Studying Historical Longevity Using Survival Analysis

Vincent Weng

Introduction

This memo investigates the factors influencing longevity among notable individuals in the BHHT dataset. The research focus is to analyze how lifespans vary across different eras of birth, geographic regions, genders, and other factors. The BHHT dataset, derived from Wikipedia biographies, contains approximately 2,291,817 rows of data. The dataset captures key demographic information, including birth and death years, regions, and notable achievements. This provides a unique lens to analyze how lifespans vary across different eras, geographic regions, genders, and other factors. By examining these patterns, we can uncover disparities in survival outcomes and the extent to which external factors influence longevity. Given its scope and focus on notable individuals, this dataset offers a valuable opportunity to examine long-term trends in survival and mortality. However, it may also be influenced by selection biases, such as historical societal norms and documentation practices that may have favored recording male achievements as opposed to “notable” females. Nonetheless, this analysis seeks to uncover trends in longevity and highlight disparities based on demographic characteristics.

Methods

In this analysis, we utilized survival analysis techniques as the main method to investigate longevity trends among notable individuals. The marginal survival function, estimated using the Kaplan-Meier method, was used to calculate the proportion of individuals surviving to a certain age. This method also accounts for right-censoring, ensuring that individuals without recorded death dates, such as those still alive in 2020, are properly adjusted in the analysis. From there, survival curves were generated to visualize differences in survival probabilities across key demographic factors, including era of birth, gender, and geographic region. These curves provide an empirical estimate of lifespan distribution and allow for comparisons between various groups over time. To further assess age-specific mortality risks, we applied marginal hazard functions, which estimate the instantaneous risk of death at a given age. Unlike survival functions, which describe the cumulative probability of survival over time, hazard functions highlight how mortality risks fluctuate throughout a lifespan. This provides a much more granular view of how mortality rates evolved overtime. Finally, marginal hazard functions were also stratified by key demographic factors, such as birth era and gender, to examine how different groups experience mortality risk at various ages.

Results

The marginal survival functions show that survival probabilities improve significantly in later eras. Individuals born in the 1900s have the highest survival rates, with 60% surviving to age 80, while those born in the 1500s dropping under 20% for the same age (Figure 1). The earlier cohorts show a steeper decline, with 1500s being steeper than 1600s and 1600s being steeper than 1700s etc (Figure 1). Gender differences are also evident across all ages, with females consistently showing higher survival rates than males. By age 80, the male survival rate is only 0.4 as opposed to the female survival rate of 0.6 at the same age (Figure 2). Regional survival analysis also reveals interesting results, with Africa and Asia showing slightly higher survival probabilities that are visually differentiable past age 80. However, the differences remain quite subtle across the five regions and may not be statistically significant without further analysis.

A similar pattern emerges when examining marginal hazard functions. Stratified by era, the marginal hazard functions reveal that individuals born in earlier periods, such as the 1500s, experience the highest hazard rates, with sharp increases starting around age 60 and peaking near age 80 (Figure 4). In contrast, those born in the 1900s exhibit significantly lower hazard rates, which increase more gradually and remain below 0.1 even at older ages (Figure 4). Furthermore, gender differences in hazard rates are again evident across all age groups, with males consistently showing higher risks than females. This difference becomes particularly pronounced after age 50, where the hazard rate for males is approximately 0.10 past age 80, compared to about 0.06 for females at that same age (Figure 5).

