DEMOGRAPHIC RESEARCH
A peer-reviewed, open-access journal of population sciences

DEMOGRAPHIC RESEARCH

VOLUME 53, ARTICLE 22, PAGES 661–704
PUBLISHED 15 OCTOBER 2025
http://www.demographic-research.org/Volumes/Vol53/22/
DOI: 10.4054/DemRes.2025.53.22

Research Article

Online obituaries as a complementary source of data for mortality in Canada

Pietro Violo

Nadine Ouellette

Contents

Demographic Research: Volume 53, Article 22
Research Article

Online obituaries as a complementary source of data for mortality in Canada

Pietro Violo[1]

Nadine Ouellette[2]

Abstract

BACKGROUND
Obituaries and death notices have existed for centuries as a form of commemoration, particularly in Western countries. With the rise of the internet, these records have become more accessible, presenting a valuable, largely untapped source for mortality research.

OBJECTIVE
We aim to collect online obituaries through web scraping and evaluate their representativeness, advantages, and limitations for use in mortality studies in Canada's two largest provinces: Quebec and Ontario.

METHODS
We web scraped 236,290 and 288,623 obituaries for Quebec and Ontario, respectively, spanning the years 2017 to 2022. Using regular expressions, a formal language for defining text-search patterns, we derived demographic variables from the text to compute mortality measures, which we then compared to a gold-standard vital statistics dataset.

RESULTS
Although obituaries in Quebec and Ontario respectively account for only half and one-third of all recorded deaths, the age and gender distributions they capture closely align with those of the general population. Infant deaths remain notably underrepresented. Life expectancy estimates derived from obituaries exceed official figures by 0.4 years for women and 0.5 for men, while the modal age at death is slightly underestimated. Despite these limitations, the timeliness and demographic representativeness of online obituaries make them a valuable supplement to conventional mortality datasets in Canada.

[1] Department of Demography, Université de Montréal, Canada. Email: pietro.violo@umontreal.ca
[2] Department of Demography, Université de Montréal, Canada

CONTRIBUTION
This study draws attention to an underused data source by leveraging Canada's bilingual context and developing methods for extracting demographic information from both French and English obituaries. We contribute to digital and computational demography by detailing techniques for web scraping, data cleaning, extraction, and validation, and by assessing coverage, age structures, gender disparities, and inherent biases in this type of textual data.

1. Introduction

For centuries, obituaries have been an integral part of many cultures around the world, serving both to announce people's passing and to commemorate the dead (Eid 2002). In the Western context, obituaries published in newspapers predominantly featured individuals who were white (Marks and Piggee 1999), male (Kastenbaum, Peyton, and Kastenbaum 1977), and of higher social status (Kearl 1987). Not only were such individuals more prone to be the subjects of obituaries, but their obituaries were also more extensive, better written, and more likely to include a photograph – an addition that was once considered expensive. However, the advent of the internet and the ease of disseminating obituaries online have diminished this exclusivity.

Until recently, newspapers served as the primary publication channel. The use of obituaries and death notices in research was therefore limited as it implied the cumbersome process of collecting newspapers clippings, scanning them, and reading each individually to extract relevant information. Due to the time-consuming nature of this manual work, sample sizes were usually small, leading to low statistical power and restricting the scope of research opportunities. Nowadays, obituaries have found their way onto the internet, making them more accessible and forever changing how we grieve (Hume and Bressers 2010; Walter et al. 2011). This pivotal change, along with advancements in data collection tools, is making this extensive body of data more available than ever before.

In this paper, we exploit an abundant compilation of online obituaries and death notices of more than 500,000 individuals whose funerals were held in Canada's two largest provinces, namely Quebec and Ontario, during the 2017–2022 period. Our goal is to examine the potential of these promising but still largely untapped sources for studying population mortality patterns while also highlighting key differences between French and English obituaries in each province. We first use web-scraping techniques to extract obituaries and death notices from the internet. Subsequently, we apply regular expressions, a formal language for defining text-matching patterns, to retrieve key information about deceased individuals in order to derive two fundamental demographic variables related to mortality: age at death and gender. After validating our data extraction process, we

compare the obituary-based data with official vital statistics death records using a variety of mortality measures. In support of open science in demography, we provide all R code used for web scraping, data cleaning, demographic variable extraction, and validation as publicly accessible replicable material. We conclude the article by discussing whether obituaries hold promise as a valuable unconventional source of data for mortality surveillance and research in the Canadian context and potentially beyond.

2. Background

2.1 Digital traces and the online deathscape

In recent years, there has been an unprecedented growth in the volume of data generated on the internet. This expansion is driven by the variety of media, such as images, videos, sound files, and texts, which are increasingly easy to produce, share, and archive. The trend of ever-increasing data creation and sharing shows no signs of slowing down, with each passing year breaking previous records (Clissa 2022). Alongside this growth, the use of digital traces, such as data gathered from social media, web search queries, or cellular networks, has also become increasingly prevalent for demographic research (Cesare et al. 2018), particularly in studying migration (e.g., Alexander, Polimis, and Zagheni 2019; Kim et al. 2020; Flores 2017) and, to some extent, in measuring fertility (e.g., Wilde, Chen, and Lohmann 2020; Wildeman, Schrijner, and Smits 2023). Mortality, however, remains a relatively poorly explored area in demographic research that leverages digital traces, or, more broadly, web-based data. Part of the reason for this gap might stem from the difficulty in distinguishing between an actual death and the cessation of online activity on social media platforms (Öhman and Watson 2019). The few studies on mortality that have taken advantage of digital traces have thus primarily focused on epidemiological surveillance. Examples include tracking influenza-like illnesses through Google search queries related to flu symptoms (Ginsberg et al. 2009) as well as correlating the incidence of common cancers with online search volumes on Google regarding cancer symptoms, diagnosis, therapies, and so forth (Phillips et al. 2018).

Demographers face substantial challenges in using internet-based data to study mortality, largely because of the limited availability of reliable sources that can truly serve as proxies for comprehensive death records. In response, a growing interest in digital data related to death has given rise to the concept of the 'online deathscape.' In this digital landscape, the primary data are similar to what internet users leave behind, often referred to as digital traces or 'breadcrumbs,' but in the context of mortality studies, these are known as 'digital remains' (Lingel 2013). Originating mainly from social media accounts, these data usually become inaccessible after the owner's death, yet they continue to exist within the digital ecosystem and remain interactive for other users. It is common

on platforms like Facebook for people to post condolence messages on the profiles of deceased individuals. Studies have shown that users rarely unfriend those who have passed away (Pennington 2013). In rare cases, these accounts are taken over by relatives, leading to the term 'digital zombies' (Bassett 2015). The volume of digital remains is expected to become increasingly abundant in the future. Öhman and Watson (2019) suggest that the number of deceased Facebook users could reach hundreds of millions in the coming decades, potentially outnumbering living users before the end of the century.

The exploration of how death intersects with our daily interactions has been the focus of thanatologists, who study the biological, medical, psychological, social, and cultural aspects of death. The term 'cyberthanatology' was introduced by Beaunoyer and Guitton (2021) to define the convergence of death studies (-thanatology) with the digital world (cyber-), highlighting the growing influence of digital technologies. This intersection is increasingly relevant in light of an aging population, which is leading to a greater dependence on digital platforms for expressions of grief, mourning, and remembrance (Beaunoyer and Guitton 2021). The most visited websites for online grief support have been categorized into five types: practical support, informational support, services, peer support, and resources (Beaunoyer et al. 2020). For instance, peer support encompasses spaces where individuals share their experiences with death or the loss of loved ones, while practical support includes information on funeral arrangements. The significance of these websites, particularly their user demographics and content, presents valuable opportunities for demographic research on mortality and bereavement. Online obituaries, which are the focus of the present article, fall under practical support and primarily serve to commemorate a person's death and inform others of the loss. When aggregated, obituary data may also reflect the characteristics of the broader population it represents.

2.2 Obituaries and death notices

As early as the 13th century in Europe, deaths were announced orally by a public officer known as a *crieur* ('town crier'). The ritual was gradually replaced around the 16th century by death notices, a transition made possible by the advent of the printing press. These notices were stapled to the doors of churches and elsewhere in the neighborhood of the deceased (Nicolet 2018). While the format has evolved over time, these notices still contain key information, such as the deceased's name, date and age at death, funeral or memorial details, and names of surviving relatives. Obituaries, on the other hand, offer detailed narratives of a person's life, highlighting achievements, family history, education, career milestones, and affiliations with social clubs.

In Canada, the type of memorial texts now shared online by funeral homes may also differ according to the language in which they are written, whether French or English. Quebec is primarily francophone and Ontario predominantly anglophone, and the stark

linguistic differences, along with other cultural particularities between the two provinces, may be reflected in the content and structure of their respective obituaries. In Quebec, these texts most often resemble death notices, combining both essential and optional details. They typically include the deceased's name, date and age at death, and funeral arrangements. In some cases, the cause of death may be indirectly suggested by mentioning a disease-related foundation, along with information about whether the deceased was predeceased by a spouse or sibling. By comparison, in Ontario, funeral home texts shared online are more often biographical, blending elements of traditional death notices with the narrative style typical of obituaries (McNeill 2004). It is also possible that these differences are not due to language alone but rather to a higher volume of obituary content produced by corporate funeral homes in Quebec compared to Ontario, as these tend to be less personalized (see Section 3). Consequently, in Canada, texts commemorating an individual's passing appear to have evolved into a hybrid form, combining features of both death notices and obituaries. Hence, in this paper, we use the terms 'obituaries' and 'death notices' interchangeably, as we develop and present a methodology for extracting demographic data from the majority of these texts.

The study of obituaries spans several academic fields, including epidemiology (Allison and St Leger 1999), literature (Barry 2008), psychology (Radtke, Hunter, and Stam 2000), thanatology (Beaunoyer and Guitton 2021), and sociology (Fowler and Bielsa 2007). Historically, most research in this area has focused on qualitative and anthropological approaches, but in recent years, obituaries have gained recognition for their potential to provide valuable demographic insights. For instance, at the onset of the COVID-19 pandemic, researchers in Italy used digitized newspaper obituaries to estimate early mortality rates (Buonanno and Puca 2021). Online obituaries taken from a regional newspaper have also been used in Austria to analyze socioeconomic inequalities in lifespan, with the size of the obituary serving as a convincing proxy for the deceased's socioeconomic status (Mayer, Berger, and Oberndorfer 2023). Since 2019, the Central Statistics Office of Ireland has reported on the use of publicly available online death notices to measure mortality trends. The findings are published through its Frontier Series, which is designed to showcase experimental methods and data sources. Results indicate that online death notices in Ireland provide high-quality mortality estimates up to 12 months earlier than official data, owing to the Irish tradition of prompt funeral arrangements (Central Statistics Office, Ireland 2023). In Canada, manually searching for obituaries on the internet has proven effective in locating individuals lost to follow-up in clinical trials, helping to reduce research biases caused by attrition (Soowamber et al. 2016). Meanwhile, a study in Germany (Schnell and Redlich 2019) analyzes web-scraped obituaries from a town's online newspapers and finds that only about 18% of deaths were represented. The data reveal demographic biases, specifically an under-representation of men and foreign-born individuals, making the sample unrepresentative of the general population. The study

concludes that online newspaper obituaries are not a reliable substitute for administrative datasets, though the degree of representativeness may vary across cultural contexts.

The purpose of this paper is to assess whether, in the Canadian context, obituaries collected online can serve as a reliable complementary source to the vital statistics system for capturing mortality data. We focus on the provinces of Quebec and Ontario, which together account for roughly 60% of the national population in recent years (Statistics Canada 2018). Our analysis mainly uses obituaries shared online by funeral homes, as previous research has documented systematic biases in newspaper obituaries. Funeral home obituaries are generally more accessible because they are not limited by the costly and restricted space of newspapers, which can introduce selection bias (Mayer, Berger, and Oberndorfer 2023). Because online obituaries have not yet undergone thorough validation as a source for capturing mortality data in Canada, we conduct a comprehensive comparison with official vital statistics to understand the strengths and limitations of this unconventional demographic data source. The country's bilingual context offers a unique opportunity to examine obituaries in both French and English, an area that remains underexplored in demography and thanatology. We present methods for extracting demographic information from obituaries in each language and highlight linguistic and structural differences, illustrating the challenges involved in processing such textual data. Our work is part of a broader effort to understand the value of digital data in demographic analysis by demonstrating how obituaries, traditionally examined in qualitative, anthropological contexts, can serve as a quantitative source of mortality information in Canada.

