

**DEPARTMENT OF ECONOMICS
UNIVERSITY OF STELLENBOSCH**

**INEQUALITY OF HEALTHCARE IN EARLY 20TH CENTURY
SOUTH AFRICA: A QUANTITATIVE STUDY OF PAARL**

by

20713479

Research assignment presented in partial fulfilment of the requirements for the degree of Bachelors of Economic and Management Sciences with Honours at the University of Stellenbosch.

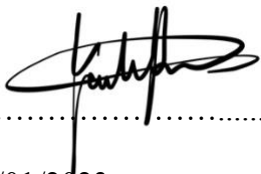
Supervisor: Prof Johan Fourie

March 2023

Declaration

I, the undersigned, hereby declare that:

- (i) the work contained in this assignment is my own work; and
- (ii) in the instance where my research assignment is based on previously submitted work, I have provided detailed information:
 - (a) regarding the nature, substance and origin of the overlap in the space below and throughout my research assignment (using standard referencing conventions or footnotes),
 - (b) regarding content that has been added to the previous submission,
 - (c) and that I understand that the evaluation of my research assignment will be primarily based on the new work.

Signature:

Date:05/01/2023.....

Acknowledgments

My thanks to WS Horn, Matthys Carstens, Stehan Malherbe, Aidan Forbes, Christiaan de Beer, Kara Botha, Madré Meyer, Vernon Simons, Gabriella Neilon and classmates for their continuous support and friendship throughout this journey. Thanks to my house mates, Emma Wiehman and Joshua Eva – your encouragement carried me through the frustration and stress while writing this dissertation.

An honorary mention to Bohemia and the SU Choir for becoming a friendly space where I could clear my head and relax after a long day working on this project.

A very special thanks to my supervisor, Johan Fourie. Not only for the invaluable advice and resources he allowed me to use, but for being patient and kind throughout this journey.

Lastly, thanks to my family and especially my parents, Resia and Nico, for their love and support. Without your patience, advice, and many snacks you carried to my desk, this dissertation would not have seen the light.

Contents

1.	Introduction	1
2.	Literature Review	2
2.1.	Socioeconomic Determinants of Health Outcomes	2
2.2.	Disease and Causes of Death	4
2.3.	Demographic Determinants of Health Outcomes	5
2.4.	Historical Context of South Africa	6
2.5.	The Town of Paarl	8
2.6.	Research Questions and Hypotheses Arising from the Literature Review	9
3.	Materials and Data	10
4.	Empirical Strategy	12
4.1.	Data Preparation and Cleaning	12
4.2.	Descriptives of Variables	15
4.3.	Plotting and Visualisation	16
4.4.	Econometric Methods and Statistical Inference	17
5.	Results	19
5.1.	Plotting and Visualization of Variables	19
5.2.	Econometric Results	27
5.2.1.	Healthcare Access	28
5.2.2.	Quality of Healthcare	29
6.	Discussion of Empirical Results	31
7.	Conclusions	31
8.	References	33

Tables

Table 1: Expanded dataset extract..... 14

Figures

Figure 1: Timeline of health-related milestones	7
Figure 2: Death certificate of Jan Pretorius.....	11
Figure 3: Number of deaths per year by race	20
Figure 4: Density plots of age at time of death	21
Figure 5: Occupations of deceased individuals by race	22
Figure 6: Prevalent causes of death by age group over time.....	23
Figure 7: Prevalent causes of death over time by race.....	24
Figure 8: Health access over time by race in Paarl	25
Figure 9: Duration of last illness by health access, grouped by race	26
Figure 10: Duration of last illness by health access, grouped by cause of death	27
Figure 11: Coefficient plot of Equation 1 results	28
Figure 12: Coefficient plot of Equation 2 results	30

1. Introduction

On the 1st of March 1900, my namesake, Jan Pretorius of Pentz Street, Paarl, died eleven months after contracting tuberculosis. He did not have the means to access healthcare, despite being employed as a mason. He left behind his wife, Maria Pretorius, and eleven children, most of whom were illiterate. Of his descendants, at least seven died of tuberculosis, though all of them had the luxury of medical care which their (grand)father did not have. The story of the Pretorius family of Pentz Street is a stark reminder of the health outcomes faced by many people of colour in early 20th century South Africa.

Human welfare is a foundation of economics. Like any other field of study, it is inextricably intertwined with the human experience and, as such, is greatly influenced by our mortality. The question of welfare and living standards has often been levied against health outcomes and mortality rates in the literature on economic history. Evidently, the aphorism that “the pleasures of life are worth nothing if one is not alive to experience them” by Cutler, Deaton & Lleras-Muney (2006: 97) rings especially true in the context of this study on mortality. In South Africa, child mortality from 1990 to 2005 has risen instead of decreased, one of only 12 nations to have this occur. The question of healthcare inequality therefore is still as relevant now as it was in the early 20th century. Hence, the aim of this thesis is to study how demographic and socioeconomic differences affected healthcare outcomes in early 20th century South Africa.

This thesis makes use of a range of empirical strategies, among which visualisation of data and econometric techniques, to test its hypothesis that people of colour had significantly worse health outcomes than whites, both in terms of access to healthcare and quality of medical care. The hypothesis was tested through the use of historical death certificates and it was found that coloureds and blacks were significantly less likely to have access to medical care and that the quality of medical care may have differed between racial groups. Additionally, the study argues that quality of healthcare received by individuals also depended on the cause of death, with disease such as tuberculosis and other respiratory diseases often receiving inferior treatment to other causes of death.

2. Literature Review

Across the globe, health outcomes have had many determinants historically among which income, race, residential environment, and access to healthcare. Furthermore, extensive literature has been written on the topic of historical mortality in the early 20th century, though in South Africa quantitative analyses of historical mortality trends remain elusive. The lack of reliable data and measurements of these determinants remain a main challenge not only for South African economic historians, but globally (Simkins & Van Heyningen, 1989: 79; Bengtsson & Van Poppel, 2011: 343; Millward & Bell, 2001: 700). Hence, the search for suitable proxies through which to measure these determinants is a main task for much of this study. Additionally, quantifying health inequality necessarily implies the need for a grasp of the historical context which one is analysing. To this end, I will conclude with a brief contextualisation of the mortality landscape of early 20th century South Africa, and particularly the Paarl/Cape Town region.

2.1.Socioeconomic Determinants of Health Outcomes

As a first socioeconomic determinant of health outcomes, I consider literacy as a measure of social class and educational attainment. To this end, literacy may be regarded as an important predictor of an individual's level of education and as such of mortality. In the literature, Atkinson, Francis, Gregory & Porter (2017: 1287) found that female literacy specifically influences infant mortality rates. This is because literacy is a good indicator of socioeconomic status, which influences access to medical services (Fourie & Jayes, 2021: 9), but also affects fertility rates and the agency of women (Atkinson, et al., 2017: 1284). However, there exists ambiguity in the literature regarding the effects of literacy on mortality and health (Green & Hamilton, 2013: 150). Jaadla, Potter, Keibek & Davenport (2020: 1017) and Edvinsson & Lindkvist (2011: 383), for example, find no evidence that literacy determines health outcomes when studying the effects of maternal literacy on infant mortality rates. However, both papers acknowledge the open-endedness of the debate, while also not accounting for heterogeneity between countries as well as racial disparities. Nevertheless, literacy is an important indicator of economic status, which is known to affect mortality.

There exists a reciprocal relationship between health and economic status. That is, health influences one's ability to work, generate an income and sustain wealth, while income and economic status are concurrently regarded as determinants of health outcomes (O'Donnell, Van Doorslaer & Ourti, 2013: 3; Eli, 2015: 476). Perhaps more importantly, health outcomes determine, to a great extent, living standards and economic progress. This is why health inequality is viewed as an imperative economic question. It is generally accepted that those who are financially better-off are also in better health due

to greater access to medical services, with many studies in the literature finding positive relationships between income and health (Baird, Friedman & Schady, 2011; O'Donnell, et al., 2013: 6; Jaadla, et al., 2020: 999-1000; Gyasi, Phillips & David, 2019: 1091).

Household income is an important economic determinant in explaining differences in the use of health services (Gyasi, Phillips & David, 2019: 1091). Importantly, it has also been found that low income is a significant barrier to healthcare access, with poor individuals delaying medical care until serious problems arise (Whitehead, Dahlgren & Evans, 2001: 834). Although reliable data on income and wages are often difficult to come by, income is likely related to occupation and place of residence with individuals of higher social and/or economic class often living and working in lower mortality environments (Jaadla, et al., 2020: 994).

Higher income not only implies greater access to, and affordability of, higher quality healthcare, but also unlocks the door to better housing, nutritious food, and better education (Birchenall, 2007: 353; Baird, et al., 2011: 847; Jaadla, et al., 2020: 1005; Eli, 2015: 468) – all of which are determinants of health. While higher income does allow one access to greater nutrition, Birchenall (2007: 353) argues that income is only partly responsible for gains in higher nutrition. This is due to the effect of disease and infection, which greatly affect an individual's ability to absorb nutrients from food (Arora, 2001: 736). Horrell & Oxley (2012: 1377), similarly find that increased income increases nutritional attainment for households, but that the physical environment in which individuals worked influenced nutritional impact. Hence, disease, labouring, and residential environment play an important role in the impact of income on generating good health outcomes.

Occupation is thus another imperative consideration in determining health outcomes and the literature points to a strong correlation between mortality and employment (Green & Hamilton, 2013: 154). In the economic history literature occupation is mainly classified according to the historical classification of occupations (HISCO) (Van Leeuwen, Maas & Miles, 2002). Mortality also varies pertaining to type of employment. For example, workers in sawmills often witnessed mutilating, sometimes fatal, accidents (Edvinsson & Lindkvist, 2011: 378). Moreover, occupation is regularly employed in the literature as an indicator of socioeconomic differences in mortality (Bengtsson & Van Poppel, 2011: 348). Fourie & Jayes (2021: 9) find that workers had varying levels of health access; with professionals & managers enjoying the greatest access to medical care. Labourers in farming and production, however, had the lowest access to healthcare. Both Edvinsson & Lindkvist (2011: 383) as well as Jaadla, et al. (2020: 994) also find that large differences exist in mortality between different occupational groups – though specifically find that agricultural workers had among the lowest mortality rates, regardless of the fact that they were some of the poorest labourers. This

produces an interesting and counterintuitive dichotomy, where the poorest workers and those with the least access to health services also consist of the lowest mortality rates. This dichotomy may best be explained by the residential environment.

A significant determinant of health and mortality resides with place of residence. Place of residence captures the long-term and generally static effects of the environment on mortality, which are typically determined by water quality, population density, medical knowledge, and food availability (Elo & Preston, 1992: 196). The literature often differentiates between urban and rural residence when discussing mortality. Historically, mortality rates were the highest in geographical areas with industry and dense populations (Woods, Williams & Galley, 1993: 41). Subsequently, the existence of the so-called urban ‘mortality penalty’ is well-known in the economic history literature, whereby high mortality plagued urban populations and rural populations had much lower mortality rates (Jaadla, et al., 2020: 994; Birchenall, 2007: 353). The rural/urban mortality divide may be attributed to the spread of infectious disease, owing to poor sanitation and hygiene in urban communities, as opposed to rural ones.

2.2.Disease and Causes of Death

Infectious diseases, specifically tuberculosis, were a leading cause of death in most developed countries during the 19th and 20th century (Elo & Preston, 1992: 187; Hanlon, Hansen & Kantor, 2021: 45). The Government of the Union of South Africa, in fact, found the spread of tuberculosis so alarming that they instituted the Tuberculosis Commission in the year 1912 to determine its risk (Collins, 1982: 785). The spread of infectious disease was likely exacerbated by poor sanitation and overcrowding, as overcrowding essentially exposes individuals to greater risk of infection due to higher contact with infected individuals (Edvinsson & Lindkvist, 2011: 387). Overcrowding also led to serious sanitary problems, often associated with sewerage and water usage (Alsan & Goldin, 2019: 587; Geruso & Spears, 2018: 147). To this end, it is clear that infectious diseases were rife in areas with overcrowding and dismal sanitary conditions in South Africa (Davenport, 1971: 7; Coovadia, Jewkes, Barron, Sanders, & McIntyre, 2009: 820).

During the first half of the 20th century the most prevalent respiratory diseases included tuberculosis, bronchitis, whooping cough, and pneumonia, all of which are mostly airborne and spread through contact with infected individuals (Hanlon, et al., 2021: 45). Importantly, respiratory disease (particularly tuberculosis) played a critical role in the difference between mortality of whites and people of colour (Meeker, 1976: 16). Gastrointestinal and digestive diseases, mostly waterborne, are often spread through faecal matter and include diarrhoea, typhoid, dysentery and cholera (Geruso &

Spears, 2018: 128; Beach, Ferrie, Saavedra & Troesken, 2016: 41; Hanlon, et al., 2021: 50). Infants are particularly susceptible to gastrointestinal illness, with diarrheal deaths accounting for almost a third of infant deaths in the 26 most populous American cities in 1910 (Anderson, Rees & Wang, 2020: 3). While not as prevalent as their respiratory and gastrointestinal counterparts, non-communicable diseases also directed trends in historical mortality, with cardiovascular disease, cirrhosis of the liver, and various cancers being the main culprits (Elo & Preston, 1992: 188; Hanlon, et al., 2021: 45).

While race and age cohort are both of critical importance, mortality also has a prevalent gendered dimension. Lives during the period under study were (and to a great extent still are) heavily gendered in terms of occupations and lifestyles. For instance, some studies have found that women are more likely to seek medical attention than men (Gyasi, et al., 2019: 1090), where others supply evidence that men often engaged in practices that exposed them to greater health risks – particularly alcohol misuse – which sees males often showing higher prevalence of cardiovascular, liver & kidney disease, as well as higher incidence of cancer (Edvinsson & Lindkvist, 2011: 386; Ross, Masters & Hummer, 2012: 1158).

2.3. Demographic Determinants of Health Outcomes

Race and gender as demographic determinants of health outcomes exist on both an institutional and individual level. On the institutional level, one sees discrimination based on gender and/or race manifesting through essential infrastructure and public service exclusion. Historically, the development of water and sanitation infrastructure has been fundamentally regarded as one of the most important public health interventions, particularly playing a central role in the reduction of infant mortality rates. Investments into water and sanitation infrastructure and treatments are known to lead to positive health outcomes for communities (Gallardo-Albarrán, 2020: 754; Chapman, 2022: 204). Alsan & Goldin (2019) researched the effects of water and sewerage infrastructure on quality of life and infant survival rates in late 19th and early 20th century Boston. They find significant positive contributions to infant health after public water and sanitation initiatives (2019: 18). Similarly, Peltola & Saaritsa (2019: 298) find that water chlorination significantly contributed to sharp declines in infant mortality in Finnish cities and towns between 1870 and 1938.

As early as 1850, major South African communities also started to develop relatively sophisticated water supply and sanitation infrastructure. Mäki (2010) studied the development of water and sanitation in four major South African urban centres: Cape Town, Grahamstown, Johannesburg, and Durban. All four urban centres had seen marked increases in population growth between 1850 and

1921 and had embarked on a mission to institute health-based water and sanitation policies to ensure access to clean water (Mäki, 2010). While whites enjoyed access to water and sanitation infrastructure, the same could not be said for people of colour.

One also observes the race-based exclusion from medical institutions as a common historical occurrence (Fourie & Jayes, 2021: 2; Hoehn-Valasco, 2018: 56). Institutional exclusion causes spill-overs to individual level health dynamics. The medical exploitation of demographic groups is a prime example of how institutional discrimination leads to negative health outcomes, with individual level repercussions resulting in reduced trust in medical services (Alsan & Wanamaker, 2018: 430). The result is that marginalised groups refuse to seek medical attention due to high incidence of institutional distrust. An important consideration are the differences between racial and ethnic groups, given the diverse demographic profile of South Africa consisting almost entirely of black Africans, white Europeans, south Asians and mixed-race individuals. Specifically, differences within race groups also exert influence on mortality outcomes. Green & Hamilton, for example, find significant differences between mixed-race blacks and non-mixed race blacks (2013: 157). One would thus expect to see not only differences between whites and people of colour, but between different ethnic groups as well.

Heightened inequalities regarding access to medical care, especially during health crises such as pandemics, have been noted in the literature as well (Blumenshine, Reingold, Egerter, Mockenhaupt, Braveman & Marks, 2008: 713). Specifically regarding South Africa, Fourie & Jayes (2021: 16) find that access to healthcare is strongly influenced by race and was exacerbated during the Spanish influenza of 1918. In fact, people of colour were often associated with disease, and blamed for the spread of numerous infectious diseases (Maylam, 1990: 61). Prejudice regarding race worsened the unequal health outcomes faced by people of colour in early 20th century South Africa.

2.4. Historical Context of South Africa

South African history is one plagued by prejudice and racialised governmental policies. Figure 1 illustrates important events and policies instituted between 1880 and 1925¹ that have been found to be influential in health outcomes. In the period under study, two specific policies warrant special mention, namely the Native Land Act of 1913 and the Public Health Act of 1919². The Native Land Act of 1913 restricted the areas of land that black South Africans were permitted to own, and the

¹ Note the Births and Deaths Registration Act of 1894, where this study finds its data source.

² For a comprehensive discussion regarding the public health and town planning legislation over 20th century South Africa, refer to Parnell (1993).

Public Health Act of 1919 imposed a number of regulations regarding the containment and spread of infectious diseases (Feinberg, 1993: 68-70; Coovadia, et al., 2009: 819; Parnell, 1993: 487; Zangel, 2017: 231-233).

