Health By Wealth: Optical character recognition and LLMs with population-level probate and administrative data uncover substantial inequalities in health over the very long run

Naomi Muggleton[1,2], Aaron Reeves[3,4], **Charles Rahal**[5,6], Paul Moore[4], Linda Li[7], and Alexandra Rottenkolber[8]

[1]Warwick Business School, University of Warwick
[2]Brasenose College, University of Oxford
[3]Department of Sociology, London School of Economics and Political Science
[4]Department of Sociology and Intervention, University of Oxofrd
[5]Oxford Population Health, University of Oxford
[6]Nuffield College, University of Oxford
[7]Department of Methodology, London School of Economics
[8]Department of Sociology, Linköping University

November 1, 2025

This is a preliminary work in progress. Please do not cite or distribute.

Abstract

Health inequalities in life expectancy between the rich and poor remain a persistent challenge. Yet, their long-run evolution has been poorly understood due to limited historical data combining socio-economic status and mortality. Existing evidence is often fragmented across time and place, leaving unresolved when and why disparities changed. We address this by constructing a novel dataset spanning 130 years, linking 16 million digitized probate records to 66 million death registrations, providing unprecedented detail on wealth and mortality in England and Wales. This enables macro-level analysis from individual microdata. Between 1860 and 1990, life expectancy inequalities narrowed markedly, particularly after 1940. The gap at age 20 between the wealthiest 1% and poorest groups declined from 17.3 years for women and 15.8 years for men to just 2–3 years in recent decades, coinciding with the emergence of the welfare state. We further link genealogical data to the same probate records to assess intergenerational effects. Contrary to earlier debate, we find clear evidence that individuals born into wealthier families lived longer before 1900. Our results show that health inequalities, though enduring, can be reduced. The remarkable rise in life expectancy after 1850 was driven largely by gains among the least wealthy, reflecting collective societal efforts—through public health initiatives, poverty reduction, and universal healthcare—to extend healthy lives to all.

Keywords: *Machine Learning*, *Social Inequality*, *Population Health*, *Mortality*

It is not contentious to hypothesize that health inequalities are both pervasive, and durable. They exist in every country for which we have data (Murtin et al., 2017), and they have been largely stable in recent years (Mackenbach et al., 2018), not least but especially in countries which have completed their demographic transition. The degree to which health inequalities are in fact durable over the long run is, however, unclear. In part this is because the data used to analyse this question only covers a relatively short period of time (a few decades at most), and health inequalities are produced over and reproduced across entire life courses. Intergenerational analysis over the long run is sincerely and especially lacking. Other historical work is almost exclusively focused on individual – often geographically unique – parts of space and time, resulting in a piecemeal understanding of this critical demographic and population health problem. Limited examples include, for example, Chadwick (1843) which estimates that the average age at death in Liverpool in 1840 for professional persons was 35, but just 15 for laborers, mechanics, and servants. Other, more recent examples come from the influential 'Whitehall Studies' (such as Marmot and Brunner, 2005) which show that each hierarchical level of civil servant job seniority experienced a higher mortality rate than the group one step higher. This results in academics having to rely on snapshots of information to draw conclusions, despite a multitude of theories which explicate the link between things like socioeconmic status, wealth, and health outcomes (Phelan et al., 2010). Indeed, data on health inequalities over the very long-run is extraordinarily uncommon, and is usually only stitched together into snapshots based on non-comparable projects which are based on diverse and disparate sources of information. We simply do not know when, if at all, and – if so – why inequalities in life expectancy have changed.