3. Sources of data

We conducted our analysis using data from two sources: namely obituaries from the Necro Canada archive (2024), and official mortality and population data from the Canadian Human Mortality Database (CHMD) (2024). In 2024, Quebec's Ministry of Health and Social Services registered 224 funeral service providers, while Ontario's Bereavement Authority licensed 546 funeral establishment operators under classes 1 and 2, licenses that typically authorize the management of funerals and the publication of obituaries. Most of these providers or businesses maintain websites that, in addition to providing services and contact information, also publish death notices for upcoming funerals. Building and maintaining individual scrapers for each of these websites would be a monumental task. Instead, specialized sites like Necro Canada collect the majority of death notices published across Canadian provinces, primarily serving as a commemoration platform. We selected Necro Canada for its accessibility, transparency, and broad coverage. During the year 2024, it published an average of 106 daily notices from 125 Quebec funeral homes and 151 notices from 313 Ontario services – representing 55.8% (125/224) and 57.3% (313/546) of all funeral services in each province, respectively. The com-

position of funeral homes varies by province, usually including a mix of independently owned businesses and larger corporate groups that operate multiple locations.

In Quebec, the top 10% of online obituary contributors in 2024 accounted for more than 35% of all provincial obituaries. Manual review showed that all 13 of these top contributors operated in multiple locations across cities or regions. Their obituaries tended to be shorter, averaging 1,108 characters, compared to 1,289 for the remaining 112 funeral homes. In Ontario, the top 10% of obituary producers accounted for 33.4% of all obituaries in 2024. Of the 32 top contributors, only 14 operated in multiple locations. Unlike in Quebec, obituary length in Ontario showed little variation between top and bottom producers, averaging 1,314 and 1,339 characters, respectively. This suggests that obituaries in Ontario are generally longer than those in Quebec, and that obituaries curated by corporate providers tend to be more concise. Given the higher proportion of corporate funeral homes in Quebec and the more diverse mix of privately owned providers in Ontario, these structural differences are reflected in obituary length. The difference in length may also reflect cultural factors, with English-language obituaries potentially placing more emphasis on personalization and narrative detail than their French-language counterparts. Regardless of whether the provider is a corporation or an independently owned business, most obituaries include the essential information required for basic demographic analyses, such as identifiable age at death and gender, as will be demonstrated in Section 4.3.

We retrieved content from the Necro Canada archive for Quebec and Ontario, covering the years 2017 to 2022, as 2017 marks the beginning of the archive and 2022 is the most recent year for which comprehensive mortality data from vital statistics are available for comparison. Details of our web-scraping method for the extraction of online obituaries are provided in Section 4.1, while Section 4.2 explains how we identified our demographic variables from obituary texts.

To evaluate the representativeness of online obituaries for demographic mortality studies in Quebec and Ontario, we compared age- and sex-specific death counts and mortality indicators derived from obituaries with the exhaustive data included in the CHMD (see Section 4.4). From the CHMD, we extracted observed death counts and estimates of population exposure by sex, single years of age, and single calendar years (2017 to 2022) for both provinces under study. The primary source of data for death counts in the CHMD is the Canadian Vital Statistics Death Database administered by Statistics Canada, which gathers information from all provincial and territorial vital statistics registries on all deaths recorded in Canada. It provides the most complete data on deaths in the country. The CHMD's estimates of population exposure are based on Statistics Canada's Demographic Estimates Program, which produces official annual estimates of population by age and sex for Canadian provinces and territories, as well as on death counts. As a satellite project of the Human Mortality Database, all calculations in the CHMD are conducted according to a state-of-the-art methods protocol (Wilmoth et al.

2021) that uses sophisticated demographic methods, in addition to maintaining high standards with regard to data quality.

4. Methods

4.1 Web scraping of online obituaries

Whereas online obituaries in Austria, Germany, and Ireland are often shared as image files (e.g., JPEGs, PNGs, or PDFs) (Mayer, Berger, and Oberndorfer 2023; Schnell and Redlich 2019; Central Statistics Office, Ireland 2023), in Canada they are more commonly published as text on individual web pages created for each deceased person. To gather these data, dedicated tools that process the hypertext markup language (HTML) from websites come into play. This technique, known as web scraping, enables us to gather data of interest from the internet for our research purposes. We built our web scraper using the `Rvest` package for `R` (version 4.2.1), and the corresponding pseudo-code is provided in the replicable material. This allows us to systematically navigate the Necro Canada archives' website, collecting obituary data for Quebec and Ontario and storing it before beginning the cleaning process.

In our cleaning process, we began by removing duplicate entries, then we standardized the text by converting all characters to lowercase, eliminating special characters, and replacing Latin characters with diacritics (e.g., é, ü, ñ) with their closest ASCII equivalents (e.g., e, u, n). Furthermore, prior to text extraction, we classified the obituaries by language using the `R` package `cld2`, an interface to Google's Compact Language Detector 2, relying on probabilistic methods to identify each text's language. This package can detect up to 80 different languages and can recognize instances where multiple languages are used, as is often the case in Canada due to its bilingual nature. This classification step allowed us to develop two distinct procedures: one for processing French obituaries in the two provinces and another for those in English. This approach facilitated the accurate extraction of our two key demographic variables of interest – namely, age at death and gender. After the cleaning process, we obtained a final dataset of 524,913 web-scraped obituaries: 236,290 for Quebec and 288,623 for Ontario, spanning the period from January 1, 2017, to December 31, 2022.

4.2 Extracting the demographic variables gender and age at death from obituary texts

The extraction of demographic variables from obituary texts is a two-step process, involving identifying gender first, followed by determining age at death. Note that for obituaries, we use the term 'gender,' rather than 'sex,' because it reflects how the person

was identified and referred to by their entourage. We later compare this to the CHMD data, which includes only sex as a characteristic. For the remainder of the paper, we will use the term 'gender,' though we acknowledge that we are not covering the entire gender spectrum. As the data evolve in the future to better reflect the reality experienced by Canada's population, one can be more inclusive.

Gender extraction varied by obituary language. English obituaries tend to follow consistent naming conventions, with the given name preceding the surname (e.g., 'John Doe'), allowing us to easily isolate the decedent's first name and submit it to the `gender R` package. The package includes a historical database of names and their associated genders, along with the probability (or historical attribution proportion) of each name being classified as masculine or feminine. To ensure greater accuracy in gender attribution, we retain only probability values greater than 80%. Naming conventions in French are much less standardized, with formats such as 'John Doe,' 'Doe John,' or 'Mr. Doe,' complicating automated gender inference based on names alone.

The remaining English death notices (with probability values $\leq 80\%$), together with all French death notices, are processed using a string-searching algorithm based on regular expressions (Thompson 1968) to extract gender. Regular expressions are sequences of characters that define search patterns, enabling the algorithm to locate specific keywords, phrases, or character combinations within the text. This allows us to infer demographic variables, including gender. The collection of expressions for each language is provided in Appendix Table A-1. In short, for French obituaries, gender is directly attributed when feminine markers, such as 'décédée' (the feminine form of 'deceased') or 'née' (indicating a woman's maiden name), are present, or when masculine markers, such as 'décédé' (masculine form of 'deceased') or honorary titles, such as 'Monsieur (Mr.)' and 'Madame (Mme),' are identified. If this approach fails, we then count other gender-specific words, such as 'il' or 'elle' ('he' or 'she'), and words like 'époux/épouse de' ('spouse of'), among others, and assign gender according to the highest frequency. A similar strategy is used for English obituaries. When the `gender R` package does not clearly indicate the gender, we look for clues like 'he' or 'she' and phrases like 'husband of' or 'wife of.'

For determining age at death, French obituaries often follow a formulaic pattern, most commonly including the phrase 'est décédé à l'âge de XXX ans,' where XXX represents the age at which the individual passed away. English obituaries exhibit more variation, with age at death appearing in sentences such as 'passed at the age of XXX,' 'at age XXX,' 'on their XXX birthday,' or 'at XXX year(s) of age.' Occasionally, these patterns refer to life events rather than death (e.g., 'At the age of 26, he met his beloved wife...'), which may lead to misclassification. To overcome this, we extract all numerical age references using regular expressions and retain only the highest value.

When age is not explicitly stated in either language, we search for birth and death year intervals (e.g., 1930–2022) and calculate the age at death, assuming the deceased had celebrated their birthday before passing. We also exclude years in parentheses (e.g.,

(1928)), which often refer to a spouse's or parent's birth or death year, as these can result in misattribution.

Another distinctive feature of English obituaries is the frequent use of the phrase 'In (his or her) XXXth year,' found in over 50,000 obituary texts. This is ambiguous, as it can either indicate that the person had completed XXX years of life or that the person's age at death was XXX-1 years. In such cases, we cross-validate with other age-at-death data included in the obituary text to determine the correct age.

The final step for determining age at death involves identifying infant deaths, which require a distinct set of instructions due to their unique circumstances. Many obituaries for newborns include terms such as 'Baby' in English, 'Bébé' in French, or 'BB' alongside the name of the deceased. Another way to identify infant deaths is by detecting identical birth and death years (e.g., 2017–2017). But in some cases, adults also appear with identical birth and death years, typically due to clerical errors made during obituary creation. To confirm that an obituary concerns an infant, it must contain specific sets of keywords, such as 'petit ange' ('little angel'), 'fils/fille de' ('son/daughter of'), or 'laisse ses parents' ('survived by his/her parents') in French, and 'his/her mother,' 'son/daughter of,' or 'his/her parents' in English. These keywords are almost always found in infant obituaries and are rarely used in those for older individuals.

4.3 Assessing the accuracy and data retention of regular expressions

To measure the accuracy of our demographic variable extraction using regular expressions, we conducted a validation exercise involving manual checks for each province and gender. Specifically, we randomly selected 100 obituaries from each of the following age groups: 0–17, 18–64, 65–99, and 100+, yielding a total of 1,600 records. Only the name of the deceased and obituary text were retained, and age at death and gender were manually extracted. These manually retrieved values were then compared with those generated from regular expressions.

In Quebec, where most obituaries are written in French, Table 1 shows that our gender extraction procedure performed exceptionally well, correctly identifying gender in 97% to 100% of the cases. The few misclassifications mostly stemmed from human errors in the obituary texts (for example, 'décédé' was occasionally missing an extra 'e' ('décédée') when referring to women). For extracting age at death, our method was nearly flawless as well, with accuracy rates ranging from 98% to 100% across all age groups, except for the 0–17 age group, which showed slightly lower rates of 92% and 94% for young women and young men, respectively. The main challenge in this group came from rare instances in which individuals who had died at an older age were incorrectly listed with identical birth and death years (e.g., 2020–2020), a format typically associated with infant deaths.

In Ontario, where the majority of obituaries are published in English, we had a similar issue with such repeated birth and death years, impacting rates for the 0–17 age group (Table 2). Additionally, some obituaries read like mini-biographies, where phrases such as 'at the age of XXX, he met his wife' could be misinterpreted as indicating age at death. We solved most of these problems by taking the highest age value from our multiple regular expression matches. Nevertheless, a few errors remained, especially in cases where no additional information (e.g., birth year) was available. The gender extraction in Ontario, mainly derived from names, was highly successful, albeit it occasionally struggled with names such as 'Jean,' which is clearly masculine in French but can be ambiguous in English.

Overall, our regular expressions have proven to be highly accurate across both provinces and languages. Errors were primarily due to inconsistencies or omissions in the original obituary texts rather than methodological flaws.

Table 1: Regular expression extraction accuracy, Quebec, 2017–2022

Age group	Gender	Total	Accuracy		
			Age (in %)	Gender (in %)	Age and gender (in %)
0–17	Woman	200	92	97	91
0–17	Man	253	94	100	94
18–64	Woman	11,363	98	100	98
18–64	Man	16,077	98	100	98
65–99	Woman	86,669	100	99	99
65–99	Man	78,311	100	100	100
100+	Woman	2,090	100	100	100
100+	Man	378	100	99	99

Source: Authors' calculations based on data from Necro Canada (2024).

Table 2: Regular expression extraction accuracy, Ontario, 2017–2022

Age group	Gender	Total	Accuracy		
			Age (in %)	Gender (in %)	Age and gender (in %)
0–17	Woman	346	92	100	92
0–17	Man	441	95	100	95
18–64	Woman	15,965	98	99	97
18–64	Man	25,029	100	99	99
65–99	Woman	91,470	99	100	99
65–99	Man	88,107	99	97	96
100+	Woman	2,600	99	100	99
100+	Man	650	97	98	95

Source: Authors' calculations based on data from Necro Canada (2024).