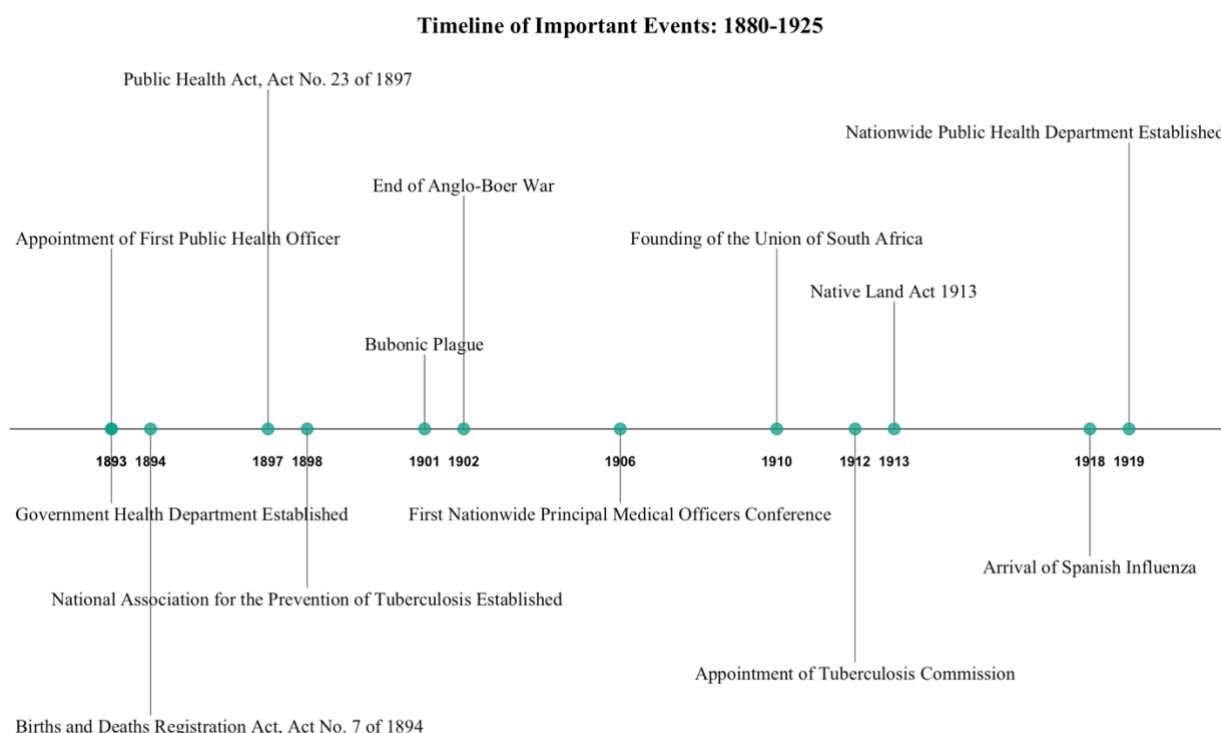


Figure 1: Timeline of health-related milestones. Source: Coovadia, et al. (2009), Zangel (2017), Tuberculosis Commission (1914); own calculations.

These policies had far reaching consequences regarding health outcomes for the black and coloured population, and were part of a broader system of racialised governmental policies that were implemented in South Africa during this period. While the Public Health Act did not explicitly limit the means of people of colour to access medical services, the main shortcoming of the Act lay in its inability to include the needs of black and coloured people³. The Native Land Act, however, did play a decisive role in relegating people of colour to overcrowded and unsanitary places of residence (Mpeta, Fourie & Inwood, 2018: 2). The legacy of these policies serve to further entrench inequality and oppression in South Africa, and their effects are still felt to this day (Coovadia, et al., 2009: 817).

Importantly, these policies reflect in health outcomes faced by different groups. For instance, the mortality rate of black and coloured South African infants was significantly higher than that of white

³ Zangel (2017) argues that the Public Health Act of 1919 was designed with the healthcare needs of specifically the white population in mind. Similarly, Parnell (1993) discusses how the Act supplemented segregationist policies founded some years after.

infants, with the same figures manifesting in the Cape as well (Marks & Andersson, 1987: 179; Mpeti, et al., 2018: 2; Simkins & Van Heyningen, 1989: 95). Additionally, people of colour were more likely to suffer from malnutrition and diseases such as tuberculosis, with markedly lower access to healthcare than whites (Andersson, 1990: 145 & 147).

Arguably, a lack of black and coloured medical practitioners also lay at the heart of many unequal health outcomes. In particular, the healthcare landscape in the Cape was dominated by white British born-and-trained doctors (Van Heyningen, 1989: 456). There was no shortage of medical men in the Cape between 1900 and 1930. In fact, the ballooning of medical practitioners in this period may have led to a notable decline in white mortality. Though since very few people of colour could neither afford nor wished to consult white doctors, the concentration of doctors in the Cape did little to reduce mortality for blacks and coloureds (Van Heyningen, 1989: 461).

To further our understanding regarding how demographic and socioeconomic differences resulted in heterogeneous health outcomes for different groups in South Africa, and in particular the Cape, this study uses the town of Paarl as a case study. Paarl is the third-oldest European settlement in South Africa and surprisingly little research has been done regarding historic mortality outcomes in this influential town. I specifically study mortality in Paarl in the period from 1900 to 1930.

2.5. The Town of Paarl⁴

Paarl is situated east of Cape Town in the Drakenstein Valley, a region characterised by a Mediterranean-like climate, fertile soil, and sources of water (Drakenstein Heritage Survey Group (DHS), 2012: 18; Pinkerton, 1814: 10-12). The first European settlements had been established by 1688 by then Dutch Cape Colony governor, Simon Van der Stel (Department of Water and Sanitation (DWS), 2004: 10; Horner & Wilson, 2008: 15). Paarl had a linear development, owing to the geographical constraints of the Berg River and Paarl Mountain, developing mainly along the Berg River, and producing a considerably long wagon track (currently the Paarl Main Road) (DHS, 2012: 27).

From the late 19th to early 20th century, Paarl had undergone industrialisation and had a mature agricultural sector. While the economy was mostly reliant on agriculture (cultivation of vines for wine production), wagon-building, wool washery, and the Paarl mill also contributed strongly (DHS, 2012: 24). One also observes racial segregation emerging in the early 20th century, with “social and

⁴ This section is an adjusted extract from Pretorius (2022: 2-3) submitted for Economic History 771

racial engineering” culminating to the Group Areas Act of 1961 where around 10,000 Paarl citizens were relocated (DHSB, 2012: 26).

The demographic profile of Paarl also points to a diverse population in the town. Fascinating work by Horner & Wilson (2008) delivers a comprehensive demographic profile of the Cape Province from the early 19th to late 20th century. By 1904 Paarl had reached an estimated population size of 30,423 people; of whom 41% were white; 52% were mixed/coloured; and 4% were black (Horner & Wilson, 2008: 73). The population grew steadily to a peak of 35,000 by 1921, where the proportion of both black and white residents decreased to 2% and 38% respectively (Horner & Wilson, 2008: 76). Additionally, only around 57% of the population lived in urban areas, which remained stable from 1900-1934 (Horner & Wilson, 2008: 73-77). Given the diverse demographic profile and importance of Paarl, it is likely that the town experienced the same racial discrimination as the rest of the country, making it an ideal case study for this thesis.

2.6. Research Questions and Hypothesis Arising from the Literature Review

This literature review points to inequality regarding access to healthcare, which may have influenced health outcomes for different race groups. Moreover, access to healthcare may have also been determined by socioeconomic indicators such as occupation and place of residence, as well as demographic indicators such as gender and race. Additionally, the literature review suggests that there exists inequality pertaining to the disease environment experienced and quality of healthcare received by individuals. The following two research questions, therefore, direct this study:

- 1.) Were there any significant differences in access to healthcare between different socioeconomic and demographic groups in Paarl?
- 2.) Does the quality of healthcare that individuals receive differ based on race and/or cause of death?

The hypothesis of this study is that people of colour had significantly worse health outcomes than whites, both in terms of access to healthcare and quality of medical care. The hypothesis was tested through the use of historical death certificates which captured many data points related to the determinants of health, as discussed throughout the literature review. The two research questions identified are directly linked to the hypothesis by disentangling individual level determinants of mortality, allowing for quantitative measures of health inequality.

3. Materials and Data

To test the hypothesis stated in the preceding section, this study made use of data transcribed from death certificates from Paarl between 1900 and 1930. The original death certificates are archived in the genealogical records of FamilySearch. The data are mostly consistent, with few missing observations and span 24,195 observations. A combination of graphic descriptions and econometric techniques were applied to study the health outcomes captured by the certificates. While I am not the first to conduct research on mortality in South Africa over this period, very few studies have made use of death certificates to conduct their research, hence this is arguably the first study of its kind.

Mortality registration was institutionalized in many countries of Europe in the early nineteenth century, covering all genders and social classes of the population (Bengtsson & Van Poppel, 2011: 349). In South Africa, deaths were recorded as required by government from 1895 after the institution of the Births and Deaths Registration Act, Act no. 7 of 1894. This was done to gain a better understanding of mortality rates and the causes of death, which would in turn help to inform public health policies (Zangel, 2017: 76) and allowed for more detailed analysis of mortality patterns.

An example of the data captured on a typical death certificate in the early 20th century, is seen in Figure 2, which displays the death certificate of Jan Pretorius. The standard certificate contains particulars, such as an individual's name & surname, sex, usual place of residence, age, race, marital status, and occupation. Other important information regarding the circumstances of an individual's death is also noted on the certificate. In particular the place of death, duration of last illness, cause of death, and date & time of death.

Importantly, section 12 of the death certificate was reserved for the signature of a 'medical man'. If signed by a medical man, a cause of death must be noted on the death certificate. In the case of Jan Pretorius, there was no medical man present and this part is left blank. Another section on the certificate is open to the 'informant' – the individual that registered the death. Maria Pretorius, widow to Jan Pretorius, was the informant in this example. The informant could either write their signature or leave a mark on the death certificate.

FORM OF INFORMATION OF A DEATH: ACT No. 7 OF 1894.

WARNING—The penalties for false statements wilfully made are the same as those for perjury.
Anyone who loses a COMPLETED registration form is liable to a penalty not exceeding £2.

DECEASED—

1. Christian Names and Surname Jan Pretorius
2. Sex Male
3. Usual place of Residence Verlatenkloof Bellington
4. Age 53 ~~45~~ years
5. Race (a) Mixed
6. Whether Single, Married, Divorced or Widowed (b) Married
7. Occupation Mason
8. Date of Death First March (1900)
9. Place of Death Verlatenkloof
10. Intended Place of Burial D. R. C. M. Cemetery
11. Causes of Death Consumption
- 11A. Duration of last Illness Eleven months
12. Medical Man's Name —

INFORMANT—

13. Original Signature [or Mark] Maria ^{ker} Pretorius
14. Qualification Deceased present at death
15. Residence Verlatenkloof

Signed in my presence on this _____ day of _____

Witness (c).

(This space intended for Rural Area reports).

The following spaces are reserved for the use of Assistants for Urban Areas, and of the Deputy Registrar.

No one else should fill them up.

When Registered 1st March 1900. Sub-district of Bellington
(Signature) Deaundene de Assist. to Deputy Registrar (Urban Areas).
When Registered 5th March 1900. District of Paarl
(Signature) Deaundene de DEPUTY REGISTRAR. No. of Entry. 15-57

N.B.—If the Certificate of a Medical Practitioner is produced, the causes of Death and duration of illness must be recorded in the Registration Book by the Deputy Registrar and Assistant to the Deputy Registrar (Urban Areas) as stated in such Certificate, which is to be attached to the form.

(a) If born outside Cape Colony, enter on the same line in addition to the Race, the name of the Country, State or Colony where born, if known.

(b) If married, divorced or widowed, state on this line 11 the total number of children deceased has had.

(c) When a Rural Assistant, Field-cornet or Police Officer writes out form for Informant, he should add the words "Form written out by me" and sign as "Rural Assistant," "Field-cornet" or "Police Officer," as the case may be.

NOTE—If Informants in Municipalities, V. M. Boards and Special Urban Areas, do not appear personally before D. R. or A. D. R., proper Declaration must be completed and attached hereto. In Rural Areas, Informants can report in three ways—one being under Declaration. Medical Certificates essential in Urban Areas.

Figure 2: Death certificate of Jan Pretorius. Source: Image 160, Paarl, 1900, South Africa, Cape Province, 'Civil Deaths, 1895-1972,' database with images, FamilySearch; National Archives Pretoria.

The registration of deaths was certainly not perfect. Zangel (2017: 76) also notes that the process had “teething problems” initially, but that the recording of deaths became more efficient as compliance rose. There also exists some inherent bias in the certificates, in that not all deaths were recorded and that some of those that were recorded were often not up to standard. Particularly, black and coloured deaths were at times underreported and some important information was only partially captured. For example, many death certificates contain a street name at place of residence, which indicates that the deceased likely lived in an urban area. Upon initial inspection, it would seem that Jan Pretorius’ death certificate has no street name recorded and that he possibly lived in a rural community. Closer analysis revealed that he and his family had in fact lived in Pentz street, urban Wellington.

Nevertheless, the individual level data captured by the death certificates remain invaluable. There is a large deficit of historical demographic data for South Africa. Hence, using death certificate data allows us to fill this void. More specifically, the data allows one to delve into demographic trends on a much more detailed level than before. Moreover, the deaths registered in the western region of the Cape, particularly Paarl and Wellington, are similar to other official records of the time (Fourie & Jayes, 2021: 4). Due to persistent racial inequalities in South Africa and lack of accurate historical data, the usefulness of death certificates allows one to quantify and understand the influence of historical mortality patterns on an individual level. Here, the research questions identified direct the methodology applied.

4. Empirical Strategy

The empirical methods contained in this section were used to answer both research questions. The two research questions pertain to access to healthcare and quality of healthcare. In order to answer these research questions, three empirical strategies were employed: data preparation, data visualisation, and econometric modelling. In the data preparation phase, the death certificate data was cleaned and organised into a useable format. Following data preparation, the data was visualised to study trends in order to determine preliminary relationships between variables and to guide econometric modelling. Finally, econometric and statistical techniques were used to model disparities in health outcomes.

4.1.Data Preparation and Cleaning

Data preparation involved cleaning the data and creating variables for analysis. The raw data has many formatting inconsistencies and therefore can only be used after comprehensive data cleaning and categorisation, as is the case with most historical data. To this end, preparing the data for analysis was the most time-consuming part of the study and required exceptional analytical accuracy. The data

preparation was documented and the methodology is explained in this section, as well as select appendices.

Data cleaning was initiated by developing and running cleaning scripts on statistical software, specifically R (see Appendix B). The data was then manually refined on Excel, as formatting differences in the data exist and preparation could therefore not be fully automated via R. Although the methodologies were similar for the two research questions, there were niche differences in the data cleaning techniques. For instance, a ‘needle-haystack’ programming technique was applied to clean the occupation, cause of death, race, and residential variables by combing each observation for specific terms and phrases; whereas simple ‘text to numeric’ scripts were used for the age at death and duration of last illness variables. A full list of tools used for data preparation and plotting in R is contained in the code document in Appendix B.

After mutating the data into a consistent format, the dataset was expanded with dummy coding, where additional columns containing 1’s and 0’s were appended to each categorical variable. An extract from the data containing the newly added dummy variables are shown in Table 1. There are eleven variables in total that were created and used, consisting of two numeric variables, namely: *duration of illness* and *age at time of death*; and nine categorical variables, namely: *sex*, *race*, *age cohort*, *informant signature*, *doctor signature*, *street name*, *occupation*, *cause of death*, and *pandemic*. Each of the variables in use require some exposition.

Table 1: Expanded dataset extract. Source: Cape Province, 'Civil Deaths, 1895-1972,' database with images, FamilySearch; National Archives Pretoria; own calculations.

Image Number	Race	Coloured	Black	White	Sex	Male	pandemic
197/1900	Coloured	1	0	0	Female	0	0
523/1916	Coloured	1	0	0	Male	1	0
438/1906	Coloured	1	0	0	Male	1	0
229/1925	White	0	0	1	Female	0	0
599/1918	Coloured	1	0	0	Female	0	0
155/1927	Coloured	1	0	0	Male	1	0
22/1911	White	0	0	1	Female	0	0
238/1911	White	0	0	1	Male	1	0
380&381/1907	White	0	0	1	Female	0	1
575/1930	Coloured	1	0	0	Male	1	0
709/1928	Black	0	1	0	Female	0	0
Image Number	Occupation	Occupation of Father	Occupation of Mother	Pensioners	Children	Farming	Duration of Last Illness
197/1900	Children			0	1	0	-0,2628
523/1916	Production			0	0	0	-0,3989
438/1906	Farming	Farming		0	0	1	0,1738
229/1925	Sales & Service			0	0	0	-0,0694
599/1918	Sales & Service			0	0	0	0,1738
155/1927	Professionals & Managers	Professionals & Managers		0	0	0	0,9845
22/1911	Professionals & Managers			0	0	0	0,4063
238/1911	Professionals & Managers	Farming		0	0	0	0,9845
380&381/1907	Sales & Service			0	0	0	-0,2282
575/1930	Production			0	0	0	1,2257
709/1928	Sales & Service			0	0	0	3,3618
Image Number	Signature or Mark	Informant Signature	Medical Man's Name	Medical Man Signature	Usual Place of Residence	Street Name	
197/1900	Mark	0	N/A	0	Fransche Rug F.C. Paardenberg Paarl	0	
523/1916	Mark	0	Dr D B Logie	1	Pentz Street Wellington	1	
438/1906	Mark	0	Dr E T F Malan	1	Jubilee Street Paarl	1	
229/1925	Signature	1	Dr De Jager	1	Church Street, Wellington, District Paarl	1	
599/1918	Signature	1	Dr De Villiers	1	Kortvlei, Kd, Paarl	0	
155/1927	Signature	1	N/A	0	School Street Lower Paarl	1	
22/1911	Signature	1	Dr AM Moll	1	Joostenberg, Achter Paarl, Paarl	0	
238/1911	Signature	1	R L Wolfe	1	De Hoop, Zuider Paarl	0	
380&381/1907	Signature	1	Dr Hammann	1	Nederburg Klein Drakenstein Paarl	0	
575/1930	Signature	1	Dr H A Hahn	1	Market Street, Wellington, District Paarl	1	
709/1928	Mark	0	D Hoffman	1	Fransch Hoek Distr Paarl	0	
Image Number	Production	Prof_and_Man	Sales_and_Service	Cause of Death	resp	TB	
197/1900	0	0	0	0 Convulsions/Indeterminate	0	0	
523/1916	1	0	0	0 Gastrointestinal Illness	0	0	
438/1906	0	0	0	0 Complications During Birth	0	0	
229/1925	0	0	0	1 Cardiovascular Illness	0	0	
599/1918	0	0	0	1 Convulsions/Indeterminate	0	0	
155/1927	0	1	0	0 Convulsions/Indeterminate	0	0	
22/1911	0	1	0	0 Respiratory Illness	1	0	
238/1911	0	1	0	0 Cardiovascular Illness	0	0	
380&381/1907	0	0	0	1 Respiratory Illness	1	0	
575/1930	1	0	0	0 Respiratory Illness	0	1	
709/1928	0	0	0	1 Sexually Transmitted Disease	0	0	
Image Number	gastro	cardio	birth	cancer	convulsions	STI	
197/1900	0	0	0	0	1	0	
523/1916	1	0	0	0	0	0	
438/1906	0	0	1	0	0	0	
229/1925	0	1	0	0	0	0	
599/1918	0	0	0	0	1	0	
155/1927	0	0	0	0	1	0	
22/1911	0	0	0	0	0	0	
238/1911	0	1	0	0	0	0	
380&381/1907	0	0	0	0	0	0	
575/1930	0	0	1	0	0	0	
709/1928	0	0	0	0	0	1	
Image Number	Age at Time of Death	0-9_Years	10-19_Years	20-34_Years	35-49_Years	50+_Years	
197/1900	2	1	0	0	0	0	
523/1916	23	0	0	1	0	0	
438/1906	1	1	0	0	0	0	
229/1925	54	0	0	0	0	1	
599/1918	43	0	0	0	1	0	
155/1927	7	1	0	0	0	0	
22/1911	12	0	1	0	0	0	
238/1911	64	0	0	0	0	1	
380&381/1907	78	0	0	0	0	1	
575/1930	51	0	0	0	0	1	
709/1928	26	0	0	1	0	0	

4.2.Descriptives of Variables

Age at time of death is the age in years that an individual has lived up until their death. *Age at time of death* was then categorised into different age categories. Five age categories were identified: *0-9 Years*, *10-19 Years*, *20-34 Years*, *35-49 Years*, and *50+ Years*; each with their own column of 1's and 0's.

