Gender Gap in Wikipedia: Are Women Vanishing Soon?

Khandaker Tasnim Huq
University of Maryland

Giovanni Luca Ciampaglia
University of Maryland

Abstract

We investigate whether biographies of women in the English Wikipedia are nominated for deletion sooner than those of men. Using survival analysis, we explore gender disparities in nomination timing from January 1, 2001, to November 3, 2023, covering the entire history of the Articles for Deletion (AfD) process. We examine factors influencing the creation-to-nomination time, including the chronological evolution of Wikipedia, whether an individual was a living person at the time of nomination, and their historical period. Our findings indicate that women are nominated for deletion faster than men, suggesting that gender strongly influences the risk of deletion consideration in AfD.

Keywords: Wikipedia, Gender Gap, Article for Deletion, Survival Analysis, Wikidata

Introduction

In Wikipedia, Article for Deletion (AfD) is a collective deliberation process to find a consensus on whether a subject should be deleted from the encyclopedia on the basis of notability. Prior work on gender and AfD indicate that notable women are nominated for deletion more frequently than men (Tripodi, 2023). However, it is unknown how quickly those biographies get nominated for deletion. Estimating deletion likelihood over the lifespan of an article holds significance for various reasons. First, articles mature over time due to the collaborative editing, which typically enriches their content and references. Notability assessments are an integral part of this editorial process, but articles may face nomination for deletion at any stage of development. Thus, because articles are continuously developed, early nominations shorten the window for further improvement. Second, in recent years community interventions like "Women in Red" have increased the share of biographies on women; but articles typically receive higher scrutiny, and thus are at greater risk of deletion, when they are new (Tripodi, 2023). Thus, our objective is to investigate: how promptly are biographies nominated for deletion in the AfD process? Specifically, we aim to determine if the biographies of women face a disadvantage in the timing between creation and nomination for deletion.[1]

Survival from Nomination for Deletion

We aim to address the main question while accounting for factors that, in addition to gender, could influence considerations for deletion. One such factor is the biographical status of individuals: notability debates are more challenging for living people due to concerns about the reliability of their entries and the need to prevent harms to their reputation. This may be especially relevant for the biographies of women. Societal awareness of gender equality has improved over the past century, leading to a higher representation of women among notable figures in encyclopedias. This could mean that women are proportionately more featured in articles about living people compared to the overall ratio of women in Wikipedia biographies.

Another factor is whether the person nominated is an historical figure. Wikipedia is affected by a well-known bias for recent events, which may lead to fewer historical figures being represented on the platform. This raises concerns about the susceptibility of female historical figures to deletion nominations in AfD.

Finally, Wikipedia has evolved significantly over the course of its history as a collaborative project, expanding its coverage and establishing stringent rules to maintain content quality and prevent vandalism. The proportion of biographies of women is still low despite efforts within the community (like the aforementioned "Women in Red" project) to reduce the gap in gender coverage. The AfD process may play a role in limiting the effectiveness of these interventions, since within these deliberations it is often debated whether gender should be taken into account when gauging the notability of a subject. If this is the case, then we should observe a higher likelihood of nomination for biographies of women created later during the history of Wikipedia.

Data

Our goal is to estimate the probability of 'survival' from nomination in AfD as a function of article 'age' (i.e., time

[1] This research was partially supported by a grant from the Wikimedia Research Fund.

since its creation). Therefore, using Quarry and the MediaWiki API, we gathered both existing and nominated biographies in English Wikipedia and their creation dates from January 1, 2001, to November 3, 2023, covering the entire history of the AfD process. We accessed the Archive table to get the creation dates of deleted biographies in AfD. In total, we collected 1,975,779 biographies (19.5% women) among which, 84,366 biographies (25% women) were nominated for deletion. Additionally, we retrieved from the SPARQL endpoint of WikiData the gender, date of birth, and date of death of each subject to determine their historical or contemporary status. Dates of birth range from 7999 B.C. to 2022 A.D. We also obtained a list of articles on living people using PetScan (`meta.wikimedia.org/wiki/PetScan/`).

Analysis of the gender information from Wikidata revealed that only a small fraction of biographies (0.09%) are labeled with genders other than 'man' and 'woman'. Thus, to prevent introducing statistical bias in our results, we only consider these two genders, while admitting that these are not the only genders where different gaps exist.