Interpretation

The analysis reveals significant trends in survival probabilities and mortality risks across time periods, genders, and regions. Survival has improved over time, with individuals born in the 1900s exhibiting the highest survival rates, likely due to advancements in healthcare and living conditions, while those born in the 1500s show a much steeper decline. Hazard modeling also revealed consistent findings, showing that individuals born in the 1900s faced significantly lower mortality risks compared to those born in earlier eras. This aligns with the observed improvements in survival probabilities over time and can similarly be attributed to the same reasons in the improvement to quality of life. Next, females consistently demonstrated higher survival probabilities and lower hazard rates than males, particularly at older ages, which may be due to biological resilience and lower engagement in high-risk behaviors. The widening mortality gap suggests that women not only live longer but also maintain lower health risks in later life, reinforcing the persistent survival advantage of females. Marginal hazard function results further indicate that individuals born in later eras face lower age-specific mortality risks, with hazard rates increasing more gradually and peaking at older ages. Furthermore, the slight visual and negligible differences in survival function stratified by regions are likely reflected by the historical variations in healthcare, economic development, and living standards. Regions with better access to medicine and sanitation saw greater longevity gains over time. However, these differences may also stem from dataset biases, as historical records may overrepresent notable individuals from certain regions.

Something noteworthy is that these results may be influenced by dataset biases, particularly in how "notability" is defined. Historically, societal norms prioritized documenting male achievements, leading to a disproportionate representation of men, particularly those in high-risk professions. This may artificially inflate male mortality rates. Conversely, women in the dataset may have been more likely to come from aristocratic or other privileged backgrounds associated with longer lifespans. This raises important questions about whether the observed survival patterns are primarily driven by geography, time period, or dataset selection biases. Nonetheless, these findings shed some light on the interplay of historical, biological, and social factors in shaping survival overtime. Additional research and analysis could examine factors like occupation and socioeconomic status to better understand these trends.