Regarding *sex*, a dummy variable with 'male' as reference category was created. The variable is denoted as *male*.

Race was categorised according to racial classifications used by the South African government. The following race groups populate the data: *black*, *coloured*, and *white*; each with their own column of 1's and 0's.

If the observation had a medical man's signature, *doctor signature* is coded to 1 (0 otherwise). This variable was used as a proxy for access to healthcare. The rationale is that it is likely that an individual had been inspected by a doctor or sought medical attention if their death certificate was signed by a medical man. Fourie & Jayes (2021) has also employed this proxy.

Where an informant had signed a death certificate instead of leaving only a mark, *informant signature* was coded to 1 (0 otherwise). This variable was employed as a proxy for literacy of the deceased.

Similarly, if usual place of residence contained "street", "road", or "lane", *street name* was coded to 1, and 0 otherwise. The variable was used as a proxy for whether an individual lived in an urban or rural environment.

A dummy variable, *pandemic*, was added to indicate whether the registered death happened during the Spanish flu pandemic of 1918 to control for the detrimental effect the Spanish flu had on healthcare. These effects were briefly discussed in the literature review.

Regarding *occupation*, six HISCO occupational categories were identified, namely: *farming*, *production*, *sales & service*, *professionals & managers*, *children*, and *pensioners*. Additionally, the data also contain information on the occupation of the deceased's father and mother. To this end, where no occupation for a deceased individual was indicated, the occupation of the individual's father (or mother) was used. This allowed for an estimation of the economic class to which an individual belonged. Furthermore, parental occupation affecting income and economic class is a reasonable assumption, given that occupational decisions historically were influenced by skills possessed by the

father of an individual (Saad, 1989: 457). The code used to group occupational titles into occupational groups is shown in Appendix B.

Cause of death consists of eight categories: respiratory illness (*resp*), gastrointestinal illness (*gastro*), cardiovascular illness (*cardio*), complications during birth (*birth*), cancer (*cancer*), convulsions/indeterminate⁵ (*convulsions*), tuberculosis (*TB*) and sexually transmitted illness (*STI*). The code document in Appendix B shows how causes of death were grouped into the abovementioned categories.

Duration of last illness is a numerical variable that captures the estimated period an individual was unwell before succumbing to the cause of their death, measured in days. As the duration of illness differs depending on the cause of death, *duration of illness* was standardised by cause of death in order to make observations comparable. This variable was used as a proxy for the quality of healthcare received by an individual if that individual had access to a medical man. The rationale is that the a longer duration implies that the medical man could keep the deceased alive over longer periods.

4.3. Plotting and Visualisation

The aim of this empirical strategy was, firstly, to explore trends in the data; and secondly, to understand how the different variables relate to each other. These variables pertain to both access to healthcare and quality of healthcare, and thus both of the research questions are explored here. The prepared data was visualised with R and the code used to produce data visualisations is contained in Appendix B.

The first relationship that was explored was how mortality rates have evolved over time per age cohort by race. This involved a simple line plot to study whether mortality rates decreased, increased or stagnated. Another visualisation was presented to study the distribution of age at time of death by race, literacy, and place of residence. Regarding literacy as a socioeconomic indicator, the study makes use of the fact that the informant that registered the deceased individual signed or marked the death certificate. The assumption is made that if the informant, which is usually a family member or friend of the deceased, had signed that certificate that they are likely literate. Given the fact that the informant was likely literate, it may also be presumed that the deceased individual was literate. These

⁵ Many causes of death were noted as convulsions. This was commonly noted as a cause of death at the time. In modern times, however, we know that there is some underlying cause of death and convulsions may be attributed to a broad range of causes. Therefore, where only convulsions were indicated as the cause of death, it was categorised as indeterminate to control for this in the model.

visualisations provide a broad overview of the differences between mortality rates of white, coloured, and black individuals.

The second relationship that was visualised was how occupations differ between race groups. The aim was to illustrate whether people of colour were more likely to be employed in lower income and/or higher mortality occupations. These relationships were presented through stacked bar plots to show proportionally how occupational groups were distributed. Additionally, occupations were compared between individuals that had a street name on their death certificates, as well as recorded deaths with informants that have signed or marked the certificate.

Next the study graphically presented how cause of death evolved over time for different age and race groups. The aim of these visualisations was to study trends in prevalent diseases in Paarl over the period of study and whether specific diseases affected different age groups and races. As infant mortality is an important indicator of welfare, there is also a focus on causes of death for children aged 0-9 years. Similar to the occupation visualisations, the study made use of stacked bar charts to present the proportion of deaths each cause of death consists in.

Finally, the study visualised trends for the two outcome variables that were employed for statistical inference: access to and quality of healthcare. Simple line plots were used to study trends in the number of death certificates signed by a medical man over time. This, I believe, likely presents how healthcare access evolved over time for different race groups. To study whether there were differences in duration of illness for different groups that had a medical man's signature and those that did not, boxplots were produced. The boxplots allowed for preliminary analysis of how duration of illness was distributed for those with or without healthcare access.

4.4.Econometric Methods and Statistical Inference

For the final empirical strategy, suitable indicators of health outcomes had to be identified. The indicators must be appropriate for the population being studied and must be able to provide accurate and reliable results on health outcomes. To this end, two response variables were identified, namely the presence of a signature from a medical man on the death certificate and the duration of last illness. I propose that these two variables were related to healthcare access and quality respectively. Explanatory variables were chosen based on factors that influence health outcomes, as discussed in the literature review, as well as the results of the data visualisations. Two specifications were produced that directly link to the research questions, using *doctor signature* and *duration of illness* interacted with *race* as response variables respectively.

For the first research question, the aim was to estimate whether there were any significant differences in access to healthcare for different demographic groups and social classes. In this model, three specifications were employed and the presence of a medical man's signature was used as response variable. The signature of a medical man is considered a proxy for access to healthcare. A logistic regression model was used and is presented in Equation 1 below. For the full model, the explanatory variables are *race*, with whites as the reference category; *sex*, with male as reference category; *street name*, a proxy for whether the deceased lived in an urban environment; *informant signature*, if the individual's informant signed the certificate with their signature; and *occupation*, with professionals & managers as reference category. A control variable, *pandemic*, was added to control for the detrimental effects of the Spanish flu on healthcare access.

Equation 1: *Specification 1 proposed model.*

$$\text{Doctor Signature} = \alpha + \beta_1 \text{Race} + \beta_2 \text{Sex} + \beta_3 \text{Street Name} + \beta_4 \text{Informant Signature} + \beta_5 \text{Occupation} + \beta_6 \text{Pandemic} + \mu$$

The second research question asks whether there were any significant differences in the quality of healthcare received by white, coloured, and black individuals by numerous indicators. Four linear regression models were specified: a full model for all races, and models for whites, coloureds, and blacks. The mathematical expression for the models are shown in Equation 2. There are six explanatory variables: *doctor signature*, in order to estimate whether access to healthcare increased duration of illness; *cause of death*, with death by convulsions/indeterminate as reference category; *age cohort*, to control for the effect that age had on shortening/prolonging duration of illness with 20-34 Years as reference age category; *street name*, to study whether an urban residential environment affected duration of illness; and *informant signature*, to isolate the effects of literacy on duration of illness.

Equation 2: *Specification 2 proposed model.*

$$\text{Duration of Illness} \times \text{Race} = \alpha + \beta_1 \text{Doctor Signature} + \beta_2 \text{Cause of Death} + \beta_3 (\text{Doctor Signature} \times \text{Cause of Death}) + \beta_4 \text{Age Cohort} + \beta_5 \text{Street Name} + \beta_6 \text{Informant Signature} + \mu$$

5. Results

The following section provides the empirical results discussed in the methodology. The study first interpreted the results of the data visualisation. The visualisations confirmed many of the health disparities that were found in the literature review. Following visualisation, the results for the econometric methods were discussed and interpreted.

5.1. Plotting and Visualization of Variables

In the literature discussing South African mortality, it has been found numerous times that mortality rates differ based on demographic and socioeconomic factors. The same trend is present in the death certificate data. Figure 3 represents the number of deaths in Paarl for white and coloured individuals by age cohort. While consisting of approximately 41% of the population in Paarl during the period under study, mortality rates for white individuals was significantly lower than that of coloured individuals, which made up around 52% of the population. Additionally, mortality for infants and children reduced dramatically for whites, whereas coloured infant and child deaths showed no clear change over the period. In fact, infant deaths from the coloured population made up the largest portion of deaths over the period, remaining significantly higher than any other age cohort.

Figure 4 contains density plots of age at time of death given different specific demographic and socioeconomic indicators, namely race, literacy, and place of residence. Characteristic to the period under study, most deaths could be attributed to infants. Infant mortality rates appeared to be much higher for individuals whose informant had not signed their certificate with a signature. Similarly, infant mortality was higher for those that had no street name appear on their death certificate. Additionally, it seems like individuals whose certificates contained an informant signature and street name tended to die at older ages than those that did not, though this was a smaller difference.

Given these trends, the data provided preliminary evidence that there were indeed differences in mortality outcomes given race, literacy, and place of residence, in congruence with the literature on early 20th century mortality. Another socioeconomic indicator of health outcomes that was discussed in the literature review was an individual's occupation.

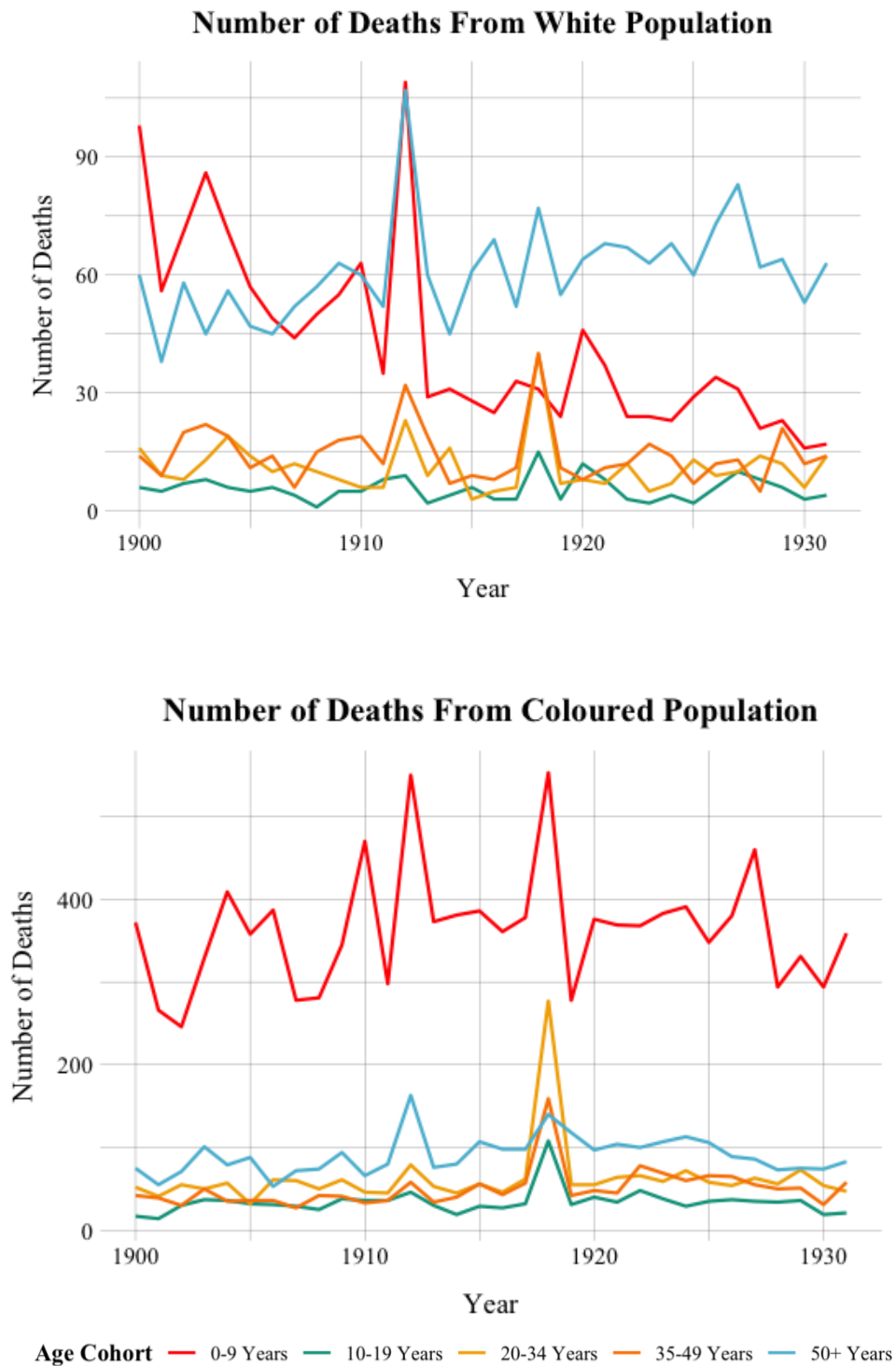


Figure 3: Number of deaths per year by race. The top panel represents the total number of death certificates for white individuals. The bottom panel represents the total number of death certificates for coloured individuals. The sample size for black individuals is too small to produce a meaningful plot. The same rhetoric is followed in most visualisations. Source: Cape Province, ‘Civil Deaths, 1895-1972,’ database with images, FamilySearch; National Archives Pretoria; own calculations.

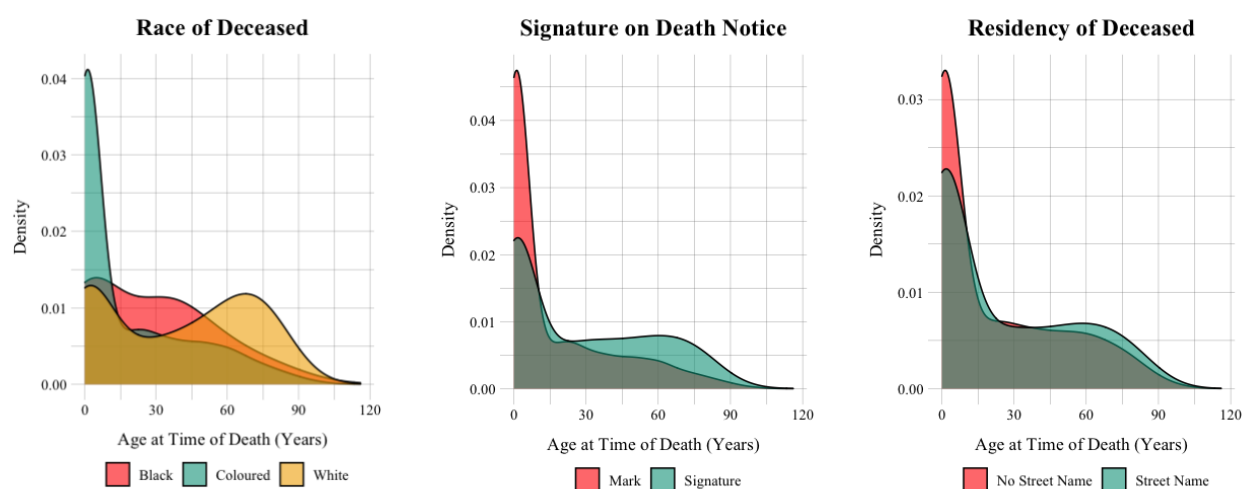


Figure 4: Density plots of age at time of death. Panel (a) illustrates age at time of death for black, coloured and white individuals. Panel (b) shows the distribution of age at time of death for individuals with informants that signed their death certificates and those that have only been marked. Panel (c) illustrates the distribution of age at time of death given that the deceased's death certificate contained a street name in the 'Usual place of residence' field. Source: Cape Province, 'Civil Deaths, 1895-1972,' database with images, FamilySearch; National Archives Pretoria; own calculations.

Figure 5 illustrates the distribution of occupations across race, based on an informant's signature or street name on the death certificate. It is clear that coloured and black individuals were mostly employed in farming and production as labourers. Interestingly, there seemed to be little difference in occupational group distribution between individuals with a mark or signature on their death certificate, suggesting that occupational status had a much stronger racial dimension than educational dimension. For certificates that contained a street name, less coloured and black individuals were employed in farming, where employment in production was higher. This is a logical outcome, given that most opportunities for employment in farming was found in more rural areas and conversely so for production.



Figure 5: Occupations of deceased individuals by race. Top panel shows occupation based on whether a signature or mark appears on the certificate. Bottom panel indicates occupation based on whether a street name appears on the certificate. Source: Cape Province, 'Civil Deaths, 1895-1972,' database with images, FamilySearch; National Archives Pretoria; own calculations.

Figure 6 shows the evolution of causes of death over the period under study for different age groups. A clear trend was the increase in cancer and cardiovascular illness as a cause of death as individuals age. Another important observation was the high prevalence of respiratory disease across all age groups. It is clear that gastrointestinal illness disproportionately affected younger individuals more.

Similarly, Figure 7 plots prevalent causes of death by race for all age groups (top panel) and ages 0-9 years (bottom panel). An important observation emerging from these plots is that respiratory disease was more prevalent under coloureds, while whites had much higher incidence of cancer and cardiovascular disease. Additionally, respiratory and gastrointestinal illness became a less prevalent cause of death over the period studied for the white population, while remaining generally consistent for coloured individuals. The sharp increase in respiratory disease in 1918 can be attributed to the arrival of the Spanish flu.