Survival Analysis

We use the Kaplan-Meier estimator (Kaplan and Meier, 1958) to estimate the probability of survival from nomination. We also employed the Cox proportional hazards model (Cox, 1972) to assess the risk of nomination considering the following three variables: *a*) Gender – if the subject is a woman (1) or man (0); *b*) Status – a variable with three levels: 1. *Historical* – if the subject was born before 1907 (cutoff estimated as the year of birth of the verified oldest living person at study time), 2. *Contemporary Alive* – if the subject was alive at the time of nomination/analysis, or 3. *Contemporary Dead* – if not alive; and *c*) Wikipedia age – the age of Wikipedia at the time of creation of the article.

Results

Figure 1 (Left) shows Kaplan-Meier curves for biographies of men and women, revealing a steeper drop for women, suggesting quicker deletion nominations than men. In Figure 2 (Left), the hazards model indicates that gender strongly influences the risk of deletion consideration in biographies, with biographies of women nominated 34% faster than those of men. We examined how gender influences premature deletion nomination risk. Figure 2 (Left) shows the result from fitting the hazards model with interaction terms between gender and status. The interaction analysis involving gender yielded statistically significant improvements over the baseline model. The interaction with 'Historical' (Figure 2 (Right)) showed a positive coefficient, indicating historical women face a deletion disadvantage compared to men. Moreover, Figures 3 illustrate the marginal effects of gender and status on the risk of nomination before and after interaction, suggesting that living women face a deletion disadvantage compared to other groups. Figure 3 (Right) also shows that historical women have higher risk of nomination than contemporary deceased men. Finally, in a retrospective analysis shown in Figure 1 (Right), we observe how factors such as gender, Wikipedia age, and status evolve over the history of Wikipedia. Early on, the influence of gender on nomination risk was negative, but it steadily increased until 2006, and remained consistently positive thereafter. Also, both historical and contemporary deceased women are at a disadvantage from the very beginning and are still at risk.

Discussion

Our research uncovers a tendency to prematurely question the notability of women in the AfD process at a higher rate than men. Furthermore, our retrospective analysis shows that gender has long been a significant factor in deletion nomination risk, even after interventions like "Women in Red". Taken together, these findings suggest that the notability of women is quickly undermined, giving them fewer opportunities to enhance their Wikipedia presence. This highlights the challenges of preserving on Wikipedia knowledge about women throughout history, where traditional narratives have often overlooked their contributions. The under-representation of women in historical and contemporary records contributes to perceptions of lower notability, leading to premature actions in the AfD process. Allowing sufficient time between article creation and deletion consideration could enable the development of biographies, especially for individuals with limited secondary references. A recommendation interface for AfD nominators could aid in assessing subject notability before deletion flagging in AfD. While the AfD deals with gauging notability, the analysis can expand to other deletion processes to uncover gender disparity. Finally, revising nomination guidelines to broaden the set of contributions and achievements for gauging notability could reduce the Wikipedia gender gap.

References

[Cox1972] David R Cox. 1972. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202.

[Kaplan and Meier1958] Edward L Kaplan and Paul Meier. 1958. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481.

[Tripodi2023] Francesca Tripodi. 2023. Ms. categorized: Gender, notability, and inequality on wikipedia. *New media & society*, 25(7):1687–1707.