Having established the reliability of our extraction methods, we now turn to data retention. Taking extensive precautions to avoid introducing further bias inevitably results in some data loss. In addition, some obituaries simply lacked enough information to allow for the successful extraction of both age at death and gender. As shown in Table 3, data retention, calculated by dividing the number of obituaries for which both gender and age at death could be successfully inferred by the total number of web-scraped obituaries, varies by province and by obituary language. In Quebec, we collected 236,290 unique obituaries through web scraping, with 92.1% (217,561) written in French. We successfully extracted age at death and gender for 82.7% (195,341) of all the web-scraped obituaries of the province. Most losses occurred during the age-at-death inference step: 90.7% (195,341+19,009) of all observations had their gender extracted, versus 84.2% (195,341+3,679) having age at death. In Ontario, we collected 288,623 unique web-scraped obituaries, 96.7% (278,906) of which were in English. Here, we retained 77.8% of the data. Once again, inferring age at death proved to be the most challenging step.

Data retention varied over time (Table 4). For French obituaries in Quebec, retention rates rose from 78% in 2017 and stabilized around 85% thereafter. In Ontario, English-language obituaries began at 86.2% in 2017 and declined to 76.2% by 2022. Minority-language obituaries consistently had lower extraction rates than majority-language ones: English obituaries in Quebec fell from 60% in 2017 to 51.4% in 2021 before returning to about 60% in 2022; French obituaries in Ontario ranged from 13.9% in 2017 to 50% to 65% in later years.

In Quebec, obituaries published between 2017 and 2022 with complete extraction of all demographic variables had a median text length of 1,130 characters, while those missing age at death or gender had a median of 399 characters. In Ontario, the corresponding medians were 1,288 characters for complete obituaries and 175 characters for incomplete ones, indicating that obituary length is strongly and positively associated with the successful extraction of our key demographic variables. In both provinces, men and women had a similar median obituary length.

Table 3: Extraction of demographic variables from French and English obituaries, by province, 2017–2022

Description	Quebec					Ontario				
	French	(%)	English	(%)	Total	French	(%)	English	(%)	Total
Age ∩ gender	184,334	(85)	11,007	(59)	195,341	4,261	(44)	220,347	(79)	224,608
Women	95,167		5,155		100,322	2,205		108,176		110,381
Men	89,167		5,852		95,019	2,056		112,171		114,227
Gender only	13,374	(6)	5,635	(30)	19,009	553	(6)	46,605	(17)	47,158
Women	7,096		2,868		9,964	304		23,275		23,579
Men	6,278		2,767		9,045	249		23,330		23,579
Age only	2,932	(1)	747	(4)	3,679	83	(<1)	3,585	(1)	3,668
No data	16,921	(8)	1,340	(7)	18,261	4,820	(50)	8,369	(3)	13,189
Web scraped	217,561	(100)	18,729	(100)	236,290	9,717	(100)	278,906	(100)	288,623

Source: Authors' calculations based on data from Necro Canada (2024).

Table 4: Data retention after demographic variable extraction, by language and province, 2017–2022

Year	Quebec				Ontario			
	French		English		French		English	
	Total	Retn (in %)	Total	Retn (in %)	Total	Retn (in %)	Total	Retn (in %)
2017	34,530	78.0	1,484	59.9	2,540	13.9	24,192	86.2
2018	39,348	85.1	2,852	65.3	1,650	51.1	62,514	82.8
2019	34,668	85.2	2,399	57.2	1,287	49.6	43,901	76.1
2020	35,218	86.4	3,001	55.5	1,612	50.2	47,003	77.2
2021	35,248	87.4	3,264	51.3	1,155	64.6	49,708	77.9
2022	38,549	85.9	5,729	61.8	1,473	59.0	51,588	76.2
Total	217,561	84.8	18,729	58.8	9,717	43.8	278,906	79.0

Source: Authors' calculations based on data from Necro Canada (2024).

4.4 Selected mortality indicators

First, we focus on the timing of deaths, assessed by the age distribution of observed death counts $D_{x,t,i,j}$, where age $x = 0, 1, \ldots, \omega$, calendar year $t = 2017, 2018, \ldots, 2022$, gender $i \in \{\text{Women, Men}\}$, and source of data $j \in \{\text{Obituaries, CHMD}\}$ in both provinces. As deaths from online obituaries constitute a subset of those recorded in official vital statistics, we examine age distributions not only in absolute numbers of deaths but also in relative terms to facilitate our comparison of their age schedule between data sources,

such as:

$$\pi_{x,t,i,j} = \frac{D_{x,t,i,j}}{\sum_{x=0}^{\omega} D_{x,t,i,j}}.$$

The latter is perhaps our most important indicator. Indeed, it offers valuable insights into potential disparities between the schedule of deaths observed in the sampled population from online obituaries and that of the broader population of each province.

To summarize the information contained in $\pi_{x,t,i,j}$, we can calculate the mean age at death in a given population, defined as $\bar{x}_{t,i,j} = \sum_{x=0}^{\omega}(x \cdot \pi_{x,t,i,j})/\sum_{x=0}^{\omega} \pi_{x,t,i,j}$, as well as the median age at death, $\tilde{x}_{t,i,j} = y : \min(\sum_{x=0}^{y} \pi_{x,t,i,j}) \geq 0.5$. It should be noted, however, that since $\bar{x}_{t,i,j}$ and $\tilde{x}_{t,i,j}$ are solely based on observed death counts, they depend on the population's age composition, which is likely to vary over space and time. For this reason, such measures are typically avoided in cross-population analysis. Nevertheless, they offer a general view of mortality trends over time when other data are unavailable.

Summary indicators suited for comparing trends and levels of mortality between populations involve the use of both $D_{x,t,i,j}$ and corresponding numbers of person-years lived (i.e., population exposures), denoted by $E_{x,t,i,j}$, for the computation of death rates. While population exposure estimates are available for the total populations of Quebec and Ontario, they are not directly available for the sample populations derived from online obituaries. To address this, we scaled the age-at-death distribution from obituaries as follows:

$$D^*_{x,t,i,\text{Obituaries}} = \pi_{x,t,i,\text{Obituaries}} \cdot \sum_{x=0}^{\omega} D_{x,t,i,\text{CHMD}}.$$

This ensures consistency between the total number of deaths reported in obituaries and those recorded in vital statistics in any given calendar year and gender, since

$$\sum_{x=0}^{\omega} D^*_{x,t,i,\text{Obituaries}} = \sum_{x=0}^{\omega} D_{x,t,i,\text{CHMD}},$$

and thus permits the computation of the following two series of death rates:

$$M_{x,t,i,\text{Obituaries}} = D^*_{x,t,i,\text{Obituaries}}/E_{x,t,i,\text{CHMD}}$$

$$M_{x,t,i,\text{CHMD}} = D_{x,t,i,\text{CHMD}}/E_{x,t,i,\text{CHMD}}.$$

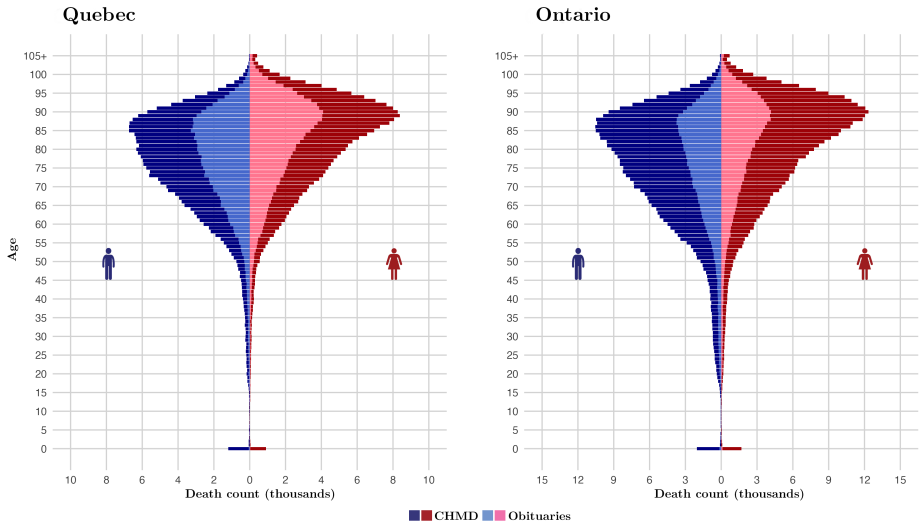
We then convert $M_{x,t,i,\text{Obituaries}}$ and $M_{x,t,i,\text{CHMD}}$ into complete period life tables for each source of data, following the Human Mortality Database methods protocol closely (Wilmoth et al. 2021: Section 7.1), except for the mortality modeling at very old ages. In addition to comparing age-at-death distributions derived from life tables, denoted by $d_{x,t,i,j}$, and survival functions $l_{x,t,i,j}$, we examined differences in life expectancy at birth $e_{0,t,i,j}$ by gender, year, and data source to assess the accuracy of online obituaries. We also complement our life expectancy measures with the modal (i.e., most frequent) age at death among adults, denoted by M. We derived M from a Poisson regression mortality model fitted through a nonparametric smoothing method based on P-splines (Ouellette and Bourbeau 2011; Horiuchi et al. 2013), using the `R` package `MortalitySmooth` (Camarda 2012). The age range over which the modeling took place spanned from 5 to 104 years old. Moreover, the smoothing parameter was determined by minimizing the Bayesian Information Criterion (BIC), which has been shown to be best suited for smoothing mortality data (Currie, Durbán, and Eilers 2004).

Before disclosing the results, it is important to note that, since obituaries contain individual-level data, we have taken deliberate measures to handle this information responsibly, as detailed in the discussion (Section 6). To ensure privacy, only aggregate values are reported throughout the analysis.

5. Results

Figure 1 compares the number of deaths by age and gender in the provinces of Quebec and Ontario during the 2017–2022 period (i.e., all years combined), by source. Lighter colors represent the number of deaths observed through obituaries, while darker colors correspond to vital statistics included in the CHMD as our exhaustive, ground-truth dataset. This color scheme remains consistent across all our figures. Between 2017 and 2022, Quebec registered a total of 424,678 deaths (213,255 women and 211,423 men) according to vital statistics, while our web scraping of online obituaries during this time frame amounted to 236,290, accounting for 55.6% of all recorded deaths. Following the demographic variable derivation, the count of usable obituaries for these years stood at 195,341 (100,322 women and 95,019 men), or 46.0% of the complete death toll. In Ontario, the coverage of all deaths was lower than in Quebec. Vital statistics recorded 676,702 deaths (347,971 men and 328,731 women) during the same period, while 288,623 online obituaries were collected, representing 42.7% of the entire death count. After extracting the gender and age-at-death information, the number of obituaries decreased to 224,608 (110,381 women and 114,227 men), covering 33.2% of total deaths.

Figure 1: Observed death counts by age, gender, and source of data, Quebec and Ontario, 2017–2022



Sources: Authors' calculations based on data from Necro Canada (2024) and the CHMD (2024).
Note: The last age interval is open-ended, and it covers the ages of 105 years and above.

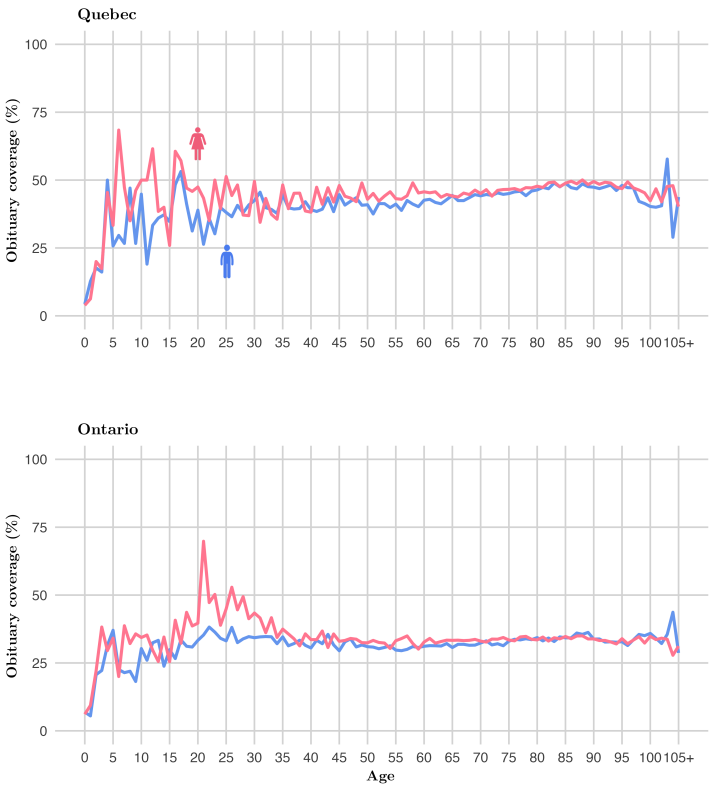
Coverage of deaths, calculated by dividing the number of obituaries by the number of observed deaths, varied across ages, as depicted in Figure 2. In Quebec, during the 2017–2022 period, only 4.2% of all infant deaths were covered in obituaries (3.9% for girls and 4.4% for boys), representing the absolute lowest coverage of all ages throughout the lifespan. Up until early adulthood, the percentages varied erratically from one age to another due to the extreme rarity of deaths at those ages, ranging from 6.3% to 68.4%. After age 25, the coverage stabilized at approximately 38% to 45%, maintaining this range until about age 70. Beyond 70 years old, where nearly three-quarters of all deaths in Quebec occurred between 2017 and 2022, the coverage increased, peaking at around 50% between ages 85 and 90. In the last age interval (105+ years), 40.2% and 43.8% of deceased women and men were represented in obituaries, respectively. Overall, women were slightly better covered in obituaries than men (45.3% vs. 43.1%). For results specific to each calendar year, please refer to Appendix Figures A-1 and A-2.