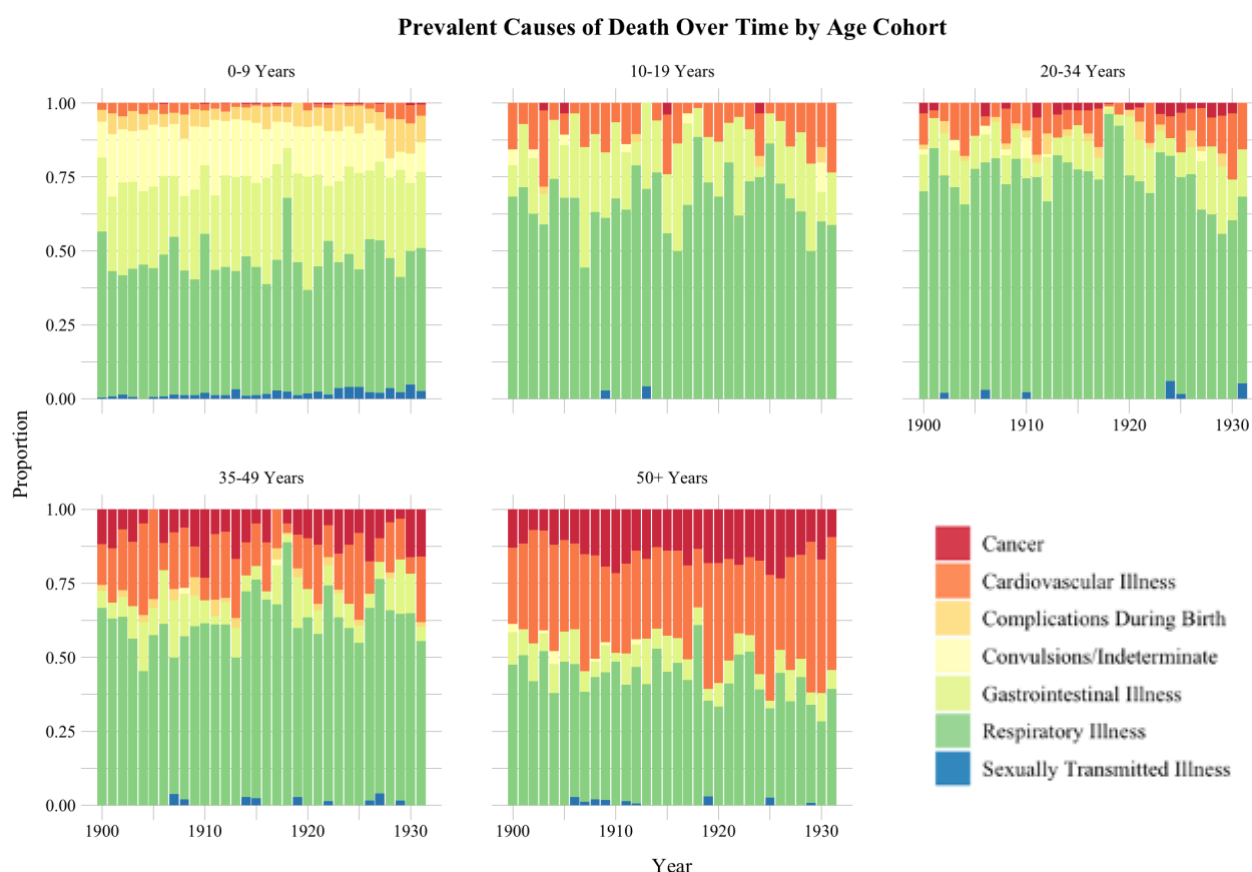


Figure 6: Prevalent causes of death by age group over time. Source: Cape Province, 'Civil Deaths, 1895-1972,' database with images, FamilySearch; National Archives Pretoria; own calculations.

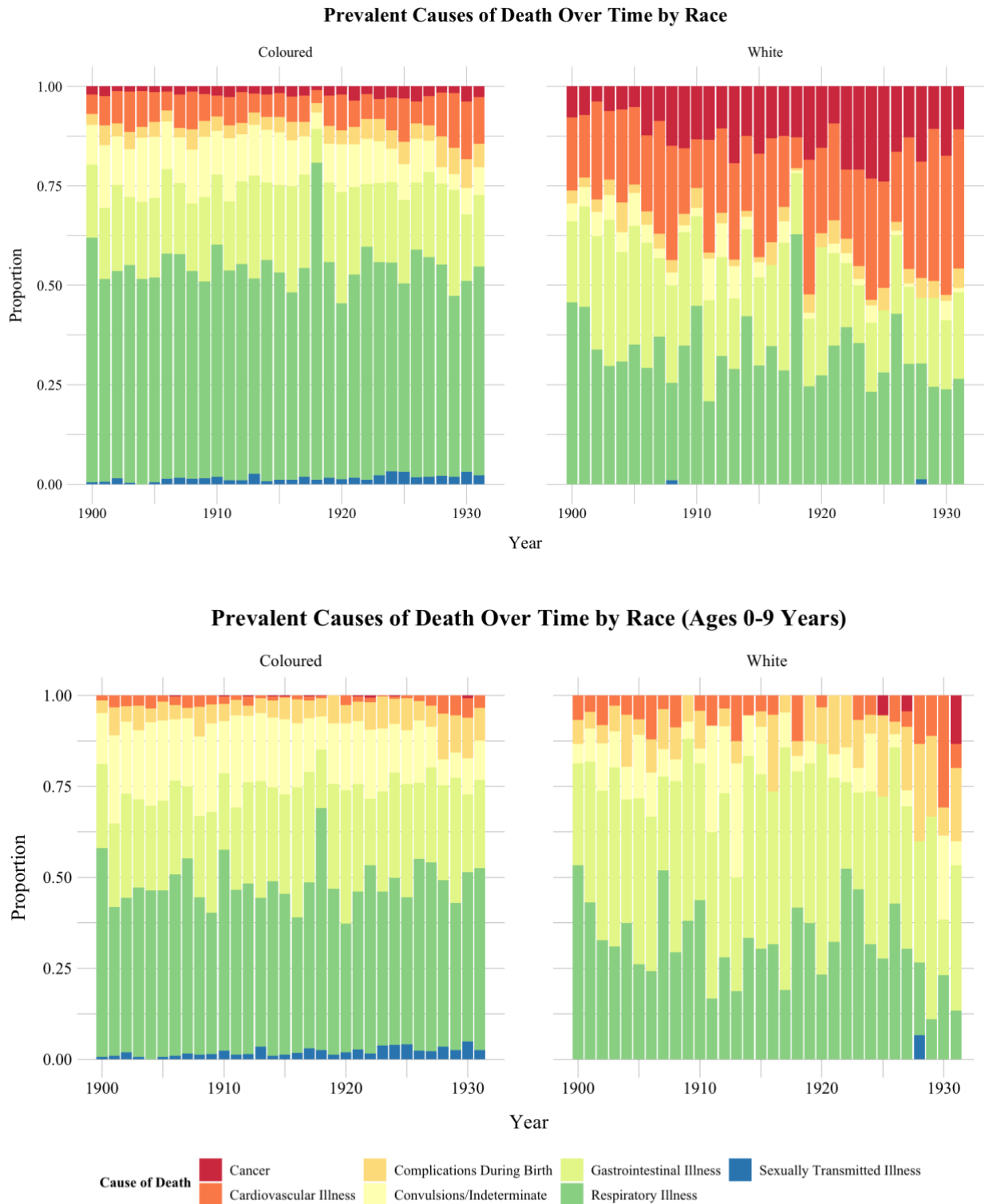


Figure 7: Prevalent causes of death over time by race. Top panel presents prevalent causes of death over all age cohorts for white and coloured individuals. Bottom panel reflects prevalent causes of death over time for white and coloured young children aged 0-9 years. Source: Cape Province, 'Civil Deaths, 1895-1972,' database with images, FamilySearch; National Archives Pretoria; own calculations.

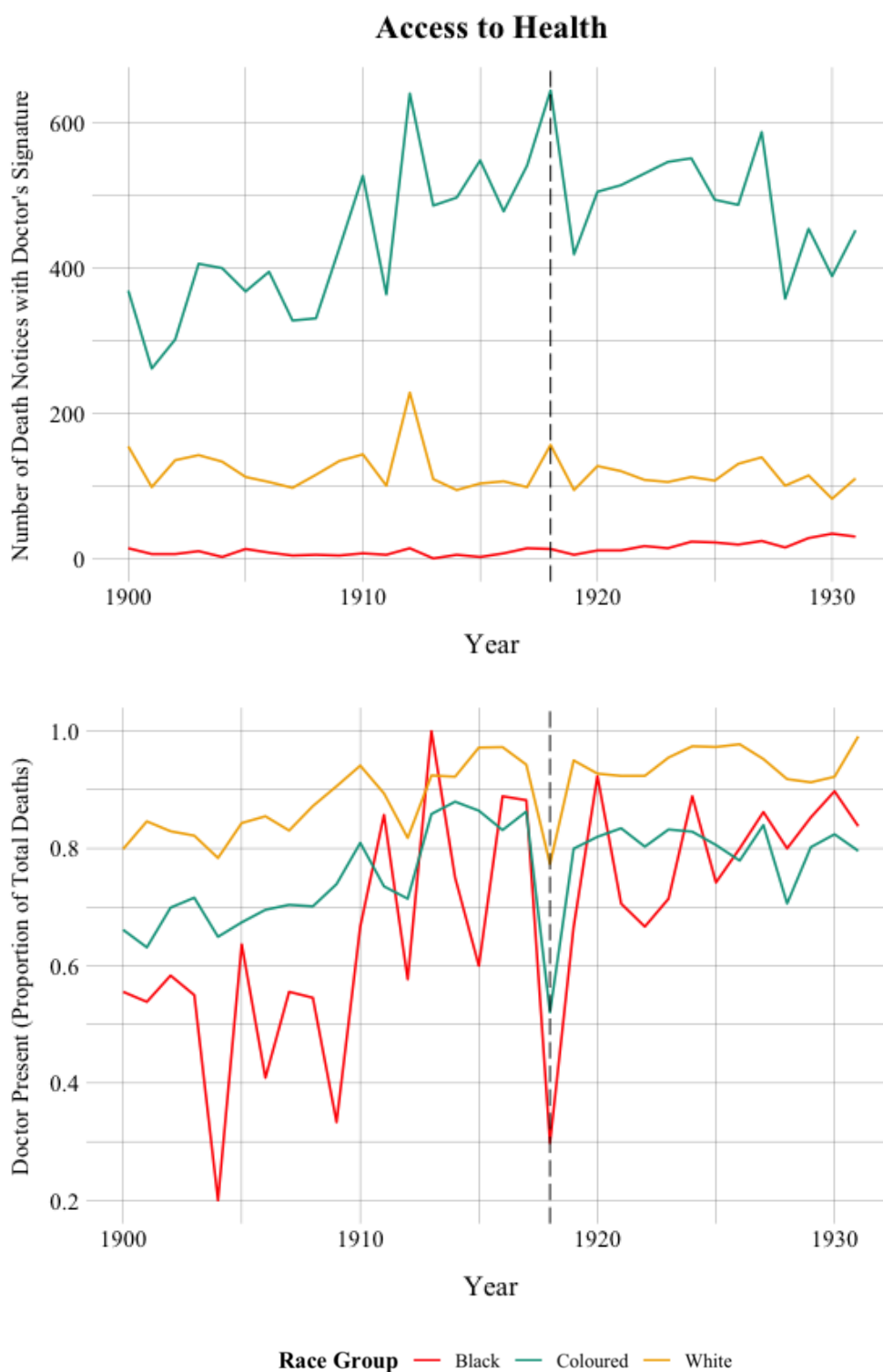


Figure 8: Health access over time by race in Paarl. The top panel represents the total number of death certificates signed by a medical man. The bottom panel represents the proportion of death notices signed by a doctor as a proportion of total deaths in a race group. Dotted line indicates the arrival of Spanish flu in 1918. Source: Cape Province, 'Civil Deaths, 1895-1972,' database with images, FamilySearch; National Archives Pretoria; own calculations.

Regarding the first research question, the study aimed to measure how access to healthcare was influenced by socioeconomic and demographic factors. As a proxy for healthcare access, I used whether a medical man's signature appeared on a death certificate or not. Figure 8 above plots the number of death certificates signed by a medical man as it evolved over time. Two important trends emerged: 1) there was a general upward trend in the proportion of death certificates signed by a medical man for all races; 2) there was a significant drop in the proportion of signed certificates in 1918 (more so for people of colour), coinciding with the arrival of the Spanish flu in South Africa.

Regarding the second research question, which studied the quality of healthcare received, Figure 9 sought to discover whether there are differences in the duration of illness for individuals' certificates signed by a medical man or not, grouped by race. For all races, the box plots revealed that access to healthcare may have indeed improved the duration of last illness. White individuals in general seemed to have shorter duration of illness until death. However, this might have been due to higher incidence of cardiovascular illness, which was much more acute than other illnesses, such as tuberculosis (which mostly affected the black and coloured population) of which individuals could suffer from for years.

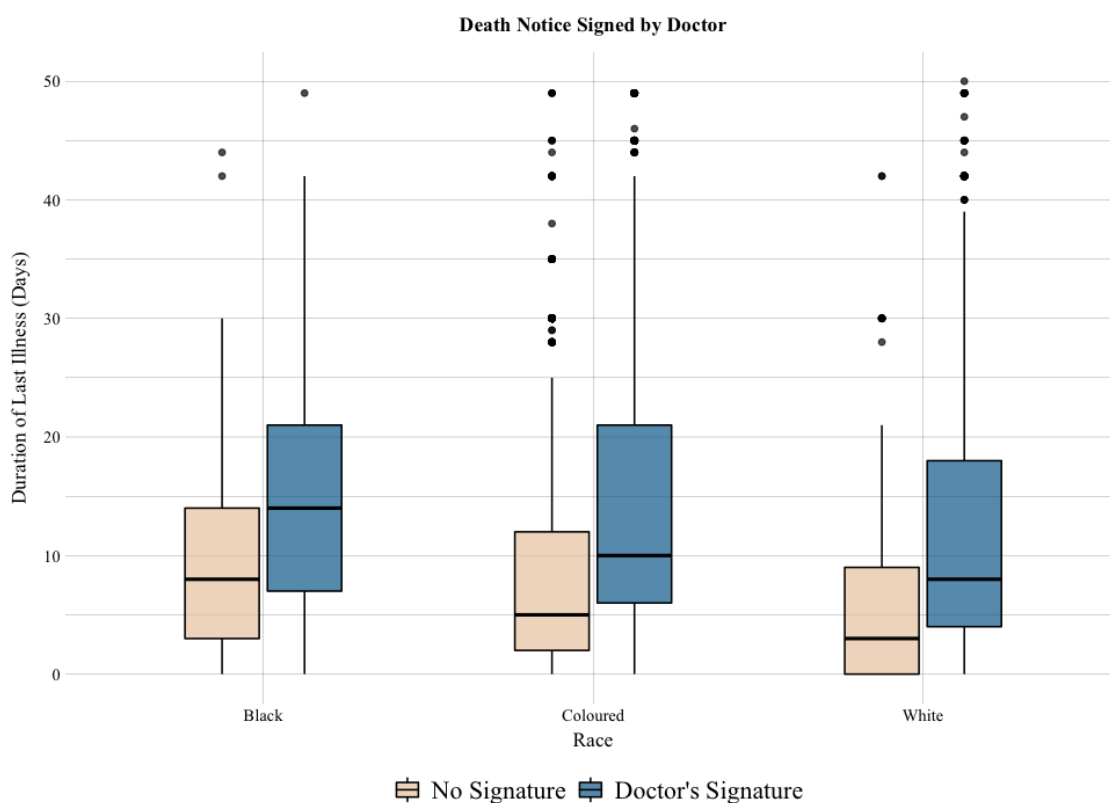


Figure 9: Duration of last illness by health access, grouped by race. These boxplots were produced with uncontrolled (not standardised) duration of illness values. Source: Cape Province, 'Civil Deaths, 1895-1972,' database with images, FamilySearch; National Archives Pretoria; own calculations.

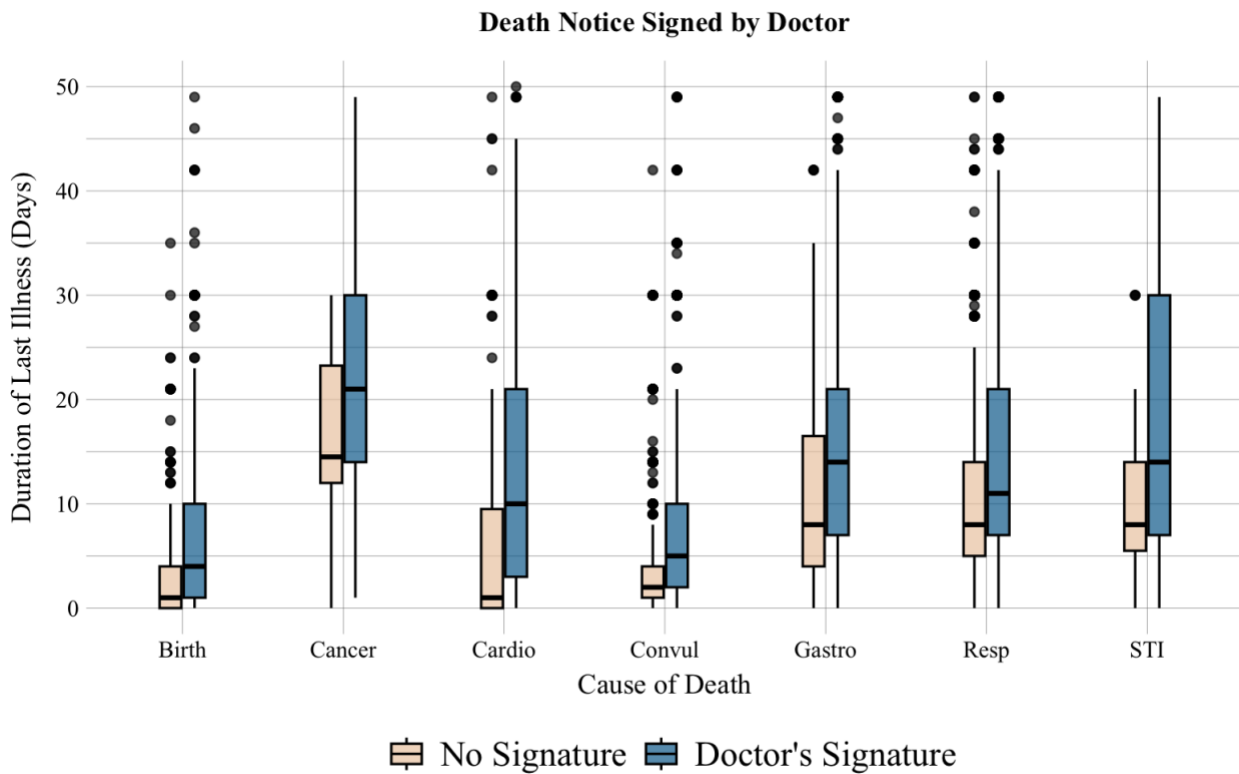


Figure 10: Duration of last illness by health access, grouped by cause of death. These boxplots were produced with uncontrolled (not standardised) duration of illness values. Source: Cape Province, 'Civil Deaths, 1895-1972,' database with images, FamilySearch; National Archives Pretoria; own calculations.