Here we present the first, original and yet unpublished results from an unusually and extraordinarily detailed and uniquely constructed dataset. While our research design is extraordinarily computational in nature, it addresses the dearth of studies as described above. We parse digitized versions of handwritten probate records using the latest techniques in Computer Vision and Optical Character Recognition. This data comes from the National Probate Calendar (created by the Probate Registry); a register of proved wills and subsequent bequeathments available following the enactment of the Court of Probate Act 1857 across administrations in England and Wales since 1858. This data is at the almost entire population level. Processing this data for demographic analysis involved a complex pipeline to make the data 'research ready'. We first set up 138 virtual machines; one for each year. We then sent requests for probate images at a rate of – on average – one every 10 seconds. We then used wildcard searches to iterate through all surnames between Aa* and Zz*, and saved the images for downstream processing. This processing – on over one million `.png` images – involved; inverting colours, binarizing grayscale, programatically removing ink blots, detecting margins with Gaussian processes, programatically removing dust, removing handwritten margin annotations, calculating pixel angulation and skew, page resizement (based on optimal extraction), the artificial creation of bounding boxes, optical character recognition, the use of large language models to detect extraction mistakes, and finally, fuzzy regular expressions to create tabular data ready for analysis. The millions of images, when stacked on their edges, would be over one hundred metres in length if they were printed in their physical version. These data encompass all wills proved in the UK above certain administrative thresholds (with rare exception, such as for members of the Royal Family), allowing us to capture the full wealth distribution of almost all of a large population. This ranges – over 16,634,470 rows of cleaned data – from the upper echelons of British society (our dataset includes, for example, Charles Darwin who left £145,911 in 1842, inflation adjusted to ∼£18.75m in 2025), to tradespeople and widows leaving a few pounds (such as John Croxton, a comparatively unknown labourer who died in December 1854 and left £5 (inflation adjusted to £625 in 2025) to a farmer. To note – and while we have full 'ethical review board' approval – our analysis is exempt from GDPR compliance restrictions because it primarily involves individuals who have sinced deceased. Based on a complex record linkage procedure which involves *extensive* manual verification and hand labeling for model training, we reconcile the probate with administrative death records for the purpose of recovering dates of birth which allows us to compute individual level longevity at scale for the first time. This comes in the form of digitised and

population wide death registration data (n=66,161,380) which involves thousands of manual transcriptors, as well as with other auxiliary sources, such as tools of gender inference and vast geneological data in order to examine the intergenerationla transmission of wealth and its subsequent effect on life expectancy. Our computational pipeline is visualised in Figure 1.

Despite temporal variation, we can at times match the vast majority of death records onto the probate due to our advanced record linkage techniques. This is done through a new, novel, and custom-built Python 3.x library which also computes a range of life expectancy based statistics, visualisations, and statistical tests. We aim to publish this library 'Open Source' (under a GNU GPL 3.0 license) in advance of the conference, so that others may gain utility from it. This library also computes an extremely large number of cohort based life tables – tens of thousands – based on the entire feasible range of assumptions which any reasonable demographer could have chosen to make. This includes not just different conceptualisations of wealth, but also a multitude of different methodological assumptions with regards to smoothing of mortality rates. While the foundations of this library are based on Preston (2000), it also leans on ideas from Caughley (1966) and a multitude of other, more recent methodological literature (Riffe et al., 2019). The existence of the 'personal effects' variable which is programatically extracted from the raw images allows us to make comparative statements about longevity by wealth at death at an unprecedented temporal span and resolution. Figure 2, for example shows simple age at death calculations across each of the fourth quartiles of the wealth at death distribution (by year) using observations on individuals found in the probate records. However, by virtue of the fact that we have the entire population register of deaths, we are able to decompose observations into those who hold an amount of wealth which would put them in the 'top 1%' of population, those who held some wealth, and those members of the population who held no wealth (or wealth below the small threshold for inclusion). This is shown in Figure 3. For both male and females, for all years in our sample, and for life expectancies at almost all ages, we observe enormous stratification by personal effects at death. This gap narrows over time – and especially since the 1940s – but in earlier parts of our sample, the gap in life expectancy at age 20 between the '1%' and those holding no or negligible wealth is as wide as 17.3 years for females, and 15.8 years for males. In addition to Figures 1-3, our analytical pipeline also outputs a large number of other, illuminating examples. Further to this, we link individuals across time in order to combat endogeneity, again showing that life expectancy is a function of wealth, but, there, of wealth held by parents. This enables an analysis of the effects of transmission of wealth before individuals have the opportunity to acrue wealth through other, non-hereditary channels.