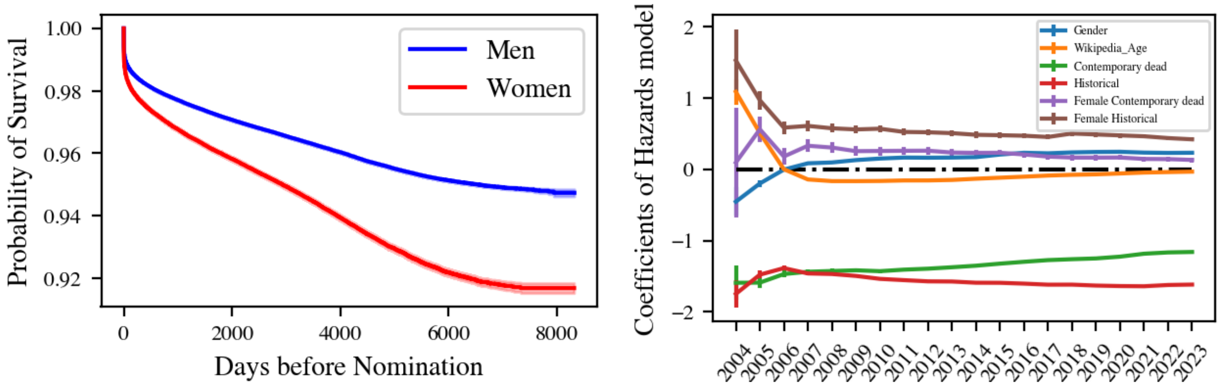

Figure 1: Left: The probability of survival of the biographies from nomination for deletion. The shaded area corresponds to the 95% confidence intervals. Right: Retrospective survival analysis. Each data point corresponds to the coefficients of the Cox proportional hazards model, fitted only on the data of articles created up to that year. Articles that were nominated after the observation window correspond to censored observations. The error bars represent robust standard errors. The black dash-dotted line corresponds to a coefficient value of zero.

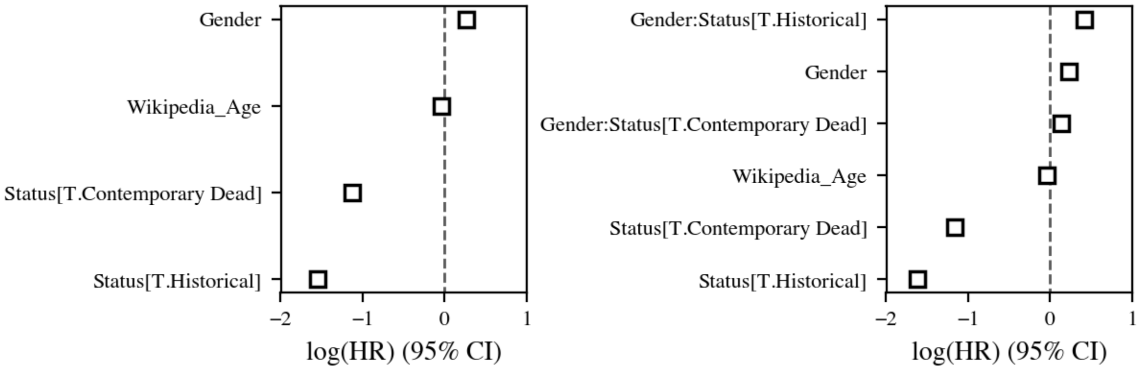

Figure 2: Results of Cox proportional hazards models on the full dataset. Left: Baseline model; Right: the model with interaction terms between gender and status. In both plots, error bars represent robust standard errors and are all smaller than the data points.

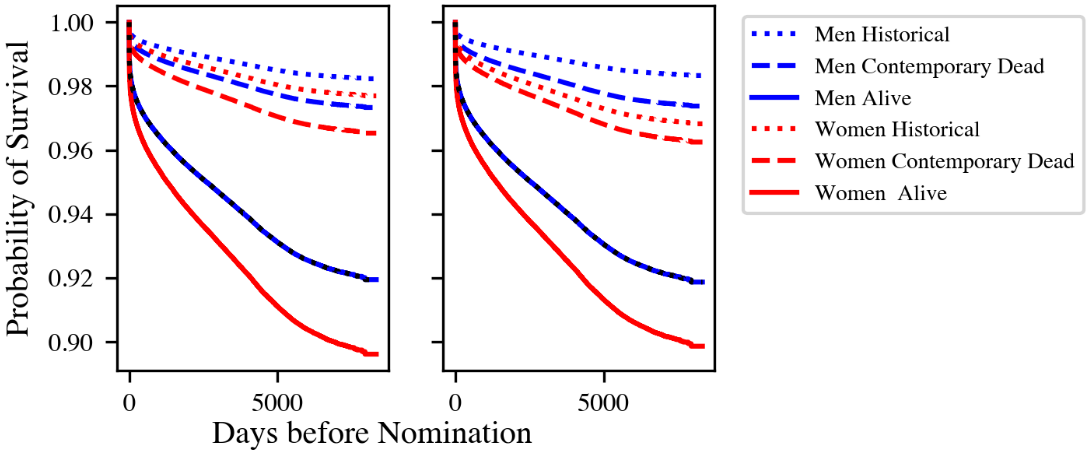

Figure 3: Marginal effects of gender and status on the full dataset. Left: Baseline model; Right: the model with interaction terms between gender and status.