Figure 10 above plots the distribution of the duration of illness for the cause of death categories identified in the study. The aim was to observe whether access to healthcare could extend the duration of illness for different disease categories. For all categories, Figure 10 shows that those individuals with a medical man's signature did on average live longer than those who had no signature. These preliminary results suggest that healthcare was less accessible for people of colour and that those who had healthcare access could see their lives extended. However, to provide concrete evidence of this hypothesis, statistical inference was necessary.

5.2.Econometric Results

Following the descriptive account of the variables as discussed above, models for the two research questions were developed and focussed on access to healthcare (Research Question 1) and quality of healthcare (Research Question 2). Figures 11 & 12 illustrate the results of the models presented in Equations 1 & 2 respectively. Tables A1 & A2 in Appendix A summarise the results.

5.2.1. Healthcare Access

A waterfall logistic regression was used to produce the first model by continuously adding more of the variables specified to the model for each specification. Specification 3, the full model, is represented by the red dots in Figure 11. The results from the race variable, with white as the reference category, indicate significant differences in healthcare access for black and coloured individuals relative to whites. The results suggest that coloureds and blacks are 0.83 and 1.10 times less likely, respectively, than whites to have their death notice signed by a medical man. These results are significant at a 0.1% significance level.

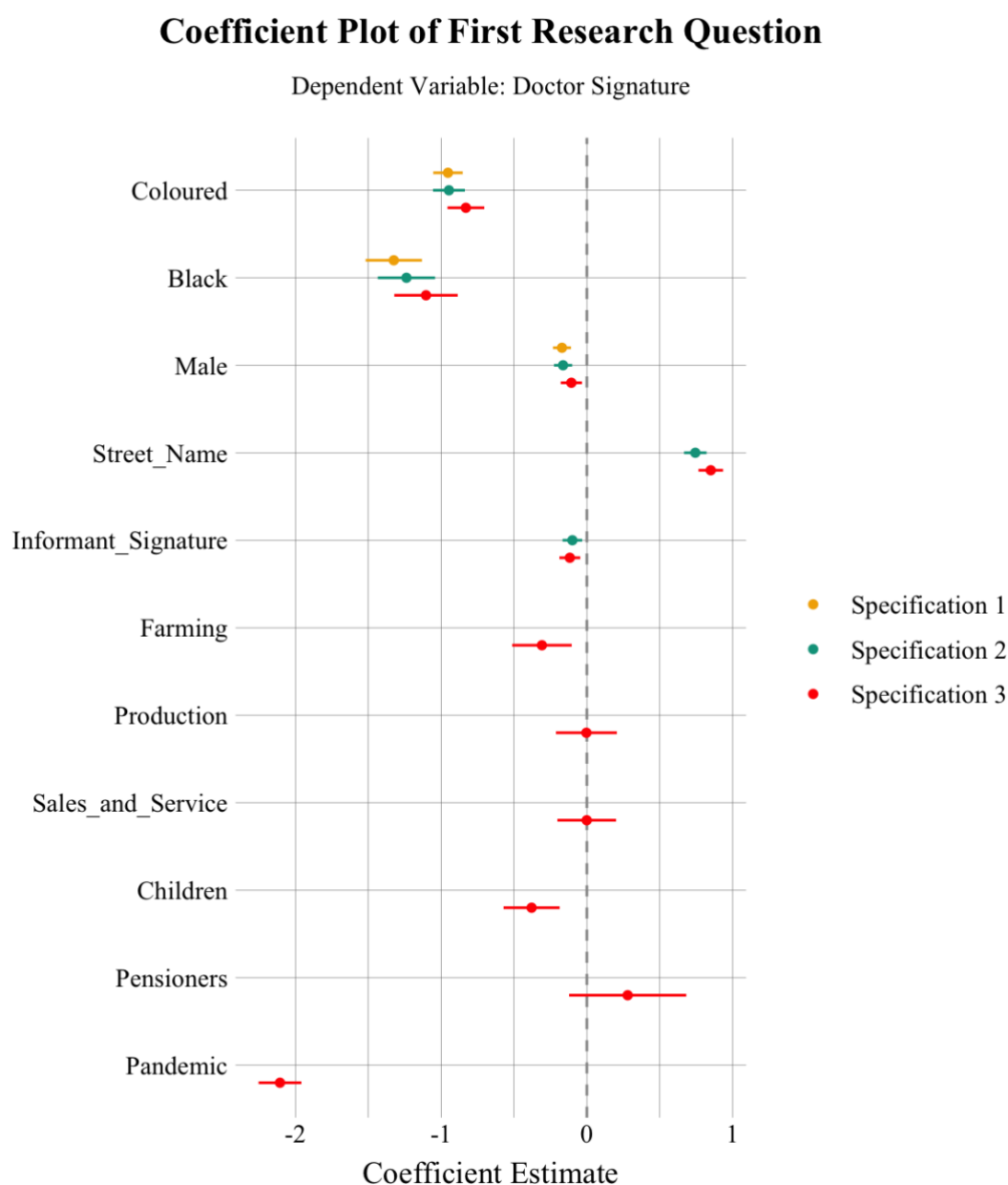


Figure 11: Coefficient plot of Equation 1 results. Source: Cape Province, 'Civil Deaths, 1895-1972,' database with images, FamilySearch; National Archives Pretoria; own calculations.

Men were also approximately 0.11 times less likely to access medical services than women. Having a street name appear on a death certificate increased the likelihood of medical access by 0.85. Interestingly, the informant signature variable had a negative coefficient, indicating a 0.12 times lower likelihood of accessing medical care than those with no informant signature. These results were significant at a 1% significance level.

The occupation variables also had some significant results, with those employed in farming around 0.31 times less likely to access medical care than those employed in professional & managerial occupations. There seemed to be no significant difference in access to healthcare for those employed in production and sales & services relative to professionals & managers. Pensioners had a 0.21 times higher likelihood of access to medical care than professionals & managers, though this result was insignificant. Furthermore, and perhaps most importantly, children were 0.38 times less likely to have had access to healthcare at a 0.1% level of significance, unless one of their parents were employed as professionals or managers.

5.2.2. Quality of Healthcare

A multiple linear regression model was employed to measure the quality of healthcare. The results are illustrated in Figure 12. The first model with all races indicate that the presence of a doctor's signature on a death certificate increased the duration of illness by 0.37 standard deviations above the mean duration at a 5% significance level. While whites and blacks showed no significant differences in duration of illness whether a medical man signed or not, coloured individuals saw an increase of 0.4 standard deviations above the mean duration of illness given that a medical man signed their certificate at a 1% level of significance.

Similarly, individuals with a street name on their death certificate saw an increased duration of illness of 0.04 standard deviations above mean duration, where coloureds had an increase of 0.03 standard deviations; both at a 5% significance level. An informant signature also indicated significant results, with whites seeing a 0.02 standard deviation increase in duration, where coloured and black individuals had a 0.04 and 0.01 standard deviation decrease in duration of illness.

All interaction terms with cause of death and doctor's signature delivered insignificant results. The cause of death reference category was STI, chosen as reference since the duration of illness for this cause of death consisted in the least variation. While none of the results are significant, an interesting trend emerged: the coefficients on all causes of death were positive for white individuals (save for cancer) and negative for coloured individuals. The results from the model may also suggest that a medical man did not make a difference in the duration of illness for all causes of death.

Coefficient Plot of Second Research Question

Dependent Variable: Duration of Last Illness*Race

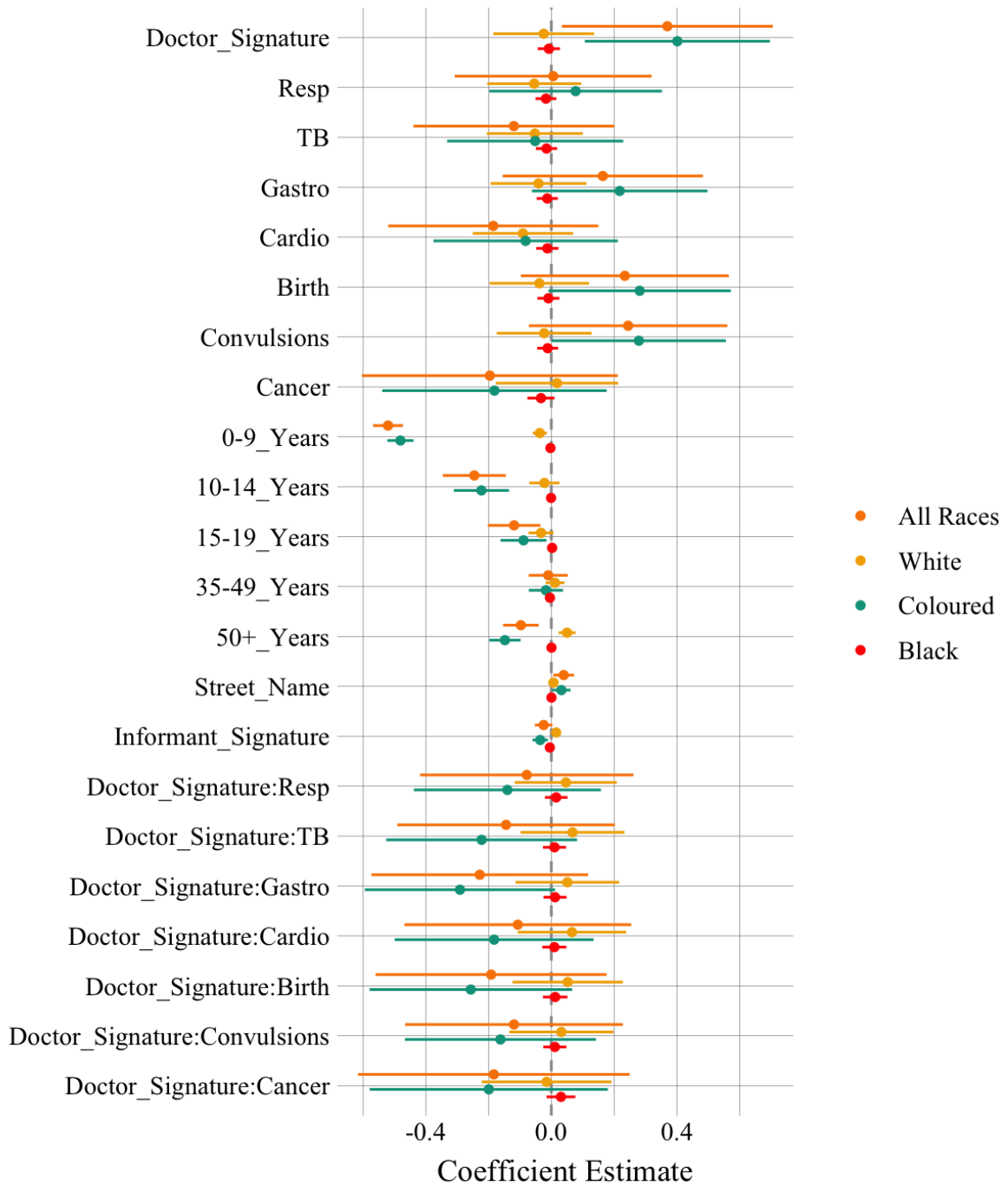


Figure 12: Coefficient plot of Equation 2 results. Source: Cape Province, 'Civil Deaths, 1895-1972,' database with images, FamilySearch; National Archives Pretoria; own calculations.

6. Discussion of Empirical Results

Following the results, this section seeks to address possible criticisms that may be levied against the empirical results contained in this study. The first regards the proxy for access to healthcare, a medical man's signature. It is possible that this is a weak measure of access to healthcare. A general assumption is that access to healthcare decreases mortality, yet mortality is the defining feature of this data set. As such, there exists inherent bias in the death certificate data in that one only takes account of those left behind – not of those who were successfully treated by a medical man and could live another day. This also then has even more far reaching consequences for the proxy for quality of healthcare, namely duration of last illness.

The logic behind using the duration of illness as a proxy for quality of medical care is that if the medical man could increase the lifespan of an individual, it may imply that they more competently treated their patient. This study then does not take account of the quality of medical care received by survivors of illness. Additionally, in the literature review it was discussed how people of colour and poorer individuals often delayed medical care until a serious problem arose. For this reason, it is possible that individuals only turned to medical consultation when illness has advanced so far as to have become untreatable. The result is that medical professionals could not extend the lifespan of individuals, even if the quality of medical care was similar to those who actively sought medical help at early stages of illness.

While these issues may be of great concern, the results of the model still provide important information into the inequality of healthcare that was present in South Africa in the early 20th century. This study does not purport to assume that selection bias is not significant, but it does argue that even at the point of death, some individuals were treated differently than others – and *that* is the core of healthcare inequality.

7. Conclusions

In 2022, the Pretorius family of Winelands Street, Paarl, is privileged enough to not have been affected by tuberculosis or other treatable diseases. They are some of the lucky few in South Africa, where most still suffer from widespread pandemics such as tuberculosis as well as healthcare inequality. While many improvements have been made – infant mortality dramatically decreased from 2005 to present – health outcomes in South Africa remain highly racialised and the root of this may be found by studying how healthcare developed in the early 20th century.

The results observed in this study are likely the product of the racialised policies of the time. It was illustrated that infant mortality rates were much higher for people of colour, those who were illiterate, and those that lived in rural areas. Additionally, it was revealed that disease prevalence was highly racialised with people of colour suffering more from respiratory diseases, and that lifestyle diseases such as cardiovascular illness and cancer were more likely to affect white individuals. Occupations were also racialised, with most people of colour being employed in lower paying jobs, and whites enjoying most of the higher paying occupations.

It was argued that black and coloured individuals were much less likely to have access to healthcare and that those who did have access could not obtain healthcare of the same quality as their white counterparts. The likelihood of healthcare access, while mostly dependent on race, was also determined to increase if an individual lived in an urban environment. Higher paying occupations also increased the chance of accessing healthcare. Importantly, it was found that consulting a medical man did increase the duration of illness, a proxy for quality of healthcare, but that these gains disappear when accounting for race and different disease categories. To this end, demographic and socioeconomic differences have in fact played an important role in healthcare outcomes in early 20th century Paarl and continue to do so in present day, democratic South Africa.

8. References

- Alsan, M. & Goldin, C. 2019. Watersheds in Child Mortality: The Role of Effective Water and Sewerage Infrastructure, 1880 to 1920. *Journal of Political Economy*, 127(2): 586-638.
- Alsan, M. & Wanamaker, M. 2018. Tuskegee and the Health of Black Men. *Quarterly Journal of Economics*, 133(1): 407-455.
- Anderson, D. M., Rees, D. I. & Wang, T. 2020. The phenomenon of summer diarrhea and its waning, 1910-1930. *Explorations in Economic History*, 78(1): 101341.
- Andersson, N. 1990. Tuberculosis and social stratification in South Africa. *International Journal of Health Services*, 20(1): 141-165.
- Arora, S. 2001. Health, Human Productivity, and Long- Term Economic Growth. *Journal of Economic History*, 61(3): 699-749.
- Atkinson, P., Francis, B., Gregory, I. & Porter, C. 2017. Patterns of infant mortality in rural England and Wales, 1850–1910. *Economic History Review*, 70(4): 1268-1290.
- Baird, S., Friedman, J. & Schady, N. 2011. Aggregate income shocks and infant mortality in the developing world. *Review of Economics and Statistics*, 93(3): 847-856.
- Beach, B., Ferrie, J., Saavedra, M. & Troesken, W. 2016. Typhoid Fever, Water Quality, and Human Capital Formation. *Journal of Economic History*, 76(1): 41-75.
- Bengtsson, T. & Van Poppel, F. 2011. Socioeconomic inequalities in death from past to present: An introduction. *Explorations in Economic History*, 48(1): 348-356.
- Birchenall, J. A. 2007. Escaping high mortality. *Journal of Economic Growth*, 12(1): 351-387.
- Blumenshine, P., Reingold, A., Egerter, S., Mockenhaupt, R., Braveman, P. & Marks, J. 2008. Pandemic influenza planning in the United States from a health disparities perspective. *Emerging Infectious Diseases*, 14(5): 709-715.
- Chapman, J. 2022. Interest Rates, Sanitation Infrastructure, and Mortality Decline in Nineteenth-Century England and Wales. *Journal of Economic History*, 82(1): 175-210.
- Collins, T. F. B. 1982. The History of Southern Africa's First Tuberculosis Epidemic. *History of Medicine*, 62(1): 780-788.

- Coovadia, H., Jewkes, R., Barron, P., Sanders, D. & McIntyre, D. 2009. The health and health system of South Africa: historical roots of current public health challenges. *Lancet*, 374(9692): 817-834.
- Cutler, D., Deaton, A. & Lleras-Muney, A. 2006. The Determinants of Mortality. *Journal of Economic Perspectives*, 20(3): 97-120.
- Davenport, T. R. H. 1971. *The beginnings of urban segregation in South Africa: The Natives (Urban Areas) Act of 1923 and its background*, Grahamstown: Institute of Social and Economic Research Rhodes University.
- Department of Water and Sanitation. 2004. *State of Rivers: Berg River*, Pretoria: Republic of South Africa.
- Drakenstein Heritage Survey Group. 2012. *Drakenstein Heritage Survey Volume 1: Heritage Survey Report*, Paarl: Drakenstein Municipality.
- Edvinsson, S. & Lindkvist, M. 2011. Wealth and health in 19th Century Sweden. A study of social differences in adult mortality in the Sundsvall region. *Explorations in Economic History*, 48(1): 376-388.
- Eli, S. 2015. Income Effects on Health: Evidence from Union Army Pensions. *Journal of Economic History*, 75(2): 448-478.
- Elo, I. T. & Preston, S. H. 1992. Effects of Early-Life Conditions on Adult Mortality: A Review. *Population Index*, 58(2): 186-212.
- Feinberg, H. M. 1993. The 1913 Natives Land Act in South Africa: Politics, Race, and Segregation in the Early 20th Century. *The International Journal of African Historical Studies*, 26(1): 65-109.
- Fourie, J. & Jayes, J. 2021. Health inequality and the 1918 influenza in South Africa. *World Development*, 141(1) 1-18.
- Gallardo-Albarrán, D. 2020. Sanitary infrastructures and the decline of mortality in Germany, 1877–1913. *Economic History Review*, 73(3): 730-757.
- Geruso, M. & Spears, D. 2018. Neighborhood Sanitation and Infant Mortality. *American Economic Journal: Applied Economics*, 10(2): 125-162.
- Green, T. L. & Hamilton, T. G. 2013. Beyond black and white: Color and mortality in post-reconstruction era North Carolina. *Explorations in Economic History*, 50(1): 148-159.