We provide the most granular picture on inequalities in life expectancy to date. This analysis elucidates when, and among which parts of the wealth distribution these changes in life expectancy occurred over more than a century of unprecedented data. Declines in health inequalities are driven by economic growth and increased government support to low-income households, and touch on important debates across various academic disciplines (demography, economic history, sociology, and population health), contributing to theories of demographic transition, the 'fundamental causes' of health inequalities, and the impact of industrialization on society. Our advanced computational techniques extend and refine classical demographic frameworks. Through the integration of probabilistic models of lifespan with detailed wealth distribution analyses, our approach addresses persistent gaps in the literature concerning the evolution of health disparities over the long run. Our custom-built Python library enables extraction of demographic parameters and facilitates the computation of robust confidence intervals, enhancing the statistical rigor of our findings. Moreover, the implementation of both deterministic and probabilistic record linkage methods significantly improves the accuracy of matching probate records with death registration. This framework solidifies our historical findings and reveals subtle interactions between socioeconomic status and mortality that have previously been overlooked. Consequently, our analysis contributes a comprehensive understanding of the determinants of health inequalities through a new computational lens, offering valuable insights for policymakers and researchers interested in mitigating the adverse effects of social and economic disparities on population health.

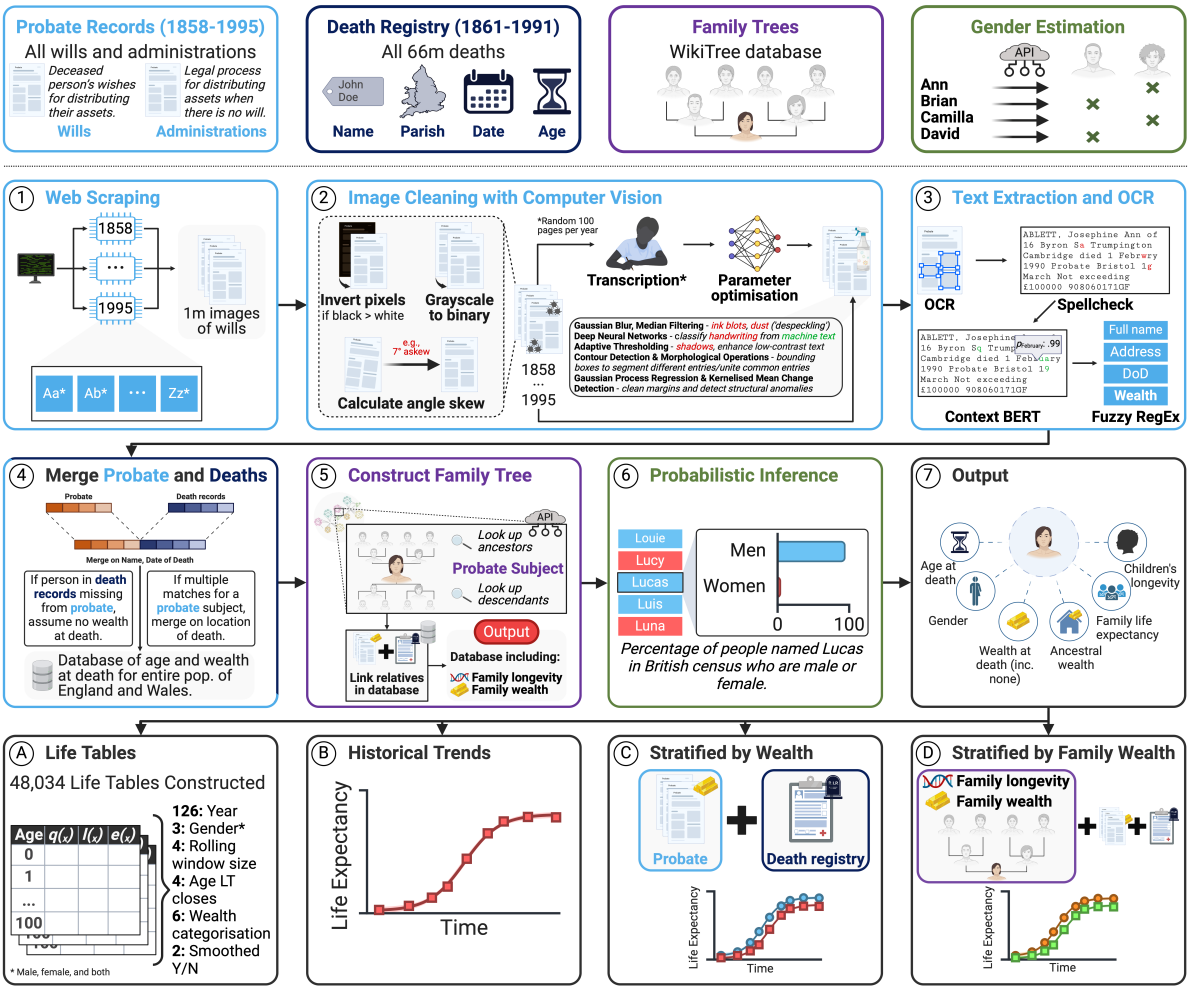

Figure 1: Our Computational Pipeline: A schematic which shows the different parts of our computational pipeline.

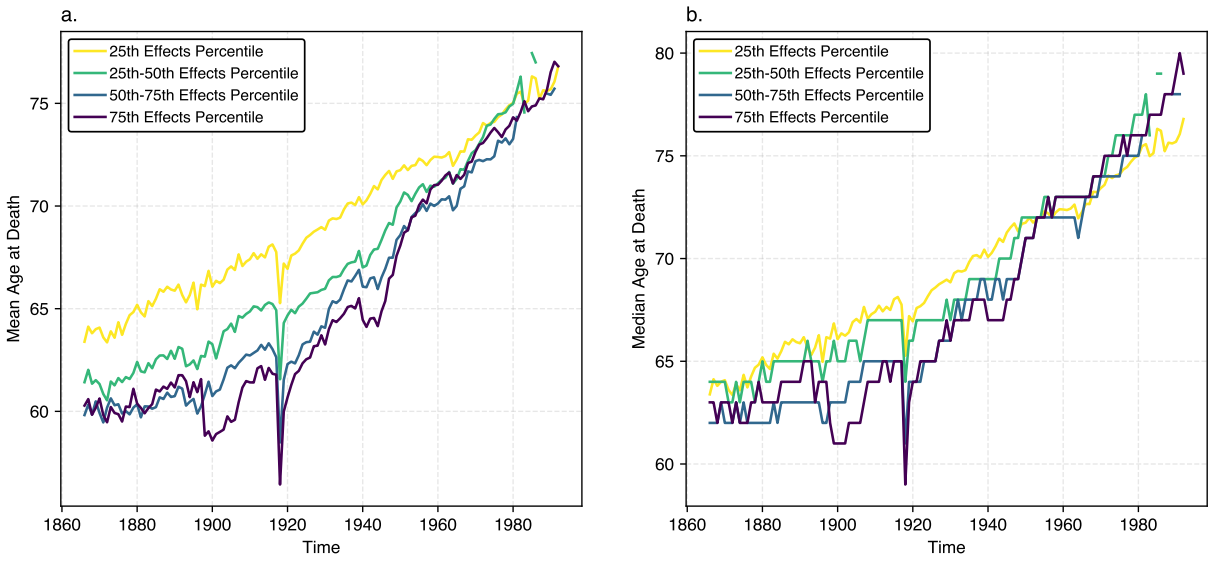

Figure 2: Age at Death by Effects: Subfigure 'a.' shows mean age at death across the four quartiles of the wealth distribution. Subfigure 'b.' shows the equivilent calculations, but instead using the median instead of the mean. Both figures are for male and females combined.