The coverage of deaths in Ontario closely mirrors trends seen in Quebec. Again, obituaries were least likely to appear for infants, with representation at just 6.9% for boys and 6.1% for girls. Coverage steadily increased through childhood and adolescence, reaching over 30% by age 10. In adulthood, rates stabilized around one-third: 30.5% for

men and 33.6% for women at age 40, with similar levels persisting into advanced age. At age 100, coverage peaked at 35.9% for men and 34.8% for women, before declining slightly in the 105+ age group. Once again, women were slightly better represented in obituaries than men (33.6% vs. 32.8%).

Overall, obituary coverage in both provinces increases with age. Women are generally slightly better represented than men except at the very end of life, where men briefly overtake the obituary coverage of women, ending in the open-ended age interval (105+) with nearly equal representation. Irregular patterns for this age group are due to the very small number of deaths at such advanced ages, particularly for men. A similar trend is observed during adolescence, where deaths are also exceedingly rare, especially among girls.

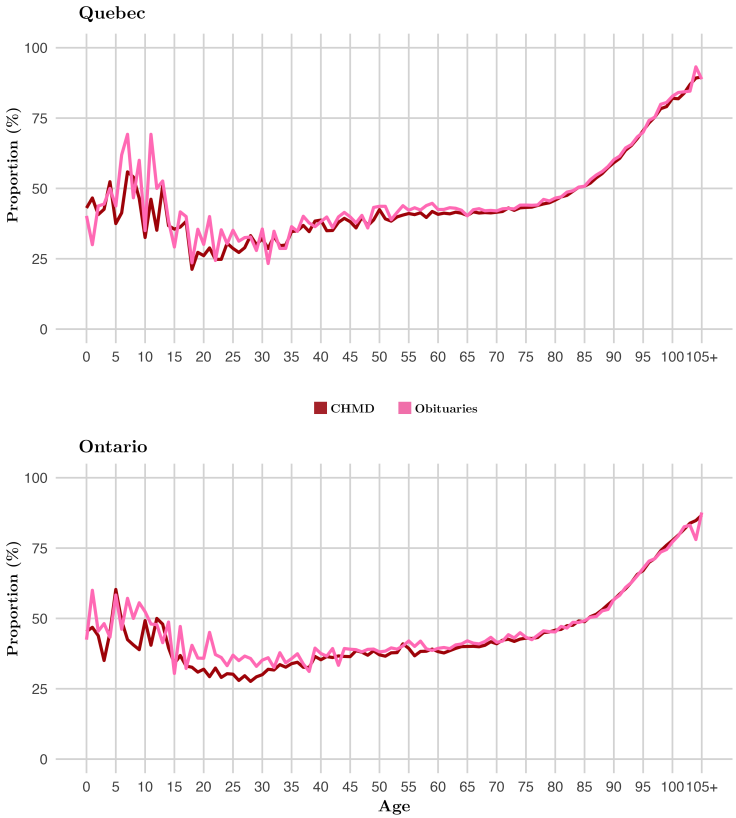
Figure 2: Obituary coverage of vital statistics deaths by age and gender, Quebec and Ontario, 2017–2022



Sources: Authors' calculations based on data from Necro Canada (2024) and the CHMD (2024).

Figure 3 displays the share of women's deaths in Quebec and Ontario, based on vital statistics and obituary records over the 2017–2022 period. Both provinces share very similar trends with slight variations. In summary, until roughly the age of 85, excluding some exceptions in childhood, the percentage of deaths among individuals identified as women is below 50% in both sources, reflecting a preponderance of masculine deaths before reaching old age. At ages 85 and above, deaths become predominantly feminine in both datasets for both provinces.

Figure 3: Proportion of women's deaths in the population by source of data, Quebec and Ontario, 2017–2022



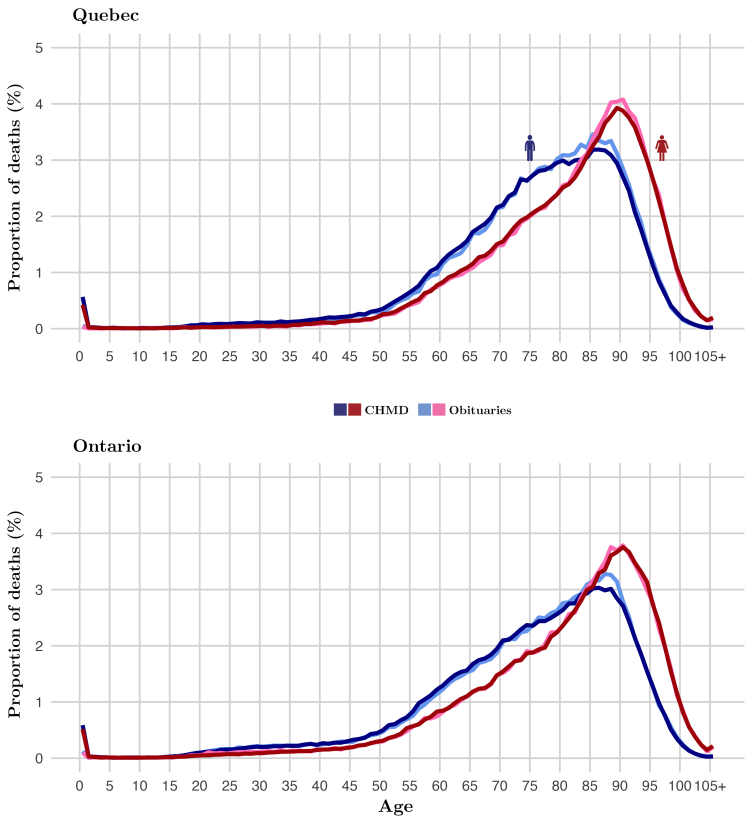
Sources: Authors' calculations based on data from Necro Canada (2024) and the CHMD (2024).

Figures A-3 and A-4 in the Appendix provide results for individual calendar years rather than for the entire 2017–2022 period. They offer additional context for minor year-to-year variations in trends and demonstrate that obituaries remain consistent with the comprehensive CHMD dataset.

Although obituaries represent only a portion (i.e., under half) of all observed deaths in Quebec and Ontario, the previous figures suggest they do capture crucial information regarding the 'death schedule' (i.e., timing of deaths or age at which people die). Figure 4 strengthens this assessment by comparing age distributions of observed deaths across the lifespan by gender and data source, noted $\pi_{x,2017-2022,i,j}$ and totaling 1 across all ages; the figure multiplies these by 100 to show percentages. For both provinces, a significant discrepancy is observed at age 0, with infant mortality being almost nonexistent in online obituaries. Beyond this, the distributions derived from obituaries and vital statistics are remarkably similar, closely aligning at nearly every age. This similarity encompasses various fluctuations that mirror each other closely, even beyond the peak of the old-age death heap (i.e., modal age). Still, the obituaries' curve tends to fall slightly below that of our exhaustive dataset up to around ages 70 to 80 for both men and women, and the trend reverses subsequently, with the most significant discrepancy since age 0 occurring near the modal age-at-death. This is perhaps mostly due to the shortage of infant deaths in obituaries, which ends up being redistributed across other ages and appears to accumulate at the peak of the distribution (see Figures A-5 and A-6 for single calendar years in the Appendix).

Prior to calculating life tables by gender and data source, we computed mortality indicators summarizing the central tendency of observed deaths. Tables 5 and 6 compare the gender- and year-specific mean and median ages at death, denoted $\bar{x}_{t,i,j}$ and $\tilde{x}_{t,i,j}$, between the two sources. In both provinces, obituaries tend to systematically overestimate the mean age at death. The difference in $\bar{x}_{t,i,j}$ between obituaries and the CHMD is greater among men than among women in every year. Regarding the median age at death ($\tilde{x}$), it does not deviate by more than one year from the vital statistics, except for one instance among men in Quebec in 2021, where the deviation reached two years. In Ontario, this metric is very close to the observed values (no difference for women and a difference of one year for men).

Figure 4: Age distribution of observed deaths by gender and source of data, Quebec and Ontario, 2017–2022



Sources: Authors' calculations based on data from Necro Canada (2024) and the CHMD (2024).

Table 5: Observed mean ($\bar{x}$) and median ($\tilde{x}$) ages at death by gender and source of data, Quebec, calendar years 2017 to 2022

Gender	Year	Mean age at death			Median age at death		
		Obituaries	CHMD	$\Delta\bar{x}$	Obituaries	CHMD	$\Delta\tilde{x}$
Women	2017	80.48	79.79	0.69	84	83	1
	2018	80.82	80.16	0.66	84	84	0
	2019	80.93	80.29	0.64	84	84	0
	2020	81.28	80.63	0.65	84	84	0
	2021	81.05	80.29	0.76	84	83	1
	2022	81.31	80.61	0.70	84	83	1
Men	2017	75.31	74.24	1.07	78	77	1
	2018	75.80	74.77	1.03	78	77	1
	2019	76.06	74.82	1.24	78	77	1
	2020	76.57	75.43	1.14	79	78	1
	2021	76.44	75.06	1.38	79	77	2
	2022	76.78	75.88	0.90	79	78	1

Sources: Authors' calculations based on data from Necro Canada (2024) and the CHMD (2024).

Table 6: Observed mean ($\bar{x}$) and median ($\tilde{x}$) ages at death by gender and source of data, Ontario, calendar years 2017 to 2022

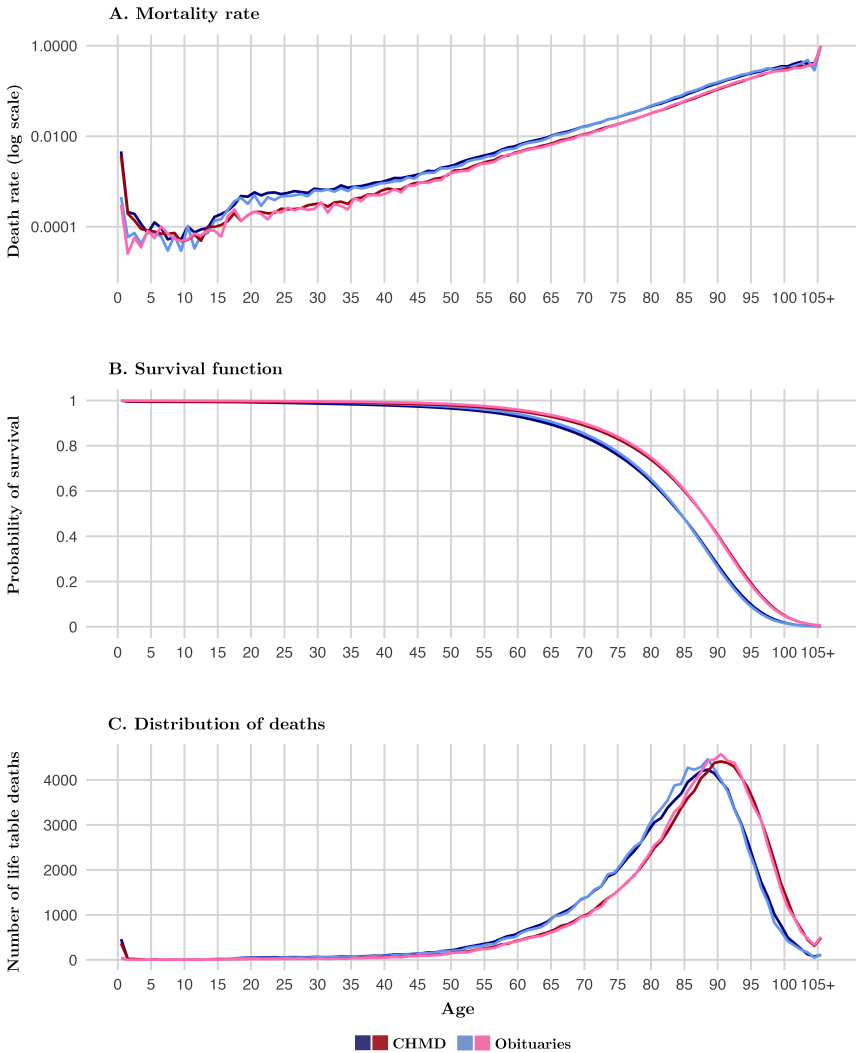
Gender	Year	Mean age at death			Median age at death		
		Obituaries	CHMD	$\Delta\bar{x}$	Obituaries	CHMD	$\Delta\tilde{x}$
Women	2017	79.64	79.26	0.38	83	83	0
	2018	79.59	79.40	0.19	83	83	0
	2019	79.63	79.53	0.10	84	84	0
	2020	79.73	79.54	0.19	83	83	0
	2021	79.29	79.15	0.14	83	83	0
	2022	79.76	79.58	0.18	83	83	0
Men	2017	74.71	73.71	1.00	78	77	1
	2018	74.45	73.83	0.62	78	77	1
	2019	74.58	73.92	0.66	78	77	1
	2020	74.41	73.72	0.69	78	77	1
	2021	74.32	73.33	0.99	78	77	1
	2022	74.85	74.09	0.76	78	77	1

Sources: Authors' calculations based on data from Necro Canada (2024) and the CHMD (2024).