Figures

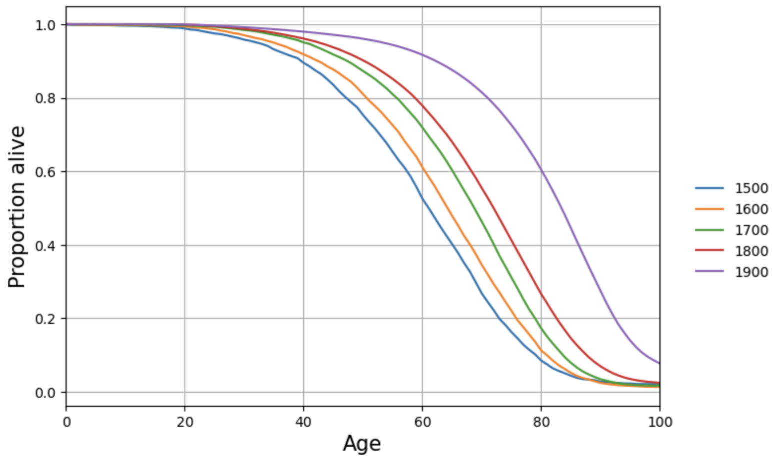

Figure 1: Marginal Survival Functions by Era

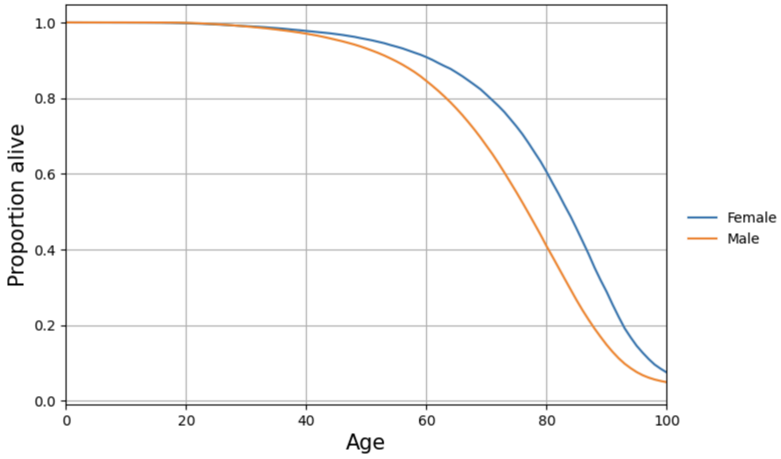

Figure 2: Marginal Hazard Function by Gender

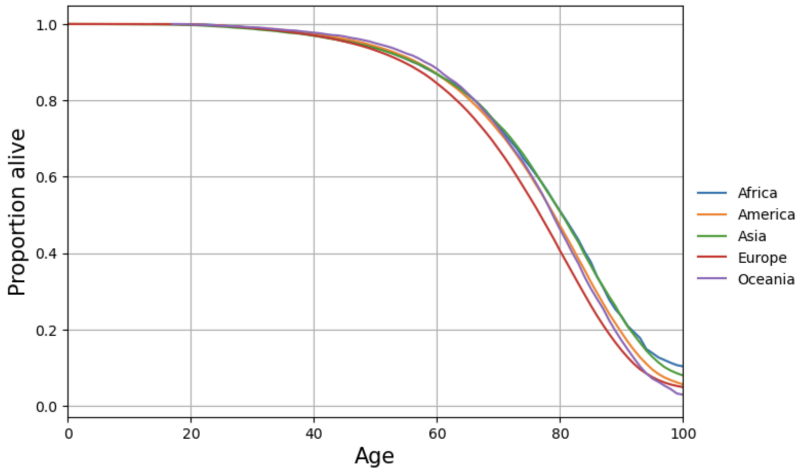

Figure 3: Marginal Survival Functions by Region

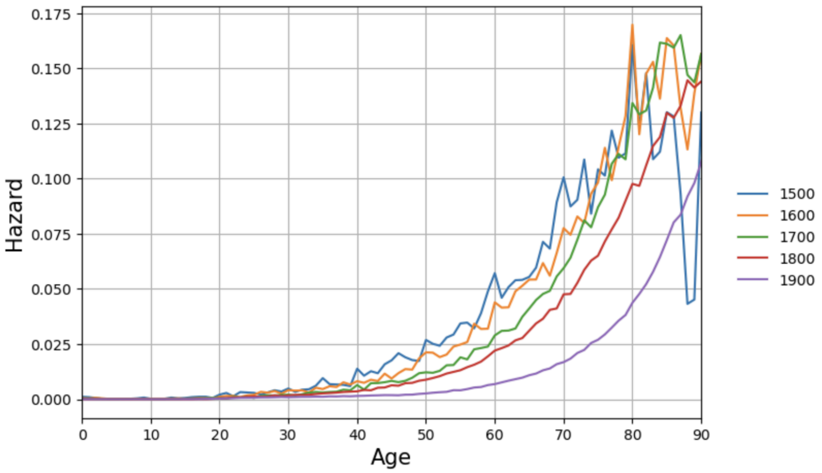

Figure 4: Marginal Hazard Function by Era

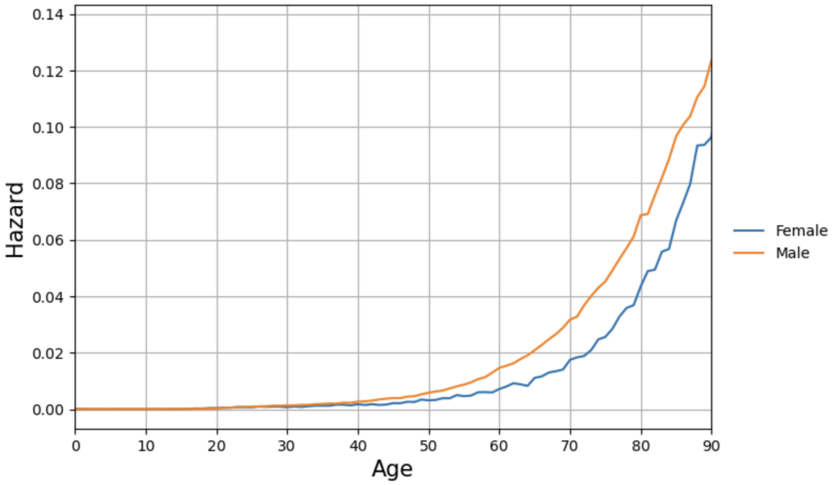

Figure 5: Marginal Hazard Function by Gender