of sibling effect, infectious diseases, and risk of atopic dermatitis during first 18 months of life. *British Medical Journal*. 2004;328(7450):1223-1226. doi:10.1136/bmj.38069.512245.fe

146. Kavanagh H, Noone C, Cahill E, English K, Locht C, Mahon BP. Attenuated Bordetella pertussis vaccine strain BPZE1 modulates allergen-induced immunity and prevents allergic pulmonary pathology in a murine model. *Clinical and Experimental Allergy*. 2010;40(6):933-941. doi:10.1111/j.1365-2222.2010.03459.x

147. Rubin K, Glazer S. The pertussis hypothesis: Bordetella pertussis colonization in the etiology of asthma and diseases of allergic sensitization. *Medical Hypotheses*. 2018;120:101-115. doi:10.1016/j.mehy.2018.08.006

148. Scudellari M. Cleaning up the hygiene hypothesis. 2017;114:1433-1436. doi:10.1073/pnas.1700688114

149. Dupuis JM, Rathkopf M. Childhood infections and the risk of asthma: A longitudinal study over 37 years. 2013;132:S32-S33. doi:10.1542/peds.2013-2294AAA

150. Kendirli SG, Yilmaz M, Bayram I, Altintas DU, Inal A, Karakoc G. Potential association between allergic diseases and pertussis infection in schoolchildren: Results of two cross-sectional studies seven years apart. *Allergologia et Immunopathologia*. 2009;37(1):21-25. doi:10.1016/S0301-0546(09)70247-2

151. Felippe MJB. Immunotherapy. *Equine Infectious Diseases: Second Edition*. Published online January 2014:584-597.e5. doi:10.1016/B978-1-4557-0891-8.00066-X

152. Andrus JK, De Quadros CA, Castillo-Solorzano C. Measles. *Tropical Infectious Diseases*. Published online January 2011:347-351. doi:10.1016/B978-0-7020-3935-5.00054-9

153. Shanks GD, Hu Z, Waller M, et al. Epidemiology in History Measles Epidemics of Variable Lethality in the Early 20th Century. *American Journal of Epidemiology*. 2014;179(4):413-422. doi:10.1093/aje/kwt282

154. Dayan GH, McLean HQ. Measles. *International Encyclopedia of Public Health*. Published online January 2017:565-569. doi:10.1016/B978-0-12-803678-5.00270-8

155. Mishra Y, Sharma L, Dhiman M, Sharma MM. Endophytic fungal diversity of selected medicinal plants and their bio-potential applications. *Fungi Bio-Prospects in Sustainable Agriculture, Environment and Nano-Technology*. Published online January 2021:227-283. doi:10.1016/B978-0-12-821394-0.00010-X

156. Kowalzik F, Faber J, Knuf M. MMR and MMRV vaccines. *Vaccine*. 2018;36(36):5402-5407. doi:10.1016/J.VACCINE.2017.07.051

157. Petrova VN, Sawatsky B, Han AX, et al. Incomplete genetic reconstitution of B cell pools contributes to prolonged immunosuppression after measles. *Science Immunology*. 2019;4(41):6125. doi:10.1126/sciimmunol.aay6125

158. Mina MJ, Kula T, Leng Y, et al. Measles virus infection diminishes preexisting antibodies that offer protection from other pathogens. *Science*. 2019;366(6465):599-606. doi:10.1126/science.aay6485

159. Landers J. Death and the Metropolis: Studies in the Demographic History of London, 1670–1830. *Death and the Metropolis*. Published online July 1993. doi:10.1017/CBO9780511895494

160. Atkins PJ. *Milk consumption and tuberculosis in Britain, 1850-1950 [Working Paper]*. University of Durham; 2012. https://dro.dur.ac.uk/10386/

161. Lange C, Aaby P, Behr MA, et al. 100 years of Mycobacterium bovis bacille Calmette-Guérin. *The Lancet Infectious Diseases*. 2022;22(1):e2-e12. doi:10.1016/S1473-3099(21)00403-5

162. Glaziou P, Floyd K, Raviglione M. Trends in tuberculosis in the UK. *Thorax*. 2018;73(8). doi:10.1136/thoraxjnl-2018-211537