- Gyasi, R. M., Phillips, D. R. & David, R. 2019. Explaining the gender gap in health services use among Ghanaian community-dwelling older cohorts. *Women & Health*, 59(10): 1089-1104.
- Hanlon, W. W., Hansen, C. W. & Kantor, J. 2021. Temperature, Disease, and Death in London: Analyzing Weekly Data for the Century from 1866 to 1965. *Journal of Economic History*, 81(1): 40-80.
- Hoehn-Valasco, L. 2018. Explaining declines in US rural mortality, 1910–1933: The role of county health departments. *Explorations in Economic History*, 70(1): 42-72.
- Horrell, S. & Oxley, D. 2012. Bringing home the bacon? Regional nutrition, stature, and gender in the industrial revolution. *Economic History Review*, 65(4): 1354-1379.
- Jaadla, H., Potter, E., Keibek, S. & Davenport, R. 2020. Infant and child mortality by socio-economic status in early nineteenth-century England. *Economic History Review*, 73(4): 991-1022.
- Marks, S. & Andersson, N. 1987. Issues in the Political Economy of Health in Southern Africa. *Journal of Southern African Studies*, 13(2): 177-186.
- Maylam, P. 1990. The Rise and Decline of Urban Apartheid in South Africa. *African Affairs*, 89(354): 57-84.
- Meeker, E. 1976. Mortality Trends of Southern Blacks, 1850-1910: Some Preliminary Findings. *Explorations in Economic History*, 13(1) 13-42.
- Millward, R. & Bell, F. 2001. Infant Mortality in Victorian Britain: The Mother as Medium. *Economic History Review*, 54(4): 699-733.
- Mpeta, B., Fourie, J. & Inwood, K. 2018. Black living standards in South Africa before democracy: New evidence from height. *South African Journal of Science*, 114(1/2): Art. #2017-0052, 8 pages.
- Mäki, H. 2010. Comparing developments in water supply, sanitation and environmental health in four South African cities, 1840–1920. *Historia*, 55(1): 90-109.
- O'Donnell, O., Van Doorslaer, E. & Van Ourti, T. 2013. *Health and Inequality*, Amsterdam: Tinbergen Institute.
- Parnell, S. 1993. Creating Racial Privilege: The Origins of South African Public Health and Town Planning Legislation. *Journal of Southern African Studies*, 19(3): 471-488.

- Peltola, J. & Saaritsa, S. 2019. Later, smaller, better? Water infrastructure and infant mortality in Finnish cities and towns, 1870–1938. *The History of the Family*, 24(2): 277-306.
- Pinkerton, J. 1814. *A General Collection of the Best and Most Interesting Voyages and Travels in All Parts of the World*. London: Longman, Hurst, Rees, and Orme.
- Pretorius, J. H. 2022. Historic drivers of mortality in Paarl: Unlocking death certificate data from 1900 – 1931. Unpublished. Cape Town: Stellenbosch University.
- Ross, C. E., Masters, R. K. & Hummer, R. A. 2012. Education and the Gender Gaps in Health and Mortality. *Demography*, 49(4): 1157-1183.
- Saad, A. I. 1989. Schooling and Occupational Choice in Nineteenth-Century America. *The Journal of Economic History*, 49(2): 454-457.
- Simkins, C. & Van Heyningen, E. 1989. Fertility, Mortality, and Migration in the Cape Colony, 1891-1904. *The International Journal of African Historical Studies*, 22(1): 79-111.
- Tuberculosis Commission. 1914. *Report of the Tuberculosis Commission*, Cape Town: Cape Times Limited. Government Printer.
- Van Heyningen, E. B. 1989. Agents of the Empire: The Medical Profession in the Cape Colony, 1880-1910. *Medical History*, 33(1): 450-471.
- Van Leeuwen, M., Maas, I. & Miles, A. 2002. *HISCO: Historical international classification of occupations*. Leuven: Leuven University Press.
- Whitehead, M., Dahlgren, G. & Evans, T. 2001. Equity and health sector reforms: can low-income countries escape the medical poverty trap?. *The Lancet*, 358(1): 833-836.
- Woods, R., Williams, N. & Galley, C. 1993. Infant mortality in England – 1550-1950: Problems in the identification of long-term trends and geographical and social variations. In: C. A. Corsini & P. P. Viazzo, eds. *The decline of infant mortality in Europe 1800-1950: Four national case studies*. Florence: UNICEF, pp. 35-50.
- Zangel, V. A. 2017. Pulmonary Tuberculosis in Cape Town and the Karoo, 1870-1920: Policy and Attitudes. Unpublished Doctoral Dissertation. Pretoria: UNISA.

Appendix A

Table A1: Results from model 1, Equation 1.

Dependent Variable: Doctor Signature			
	Specification 1	Specification 2	Specification 3
(Intercept)	2.187 *** (0.052)	2.037 *** (0.062)	2.227 *** (0.108)
Coloured	-0.953 *** (0.052)	-0.947 *** (0.056)	-0.830 *** (0.064)
Black	-1.325 *** (0.099)	-1.238 *** (0.100)	-1.104 *** (0.111)
Male	-0.171 *** (0.031)	-0.163 *** (0.032)	-0.106 ** (0.037)
Street_Name		0.744 *** (0.039)	0.850 *** (0.043)
Informant_Signature		-0.099 ** (0.034)	-0.117 ** (0.037)
Farming			-0.309 ** (0.104)
Production			-0.003 (0.107)
Sales_and_Service			-0.001 (0.103)
Children			-0.379 *** (0.098)
Pensioners			0.280 (0.205)
Pandemic			-2.106 *** (0.075)
N	24189	24189	22399
AIC	25029.178	24639.680	22020.906
BIC	25061.552	24688.242	22117.107
Pseudo R2	0.030	0.054	0.110

*** p < 0.001; ** p < 0.01; * p < 0.05.

Table A2: Results from model 2, Equation 2.

Dependent variable: Duration of Illness * Race				
	All Races	White	Coloured	Black
(Intercept)	0.116 (0.160)	0.047 (0.076)	0.058 (0.140)	0.012 (0.017)
Doctor_Signature	0.370 * (0.171)	-0.024 (0.082)	0.402 ** (0.150)	-0.008 (0.018)
Resp	0.006 (0.160)	-0.054 (0.076)	0.077 (0.140)	-0.017 (0.017)
TB	-0.119 (0.163)	-0.053 (0.078)	-0.051 (0.143)	-0.015 (0.017)
Gastro	0.164 (0.163)	-0.041 (0.078)	0.218 (0.143)	-0.013 (0.017)
Cardio	-0.185 (0.171)	-0.090 (0.082)	-0.082 (0.150)	-0.013 (0.018)
Birth	0.234 (0.169)	-0.038 (0.081)	0.282 (0.148)	-0.009 (0.018)
Convulsions	0.245 (0.161)	-0.023 (0.077)	0.279 * (0.141)	-0.012 (0.017)
Cancer	-0.196 (0.208)	0.018 (0.099)	-0.181 (0.182)	-0.033 (0.022)
`0-9_Years`	-0.520 *** (0.024)	-0.037 ** (0.012)	-0.481 *** (0.021)	-0.003 (0.003)
`10-14_Years`	-0.245 *** (0.051)	-0.022 (0.024)	-0.223 *** (0.045)	-0.001 (0.005)
`15-19_Years`	-0.119 ** (0.043)	-0.033 (0.020)	-0.089 * (0.037)	0.003 (0.005)
`35-49_Years`	-0.010 (0.032)	0.012 (0.015)	-0.017 (0.028)	-0.004 (0.003)
`50+_Years`	-0.097 *** (0.029)	0.050 *** (0.014)	-0.148 *** (0.025)	0.001 (0.003)
Street_Name	0.039 * (0.017)	0.007 (0.008)	0.032 * (0.015)	0.001 (0.002)
Informant_Signature	-0.024 (0.014)	0.016 * (0.007)	-0.035 ** (0.013)	-0.005 ** (0.002)
Doctor_Signature:Resp	-0.078 (0.173)	0.046 (0.083)	-0.140 (0.152)	0.016 (0.018)
Doctor_Signature:TB	-0.144 (0.177)	0.068 (0.084)	-0.222 (0.155)	0.010 (0.019)
Doctor_Signature:Gastro	-0.228 (0.176)	0.051 (0.084)	-0.291 (0.155)	0.012 (0.019)
Doctor_Signature:Cardio	-0.107 (0.184)	0.066 (0.088)	-0.182 (0.162)	0.010 (0.020)
Doctor_Signature:Birth	-0.192 (0.188)	0.052 (0.090)	-0.256 (0.165)	0.012 (0.020)
Doctor_Signature:Convulsions	-0.119 (0.177)	0.032 (0.085)	-0.162 (0.155)	0.011 (0.019)
Doctor_Signature:Cancer	-0.183 (0.221)	-0.015 (0.105)	-0.199 (0.194)	0.031 (0.023)
N	17700	17700	17700	17700
R2	0.053	0.007	0.055	0.002

*** p < 0.001; ** p < 0.01; * p < 0.05.

Appendix B

Code Used for Study

JH Pretorius

2022-12-19

Purpose

The purpose of this code is to clean, visualise, and model data which was transcribed from death notices from Paarl in South Africa between 1895 and 1930.

Setup

Load in tools and packages used

```
rm(list = ls()) # Clean environment:

library(pacman)
p_load(ggplot2, dplyr, hrbrthemes, viridis, writexl, tidyr, ggribes,
       showtext, gfonts, ggrepel, stringr, tidytext, tm, htmlwidgets,
       webshot, wesanderson, dotwhisker, ivreg, jtools, GGally,
       broom.helpers, RColorBrewer, ivprobit, flextable, sf, scales, lubridate,
       tidyverse, readxl, RCurl, stringdist, naniar, glue)
```

Load Data

Cleaning starts

Clean the Age variable

```
#Special thanks to Jonathan Jayes for sharing his cleaning scripts for this variable (Age). Adjusted to suit my own naming conventions and data

data <- data %>%
  mutate(Age = trimws(Age),
         # before we get rid of punctuation
         Age = str_replace(string = Age, pattern = "1/2", replacement = ".5")
  ,
         Age = gsub('[:punct:] +', ' ', Age))

data <- data %>%
  mutate(Age = tolower(Age))

# Formatting errors in the data. Need to scour data set for words that appear in the column

words_age <- data %>%
  select(Age) %>%
```

```

unnest_tokens(word, Age) %>%
mutate(word = gsub('[:digit:]]+', '', word)) %>%
count(word, sort = T)

# Remove and replace common words
data <- data %>%
  mutate(Age = str_replace_all(string = Age, pattern = "and", replacement = ""))
) %>%
  mutate(Age = str_replace_all(string = Age, pattern = "en", replacement = ""))
%>%
  mutate(Age = str_replace(string = Age, pattern = "about", replacement = ""))
%>%
  mutate(Age = str_replace(string = Age, pattern = "abt", replacement = "")) %>
%
  mutate(Age = str_replace(string = Age, pattern = "omtrent", replacement = ""))
) %>%
  mutate(Age = str_replace(string = Age, pattern = "omtrant", replacement = ""))
) %>%
  mutate(Age = str_replace(string = Age, pattern = "approx", replacement = ""))
%>%
  mutate(Age = str_replace(string = Age, pattern = "an", replacement = "")) %>%
  mutate(Age = str_replace(string = Age, pattern = "jaar", replacement = "years
")) %>%
  mutate(Age = str_replace(string = Age, pattern = "jaaren", replacement = "yea
rs")) %>%
  mutate(Age = str_replace(string = Age, pattern = "jare", replacement = "years
")) %>%
  mutate(Age = str_replace(string = Age, pattern = "jaren", replacement = "year
s")) %>%
  mutate(Age = str_replace(string = Age, pattern = "jahren", replacement = "yea
rs")) %>%
  mutate(Age = str_replace(string = Age, pattern = "yr", replacement = "years")
) %>%
  mutate(Age = str_replace(string = Age, pattern = "yeara", replacement = "year
s")) %>%
  mutate(Age = str_replace(string = Age, pattern = "yeas", replacement = "years
")) %>%
  mutate(Age = str_replace(string = Age, pattern = "yeaers", replacement = "yea
rs")) %>%
  mutate(Age = str_replace(string = Age, pattern = "yeares", replacement = "yea
rs")) %>%
  mutate(Age = str_replace(string = Age, pattern = "yers", replacement = "years
")) %>%
  mutate(Age = str_replace(string = Age, pattern = "yrs", replacement = "years"
)) %>%
  mutate(Age = str_replace(string = Age, pattern = "yearsens", replacement = "ye
ars")) %>%
  mutate(Age = str_replace(string = Age, pattern = "maand", replacement = "mont
hs")) %>%
  mutate(Age = str_replace(string = Age, pattern = "maanden", replacement = "mo
nths")) %>%
  mutate(Age = str_replace(string = Age, pattern = "maande", replacement = "mon
ths")) %>%
  mutate(Age = str_replace(string = Age, pattern = "montha", replacement = "mon
ths")) %>%
  mutate(Age = str_replace(string = Age, pattern = "dagen", replacement = "days

```

```

")) %>%
  mutate(Age = str_replace(string = Age, pattern = "blank", replacement = ""))
%>%
  mutate(Age = str_replace(string = Age, pattern = "unknown", replacement = ""))
) %>%
  mutate(Age = str_replace(string = Age, pattern = "half", replacement = ".5"))
%>%
  mutate(Age = str_replace(string = Age, pattern = "old", replacement = "")) %>
%
  mutate(Age = str_replace(string = Age, pattern = "few", replacement = "5")) %
>%
  mutate(Age = str_replace(string = Age, pattern = "onbekend", replacement = ""
)) %>%
  mutate(Age = str_replace(string = Age, pattern = "tears", replacement = "year
s")) %>%
  mutate(Age = str_replace(string = Age, pattern = "eyars", replacement = "year
s")) %>%
  mutate(Age = str_replace(string = Age, pattern = "dae", replacement = "days"))
) %>%
  mutate(Age = str_replace(string = Age, pattern = "enige", replacement = "one"
)) %>%
  mutate(Age = str_replace(string = Age, pattern = "almost", replacement = ""))
%>%
  mutate(Age = str_replace(string = Age, pattern = "estimated", replacement = "
")) %>%
  mutate(Age = str_replace(string = Age, pattern = "omtren", replacement = ""))
%>%
  mutate(Age = str_replace(string = Age, pattern = "probably", replacement = ""
))

# Up to here most of the column should be clean. Export data to excel and
# manually clean

data <- data %>%
  mutate(Age = str_replace(string = Age, pattern = "one", replacement = "1")) %
>%
  mutate(Age = str_replace(string = Age, pattern = "two", replacement = "2")) %
>%
  mutate(Age = str_replace(string = Age, pattern = "three", replacement = "3"))
%>%
  mutate(Age = str_replace(string = Age, pattern = "four", replacement = "4"))
%>%
  mutate(Age = str_replace(string = Age, pattern = "five", replacement = "5"))
%>%
  mutate(Age = str_replace(string = Age, pattern = "six", replacement = "6")) %
>%
  mutate(Age = str_replace(string = Age, pattern = "seven", replacement = "7"))
%>%
  mutate(Age = str_replace(string = Age, pattern = "eight", replacement = "8"))
%>%
  mutate(Age = str_replace(string = Age, pattern = "nine", replacement = "9"))
%>%
  mutate(Age = str_replace(string = Age, pattern = "ten", replacement = "10"))
  mutate(Age = str_replace(string = Age, pattern = "twenty", replacement = "20"
)) %>%

```

```

mutate(Age = str_replace(string = Age, pattern = "thirty", replacement = "30"
)) %>%
mutate(Age = str_replace(string = Age, pattern = "forty", replacement = "40")
) %>%
mutate(Age = str_replace(string = Age, pattern = "fifty", replacement = "50")
) %>%
mutate(Age = str_replace(string = Age, pattern = "sixty", replacement = "60")
) %>%
mutate(Age = str_replace(string = Age, pattern = "seventy", replacement = "70
")) %>%
mutate(Age = str_replace(string = Age, pattern = "eighty", replacement = "80"
)) %>%
mutate(Age = str_replace(string = Age, pattern = "ninety", replacement = "90"
))

# Test data - export data to excel and scan through

write_xlsx(data, "{DESTINATION+NAME}")

# Import data with script above.

# Create dummy variables and labels

for(i in 1:length(data$Age_Years)){
  if (data$Age_Years[i] < 5){
    data$Cat[i] = "0-4 Years"
  } else if (data$Age_Years[i] < 10){
    data$Cat[i] = "5-9 Years"
  } else if (data$Age_Years[i] < 15) {
    data$Cat[i] = "10-14 Years"
  } else if (age_calc$Age_Years[i] < 20) {
    data$Cat[i] = "15-19 Years"
  } else if (age_calc$Age_Years[i] < 25) {
    data$Cat[i] = "20-24 Years"
  } else if (age_calc$Age_Years[i] < 35) {
    data$Cat[i] = "25-34 Years"
  } else if (age_calc$Age_Years[i] < 50) {
    data$Cat[i] = "35-49 Years"
  } else if (age_calc$Age_Years[i] < 70) {
    data$Cat[i] = "50-69 Years"
  } else {
    data$Cat[i] = "70+ Years"
  }
}

#Create dummies for categories

age_calc$`0-4_Years` = ifelse(age_calc$Cat == "0-4 Years", 1, 0)
age_calc$`5-9_Years` = ifelse(age_calc$Cat == "5-9 Years", 1, 0)
age_calc$`10-14_Years` = ifelse(age_calc$Cat == "10-14 Years", 1, 0)
age_calc$`15-19_Years` = ifelse(age_calc$Cat == "15-19 Years", 1, 0)
age_calc$`20-24_Years` = ifelse(age_calc$Cat == "20-24 Years", 1, 0)