Our second set of mortality indicators involves the construction of period life tables, based on observed death counts and population exposures, thereby providing further insights into the potential biases in obituaries compared to vital statistics. Figures 5 and 6 illustrate age-specific trends by gender and data source for the three main life table variables – namely, death rates ($M_{x,t,i,j}$), survivors ($l_{x,t,i,j}$), and deaths ($d_{x,t,i,j}$) for the 2017–2022 period in each province. As anticipated, death rates in the first panel of the figure show how obituaries underestimate mortality substantially at the beginning of life for both genders, especially at age 0, compared to vital statistics. This gap narrows quickly, and by around age 5, the curves for the two data sources align nearly perfectly. The second panel translates death rates into proportions of remaining survivors at each age, again showing well-aligned curves. Thanks to the remarkable stability of the survival curves – owing to their cumulative nature – slightly higher shares of survivors in obituaries are clearly noticeable up to ages 85 to 90, after which the trend reverses. Finally, the third panel, illustrating the life table age-at-death distribution, confirms earlier findings about the slightly differing death schedules between the two sources of data.

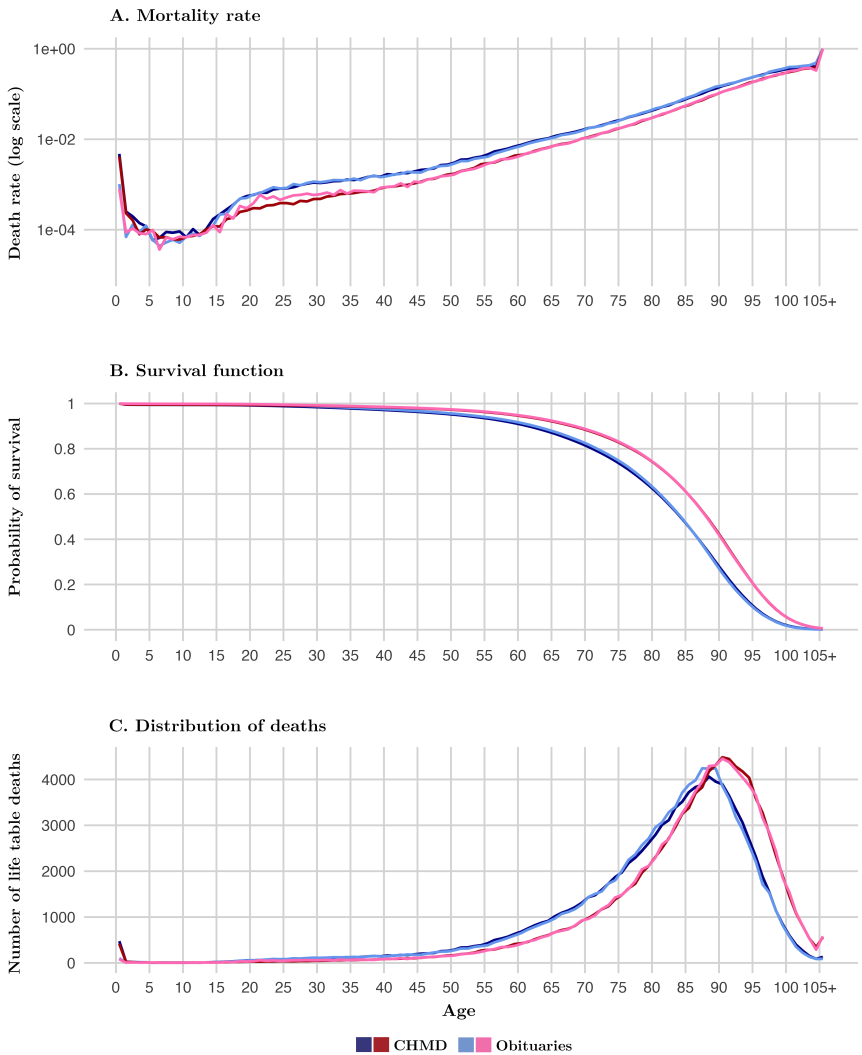
To summarize the detailed information contained in Figures 5 and 6, Tables 7 and 8 compare life expectancy at birth as well as modal age at death estimates by data source among women and men for years 2017 to 2022 in the two provinces. In Quebec, life expectancy estimates based on obituaries are about 0.4 years higher for women and 0.5 years higher for men in comparison to those based on vital statistics. In Ontario, these differences are smaller: a difference of 0.2 for women and 0.4 for men. This discrepancy in both provinces largely stems from the almost complete absence of infant mortality deaths in obituaries. In contrast, the modal age at death derived from obituaries generally falls below that observed in the broader population. This aligns with the underestimated survival functions at older ages seen in Figures 5 and 6 for obituary data.

Figure 5: Comparison of the three main life table functions by gender and source of data, Quebec, 2017–2022



Sources: Authors' calculations based on data from Necro Canada (2024) and the CHMD (2024).

Figure 6: Comparison of the three main life table functions by gender and source of data, Ontario, 2017–2022



Sources: Authors' calculations based on data from Necro Canada (2024) and the CHMD (2024).

Table 7: Life expectancy at birth (e_0) and modal age at death (M) by gender and source of data, Quebec, calendar years 2017 to 2022

Gender	Year	Life expectancy at birth			Modal age at death		
		Obituaries	CHMD	Δe_0	Obituaries	CHMD	ΔM
Women	2017	84.65	84.29	0.36	90.91	91.22	–0.31
	2018	84.57	84.18	0.39	90.08	90.63	–0.55
	2019	85.08	84.73	0.35	90.65	91.07	–0.42
	2020	84.41	84.00	0.41	90.10	90.44	–0.34
	2021	85.36	84.96	0.40	90.81	91.20	–0.39
	2022	84.58	84.14	0.44	90.27	90.55	–0.28
Men	2017	81.07	80.54	0.53	87.58	87.97	–0.39
	2018	81.09	80.61	0.48	87.53	88.19	–0.66
	2019	81.51	80.95	0.56	88.09	88.96	–0.87
	2020	81.01	80.45	0.56	87.94	88.11	–0.17
	2021	81.70	81.07	0.63	88.10	88.52	–0.42
	2022	81.04	80.55	0.49	87.29	88.15	–0.86

Sources: Authors' calculations based on data from Necro Canada (2024) and the CHMD (2024).

Table 8: Life expectancy at birth (e_0) and modal age at death (M) by gender and source of data, Ontario, calendar years 2017 to 2022

Gender	Year	Life expectancy at birth			Modal age at death		
		Obituaries	CHMD	Δe_0	Obituaries	CHMD	ΔM
Women	2017	84.62	84.44	0.18	92.86	91.30	1.56
	2018	84.56	84.37	0.19	90.93	91.22	–0.29
	2019	84.88	84.74	0.14	90.83	91.51	–0.68
	2020	84.51	84.34	0.17	90.76	91.28	–0.52
	2021	84.56	84.38	0.18	90.23	91.37	–1.14
	2022	84.27	84.08	0.19	90.27	91.06	–0.79
Men	2017	80.58	80.20	0.38	88.18	88.37	–0.19
	2018	80.50	80.20	0.30	87.97	88.48	–0.51
	2019	80.73	80.40	0.33	88.24	88.64	–0.40
	2020	80.08	79.73	0.35	88.65	88.69	–0.04
	2021	80.06	79.61	0.45	88.55	88.71	–0.16
	2022	79.99	79.58	0.41	88.56	88.40	0.16

Sources: Authors' calculations based on data from Necro Canada (2024) and the CHMD (2024).

6. Discussion

Memorial services, funerals, obituaries, and death notices constitute integral rituals to commemorate the deceased, especially in Western societies with significant Catholic and/or Protestant historical influences (Nicolet 2018). In this study, we use Canadian obituaries shared online, mainly by funeral homes, to contribute to the growing literature on digital data sources in mortality research, with a focus on their inherent biases. By automating the collection of over 500,000 online obituaries and demonstrating how to use regular expressions to extract demographic variables (gender and age at death) from the text, we compared this data source – originally not intended for demographic analysis – with official mortality records to assess how accurately it reflects the mortality profile of Canada's two most populous provinces. Additionally, we offer unique insights into the differences between obituaries written in French and English within the same country. We demonstrate that analyzing obituary texts offers valuable information for population-level mortality data. In the past, obituaries were considered unsuitable for research due to their difficult collection and glaring selection bias. Women, ethnic minorities, and economically disadvantaged groups were consistently underrepresented in obituaries. This discrepancy persists with newspaper-based obituaries (Schnell and Redlich 2019), possibly due to constrained space and publishing costs. In the Canadian context, obituaries are primarily a byproduct of funeral arrangements. Given their sheer volume and accessibility online, they form an emerging dataset with the potential to complement official mortality statistics.

Our analysis of online obituaries from Quebec and Ontario between 2017 and 2022 reveals a notable trend: Although our obituary dataset with identified gender and age-at-death variables covers just under half of deaths in Quebec and about one-third in Ontario, the age-at-death distribution closely mirrors that of the general population, except for a substantial underrepresentation of infant mortality. A similar bias has been observed in historical crowd-sourced online genealogies (Chong et al. 2022). A recent study in the United States found notable differences in the form, content, and linguistic devices used in obituaries for children compared to those for other ages (Carmon 2023), indirectly suggesting that processing obituaries for the youngest decedents separately could lead to better coverage. Evidence from our own analysis supports this approach, showing that certain linguistic markers, such as the inclusion of 'Baby' in the name or keywords like 'little angel,' 'premature,' or 'will be missed by parents,' are specific to obituaries for infants and young children. However, the primary reason for the underrepresentation of infant deaths in obituary data appears to be that in Quebec and Ontario, as in other low-mortality regions, such deaths are predominantly concentrated in the earliest stages of infancy. Besides the tragic experience of child loss, obituaries are systematically linked to funerals, which can involve significant expenses and might not be a priority for grieving parents. We also observed that young adults who are victims of accidental deaths

frequently have obituaries with distinct formats compared to those who pass away at an older age or from 'natural causes.' For instance, phrases like 'sudden passing,' 'sudden loss,' 'left too soon,' and 'short time on earth' are commonly used as euphemisms to indicate an untimely death. These obituaries tend to be more personal and sometimes include appeals for donations to specific foundations or services, which can provide clues about the cause of death (e.g., suicide prevention helplines, substance abuse recovery programs, or causes related to specific diseases).

To analyze obituaries and extract data on gender and age at death, we created two sets of regular expressions: one for English obituaries and one for French. During this process, we noticed key differences between obituaries in these two languages. French obituaries posted by funeral homes tend to follow a strict format. They usually include the dates of birth and death, details about the funeral, and a list of surviving family members. English obituaries, on the other hand, are more personal and biographical. While they provide the same basic information, they sometimes include highlights from the person's life, such as when they got married, started a career, or contributed to the community. These narrative elements can introduce ambiguity for regular expression-based extraction, so we carefully designed our regular expressions (see Table A-1) to account for the different ways demographic data appear in each language. Furthermore, English obituaries typically follow consistent naming conventions, with the given name appearing before the surname, which facilitates automated gender attribution using name-based databases. In French, naming practices are more variable, making it more difficult to reliably infer gender from the first name. Instead, we have found that for gender-inflected languages like French (in which nouns, pronouns, adjectives, and sometimes verbs are inflected to reflect gender), using regular expressions to deduce gender from text is sufficiently accurate. We anticipate comparable accuracy for languages such as Italian, Spanish, and German, though further research is needed to confirm this. Future studies should likewise assess the accuracy and retention rates of these methods across a broader range of linguistic typologies, including genderless and natural gender languages (Stahlberg et al. 2007).