163. Chakaya J, Khan M, Ntoumi F, et al. Global Tuberculosis Report 2020 – Reflections on the Global TB burden, treatment and prevention efforts. *International Journal of Infectious Diseases*. Published online March 2021. doi:10.1016/J.IJID.2021.02.107

164. Ravimohan S, Kornfeld H, Weissman D, Bisson GP. Tuberculosis and lung damage: from epidemiology to pathophysiology. *European Respiratory Review*. 2018;27(147):170077. doi:10.1183/16000617.0077-2017

165. Gennaris A, Collet J-F. The 'captain of the men of death', Streptococcus pneumoniae, fights oxidative stress outside the 'city wall'. *EMBO molecular medicine*. 2013;5(12):1798-1800. doi:10.1002/emmm.201303482

166. Bhalotra SR, Venkataramani A. The Captain of the Men of Death and His Shadow: Long-Run Impacts of Early Life Pneumonia Exposure. *SSRN Electronic Journal*. Published online 2011:81. doi:10.2139/ssrn.1940725

167. Haller JS, Bliss M. William Osler: A Life in Medicine. *The Journal of American History*. 2000;87(3):1060. doi:10.2307/2675359

168. Janoff EN, Musher DM. Streptococcus pneumoniae. *Mandell, Douglas, and Bennett's Principles and Practice of Infectious Diseases*. 2015;2:2310-2327.e5. doi:10.1016/B978-1-4557-4801-3.00201-0

169. Maiden MCJ. The impact of protein-conjugate polysaccharide vaccines: an endgame for meningitis? *Philosophical Transactions of the Royal Society B: Biological Sciences*. 2013;368(1623). doi:10.1098/RSTB.2012.0147

170. Shaheen SO, Barker DJP, Shiell AW, Crocker FJ, Wield GA, Holgate ST. The relationship between pneumonia in early childhood and impaired lung function in late adult life. *https://doiorg/101164/ajrccm14938118627*. 1994;149(3):616-619. doi:10.1164/AJRCCM.149.3.8118627

171. Zyberi G. The Transitional Justice Process in the Former Yugoslavia: Long Transition, Yet Not Enough Justice. *SSRN Electronic Journal*. Published online 2012. doi:10.2139/ssrn.2067355

172. Harrison M, Wolf N. The Frequency of Wars. *The Economics of Coercion and Conflict*. Published online December 2014:121-149. doi:10.1142/9789814583343_0005

173. Fogger SA, Moore R, Pickett L. Posttraumatic Stress Disorder and Veterans: Finding Hope and Supporting Healing. *The Journal for Nurse Practitioners*. 2016;12(9):598-604. doi:10.1016/J.NURPRA.2016.07.014

174. Yehuda R, Daskalakis NP, Bierer LM, et al. Holocaust Exposure Induced Intergenerational Effects on FKBP5 Methylation. *Biological Psychiatry*. 2016;80(5):372-380. doi:10.1016/J.BIOPSYCH.2015.08.005

175. Blomvall L. Bombing Britain - War, State and Society. Published online 2020. Accessed May 12, 2020. http://www.warstateandsociety.com/Bombing-Britain

176. Field G. Nights Underground in Darkest London: The Blitz, 1940–1941. *International Labor and Working-Class History*. 2002;62(1):11-49. doi:10.1017/S0147547902000194

177. Janis IL. *Air War and Emotional Stress: Psychological Studies of Bombing and Civilian Defense*. Greeenwood; 1976.

178. Qu F, Wu Y, Zhu Y-H, et al. The association between psychological stress and miscarriage: A systematic review and meta-analysis OPEN. *Scientific Reports*. 2017;7(1731). doi:10.1038/s41598-017-01792-3

179. Black SE, Devereux PJ, Salvanes KG. Does Grief Transfer across Generations? Bereavements during Pregnancy and Child Outcomes. *American Economic Journal: Applied Economics*. 2016;8(1):193-223. doi:10.1257/APP.20140262

180. Persson P, Rossin-Slater M, Alsan M, et al. Family Ruptures, Stress, and the Mental Health of the Next Generation †. *American Economic Review*. 2018;108(5):1214-1252. doi:10.1257/aer.20141406

181. Demakakos P, Linara-Demakakou E, Mishra GD. Adverse childhood experiences are associated with increased risk of miscarriage in a national population-based cohort study in England. *Human Reproduction*. 2020;35(6):1451-1460. doi:10.1093/humrep/deaa113