```



```
age_calc$`25-34_Years` = ifelse(age_calc$Cat == "25-34 Years", 1, 0)
age_calc$`35-49_Years` = ifelse(age_calc$Cat == "35-49 Years", 1, 0)
age_calc$`50-69_Years` = ifelse(age_calc$Cat == "50-69 Years", 1, 0)
age_calc$`70+_Years` = ifelse(age_calc$Cat == "70+ Years", 1, 0)

#At first I believed the categories to be suitable.
#But I opted to reduce the number of age categories to the five contained in the document
#I reduced the categories manually through excel by making use of the newly instantiated Cat column
```

Create dummy variable for sex

```
data$Male = ifelse(data$Sex == "Male", 1, 0)
```

Create dummy for informant signature

```
age_calc$informant_signature = ifelse(data$`Signed, Yes Or No?` == "Yes", 1, 0)
```

Create dummy for doctor signature

```
data$doctor_signature = ifelse(data$Doctor == "N/A", 0, 1)
```

Determine whether an individual has a street name in 'Place of Residence'

```
for(i in 1:length(data$`Usual Place of Residence`)){
  data$CoD[i] <- tolower(data$`Usual Place of Residence`[i])
}

for(i in 1:length(data$`Usual Place of Residence`)){
  data$street_name[i] <- grepl("street|str|straat|st|road|laan|rd|lane",
    data$`Usual Place of Residence`[i])
}
```

Causes of Death

#Note: CoD = Cause of Death column

#Resp dummies

```
for(i in 1:length(data$CoD)){
  data$CoD[i] <- tolower(data$CoD[i])
}

for(i in 1:length(data$CoD)){
  data$resp[i] <- grepl("pneumonia|lungs|lung|bronchitis|consumption|pulmonary|
broncho|cough|influenza|tuberculosis|pertussis|emphysema|emphyzema|long|phtithi
s|flu|croup|griep|diphtheria|tubercular",
```

```

        data$CoD[i])
}

#Gastro

for(i in 1:length(data$CoD)){
  data$gastro[i] <- grepl("enteric|diarrhoea|gastritis|bowels|maagkoors|cholera
|gastro|vomiting|gastroenteritis|typhoid|intestinal|maag|peritonitis|enterocoli
tis|appendicitis|stomach|enteritis|entiritis",
        data$CoD[i])
}

#Cardio

for(i in 1:length(data$CoD)){
  data$cardio[i] <- grepl("heart|cardiovascular|cardiac|carditis|myocarditis|dr
opsy|cardia",
        data$CoD[i])
}

#Birth

for(i in 1:length(data$CoD)){
  data$birth[i] <- grepl("premature|birth|geboorte|gebore|prematurity",
        data$CoD[i])
}

#Cancer

for(i in 1:length(data$CoD)){
  data$cancer[i] <- grepl("cancer|carcinoma|kanker|cancerous|sarcoma|growth|tum
our",
        data$CoD[i])
}

#Convulsions

for(i in 1:length(data$CoD)){
  data$convulsions[i] <- grepl("convulsions",
        data$CoD[i])
}

#STI

for(i in 1:length(data$CoD)){
  data$sti[i] <- grepl("syphilis|hepatitis|gonorrhoea",
        data$CoD[i])
}

# This is as clean as it gets through automation. The rest of the data was clea
ned manually in Excel

```

Occupation cleaning scripts

```
#Farming
```

```
#Occupation
```

```

for(i in 1:length(data$Occupation)){
  data$Farming[i] <- grepl("farm-worker|peasant|agriculturalist|shepherd|farm w
orker|herd|gardener|groom|wood cutter|stable boy|cowherd|cattleman|plaa
s",
    data$Occupation[i])
}

#Occupation Father
for(i in 1:length(data$Occupation_Father)){
  data$Farming_Father[i] <- grepl("farm-worker|peasant|agriculturalist|shepherd
|farm worker|herd|gardener|groom| wood cutter|stable boy|cowherd|cattleman|plaa
s",
    data$Occupation_Father[i])
}

#Occupation Mother
for(i in 1:length(data$Occupation_Mother)){
  data$Farming_Mother[i] <- grepl("farm-worker|peasant|agriculturalist|shepherd
|farm worker|herd|gardener|groom| wood cutter|stable boy|cowherd|cattleman|plaa
s",
    data$Occupation_Mother[i])
}

#Production

#Occupation
for(i in 1:length(data$Occupation)){
  data$Production[i] <- grepl("labourer|manufacturer|mason|painter|factory|carp
enter|blacksmith|shoemaker|driver|wagon|arbeider|maker|boot|builder|building|pa
inter",
    data$Occupation[i])
}

#Occupation Father
for(i in 1:length(data$Occupation_Father)){
  data$Production_Father[i] <- grepl("labourer|manufacturer|mason|painter|facto
ry|carpenter|blacksmith|shoemaker|driver|wagon|arbeider|maker|boot|builder|buil
ding|painter",
    data$Occupation_Father[i])
}

#Occupation Mother
for(i in 1:length(data$Occupation_Mother)){
  data$Production_Mother[i] <- grepl("labourer|manufacturer|mason|painter|facto
ry|carpenter|blacksmith|shoemaker|driver|wagon|arbeider|maker|boot|builder|buil
ding|painter",
    data$Occupation_Mother[i])
}

#Professionals & Managers

#Occupation
for(i in 1:length(data$Occupation)){
  data$Prof_and_Man[i] <- grepl("farmer|scholar|teacher|clerk|manager|ganger|ov
erseer|engineer|nurse|minister|doctor|attorney|law|legal|bookkeeper|accountant|
biddle",
    data$Occupation[i])
}

```

```

#Occupation Father
for(i in 1:length(data$Occupation_Father)){
  data$Prof_and_Man_Father[i] <- grepl("farmer|scholar|teacher|clerk|manager|ga
nger|overseer|engineer|nurse|minister|doctor|attorney|law|legal|bookkeeper|acco
untant|biddle",
                                     data$Occupation_Father[i])
}

#Occupation Mother
for(i in 1:length(data$Occupation_Mother)){
  data$Prof_and_Man_Mother[i] <- grepl("farmer|scholar|teacher|clerk|manager|ga
nger|overseer|engineer|nurse|minister|doctor|attorney|law|legal|bookkeeper|acco
untant|biddle",
                                     data$Occupation_Mother[i])
}

#Sales & Service

#Occupation
for(i in 1:length(data$Occupation)){
  data$Sales_and_Service[i] <- grepl("housewife|domestic servant|servant|washin
g work|cook|baker|housekeeper|washer woman|general dealer|maid|house|domestic|w
asher",
                                     data$Occupation[i])
}

#Occupation Father
for(i in 1:length(data$Occupation_Father)){
  data$Sales_and_Service_Father[i] <- grepl("housewife|domestic servant|servant
|washing work|cook|baker|housekeeper|washer woman|general dealer|maid|house|dom
estic|washer",
                                     data$Occupation_Father[i])
}

#Occupation Mother
for(i in 1:length(data$Occupation_Mother)){
  data$Sales_and_Service_Mother[i] <- grepl("housewife|domestic servant|servant
|washing work|cook|baker|housekeeper|washer woman|general dealer|maid|house|dom
estic|washer",
                                     data$Occupation_Mother[i])
}

#Pensioners

#Occupation
for(i in 1:length(data$Occupation)){
  data$Sales_and_Service[i] <- grepl("pensioner|pension|retired",
                                     data$Occupation[i])
}

#Occupation Father
for(i in 1:length(data$Occupation_Father)){
  data$Sales_and_Service_Father[i] <- grepl("pensioner|pension|retired",
                                     data$Occupation_Father[i])
}

```

```
#Occupation Mother
for(i in 1:length(data$Occupation_Mother)){
  data$Sales_and_Service_Mother[i] <- grepl("pensioner|pension|retired",
    data$Occupation_Mother[i])
}
```

Cleaning duration of last illness

#Special thanks to Jonathan Jayes for sharing his cleaning scripts for this variable (duration). Adjusted to suit my own naming conventions and data

```
data <- data %>%
  mutate(`Duration of last illness` = trimws(`Duration of last illness`),
    `Duration of last illness` = str_replace(string = `Duration of last illness`, pattern = "1/2", replacement = ".5"),
    `Duration of last illness` = gsub('[:punct: ]+', ' ', `Duration of last illness`))

data <- data %>%
  mutate(`Duration of last illness` = tolower(`Duration of last illness`))

data_word <- data %>%
  select(`Duration of last illness`) %>%
  unnest_tokens(word, `Duration of last illness`) %>%
  mutate(word = gsub('[:digit: ]+', ' ', word)) %>%
  count(word, sort = T)

# removing and replacing mistakes in transcription
data <- data %>%
  mutate(`Duration of last illness` = str_replace(string = `Duration of last illness`, pattern = "one", replacement = "1")) %>%
  mutate(`Duration of last illness` = str_replace(string = `Duration of last illness`, pattern = "sudden", replacement = "5 hours")) %>%
  mutate(`Duration of last illness` = str_replace(string = `Duration of last illness`, pattern = "blank", replacement = "")) %>%
  mutate(`Duration of last illness` = str_replace(string = `Duration of last illness`, pattern = "none", replacement = "")) %>%
  mutate(`Duration of last illness` = str_replace(string = `Duration of last illness`, pattern = "several", replacement = "5")) %>%
  mutate(`Duration of last illness` = str_replace(string = `Duration of last illness`, pattern = "dagen", replacement = "days")) %>%
  mutate(`Duration of last illness` = str_replace(string = `Duration of last illness`, pattern = "some", replacement = "5")) %>%
  mutate(`Duration of last illness` = str_replace(string = `Duration of last illness`, pattern = "abt", replacement = "")) %>%
  mutate(`Duration of last illness` = str_replace(string = `Duration of last illness`, pattern = "unknown", replacement = "")) %>%
  mutate(`Duration of last illness` = str_replace(string = `Duration of last illness`, pattern = "indefinite", replacement = "")) %>%
  mutate(`Duration of last illness` = str_replace(string = `Duration of last illness`, pattern = "few", replacement = "5")) %>%
  mutate(`Duration of last illness` = str_replace(string = `Duration of last illness`, pattern = "suddenly", replacement = "5 hours")) %>%
  mutate(`Duration of last illness` = str_replace(string = `Duration of last illness`, pattern = "maanden", replacement = "months")) %>%
  mutate(`Duration of last illness` = str_replace(string = `Duration of last il
```

```

lness`, pattern = "omtrekt", replacement = "") %>%
  mutate(`Duration of last illness` = str_replace(string = `Duration of last il
lness`, pattern = "approx", replacement = "")) %>%
  mutate(`Duration of last illness` = str_replace(string = `Duration of last il
lness`, pattern = "enige", replacement = "1")) %>%
  mutate(`Duration of last illness` = str_replace(string = `Duration of last il
lness`, pattern = "weken", replacement = "weeks")) %>%
  mutate(`Duration of last illness` = str_replace(string = `Duration of last il
lness`, pattern = "weken", replacement = "weeks")) %>%
  mutate(`Duration of last illness` = str_replace(string = `Duration of last il
lness`, pattern = "dae", replacement = "days")) %>%
  mutate(`Duration of last illness` = str_replace(string = `Duration of last il
lness`, pattern = "maande", replacement = "months")) %>%
  mutate(`Duration of last illness` = str_replace(string = `Duration of last il
lness`, pattern = "fortnight", replacement = "2 weeks")) %>%
  mutate(`Duration of last illness` = str_replace(string = `Duration of last il
lness`, pattern = "maand", replacement = "months")) %>%
  mutate(`Duration of last illness` = str_replace(string = `Duration of last il
lness`, pattern = "instantaneous", replacement = "5 hours")) %>%
  mutate(`Duration of last illness` = str_replace(string = `Duration of last il
lness`, pattern = "dag", replacement = "day")) %>%
  mutate(`Duration of last illness` = str_replace(string = `Duration of last il
lness`, pattern = "nil ", replacement = "")) %>%
  mutate(`Duration of last illness` = str_replace(string = `Duration of last il
lness`, pattern = "half", replacement = ".5")) %>%
  mutate(`Duration of last illness` = str_replace(string = `Duration of last il
lness`, pattern = "immediately", replacement = "5 hours")) %>%
  mutate(`Duration of last illness` = str_replace(string = `Duration of last il
lness`, pattern = "three", replacement = "3")) %>%
  mutate(`Duration of last illness` = str_replace(string = `Duration of last il
lness`, pattern = "two", replacement = "2")) %>%
  mutate(`Duration of last illness` = str_replace(string = `Duration of last il
lness`, pattern = "jaar", replacement = "years")) %>%
  mutate(`Duration of last illness` = str_replace(string = `Duration of last il
lness`, pattern = "night", replacement = "day")) %>%
  mutate(`Duration of last illness` = str_replace(string = `Duration of last il
lness`, pattern = "uur", replacement = "year")) %>%
  mutate(`Duration of last illness` = str_replace(string = `Duration of last il
lness`, pattern = "circa", replacement = "")) %>%
  mutate(`Duration of last illness` = str_replace(string = `Duration of last il
lness`, pattern = "een", replacement = "1")) %>%
  mutate(`Duration of last illness` = str_replace(string = `Duration of last il
lness`, pattern = "four", replacement = "4")) %>%
  mutate(`Duration of last illness` = str_replace(string = `Duration of last il
lness`, pattern = "acht", replacement = "8")) %>%
  mutate(`Duration of last illness` = str_replace(string = `Duration of last il
lness`, pattern = "atb", replacement = "")) %>%
  mutate(`Duration of last illness` = str_replace(string = `Duration of last il
lness`, pattern = "couple", replacement = "5")) %>%
  mutate(`Duration of last illness` = str_replace(string = `Duration of last il
lness`, pattern = "dage", replacement = "days")) %>%
  mutate(`Duration of last illness` = str_replace(string = `Duration of last il
lness`, pattern = "jare", replacement = "years")) %>%
  mutate(`Duration of last illness` = str_replace(string = `Duration of last il
lness`, pattern = "jare", replacement = "years")) %>%
  mutate(`Duration of last illness` = str_replace(string = `Duration of last il

```

```

lness`, pattern = "jaren", replacement = "years")) %>%
  mutate(`Duration of last illness` = str_replace(string = `Duration of last il
lness`, pattern = "monts", replacement = "months")) %>%
  mutate(`Duration of last illness` = str_replace(string = `Duration of last il
lness`, pattern = "twenty", replacement = "20")) %>%
  mutate(`Duration of last illness` = str_replace(string = `Duration of last il
lness`, pattern = "uncertain", replacement = "")) %>%
  mutate(`Duration of last illness` = str_replace(string = `Duration of last il
lness`, pattern = "moths", replacement = "months")) %>%
  mutate(`Duration of last illness` = str_replace(string = `Duration of last il
lness`, pattern = "probably", replacement = "")) %>%
  mutate(`Duration of last illness` = str_replace(string = `Duration of last il
lness`, pattern = "about", replacement = ""))%>%
  mutate(`Duration of last illness` = str_replace(string = `Duration of last il
lness`, pattern = "montha", replacement = "months"))

# digits
data <- data %>%
  mutate(`Duration of last illness` = str_replace(string = `Duration of last il
lness`, pattern = "%", replacement = ".5")) %>%
  mutate(`Duration of last illness` = str_replace(string = `Duration of last il
lness`, pattern = "%", replacement = ".25"))

# we mutate age to be a uniform format, creating age in days and in years
data <- data %>%
  mutate(data_days = round(as.numeric(as.period(`Duration of last illness`),
"days"), 0),
         data_years = round(as.numeric(as.period(`Duration of last illness`),
"years"), 2))

data <- data %>%
  mutate(data_days = ifelse(str_detect(`Duration of last illness`, pattern = "b
irth"), age_days, data_days),
         data_years = ifelse(str_detect(`Duration of last illness`, pattern = "
birth"), age_years, data_years))

```

Cleaning has been completed

Stage 2: Plotting of data

Scripts for Age plots

```

#Line plots (figure 3)

data = data %>% count(Race, Cat, Year)

data <- subset(data, Race != "Black") #the sample size for black individuals to
o small to make meaningful plot

ggplot(data=data, aes(x=Year, y=n, group=Cat, color=Cat))+
  geom_line(size=.7)+
  #Theme specifications
  theme_ipsum() +

```



```

scale_color_manual(values = wes_palette(5, name = "Darjeeling1", type = "discrete"), name = "Age Cohort")+
  theme(legend.position = "bottom",
        panel.background = element_blank(),
        plot.background = element_rect(fill = "White", color = "White"),
        panel.grid.major = element_line(color="#808080", size = 0.1),
        panel.grid.minor = element_line(color="#808080", size = 0.1),
        axis.title.x=element_text(colour="black", size = 12,family = "Times New Roman", vjust=-2,hjust=0.5),
        axis.title.y=element_text(colour="black", size = 12,family = "Times New Roman",vjust = 3,hjust=0.5),
        axis.text.y=element_text(colour = "black", size = 10, family = "Times New Roman"),
        axis.text.x=element_text(colour="black", size = 10,family = "Times New Roman"),
        plot.margin = unit(c(0.5,0.5,0.5,0.5), "cm"),
        plot.title = element_text(colour="black", size = 14,family = "Times New Roman",hjust=0.5),
        plot.subtitle = element_text(colour="black", size = 9,family = "Times New Roman"),
        plot.caption = element_text(colour="black", size = 7,family = "Times New Roman"),
        legend.text = element_text(colour="black", size = 10,family = "Times New Roman"),
        legend.title = element_text(colour="black", size = 12,family = "Times New Roman", hjust=0, face = "bold"))+
  #add labels
  labs(
    y = "Number of Deaths",
    x = "Year",
    title = bquote("Number of Deaths From Coloured Population")
  )

#Proportion plots were done with the same script

# Density plots (from figure 4)

data$informant_signature <- as.character(data$informant_signature) #numeric to string

data$street_name <- as.character(data$street_name)

ggplot(data, aes(x = Age_Years, group = informant_signature, fill = informant_signature)) +
  geom_density(alpha=0.6, adjust = 2) +
  theme_ridges() +
  theme_ipsum() +
  #facet_wrap(~Race)+
  scale_fill_manual(values = wes_palette(3, name = "Darjeeling1", type = "discrete"), name = "", labels = c("Mark", "Signature"))+
  theme(legend.position = "bottom",
        panel.background = element_blank(),
        plot.background = element_rect(fill = "White", color = "White"),
        panel.grid.major = element_line(color="#808080", size = 0.1),
        panel.grid.minor = element_line(color="#808080", size = 0.1),