For the 2017–2022 period, where comprehensive mortality and exposure data were accessible from the CHMD, we derived and compared age-specific death rates, probability of survival, life table deaths, and life expectancy at birth and modal age at death estimates from obituaries and vital statistics. Our findings highlight a slight overestimation in life expectancy at birth as reported in obituaries, by about one-third of a year for women and half a year for men. In contrast, the modal age at death is underestimated in obituaries, with greater year-to-year variation.

Another bias lies in understanding by whom obituaries are written. De Vries and Rutherford (2004) examine a sample of online memorials and find they are typically authored by the children (33%), friends (15%), grandchildren (11%), parents (10%), siblings (8%), and spouses (4%) of the deceased. This finding underlines the importance of

social networks in remembrance, both in real and digital realms. For this reason, obituaries might overlook individuals without social support at death, potentially leading to a selection bias where the least healthy individuals are excluded, as loneliness could play an important role in health disparities (Raymo and Wang 2022). In our online obituaries dataset for Quebec and Ontario, we also observed a slight underrepresentation of men at almost every age. It has been documented that men tend to have smaller social networks in later life compared to women (McLaughlin et al. 2010), thus potentially having fewer people to ensure their remembrance. Conversely, it could be due to women dying at older ages than men on average, when the likelihood of obituary publication is higher. Earlier literature reported that men were more likely to be covered in obituaries and that their obituaries tended to be longer (Kastenbaum, Peyton, and Kastenbaum 1977). Our data show that obituary lengths for men and women in both provinces were very similar. However, this does not preclude potential differences within specific subgroups, such as those defined by age or marital status, which deserve further investigation.

While it is reasonable to question whether obituaries might become obsolete in the context of increasing immigration, especially since obituaries are less common outside of Europe, the Americas, and former European colonies, and ongoing secularization, this outcome appears unlikely in the near future. The majority of older adults continue to identify with religious or culturally rooted funeral traditions. In Quebec, 90.0% of those aged 75 and older identify as Christian, compared to 75.4% in Ontario. More broadly, 64.8% of Quebecois and 52.1% of Ontarians identify as Christian, while 27.9% and 31.6%, respectively, identify as atheist (Statistics Canada 2023). Combined with our earlier findings, this suggests that even in age groups where religious affiliation is lower, individuals continue to be well represented in obituaries. Institutional support likely also helps sustain these practices. For example, Quebec provides a $2,500 death benefit through its public pension system to help cover funeral costs. As the number of annual deaths increases with population aging, obituaries are likely to remain a prevalent and culturally meaningful form of commemoration. Even in years with notable undercount, such as in Ontario in 2017 when only 25,000 obituaries were collected compared to 50,000 to 60,000 in subsequent years, obituary-based deaths still yielded reliable estimates of the total population's age-specific mortality profile (see Figure A-6).

This limitation is consistent with our decision to rely on a single source for obituary data, namely the Necro Canada archive (2024). As previously noted, this is one of the most popular and transparent obituary aggregators in Canada. We undertook a comparative effort to collect obituaries from additional archival websites, such as Echovita (2025), Le Nécrologue (2025), and the Corporation des thanatologues du Québec (2025), in an attempt to construct a more exhaustive dataset. However, this approach yielded limited additional value, since the vast majority of obituaries are duplicated across platforms. In addition, the substantial effort required to detect and eliminate these duplicates, a process that can easily introduce bias if not done carefully, outweighs the marginal benefits

of using more than one sufficiently representative source. Ideally, a centralized, continuously updated database covering the daily outputs of all funeral homes across Canada would be available. In the absence of such infrastructure, our approach represents an open, accessible, and pragmatic alternative.

For optimal results, Canadian mortality studies that do not focus on infant mortality or on marginalized populations are best positioned to benefit from obituaries. Obituaries are most valuable when used as a complementary data source to traditional demographic data, or at least when the main biases inherent in obituary data are acknowledged and addressed. Already, as demographers increasingly use internet-derived datasets, identifying and correcting for biases has become an essential preliminary step in the analysis process. A growing body of research is dedicated to developing methods that correct these inaccuracies. It is commonly advised that such datasets be paired with ground-truth data, allowing for the calibration of observed discrepancies (Zagheni and Weber 2015; Chong et al. 2022). When no other data are available, a difference-in-differences method has proven valuable in studying relative trends (Zagheni and Weber 2015).

We believe that online obituaries can serve as a valuable resource for a wide range of demographic research. In fields such as kinship, social networks, and genealogy, obituaries have the potential to provide complementary data that may not be available through other sources. Based on our observations, 78.2% of Quebec obituaries and 75.1% of Ontario obituaries enumerate immediate family members, such as partners, parents, siblings, children, grandchildren, great-grandchildren, cousins, and close friends, usually listed in the 'left in bereavement' part of the text at the end of the obituary. Additionally, obituaries explicitly indicate whether the deceased was predeceased by a loved one, which may make them useful for research on kin loss, such as cases in which children pass away before their parents (Alburez-Gutierrez, Basellini, and Zagheni 2025).

While online obituaries cannot fully replace comprehensive, ground-truth datasets in most settings, we argue that their most significant benefit is undoubtedly their timeliness, particularly during crises (e.g., heat waves, emerging infectious diseases, epidemics, natural disasters). This advantage was clearly demonstrated when obituaries were used to 'nowcast' COVID-19 deaths in the Italian provinces of Bergamo and Brescia (Buonanno and Puca 2021). Their near-instant availability can be helpful for policymakers and health officials in making better-informed decisions and enacting interventions promptly. In the Canadian context, mortality data from official statistical and public health agencies, such as the Institut de la statistique du Québec (ISQ), Institut national de santé publique du Québec (INSPQ), and Statistics Canada (SC), remain the gold standard for mortality and epidemiological data. Unfortunately, their dissemination schedules and data aggregation rules present challenges for real-time surveillance. Aside from INSPQ's cause-specific COVID-19 dashboard, which is released seven days after the reference date and with limited granularity (10-year age groups or gender but not both), none of the agencies listed above publishes public daily death counts for all-cause mortality. The best alternative is

weekly aggregated mortality data, typically released with a delay of approximately one to four months after the date of death. Moreover, public datasets are regularly aggregated into broad age intervals (ISQ: 0–49, 50–59, . . ., 80–89, 90+ years; SC: 0–44, 45–64, 65–84, 85+ years) to protect confidentiality and ensure data quality. This limits the data's usefulness for detecting rapid changes in age-specific mortality during crises. For more detailed age breakdowns (e.g., five-year age groups with an open age group at 100+), the wait can exceed a year or two before such information becomes publicly available. Obituaries, by contrast, offer a markedly faster alternative with the finest granularity possible. They consist of individual-level records that can be aggregated by age, gender, and time in flexible formats, including weekly or monthly intervals, and often include additional contextual information, such as marital status and kinship ties. In Quebec in 2022, the average time between the date of death and the obituary posting was 4.5 days; in Ontario, it was even shorter at 2.9 days. In some countries, such as Ireland, this interval is as short as 1.3 days (Central Statistics Office, Ireland 2023). Such rapid availability meets the need for high-frequency mortality data, which is essential not only for effective surveillance but also for relating specific events to periodic measures like economic indicators (Karlinsky 2024).

In closing, research involving internet-based data, such as online obituaries, must pay careful attention to ethical considerations, particularly those regarding privacy and data management. In recent years, population specialists have been harnessing the power of digital traces, finding use for them in a multitude of studies that were simply not feasible with traditional demographic data sources (e.g., population censuses, surveys, administrative datasets). As internet information becomes increasingly pervasive, the responsible use of these data is paramount. In this project, we adhered to the terms and agreements of the Necro Canada website and other ethical practices while web scraping, for example, by pacing the number of requests on the server. After the collection of obituaries, we extracted only necessary information and discarded any details that could potentially identify individuals. The resulting database is stored on a password-protected local hard drive, and is never to be shared on a cloud. By respecting these protective measures, this research project gained approval from our institution's Research Ethics Committee for Society and Culture. Through these measures, we are working toward a future where web data, beyond what is shared through application programming interfaces (APIs), can provide unprecedented insights into population-based research.

7. Acknowledgments

This project was funded by the Social Sciences and Humanities Research Council of Canada (grants 435-2023-0951 and 430-2018-01107, and graduate scholarships), as well as the Fonds de recherche du Québec – Société et culture (grant 2019-NP-254254, graduate scholarship). The authors are grateful to the participants of the 2023 and 2024 Annual Meetings of the Population Association of America, and to the Université de Montréal Longevity and Mortality team, especially Alexis Nadeau and Golden Welsh, for their valuable comments and suggestions.

References

Alburez-Gutierrez, D., Basellini, U., and Zagheni, E. (2025). When do mothers bury a child? Heterogeneity in maternal age at offspring loss. *Population Studies* 79(1): 45–57. doi:10.1080/00324728.2024.2345075.

Alexander, M., Polimis, K., and Zagheni, E. (2019). The impact of Hurricane Maria on out-migration from Puerto Rico: Evidence from Facebook data. *Population and Development Review* 45(3): 617–630. doi:10.1111/padr.12289.

Allison, T. and St Leger, S. (1999). The life span of Methodist ministers: An example of the use of obituaries in epidemiology. *Journal of Epidemiology and Community Health* 53(4): 253–254. doi:10.1136/jech.53.4.253.

Barry, E. (2008). From epitaph to obituary: Death and celebrity in eighteenth-century British culture. *International Journal of Cultural Studies* 11(3): 259–275. doi:10.1177/1367877908092584.

Bassett, D.J. (2015). Who wants to live forever? Living, dying and grieving in our digital society. *Social Sciences* 4(4): 1127–1139. doi:10.3390/socsci4041127.

Beaunoyer, E. and Guitton, M.J. (2021). Cyberthanathology: Death and beyond in the digital age. *Computers in Human Behavior* 122: 106849. doi:10.1016/j.chb.2021.106849.

Beaunoyer, E., Hiracheta Torres, L., Maessen, L., and Guitton, M.J. (2020). Grieving in the digital era: Mapping online support for grief and bereavement. *Patient Education and Counseling* 103(12): 2515–2524. doi:10.1016/j.pec.2020.06.013.

Buonanno, P. and Puca, M. (2021). Using newspaper obituaries to "nowcast" daily mortality: Evidence from the Italian COVID-19 hot-spots. *Health Policy* 125(4): 535–540. doi:10.1016/j.healthpol.2021.01.006.

Camarda, C.G. (2012). MortalitySmooth: An R package for smoothing poisson counts with P-splines. *Journal of Statistical Software* 50(1): 1–24. doi:10.18637/jss.v050.i01.

Carmon, A.F. (2023). So much life in such a short time: An exploration of children's obituaries. *OMEGA – Journal of Death and Dying* 87(4): 1323–1340. doi:10.1177/00302228211038800.

Central Statistics Office, Ireland (2023). Measuring mortality using public data sources 2019–2023 (October 2019 – June 2023). Cork: Central Statistics Office, Ireland. https://www.cso.ie/en/releasesandpublications/fp/fp-mpds/measuringmortalityusingpublicdatasources2019-2023october2019-june2023/.

Cesare, N., Lee, H., McCormick, T., Spiro, E., and Zagheni, E. (2018). Promises and pitfalls of using digital traces for demographic research. *Demography* 55(5): 1979–1999. doi:10.1007/s13524-018-0715-2.

CHMD (2024). Canadian Human Mortality Database. Université de Montréal (Canada), University of California, Berkeley (United States) and Max Planck Institute for Demographic Research (Germany). www.bdlc.umontreal.ca.

Chong, M., Alburez-Gutierrez, D., Del Fava, E., Alexander, M., and Zagheni, E. (2022). Identifying and correcting bias in big crowd-sourced online genealogies. Rostock: Max Planck Institute for Demographic Research (MPIDR working paper WP-2022-005). doi:10.4054/MPIDR-WP-2022-005.

Clissa, L. (2022). Survey of Big Data sizes in 2021. Ithaca, NY: Cornell University (arXiv 2202.07659). doi:10.48550/arXiv.2202.07659.

Corporation des thanatologues du Québec (2025). Corporation des thanatologues du Québec archives. Québec: Corporation des thanatologues du Québec. https://www.domainefuneraire.com (data downloaded in March 2025).

Currie, I., Durbán, M., and Eilers, P. (2004). Smoothing and forecasting mortality rates. *Statistical Modelling* 4(4): 279–298. doi:10.1191/1471082X04st080oa.

de Vries, B. and Rutherford, J. (2004). Memorializing loved ones on the World Wide Web. *OMEGA – Journal of Death and Dying* 49(1): 5–26. doi:10.2190/DR46-RU57-UY6P-NEWM.

Echovita (2025). Echovita archives. Quebec: Echovita. https://echovita.com (data downloaded in March 2025).