182. Afifi TO. Considerations for expanding the definition of ACEs. *Adverse Childhood Experiences: Using Evidence to Advance Research, Practice, Policy, and Prevention*. Published online January 2020:35-44. doi:10.1016/B978-0-12-816065-7.00003-3

183. Smallman-Raynor MR, Cliff AD. Operation Pied Piper: A geographical reappraisal of the impact of wartime evacuation on scarlet fever and diphtheria rates in England and Wales, 1939-1945. *Epidemiology and Infection*. 2015;143(14):2923-2938. doi:10.1017/S0950268815000175

184. NOMIS. 1961 Census - Census of Population - Data Sources - home - Nomis - Official Labour Market Statistics. Published online 2021. Accessed October 14, 2021. https://www.nomisweb.co.uk/sources/census{_}1961

185. UK Data Service J. Casweb. Published online March 2013.

186. Lawlor DA, Smith GD, Mitchell R, Ebrahim S. Adult blood pressure and climate conditions in infancy: A test of the hypothesis that dehydration in infancy is associated with higher adult blood pressure. *American Journal of Epidemiology*. 2006;163(7):608-614. doi:10.1093/aje/kwj085

187. Glynn JR, Moss PAH. Systematic analysis of infectious disease outcomes by age shows lowest severity in school-age children. *Scientific Data 2020 7:1*. 2020;7(1):1-13. doi:10.1038/s41597-020-00668-y

188. Kang S-J, Jung SI. Age-Related Morbidity and Mortality among Patients with COVID-19. *Infection & Chemotherapy*. 2020;52(2):154. doi:10.3947/IC.2020.52.2.154

189. Office of National Statistics. [ARCHIVED CONTENT] Release Edition Reference Tables - ONS. Published online 2011. Accessed November 25, 2021. https://webarchive.nationalarchives.gov.uk/ukgwa/20150908090558/http://www.ons.gov.uk/ons/publications/re-reference-tables.html?edition=tcm{\%}3A77-215593

190. Relly SJ. The political rhetoric of parity of esteem. *https://doiorg/101080/0305498520201866522*. Published online 2021. doi:10.1080/03054985.2020.1866522

191. Popham F, Iannelli C. Does comprehensive education reduce health inequalities? *SSM - Population Health*. 2021;15:100834. doi:10.1016/J.SSMPH.2021.100834

192. Gorard S, Siddiqui N. Grammar schools in England: a new analysis of social segregation and academic outcomes. *https://doiorg/101080/0142569220181443432*. 2018;39(7):909-924. doi:10.1080/01425692.2018.1443432

193. Parliament U. Direct Grant Schools (Hansard, 14 December 1966). Published online 1966. Accessed October 14, 2021. https://api.parliament.uk/historic-hansard/written-answers/1966/dec/14/direct-grant-schools

194. Endocrinology TLD&. Obesity in China: time to act. *The Lancet Diabetes & Endocrinology*. 2021;9(7):407. doi:10.1016/S2213-8587(21)00150-9

195. Cohen AK, Rai M, Rehkopf DH, Abrams B. Educational attainment and obesity: a systematic review. *Obesity Reviews*. 2013;14(12):989-1005. doi:10.1111/OBR.12062

196. Böckerman P, Viinikainen J, Pulkki-Råback L, et al. Does higher education protect against obesity? Evidence using Mendelian randomization. *Preventive Medicine*. 2017;101:195-198. doi:10.1016/J.YPMED.2017.06.015

197. Davies NM, Dickson M, Smith GD, Berg GJ van den, Windmeijer F. The causal effects of education on health outcomes in the UK Biobank. *Nature Human Behaviour 2017 2:2*. 2018;2(2):117-125. doi:10.1038/s41562-017-0279-y

198. CD A, BB B. Can increasing years of schooling reduce type 2 diabetes (T2D)?: Evidence from a Mendelian randomization of T2D and 10 of its risk factors. *Scientific reports*. 2020;10(1). doi:10.1038/S41598-020-69114-8

199. M C, B C. Association of Educational Attainment With Adiposity, Type 2 Diabetes, and Coronary Artery Diseases: A Mendelian Randomization Study. *Frontiers in public health*. 2020;8. doi:10.3389/FPUBH.2020.00112

