Occupational Divides in Survival: Evidence from Notable People Data

1. Introduction

This study examines how occupational affiliation shaped survival outcomes among notable individuals born between 1600 and 1960. By applying marginal hazards function and proportional hazards regression models, we investigated not only whether occupations differed in late-life mortality, but also how their survival trajectories interacted with historical gains in longevity.

2. Data Description and Preprocessing

The data used in this study comes from the Brief History of Human Time (BHHT) Cross-verified dataset, which contains 2.2 million notable individuals sourced from several Wikipedia editions and Wikidata. We restricted the sample to individuals born between 1600 and 1960, for both historical and statistical considerations. Before 1600, biographical records were sparse and highly uneven across regions - for example, Oceania contributes only one individual out of 60,000. After 1960, a large proportion of individuals are still alive, so their survival outcomes are greatly censored and provide limited information. As shown in Figure 1, individuals born after 1961 exhibit only a 2% death rate, producing limited, reliable inference on lifespan outcomes. Focusing on 1600-1960 (highlighted between the two green lines in Figure 1) provides the majority of the data (about 65%), ensures more complete lifespans, and avoids excessive censoring.

The survival outcome is the observed survival time of each individual, where the censoring time is 2020 (the last year when anyone died in the dataset). Individuals for whom a death year is recorded contribute complete lifespans; those still alive by 2020 are right-censored at their age in that year. Occupations are categorized into four broad domains: Culture, Science and Discovery, Leadership, and Sports/Games. Two residual categories (Other and Missing) were excluded. To mitigate potential biases from truncation and delayed entry, the models also condition on covariates for birth year and region attachment, since both of them could be determinants of mortality hazard.

The final analytic dataset, therefore, consists of a survival outcome (lifespan with censoring) and covariates for birth year, occupation, and region, encompassing a total of 1,293,384 individuals.

3. Methods

To investigate how occupational affiliation relates to survival, we employed a sequence of survival analysis techniques that progress from descriptive comparisons to more formal modeling.

To establish descriptive differences across occupations, we first estimated marginal hazard functions that capture the risk of death at each age without conditioning on covariates. For each

age, the hazard is calculated as the ratio of the number of deaths to the number of individuals still at risk immediately preceding that age. This corresponds to the discrete-time analogue of the derivative of the Nelson-Aalen cumulative hazard estimator. Then we plotted the hazard across ages 0-90 (See Figure 2) to compare the age-specific hazards across each occupation.

Next, to assess how birth cohort influences survival outcomes across occupations, we fitted a proportional hazards (PH) regression model with interaction terms between birth years and occupations. The PH model assumes that individual hazard functions are proportional over time, differences across groups can be summarized through hazard ratios, while leaving the baseline hazard unspecified. We modeled the birth year effect using cubic spline functions to capture the non-linear improvements in lifespan, and interacted these splines with occupational categories. The interaction terms test whether the effect of year of birth differs by occupation. The region of attachment is included as a control to account for geographic variation in mortality that could confound occupational comparisons. To evaluate the added power of the interaction terms, we tested the significance of the interaction terms using a log-likelihood ratio test, comparing the model with and without interactions. The resulting partial effects are visualized in Figure 3, which plots the contribution of birth year to the log hazard separately for each occupation.

4. Results and Interpretation

Based on the provided survival analysis methods, we have the following results:

The marginal hazard plot (Figure 2) reveals clear differences across occupations. Hazards remain close to zero through early life (before age 50), then rise steadily from midlife (ages 50-60), with variation in slope and magnitude across groups. Leadership occupations (green curve) display a sharper increase in hazard after age 60, indicating higher late-life mortality risk. Culture (blue curve) and Discovery/Science (orange curve) follow a similar trajectory, rising more gradually but converging with Leadership by age 90. By contrast, Sports/Games (red curve)consistently exhibit lower hazard across older ages. These results suggest that occupation is associated with differences in survival.