Eid, M. (2002). *The world of obituaries: Gender across cultures and over time*. Detroit: Wayne State University Press. doi:10.1017/S0020743804301078.

Flores, R.D. (2017). Do anti-immigrant laws shape public sentiment? A study of Arizona's SB 1070 using Twitter data. *American Journal of Sociology* 123(2): 333–384. doi:10.1086/692983.

Fowler, B. and Bielsa, E. (2007). The lives we choose to remember: A quantitative analysis of newspaper obituaries. *The Sociological Review* 55(2): 203–226. doi:10.1111/j.1467-954X.2007.00702.x.

Ginsberg, J., Mohebbi, M.H., Patel, R.S., Brammer, L., Smolinski, M.S., and Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature* 457(7232): 1012–1014. doi:10.1038/nature07634.

Horiuchi, S., Ouellette, N., Cheung, S.L.K., and Robine, J.M. (2013). Modal age at death: Lifespan indicator in the era of longevity extension. *Vienna Yearbook of Population Research* 11: 37–69. doi:10.1553/populationyearbook2013s37.

Hume, J. and Bressers, B. (2010). Obituaries online: New connections with the living – and the dead. *OMEGA – Journal of Death and Dying* 60(3): 255–271. doi:10.2190/OM.60.3.d.

Karlinsky, A. (2024). International completeness of death registration. *Demographic Research* 50(38): 1151–1170. doi:10.4054/DemRes.2024.50.38.

Kastenbaum, R., Peyton, S., and Kastenbaum, B. (1977). Sex discrimination after death. *OMEGA – Journal of Death and Dying* 7(4): 351–359. doi:10.2190/E3WV-EGHE-P785-UPTW.

Kearl, M.C. (1987). Death as a measure of life: A research note on the Kastenbaum-Spilka strategy of obituary analyses. *OMEGA - Journal of Death and Dying* 17(1): 65–78. doi:10.2190/R9D1-QJQ9-NTVJ-54T6.

Kim, J., Sîrbu, A., Giannotti, F., and Gabrielli, L. (2020). Digital footprints of international migration on Twitter. In: Berthold, M.R., Feelders, A., and Krempl, G. (eds.). *Advances in intelligent data analysis XVIII*. Cham: Springer International Publishing: 274–286. doi:10.1007/978-3-030-44584-3_22.

Le Nécrologue (2025). Le Nécrologue archives. Quebec: lenecrologue.com. https://lenecrologue.com (data downloaded in March 2025).

Lingel, J. (2013). The digital remains: Social media and practices of online grief. *The Information Society* 29(3): 190–195. doi:10.1080/01972243.2013.777311.

Marks, A. and Piggee, T. (1999). Obituary analysis and describing a life lived: The impact of race, gender, age, and economic status. *OMEGA – Journal of Death and Dying* 38(1): 37–57. doi:10.2190/K8V4-693D-EF8L-VBVA.

Mayer, S., Berger, M., and Oberndorfer, M. (2023). Equal in ashes? Exploring socioeconomic inequalities in lifespan based on obituary data in Austria. *SSM – Population Health* 24: 101550. doi:10.1016/j.ssmph.2023.101550.

McLaughlin, D., Vagenas, D., Pachana, N.A., Begum, N., and Dobson, A. (2010). Gender differences in social network size and satisfaction in adults in their 70s. *Journal of Health Psychology* 15(5): 671–679. doi:10.1177/1359105310368177.

McNeill, L. (2004). Generic subjects: Reading Canadian death notices as life writing. *Life Writing* 1(2): 151–166. doi:10.1080/10408340308518264.

Necro Canada (2024). Necro Canada archives. Québec: necrocanada.com. https://necrocanada.com (data downloaded in February 2024).

Nicolet, M. (2018). Quand l'entourage annonce la mort d'un proche: Les avis de décès comme révélateurs de nos représentations de la mort dans la vieillesse [PhD thesis]. Geneva: University of Geneva. doi:10.13097/archive-ouverte/unige:110931.

Öhman, C.J. and Watson, D. (2019). Are the dead taking over Facebook? A Big Data approach to the future of death online. *Big Data & Society* 6(1). doi:10.1177/2053951719842540.

Ouellette, N. and Bourbeau, R. (2011). Changes in the age-at-death distribution in four low mortality countries: A nonparametric approach. *Demographic Research* 25(19): 595–628. doi:10.4054/DemRes.2011.25.19.

Pennington, N. (2013). You don't de-friend the dead: An analysis of grief communication by college students through Facebook profiles. *Death Studies* 37(7): 617–635. doi:10.1080/07481187.2012.673536.

Phillips, C.A., Barz Leahy, A., Li, Y., Schapira, M.M., Bailey, L.C., and Merchant, R.M. (2018). Relationship between state-level Google online search volume and cancer incidence in the United States: Retrospective study. *Journal of Medical Internet Research* 20(1): e6. doi:10.2196/jmir.8870.

Radtke, H.L., Hunter, M., and Stam, H.J. (2000). In memoriam as in life: Gender and psychology in the obituaries of eminent psychologists. *Canadian Psychology / Psychologie canadienne* 41(4): 213–229. doi:10.1037/h0086870.

Raymo, J.M. and Wang, J. (2022). Loneliness at older ages in the United States: Lonely life expectancy and the role of loneliness in health disparities. *Demography* 59(3): 921–947. doi:10.1215/00703370-9937606.

Schnell, R. and Redlich, S. (2019). Web scraping online newspaper death notices for the estimation of the local number of deaths. In: Moucek, R., Fred, A., and Gamboa, H. (eds.). *Proceedings of the 12th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2019)*. Prague: SciTePress: 319–325. doi:10.5220/0007382603190325.

Soowamber, M.L., Granton, J.T., Bavaghar-Zaeimi, F., and Johnson, S.R. (2016). Online obituaries are a reliable and valid source of mortality data. *Journal of Clinical Epidemiology* 79: 167–168. doi:10.1016/j.jclinepi.2016.05.012.

Stahlberg, D., Braun, F., Irmen, L., and Sczesny, S. (2007). Representation of the sexes in language. In: Fiedler, K. (ed.). *Social Communication*. New York: Psychology Press: 163–187. doi:10.4324/9780203837702.

Statistics Canada (2018). Table 17-10-0005-01 Population estimates on July 1st, by age and sex. Ottawa: Statistics Canada. doi:10.25318/1710000501-eng.

Statistics Canada (2023). Table 98-10-0353-01 Religion by gender and age: Canada, provinces and territories. Ottawa: Statistics Canada. doi:10.25318/9810035301-eng.

Thompson, K. (1968). Programming techniques: Regular expression search algorithm. *Communications of the ACM* 11(6): 419–422. doi:10.1145/363347.363387.

Walter, T., Hourizi, R., Moncur, W., and Pitsillides, S. (2011). Does the internet change how we die and mourn? Overview and analysis. *OMEGA – Journal of Death and Dying* 64(4): 275–302. doi:10.2190/OM.64.4.a.

Wilde, J., Chen, W., and Lohmann, S. (2020). COVID-19 and the future of US fertility: What can we learn from Google? Bonn: IZA – Institute of Labor Economics (Discussion paper 13776).

Wildeman, J., Schrijner, S., and Smits, J. (2023). Fertility rates and social media usage in sub-Saharan Africa. *Population, Space and Place* 29(4): e2635. doi:10.1002/psp.2635.

Wilmoth, J., Andreev, K., Jdanov, D., Glei, D., and Riffe, T. (2021). Methods protocol for the Human Mortality Database, version 6. Berkeley: University of California. https://www.mortality.org/File/GetDocument/Public/Docs/MethodsProtocolV6.pdf.

Zagheni, E. and Weber, I. (2015). Demographic research with non-representative internet data. *International Journal of Manpower* 36(1): 13–25. doi:10.1108/IJM-12-2014-0261.

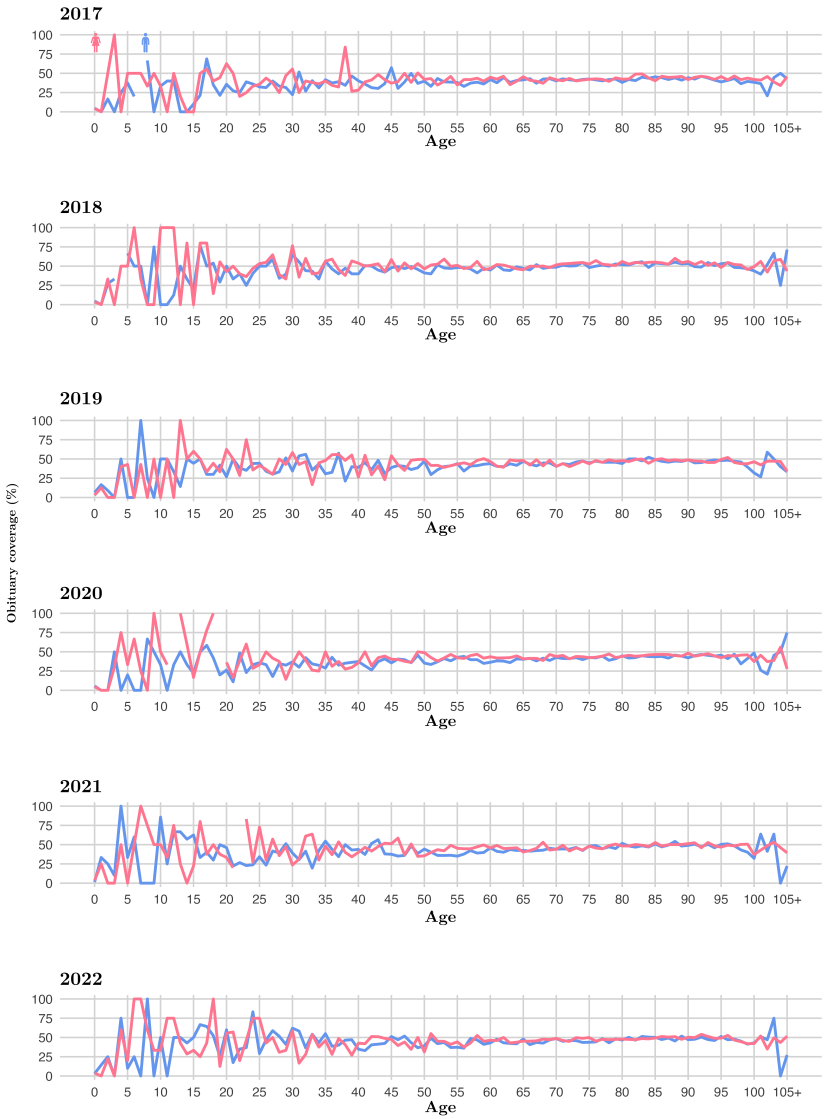
Appendix

Table A-1: Regular expressions used to extract demographic variables 'gender' and 'age at death' from obituary texts, by language

Variable	French	English
Gender – women	est decedee? (mme|madame) | est decedee | le deces de madame | elle laisse dans le deuil | epouse de | elle etait | fille de feu | deces de mme | \bnee\b	she is | wife (of|to) | daughter (of|to) | her children | mother (of|to) | she was | sister (of|to) | by her | mom (of|to) | her daughter | her son | at her | miss her | she | her | grandmother (of|to)
Gender – men	est decede (mr.|monsieur) | est decede | le deces de monsieur | il laisse dans le deuil | epoux de | il etait | fils de feu | deces de mr.	he is | husband (of|to) | son (of|to) | his children | father (of|to) | he was | brother (of|to) | by his | dad (of|to) | his daughter | his son | at his | miss him | he | him | grandfather (of|to) | his family
Age at death	agee?(de)? \d{1,3} ans?	at(the)? age(of)? \d{1,3} | at \d{1,3} years? of age
	\b(19[0-9][0-9]|200[0-9]|201[0-7]|2022)\s*-?\s*(19[0-9][0-9]|200[0-9]|201[0-7]|2022)\b	\b(19[0-9][0-9]|200[0-9]|201[0-7]|2022)\s*-?\s*(19[0-9][0-9]|200[0-9]|201[0-7]|2022)\b
	\bbebe\b | \bbb\b | petit ange | fils de | fille de | ses parents | ses grands-parents | enfant de	\bbaby\b | his mother | her mother | son of | daughter of | grandson of | granddaughter of | baby | infant | premature | their son | their daughter | by parents | his \w+ parents | her \w+ parents

Note: The regular expressions in this table do not include diacritics nor special characters. These are removed in the cleaning process described in 4.1 to facilitate extraction.