200. Merlo J, Wagner P, Leckie G. A simple multilevel approach for analysing geographical inequalities in public health reports: The case of municipality differences in obesity. *Health & Place*. 2019;58:102145. doi:10.1016/J.HEALTHPLACE.2019.102145

201. Morris R, Carstairs V. Which deprivation? A comparison of selected deprivation indexes. *Journal of Public Health*. 1991;13(4):318-326. doi:10.1093/OXFORDJOURNALS.PUBMED.A042650

202. Townsend P, Phillimore P, Beattie A. *Health and Deprivation: Inequality and The North*. Croom Helm; 1988.

203. Joyce R, Webb R, Peacock JL, Stirland H. Which is the best deprivation predictor of foetal and infant mortality rates? *Public Health*. 2000;114(1):21-24. doi:10.1038/sj.ph.1900597

204. Office HMS. *[Series] Ministry of Labour Gazette*.; 1949.

205. Allen N, Sudlow C, Downey P, et al. UK Biobank: Current status and what it means for epidemiology. *Health Policy and Technology*. 2012;1(3):123-126. doi:10.1016/j.hlpt.2012.07.003

206. Al-Sheyyab M, Batieha A, El-Shanti H, Daoud A. Henoch-Schonlein purpura and streptococcal infection: A prospective case-control study. *Annals of Tropical Paediatrics*. 1999;19(3):253-255. doi:10.1080/02724939992329

207. Johnson DR, Kurlan R, Leckman J, Kaplan EL. The Human Immune Response to Streptococcal Extracellular Antigens: Clinical, Diagnostic, and Potential Pathogenetic Implications. *Clinical Infectious Diseases*. 2010;50(4):481-490. doi:10.1086/650167

208. Sela U, Euler CW, Correa da Rosa J, Fischetti VA. Strains of bacterial species induce a greatly varied acute adaptive immune response: The contribution of the accessory genome. *PLoS Pathogens*. 2018;14(1). doi:10.1371/journal.ppat.1006726

209. Youn JC, Jung MK, Yu HT, et al. Increased frequency of CD4+CD57+ senescent T cells in patients with newly diagnosed acute heart failure: exploring new pathogenic mechanisms with clinical relevance. *Scientific Reports 2019 9:1*. 2019;9(1):1-10. doi:10.1038/s41598-019-49332-5

210. Laroumanie F, Douin-Echinard V, Pozzo J, et al. CD4+ T cells promote the transition from hypertrophy to heart failure during chronic pressure overload. *Circulation*. 2014;129(21):2111-2124. doi:10.1161/CIRCULATIONAHA.113.007101

211. D'Avila JC, Siqueira LD, Mazeraud A, et al. Age-related cognitive impairment is associated with long-term neuroinflammation and oxidative stress in a mouse model of episodic systemic inflammation. *Journal of Neuroinflammation*. 2018;15(1). doi:10.1186/S12974-018-1059-Y

212. Sulkowski ML, Jordan C, Dobrinsky SR, Mathews RE. OCD in School Settings. *The Clinician's Guide to Cognitive-Behavioral Therapy for Childhood Obsessive-compulsive Disorder*. Published online January 2018:225-241. doi:10.1016/B978-0-12-811427-8.00012-5

213. Mahara G, Wang C, Yang K, et al. The association between environmental factors and scarlet fever incidence in Beijing Region: Using gis and spatial regression models.

International Journal of Environmental Research and Public Health. 2016;13(11). doi:10.3390/ijerph13111083

214. Lu Q, Wu H, Ding Z, Wu C, Lin J. Analysis of epidemiological characteristics of scarlet fever in Zhejiang Province, China, 2004–2018. *International Journal of Environmental Research and Public Health*. 2019;16(18). doi:10.3390/ijerph16183454

215. Carter AR, Gill D, Davies NM, et al. Understanding the consequences of education inequality on cardiovascular disease: Mendelian randomisation study. *The BMJ*. 2019;365. doi:10.1136/bmj.l1855

216. Cornelis MC, Wang Y, Holland T, Agarwal P, Weintraub S, Morris MC. Age and cognitive decline in the UK Biobank. *PLoS ONE*. 2019;14(3). doi:10.1371/JOURNAL.PONE.0213948