The proportional hazards model with interaction terms between birth year (modeled with cubic splines) and occupations reveals that both cohort and occupation jointly shape survival. Figure 3 visualizes the effect of birth year on the log hazard for each occupation, holding region fixed at Europe (representing nearly half of the sample). Across all groups, late birth cohorts are associated with lower log hazards, reflecting historical improvements in longevity. Between 1600 and 1700, Sports/Games (red curve) exhibits sharper early declines in log hazard relative to the gradual improvements in other groups. From 1700 to 1850, Sports/Games hazards partially rebound, whereas other occupations continue on a slow but steady decline. After 1850, all groups converged toward substantially lower hazards, consistent with the onset of modern improvements in health and survival.

The log-likelihood ratio test yields a test statistic of 2722 (df = 12, p-value < 0.001), rejecting the null hypothesis of homogeneous cohort effects, and confirming that the effect of birth year on survival varies by occupation. These results suggest that occupation not only correlates with overall hazard level, but also conditions the timing and magnitude of historical improvements in survival.

5. Conclusion

By combining the results of the marginal hazard function with proportional hazard regression, the analysis shows differences in mortality risk across occupations. Leadership occupation faces a higher late-life mortality hazard, while Sports/Games exhibits a lower risk. Also, the interaction model reveals that individuals with different occupations experience different survival benefits of late birth. This suggests that occupational domains are linked to different exposures, opportunities, and vulnerabilities that accumulate across the life. The results underscore the value of situating survival analysis within both social and historical context.

Figure 1

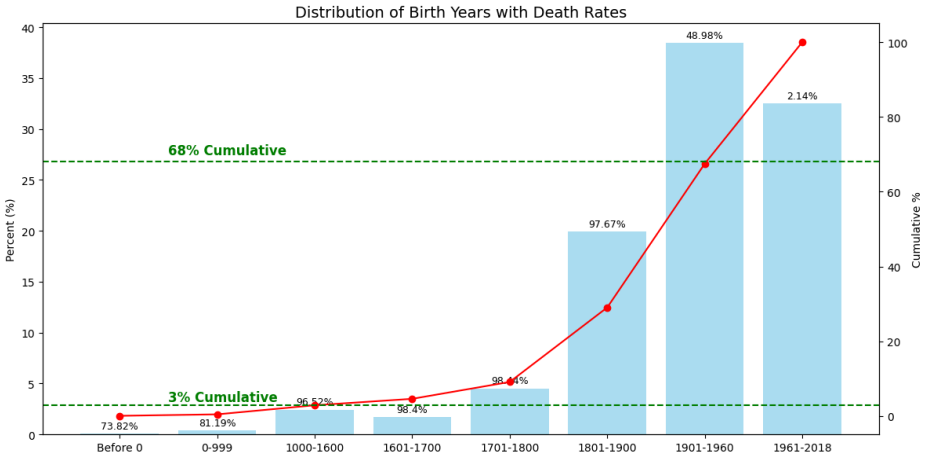

Figure 2

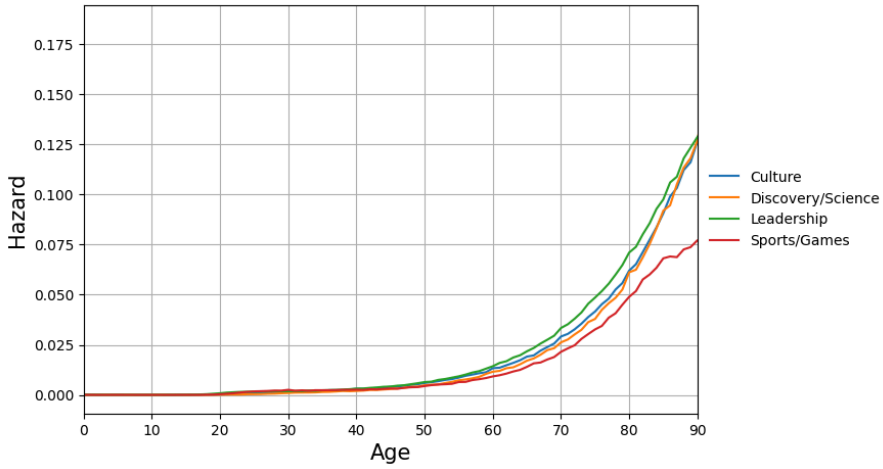

Figure 3