Figure A-1: Obituary coverage of vital statistics deaths by age and gender, Quebec, 2017–2022

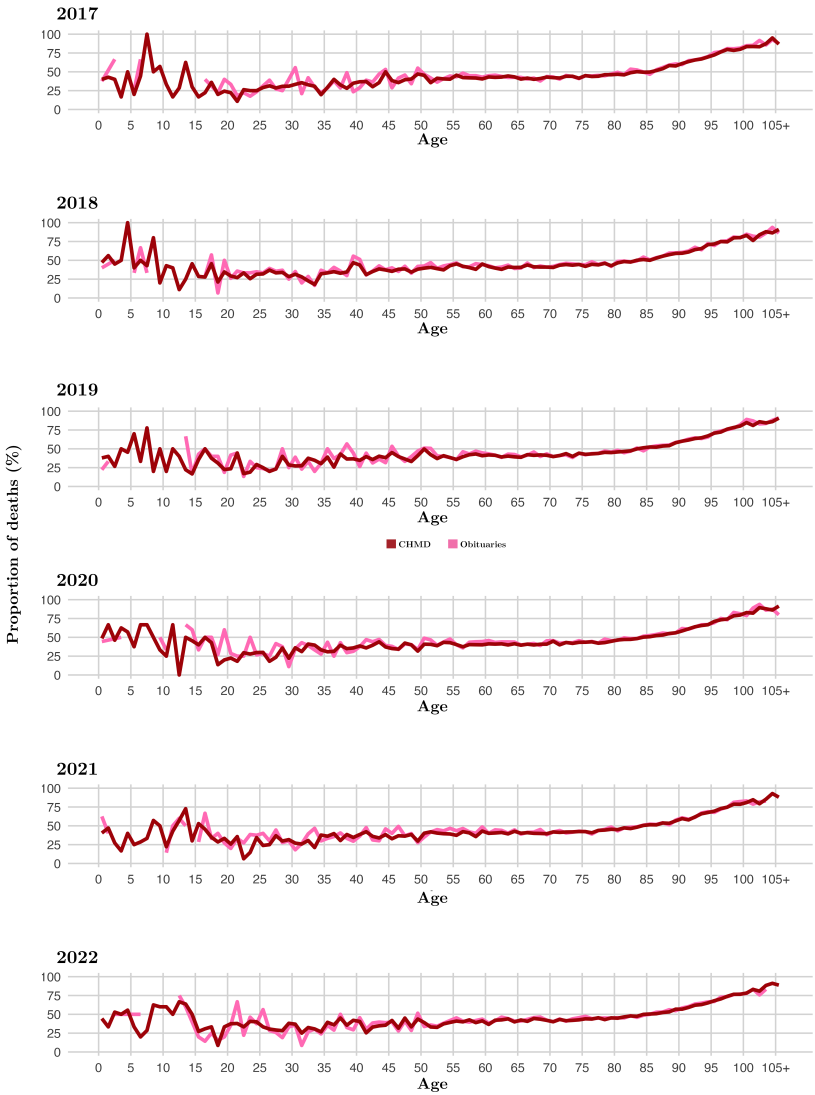


Sources: Authors' calculations based on data from Necro Canada (2024) and the CHMD (2024).

Figure A-2: Obituary coverage of vital statistics deaths by age and gender, Ontario, 2017–2022

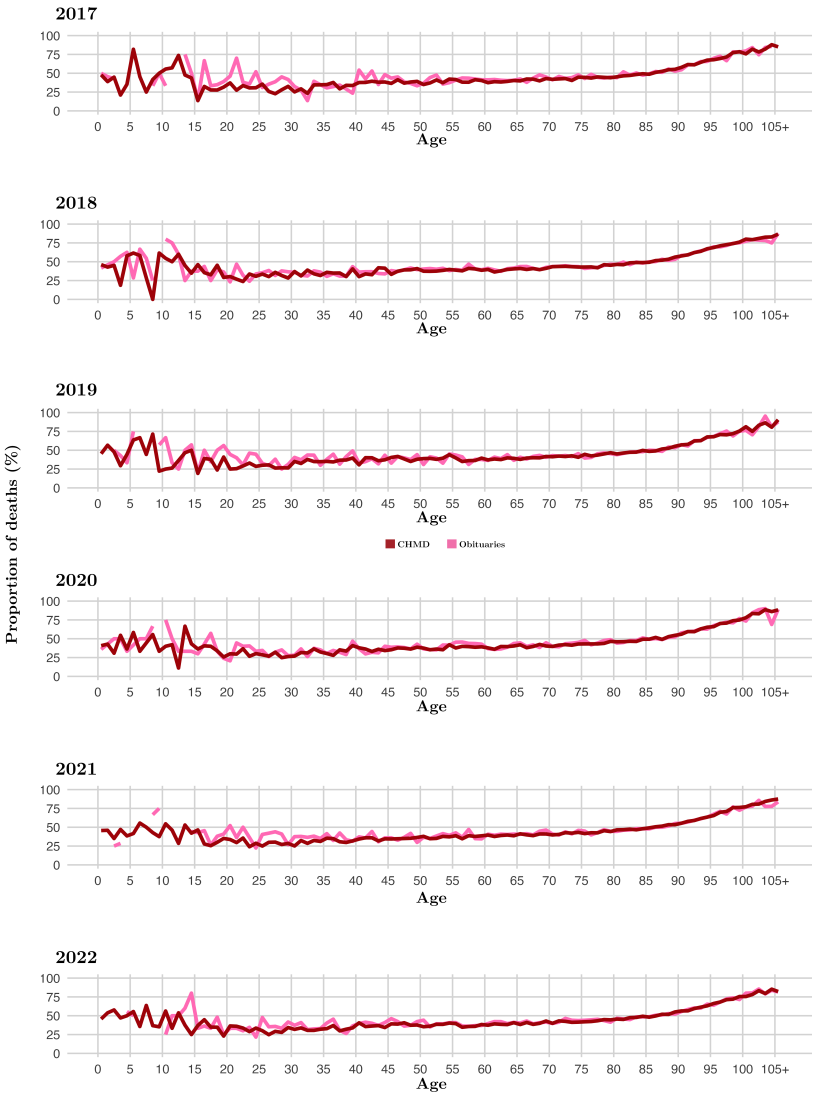
Sources: Authors' calculations based on data from Necro Canada (2024) and the CHMD (2024).

Figure A-3: Proportion of women's deaths in the population by year, Quebec, 2017–2022



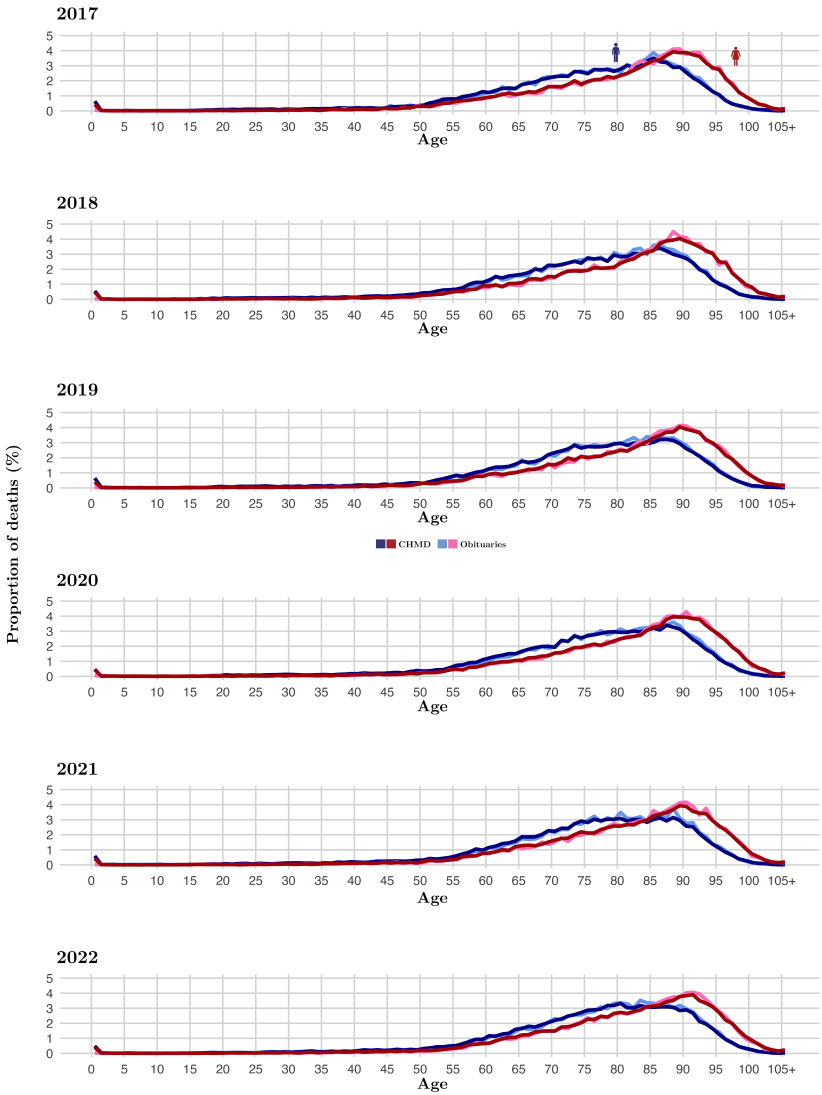
Sources: Authors' calculations based on data from Necro Canada (2024) and the CHMD (2024).

Figure A-4: Proportion of women's deaths in the population by year, Ontario, 2017–2022



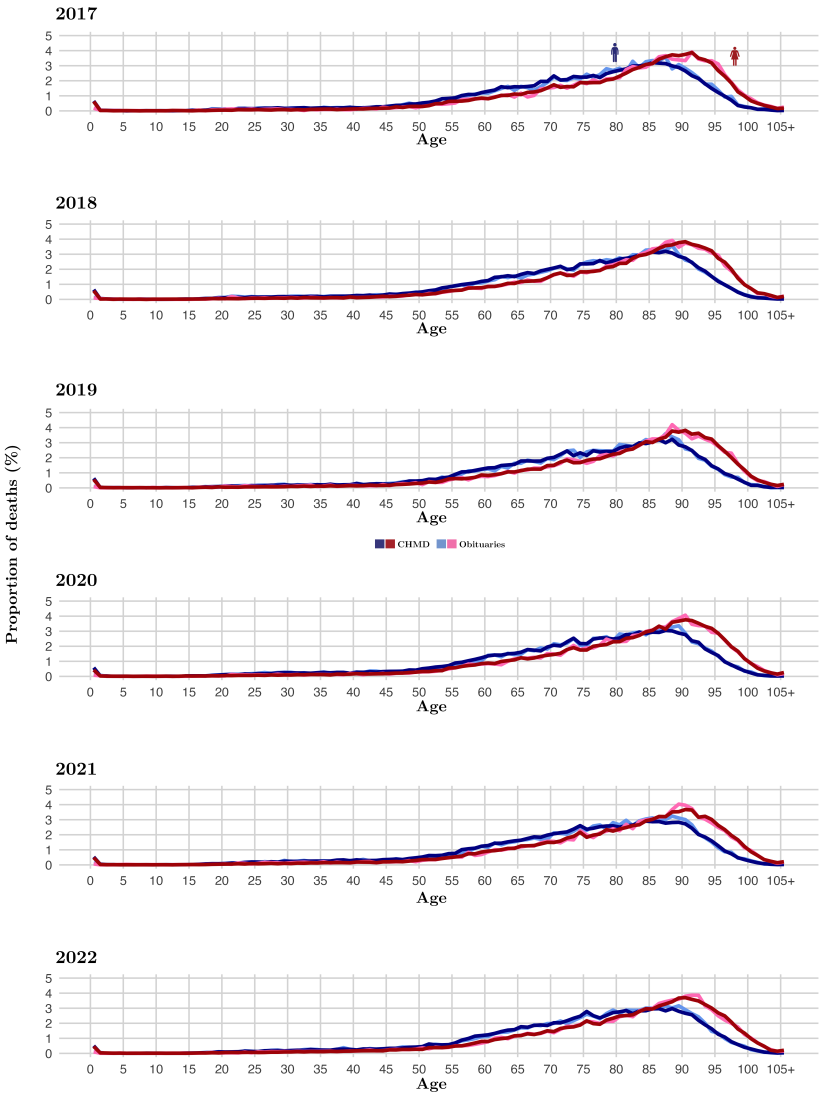
Sources: Authors' calculations based on data from Necro Canada (2024) and the CHMD (2024).

Figure A-5: Age distribution of observed deaths by gender and source of data, Quebec, calendar years 2017–2022



Sources: Authors' calculations based on data from Necro Canada (2024) and the CHMD (2024).

Figure A-6: Age distribution of observed deaths by gender and source of data, Ontario, calendar years 2017–2022



Sources: Authors' calculations based on data from Necro Canada (2024) and the CHMD (2024).