217. Okbay A, Beauchamp JP, Fontana MA, et al. Genome-wide association study identifies 74 loci associated with educational attainment. *Nature*. 2016;533(7604):539-542. doi:10.1038/nature17671

218. Lamagni T, Guy R, Chand M, et al. Resurgence of scarlet fever in England, 2014–16: a population-based surveillance study. *The Lancet Infectious Diseases*. 2018;18(2):180-187. doi:10.1016/S1473-3099(17)30693-X

219. Krishnan KC, Mukundan S, Alagarsamy J, Laturnus D, Kotb M. Host Genetic Variations and Sex Differences Potentiate Predisposition, Severity, and Outcomes of Group A Streptococcus-Mediated Necrotizing Soft Tissue Infections. *Infection and Immunity*. 2016;84(2):416. doi:10.1128/IAI.01191-15

220. Falkingham J, Sage J, Stone J, Vlachantoni A. Residential mobility across the life course: Continuity and change across three cohorts in Britain. *Advances in Life Course Research*. 2016;30:111-123. doi:10.1016/j.alcr.2016.06.001

221. Baker S. weightGIS: Weight ESRI shapefiles attributes. Published online 2020. Accessed August 13, 2020. https://github.com/sbaker-dev/weightGIS

222. RiossAvila F. Feasible Estimation of Linear Models with N-Fixed Effects. *SSRN Electronic Journal*. Published online 2013. doi:10.2139/ssrn.2366943

223. Baker S. pyBlendFigures PyPI. Published online 2021. Accessed November 11, 2021. https://pypi.org/project/pyBlendFigures/

224. Gardiner S. Penicillin: promise, problems and practice in wartime Edinburgh. *Journal of the Royal College of Physicians of Edinburgh*. 2016;46:198-205. doi:10.4997/JRCPE.2016.312

225. Public Health England. Notifiable diseases: historic annual totals - GOV.UK. Published online 2019. Accessed May 12, 2020. https://www.gov.uk/government/publications/notifiable-diseases-historic-annual-totals

226. Quaranta L. Scarred for Life. How conditions in early life affect socioeconomic status, reproduction and mortality in Southern Sweden, 1813-1968. Published online 2013. https://lup.lub.lu.se/search/publication/26071b90-1bbf-4bf9-90df-777f248788ae

227. McKeown T, Record RG. Relationship between childhood infections and measured intelligence. *Journal of Epidemiology & Community Health*. 1976;30(2):101-106. doi:10.1136/jech.30.2.101

228. Fry A, Littlejohns TJ, Sudlow C, et al. Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. *American Journal of Epidemiology*. 2017;186(9):1026-1034. doi:10.1093/AJE/KWX246

229. Alten S van, Domingue BW, Galama T, Marees AT. Reweighting the UK Biobank to reflect its underlying sampling population substantially reduces pervasive selection bias due to volunteering. Published online May 2022:2022.05.16.22275048. doi:10.1101/2022.05.16.22275048

230. Guy R, Sharp Ashley, Coelho J, Brown C, Lamagni T. Group A streptococcal infections: update on seasonal activity in England, 2021 to 2022 - GOV.UK. Published online September 2022. Accessed October 24, 2022. https://www.gov.uk/government/publications/group-a-streptococcal-infections-activity-during-the-2021-to-2022-season/group-a-streptococcal-infections-update-on-seasonal-activity-in-england-2021-to-2022

231. Asher I, Pearce N. Global burden of asthma among children. *International Journal of Tuberculosis and Lung Disease*. 2014;18(11):1269-1278. doi:10.5588/ijtld.14.0170

232. Butland BK, Strachan DP, Crawley-Boevey EE, Anderson HR. Childhood asthma in South London: Trends in prevalence and use of medical services 1991-2002. 2006;61:383-387. doi:10.1136/thx.2005.043646

233. Ho NT, Li F, Lee-Sarwar KA, et al. Meta-analysis of effects of exclusive breastfeeding on infant gut microbiota across populations. *Nature Communications*. 2018;9(1):1-13. doi:10.1038/s41467-018-06473-x