```



```

    #panel.border = element_blank(),
    axis.title.x=element_text(colour="black", size = 12,family = "Times New
Roman", vjust=-2,hjust=0.5),
    axis.title.y=element_text(colour="black", size = 12,family = "Times New
Roman",vjust = 3,hjust=0.5),
    axis.text.y=element_text(colour = "black", size = 10, family = "Times Ne
w Roman"),
    axis.text.x=element_text(colour="black", size = 10,family = "Times New
Roman"),
    plot.margin = unit(c(0.5,0.5,0.5,0.5), "cm"),
    plot.title = element_text(colour="black", size = 16,family = "Times New
Roman",hjust=0.5),
    plot.subtitle = element_text(colour="black", size = 9,family = "Times N
ew Roman"),
    plot.caption = element_text(colour="black", size = 7,family = "Times Ne
w Roman"),
    legend.text = element_text(colour="black", size = 10,family = "Times Ne
w Roman"),
    legend.title = element_text(colour="black", size = 12,family = "Times N
ew Roman", hjust=3, face = "bold"),
    strip.text = element_text(colour="black", size = 12,family = "Times New
Roman",hjust=0.5))+
  labs(
    y = "Density",
    x = "Age at Time of Death (Years)",
    title = bquote("Death Notice Signed or Marked")
  )

```

#Total number of death notices with doctor present

```

ggplot(data_deaths, aes(x=Year, y=total, group=Race, color=Race))+
  geom_line()+
  theme_ipsum() +
  scale_color_manual(values = wes_palette(3, name = "Darjeeling1", type = "disc
rete"), name = "Race Group")+
  theme(legend.position = "bottom",
    panel.background = element_blank(),
    plot.background = element_rect(fill = "#E7E6E2", color = "#E7E6E2"),
    panel.grid.major = element_line(color="#808080", size = 0.1),
    panel.grid.minor = element_line(color="#808080", size = 0.1),
    axis.title.x=element_text(colour="black", size = 10,family = "Times New
Roman", vjust=-2,hjust=0.5),
    axis.title.y=element_text(colour="black", size = 10,family = "Times New
Roman",vjust = 3,hjust=0.5),
    axis.text.y=element_text(colour = "black", size = 8, family = "Times New
Roman"),
    axis.text.x=element_text(colour="black", size = 8,family = "Times New R
oman"),
    plot.margin = unit(c(0.5,0.5,0.5,0.5), "cm"),
    plot.title = element_text(colour="black", size = 12,family = "Times New
Roman",hjust=0.5),
    plot.subtitle = element_text(colour="black", size = 9,family = "Times N
ew Roman"),
    plot.caption = element_text(colour="black", size = 7,family = "Times Ne
w Roman"),
    legend.text = element_text(colour="black", size = 9,family = "Times New

```

```

Roman")),
  legend.title = element_text(colour="black", size = 10,family = "Times N
ew Roman", hjust=3, face = "bold"))+
  labs(
    y = "Number of Death Notices with Doctor Present",
    x = "Year",
    title = bquote("Access to Health")
  )+
  geom_vline(
    aes(xintercept = 1918),
    color = "black",
    linetype = "longdash",
    size = .3
  )
)

```

#Access to Health + CoD

#Get the data ready

```

data_access = data %>% count(Race, doctor_signed, Year)

data <- subset(data, Race == "White")

ggplot(data, aes(x=Year, y=n, group=doctor_signed, color=doctor_signed))+
  geom_line()+
  theme_ipsum() +
  scale_color_manual(values = wes_palette(3, name = "FantasticFox1", type = "di
crete"), name = "Death Notice Signed by Doctor?", labels = c("No", "Yes"))+
  theme(legend.position = "bottom",
    panel.background = element_blank(),
    plot.background = element_rect(fill = "#E7E6E2", color = "#E7E6E2"),
    panel.grid.major = element_line(color="#808080", size = 0.1),
    panel.grid.minor = element_line(color="#808080", size = 0.1),
    #panel.border = element_blank(),
    axis.title.x=element_text(colour="black", size = 10,family = "Times New
Roman", vjust=-2,hjust=0.5),
    axis.title.y=element_text(colour="black", size = 10,family = "Times New
Roman",vjust = 3,hjust=0.5),
    axis.text.y=element_text(colour = "black", size = 8, family = "Times New
Roman"),
    axis.text.x=element_text(colour="black", size = 8,family = "Times New R
oman"),
    plot.margin = unit(c(0.5,0.5,0.5,0.5), "cm"),
    plot.title = element_text(colour="black", size = 12,family = "Times New
Roman",hjust=0.5),
    plot.subtitle = element_text(colour="black", size = 9,family = "Times N
ew Roman"),
    plot.caption = element_text(colour="black", size = 7,family = "Times Ne
w Roman"),
    legend.text = element_text(colour="black", size = 9,family = "Times New
Roman"),
    legend.title = element_text(colour="black", size = 10,family = "Times N
ew Roman", hjust=3, face = "bold"))+
  labs(
    y = "Number of Deaths",
    x = "Year",

```

```

    title = bquote("Access to Health: Whites that are Literate")
  )

```

Cause of Death

#Create new data set to calculate proportions

```
data_deaths = data %>% count(Year, cod, Race)
```

data <- subset(data, Race != "Black") #sample size for blacks to small to make meaningful plot

```
data <- subset(data, age == "0-9 Years")
```

```

ggplot(data, aes(x=Year, y=n, group=cod, fill=cod)) +
  geom_bar(position="fill", stat="identity", alpha=1)+
  theme_ipsum() +
  facet_wrap(~Race)+
  scale_fill_brewer(palette = "Spectral", name = "Cause of Death")+
  theme(legend.position = "none",
        panel.background = element_blank(),
        plot.background = element_rect(fill = "White", color = "White"),
        panel.grid.major = element_line(color="#808080", size = 0.1),
        panel.grid.minor = element_line(color="#808080", size = 0.1),
        axis.title.x=element_text(colour="black", size = 12,family = "Times New
Roman", vjust=-2,hjust=0.5),
        axis.title.y=element_text(colour="black", size = 12,family = "Times New
Roman",vjust = 3,hjust=0.5),
        axis.text.y=element_text(colour = "black", size = 10, family = "Times Ne
w Roman"),
        axis.text.x=element_text(colour="black", size = 10,family = "Times New
Roman"),
        plot.margin = unit(c(0.5,0.5,0.5,0.5), "cm"),
        plot.title = element_text(colour="black", size = 14,family = "Times New
Roman",hjust=0.5),
        plot.subtitle = element_text(colour="black", size = 9,family = "Times N
ew Roman"),
        plot.caption = element_text(colour="black", size = 7,family = "Times Ne
w Roman"),
        legend.text = element_text(colour="black", size = 10,family = "Times Ne
w Roman"),
        legend.title = element_text(colour="black", size = 10,family = "Times N
ew Roman", hjust=3, face = "bold"),
        strip.text = element_text(colour="black", size = 10,family = "Times New
Roman", vjust=1,hjust=0.5))+
  labs(
    y = "Proportion",
    x = "Year",
    title = bquote("Prevalent Causes of Death Over Time by Race")
  )

```

#Duration of illness plots

```
data %>%
```

```

  ggplot(aes(y=dur_ill_days, x=cod, fill=doctor_signature)) +
  geom_boxplot(width=0.3, color="black", alpha=0.7) +
  theme_ipsum() +

```

```

ylim(0,50) +
  #facet_wrap(~Year)+
  scale_fill_manual(values = wes_palette(4, name = "Darjeeling2", type = "discrete"), name="Death Notice Signed by Doctor?", labels=c("No Signature", "Doctor's Signature"))+
  theme(legend.position = "bottom",
        panel.background = element_blank(),
        plot.background = element_rect(fill = "White", color = "White"),
        panel.grid.major = element_line(color="#808080", size = 0.1),
        panel.grid.minor = element_line(color="#808080", size = 0.1),
        axis.title.x=element_text(colour="black", size = 12,family = "Times New Roman", vjust=0,hjust=0.5),
        axis.title.y=element_text(colour="black", size = 12,family = "Times New Roman",vjust = 3,hjust=0.5),
        axis.text.y=element_text(colour = "black", size = 10, family = "Times New Roman"),
        axis.text.x=element_text(colour="black", size = 10,family = "Times New Roman"),
        plot.margin = unit(c(0.5,0.5,0.5,0.5), "cm"),
        plot.title = element_text(colour="black", size = 12,family = "Times New Roman",hjust=0.5),
        plot.subtitle = element_text(colour="black", size = 10,family = "Times New Roman", hjust=0.5),
        plot.caption = element_text(colour="black", size = 10,family = "Times New Roman"),
        legend.text = element_text(colour="black", size = 15,family = "Times New Roman"),
        legend.title = element_blank(),
        strip.text = element_text(colour="black", size = 10,family = "Times New Roman", vjust=1,hjust=0.5))+
  labs(
    y = "Duration of Last Illness (Days)",
    x = "Cause of Death",
    title = "Death Notice Signed by Doctor"
  )

```

Occupation plots

```

#Prep data
occu = data %>% count(Race, Occupation, street_name)

data <- subset(data, Occupation != "Children")

ggplot(data, aes(fill=Occupation, y=n, x=Race)) +
  geom_bar(position="fill", stat="identity", alpha=0.8)+
  facet_wrap(~street_name)+
  theme_ipsum()+
  scale_fill_manual(values = wes_palette(5, name = "Darjeeling1", type = "discrete"), name="Occupational Group")+
  theme(legend.position = "bottom",
        panel.background = element_blank(),
        plot.background = element_rect(fill = "White", color = "White"),
        panel.grid.major = element_line(color="#808080", size = 0.1),
        panel.grid.minor = element_line(color="#808080", size = 0.1),
        axis.title.x=element_text(colour="black", size = 12,family = "Times New

```

```

Roman", vjust=0,hjust=0.5),
  axis.title.y=element_text(colour="black", size = 12,family = "Times New
Roman",vjust = 3,hjust=0.5),
  axis.text.y=element_text(colour = "black", size = 10, family = "Times Ne
w Roman"),
  axis.text.x=element_text(colour="black", size = 10,family = "Times New
Roman"),
  axis.ticks = element_blank(),
  plot.margin = unit(c(0.5,0.5,0.5,0.5), "cm"),
  plot.title = element_text(colour="black", size = 14,family = "Times New
Roman",hjust=0.5, face="bold"),
  plot.subtitle = element_text(colour="black", size = 10,family = "Times
New Roman", hjust=0.5),
  plot.caption = element_text(colour="black", size = 10,family = "Times N
ew Roman"),
  legend.text = element_text(colour="black", size = 10,family = "Times Ne
w Roman"),
  legend.title = element_text(colour="black", size = 12,family = "Times N
ew Roman", face = "bold"),
  strip.text = element_text(colour="black", size = 10,family = "Times New
Roman", vjust=1,hjust=0.5))+
  labs(
    y = "Proportion",
    x = "Race",
    title = "Street Name on Death Certificate"
  )

```

Stage 2, plotting, complete

Stage 3: model the data

#Spec 1 - Health Access

```

spec1_1 <- glm(Doctor_Signature ~ Coloured + Black + Male,
  family = binomial(link='logit'),
  data = data)

spec1_2 <- glm(Doctor_Signature ~ Coloured + Black + Male +
  Street_Name + Informant_Signature,
  family = binomial(link='logit'),
  data = data)

spec1_3 <- glm(Doctor_Signature ~ Coloured + Black + Male +
  Street_Name + + Informant_Signature +
  Farming + Production + Sales_and_Service +
  Pandemic,
  family = binomial(link='logit'),
  data = data)

```

#View model in R

```
export_summs(spec1_1,spec1_2,spec1_3, digits = 3)
```

#Export model to formatted table for Word

```
export_summs(spec1_1,spec1_2,spec1_3, digits = 3, to.file = "docx", file.name =
"{NAME}")
```

```
# Coefficient plot for first model
```

```
dwplot(list(spec1_1, spec1_2, spec1_3),
  vline = geom_vline(
    xintercept = 0,
    colour = "grey60",
    linetype = 2
  ), dodge_size = 0.6,
  ci_method="wald",
  errorbar_height = .2)+
  theme_ipsum()+
  scale_color_manual(values = wes_palette(3, name = "Darjeeling1", type = "discrete"), name="", labels=c("Specification 3","Specification 2","Specification 1"))+
  theme(legend.position = "right",
    panel.background = element_blank(),
    plot.background = element_rect(fill = "White", color = "White"),
    panel.grid.major = element_line(color="#808080", size = 0.1),
    panel.grid.minor = element_line(color="#808080", size = 0.1),
    axis.title.x=element_text(colour="black", size = 12,family = "Times New Roman", vjust=0,hjust=0.5),
    axis.title.y=element_text(colour="black", size = 12,family = "Times New Roman",vjust = 3,hjust=0.5),
    axis.text.y=element_text(colour = "black", size = 10, family = "Times New Roman"),
    axis.text.x=element_text(colour="black", size = 10,family = "Times New Roman"),
    axis.ticks = element_blank(),
    plot.margin = unit(c(0.5,0.5,0.5,0.5), "cm"),
    plot.title = element_text(colour="black", size = 14,family = "Times New Roman",hjust=0.5, face="bold"),
    plot.subtitle = element_text(colour="black", size = 10,family = "Times New Roman", hjust=0.5),
    plot.caption = element_text(colour="black", size = 10,family = "Times New Roman"),
    legend.text = element_text(colour="black", size = 10,family = "Times New Roman"),
    legend.title = element_text(colour="black", size = 12,family = "Times New Roman"),
    strip.text = element_text(colour="black", size = 10,family = "Times New Roman", vjust=1,hjust=0.5))+
  labs(
    y = "",
    x = "Coefficient Estimate",
    title = "Coefficient Plot of First Research Question",
    subtitle = "Dependent Variable: Doctor Signature"
  )
```

```
#Spec 2 - Duration of Illness
```

```
spec2_1 <- lm(Duration_of_Illness ~ Doctor_Signature*(Resp + TB + Gastro + Cardio + Birth + STI + Cancer)
  + `0-9_Years` + `10-14_Years` + `15-19_Years` + `35-49_Years`
  + `50+_Years` +
  Street_Name + Informant_Signature,
  data = data)
```

```

spec2_2 <- lm(Duration_of_Illness*White ~ Doctor_Signature*(Resp + TB + Gastro
+ Cardio + Birth + STI + Cancer)
              + `0-9_Years` + `10-14_Years` + `15-19_Years` + `35-49_Years`
+ `50+_Years` +
              Street_Name + Informant_Signature,
              data = data)

spec2_3 <- lm(Duration_of_Illness*Coloured ~ Doctor_Signature*(Resp + TB + Gastro
+ Cardio + Birth + STI + Cancer)
              + `0-9_Years` + `10-14_Years` + `15-19_Years` + `35-49_Years`
+ `50+_Years` +
              Street_Name + Informant_Signature,
              data = data)

spec2_4 <- lm(Duration_of_Illness*Black ~ Doctor_Signature*(Resp + TB + Gastro
+ Cardio + Birth + STI + Cancer)
              + `0-9_Years` + `10-14_Years` + `15-19_Years` + `35-49_Years` +
`50+_Years` +
              Street_Name + Informant_Signature,
              data = data)

export_summs(spec2_1,spec2_2, spec2_3, spec2_4,digits = 3)

#Second model dw plot

dwplot(list(spec2_1, spec2_2, spec2_3, spec2_4),
        vline = geom_vline(
          xintercept = 0,
          colour = "grey60",
          linetype = 2
        ), dodge_size = 0.6,
        ci_method="wald",
        errorbar_height = .2)+
theme_ipsum()+
scale_color_manual(values = wes_palette(4, name = "Darjeeling1", type = "discrete"), name="", labels=c("Black","Coloured","White","All Races"))+
theme(legend.position = "right",
      panel.background = element_blank(),
      plot.background = element_rect(fill = "White", color = "White"),
      panel.grid.major = element_line(color="#808080", size = 0.1),
      panel.grid.minor = element_line(color="#808080", size = 0.1),
      axis.title.x=element_text(colour="black", size = 12,family = "Times New Roman", vjust=0,hjust=0.5),
      axis.title.y=element_text(colour="black", size = 12,family = "Times New Roman",vjust = 3,hjust=0.5),
      axis.text.y=element_text(colour = "black", size = 10, family = "Times New Roman"),
      axis.text.x=element_text(colour="black", size = 10,family = "Times New Roman"),
      axis.ticks = element_blank(),
      plot.margin = unit(c(0.5,0.5,0.5,0.5), "cm"),
      plot.title = element_text(colour="black", size = 14,family = "Times New Roman",hjust=0.5, face="bold"),
      plot.subtitle = element_text(colour="black", size = 10,family = "Times New Roman", hjust=0.5),
      plot.caption = element_text(colour="black", size = 10,family = "Times New Roman")),

```

```

        legend.text = element_text(colour="black", size = 10,family = "Times New Roman"),
        legend.title = element_text(colour="black", size = 12,family = "Times New Roman"),
        strip.text = element_text(colour="black", size = 10,family = "Times New Roman", vjust=1,hjust=0.5))+
    labs(
        y = "",
        x = "Coefficient Estimate",
        title = "Coefficient Plot of Second Research Question",
        subtitle = "Dependent Variable: Duration of Last Illness*Race"
    )

```

END

Appendix C

Declaration on overlap with module essay

To whom it may concern,

A previous assignment submitted for evaluation to the Department of Economics, Stellenbosch University, had made use of the same data used in this thesis. The previous assignment was submitted as part of coursework for the module Economic History 771. While some of the data cleaning was done for the previously submitted assignment, it was not nearly as extensive as for this thesis. To this end, all the data used in this thesis had been newly cleaned through R and no previous data work from the previously submitted assignment was used.

The only overlap present in this thesis from the previously submitted assignment is an adjusted extract on a brief overview of Paarl from 1900 to 1930 and has been indicated clearly in this thesis. The overlap between this thesis and previous work is not substantive and merely played the role of background work. Hence, this thesis consists entirely of new work and expands greatly on previous work submitted.

A handwritten signature in black ink, appearing to read 'Jan-Hendrik Pretorius', is written over a horizontal line.

Jan-Hendrik Pretorius

(Signed 06/01/2023)