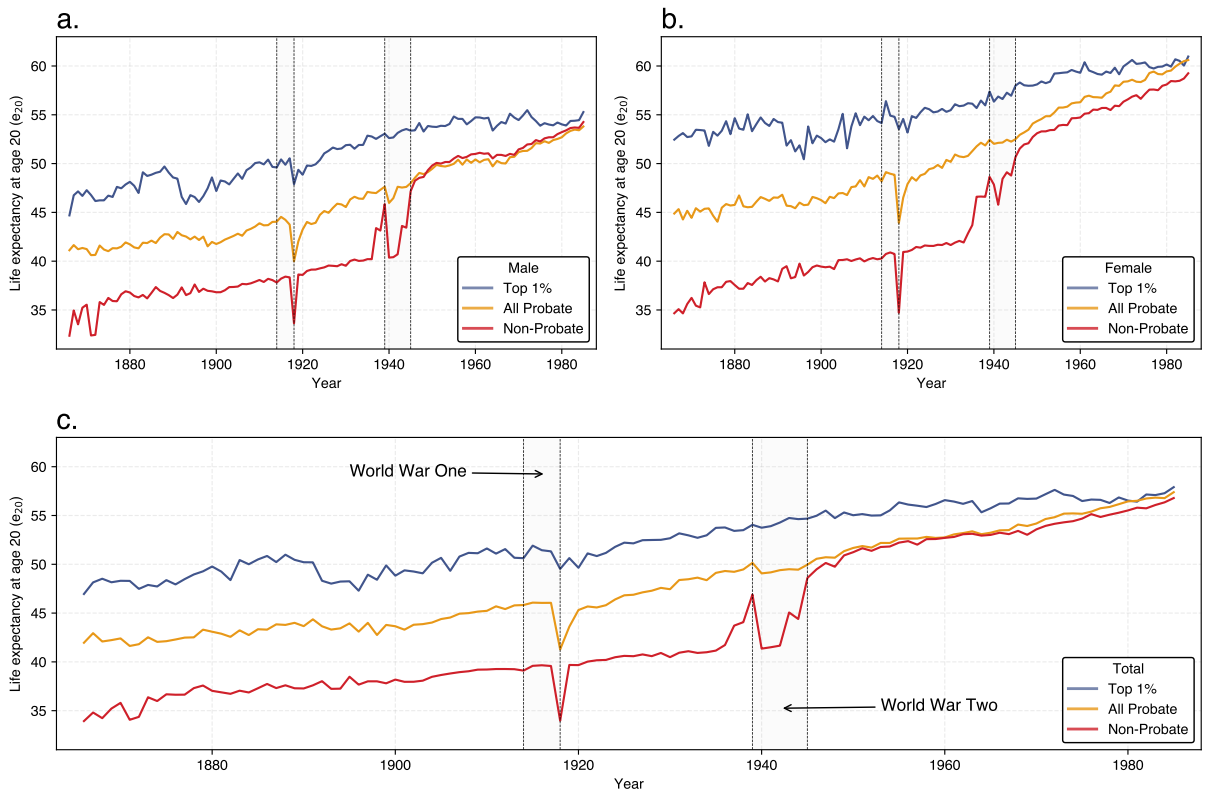

Figure 3: Life Expectancy at 20 years old: Subfigure 'a.' shows life expectancy at 20 years old for men, subfigure 'b.' for women, and both combined for subfigure 'c.'. Lines denote the wealth ('personal effects') of those estimated to be in the 'top 1%' of the population's wealth distribution, all people who have left *some* effects (on the probate register), and all other members of the general population.

References

Caughley, G. (1966). Mortality patterns in mammals. *Ecology 47*(6), 906–918.

Chadwick, E. (1843). *Report on the sanitary conditions of the labouring population of Great Britain: A supplementary report on the results of a special inquiry into the practice of interment in towns. Made at the request of Her Majesty's principal secretary of state for the Home department*. W. Clowes and sons.

Mackenbach, J. P., J. R. Valverde, B. Artnik, M. Bopp, H. Brønnum-Hansen, P. Deboosere, R. Kalediene, K. Kovács, M. Leinsalu, P. Martikainen, et al. (2018). Trends in health inequalities in 27 european countries. *Proceedings of the National Academy of Sciences 115*(25), 6440–6445.

Marmot, M. and E. Brunner (2005). Cohort profile: the whitehall ii study. *International journal of epidemiology 34*(2), 251–256.

Murtin, F., J. Mackenbach, D. Jasilionis, and M. M. d'Ercole (2017). Inequalities in longevity by education in oecd countries: Insights from new oecd estimates. *OECD Statistics*.

Phelan, J. C., B. G. Link, and P. Tehranifar (2010). Social conditions as fundamental causes of health inequalities: theory, evidence, and policy implications. *Journal of health and social behavior 51*(1_suppl), S28–S40.

Preston, S. (2000). Demography: Measuring and modeling population processes. *(No Title)*.

Riffe, T., J. Aburto, M. Alexander, S. Fennell, I. Kashnitsky, M. Pascariu, and P. Gerland (2019). Demotools: An r package of tools for aggregate demographic analysis.