234. Kalbermatter C, Fernandez Trigo N, Christensen S, Ganal-Vonarburg SC. Maternal Microbiota, Early Life Colonization and Breast Milk Drive Immune Development in the Newborn. *Frontiers in Immunology*. 2021;12. doi:10.3389/fimmu.2021.683022

235. Dierikx TH, Visser DH, Benninga MA, et al. The influence of prenatal and intrapartum antibiotics on intestinal microbiota colonisation in infants: A systematic review. 2020;81:190-204. doi:10.1016/j.jinf.2020.05.002

236. Russell SL, Gold MJ, Hartmann M, et al. Early life antibiotic-driven changes in microbiota enhance susceptibility to allergic asthma. *EMBO Reports*. 2012;13(5):440-447. doi:10.1038/embor.2012.32

237. Ma Y, Zhao J, Han ZR, Chen Y, Leung TF, Wong GWK. Very low prevalence of asthma and allergies in schoolchildren from rural Beijing, China. *Pediatric Pulmonology*. 2009;44(8):793-799. doi:10.1002/ppul.21061

238. Nermes M, Niinivirta K, Nylund L, et al. Perinatal Pet Exposure, Faecal Microbiota, and Wheezy Bronchitis: Is There a Connection? *ISRN Allergy*. 2013;2013:1-6. doi:10.1155/2013/827934

239. Oluwole O, Rennie DC, Senthilselvan A, et al. The association between endotoxin and beta-(1 → 3)-D-glucan in house dust with asthma severity among schoolchildren. *Respiratory Medicine*. 2018;138:38-46. doi:10.1016/j.rmed.2018.03.015

240. Sozańska B. Raw Cow's Milk and Its Protective Effect on Allergies and Asthma. 2019;11. doi:10.3390/nu11020469

241. Feng M, Yang Z, Pan L, et al. Associations of early life exposures and environmental factors with asthma among children in rural and urban areas of Guangdong, China. *Chest*. 2016;149(4):1030-1041. doi:10.1016/j.chest.2015.12.028

242. Yen YC, Yang CY, Ho CK, et al. Indoor ozone and particulate matter modify the association between airborne endotoxin and schoolchildren's lung function. *Science of the Total Environment*. 2020;705:135810. doi:10.1016/j.scitotenv.2019.135810

243. Lødrup Carlsen KC, Roll S, Carlsen KH, et al. Does Pet Ownership in Infancy Lead to Asthma or Allergy at School Age? Pooled Analysis of Individual Participant Data from 11 European Birth Cohorts. *PLoS ONE*. 2012;7(8). doi:10.1371/journal.pone.0043214

244. Carucci L, Coppola S, Nocerino R, Paparo L, Di Scala C, Berni Canani R. Commentary: Raw Cow Milk Consumption and Atopic March. *Frontiers in Pediatrics*. 2021;9:684662. doi:10.3389/fped.2021.684662

245. Illi S, Mutius E von, Lau S, et al. Early childhood infectious diseases and the development of asthma up to school age: A birth cohort study. *British Medical Journal*. 2001;322(7283):390-395. doi:10.1136/bmj.322.7283.390

246. Thomsen SF. Genetics of asthma: an introduction for the clinician. *European Clinical Respiratory Journal*. 2015;2(1):24643. doi:10.3402/ECRJ.V2.24643

247. Yeatts K, Sly P, Shore S, et al. A Brief Targeted Review of Susceptibility Factors, Environmental Exposures,Asthma Incidence, and Recommendations for Future Asthma Incidence Research. *Environmental Health Perspectives*. 2006;114(4):634-640. doi:10.1289/ehp.8381

248. Charles A Janeway J, Travers P, Walport M, Shlomchik MJ. The distribution and functions of immunoglobulin isotypes. Published online 2001. https://www.ncbi.nlm.nih.gov/books/NBK27162/

249. Wu L-P, Wang N-C, Chang Y-H, et al. Duration of Antibody Responses after Severe Acute Respiratory Syndrome. *Emerging Infectious Diseases*. 2007;13(10):1562. doi:10.3201/EID1310.070576

250. Novakova P, Tiotiu A, Baiardini I, Krusheva B. Herberto Chong-Neto & Silviya Novakova (2021) Allergen immunotherapy in asthma: current evidence. *Journal of Asthma*. 2019;58(2):223-230. doi:10.1080/02770903.2019.1684517

251. Bijanzadeh M, Mahesh, Padukudru A, Ramachandra, Nallur B. Genetics of asthma: an introduction for the clinician. *Indian Journal of Medical Research*. 2011;134(1):149-161.

252. Flaherty D. Vaccine-Preventable Diseases. In: *Immunology for Pharmacy*. Mosby; 2012:197-213. doi:10.1016/B978-0-323-06947-2.10025-2

253. Daniel TM. The history of tuberculosis. *Respiratory Medicine*. 2006;100(11):1862-1870. doi:10.1016/J.RMED.2006.08.006

254. Booth SJ. Bordetella Pertussis Infections☆. In: *Reference Module in Biomedical Sciences*. Elsevier; 2014. doi:10.1016/b978-0-12-801238-3.04880-7

255. Wang K, Fry NK, Campbell H, et al. Whooping cough in school age children presenting with persistent cough in UK primary care after introduction of the preschool pertussis booster vaccination: Prospective cohort study. *BMJ (Online)*. 2014;348. doi:10.1136/bmj.g3668

256. Vilhjálmsson BJ, Yang J, Finucane HK, et al. Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *American Journal of Human Genetics*. 2015;97(4):576. doi:10.1016/J.AJHG.2015.09.001

257. Demenais F, Margaritte-Jeannin P, Barnes KC, et al. Multiancestry association study identifies new asthma risk loci that colocalize with immune-cell enhancer marks. *Nature Genetics*. 2018;50(1):42-50. doi:10.1038/s41588-017-0014-7

258. Hendriks J, Blume S. Measles vaccination: Before the measles-mumps-rubella vaccine. *American Journal of Public Health*. 2013;103(8):1393-1401. doi:10.2105/AJPH.2012.301075

259. Hitz DA, Tewald F, Eggers M. Seasonal Bordetella pertussis pattern in the period from 2008 to 2018 in Germany. *BMC Infectious Diseases 2020 20:1*. 2020;20(1):1-6. doi:10.1186/S12879-020-05199-W

260. Martinez ME. The calendar of epidemics: Seasonal cycles of infectious diseases. *PLOS Pathogens*. 2018;14(11):e1007327. doi:10.1371/JOURNAL.PPAT.1007327

261. Uphoff E, Cabieses B, Pinart M, Valdés M, Maria Antó J, Wright J. A systematic review of socioeconomic position in relation to asthma and allergic diseases. *European Respiratory Journal*. 2015;46(2):364-374. doi:10.1183/09031936.00114514

262. Aryee E, Perrin JM, Iannuzzi D, Kuhlthau KA, Oreskovic NM. Association of Neighborhood Characteristics with Pediatric Asthma. *Academic Pediatric*. Published online 2022. https://doi.org/10.1016/j.acap.2022.01.001

263. Swedlund AC, Donta AK. Scarlet fever epidemics of the nineteenth century: a case of evolved pathogenic virulence? In: *Human Biologists in the Archives*. Cambridge University Press; 2009:159-177. doi:10.1017/cbo9780511542534.009

264. Vargas MH. Ecological association between scarlet fever and asthma. *Respiratory Medicine*. 2006;100(2):363-366. doi:10.1016/j.rmed.2005.04.027

265. British Lung Foundation. Asthma statistics | British Lung Foundation. Published online 2022. Accessed October 11, 2022. https://statistics.blf.org.uk/asthma

266. Horton R, Lucassen A. Ethical Considerations in Research with Genomic Data. Published online 2022. doi:10.1080/20502877.2022.2060590

267. Chan V, Gherardini PF, Krummel MF, Fragiadakis GK. A 'data sharing trust' model for rapid, collaborative science. *Cell*. 2021;184(3):566-570. doi:10.1016/J.CELL.2021.01.006

268. Cai L, Zhu Y. The challenges of data quality and data quality assessment in the big data era. *Data Science Journal*. 2015;14(0). doi:10.5334/DSJ-2015-002/METRICS/

269. Poole N. *The Cost of Digitising Europe's Cultural Heritage: A Report for the Comité des Sages of the European Commission References and Acknowledgements*. The collections Trust; 2010. https://nickpoole.org.uk/wp-content/uploads/2011/12/digiti{_}report.pdf