Inventor Life Course: A Verified Mortality Dataset for U.S. Inventors

Guilherme Junqueira
University of Florida
gjunqueira@ufl.edu

September 2026

Deaths of inventors are widely used as shocks in research on innovation and organizations, yet no public dataset of U.S. inventor deaths is both dated and verified. This proposal describes an open dataset that fills that gap. For 30,457 inventors on USPTO PatentsView who died between 1995 and 2015, I record the date of death (exact for 24,708), the place and age at death, the source that confirms the identity of the deceased, the checks that could have refuted it, and, where an obituary states one, the cause of death and whether the death was anticipated. For 18,982 inventors I further record the dates of death of both parents. All 72,529 candidate deaths that entered the procedure are released with their adjudication status, so that users can distinguish confirmed, refuted, unresolved, and not-yet-searched cases. Two tests validate the evidence grades. Against records whose correct answer is known by construction, 94.8% of confirmed matches identify the right person; and confirmed decedents stop filing patents far more reliably than refuted ones do. By comparison, the one large public file carries more than one candidate date of death for two inventors in three and reports no false-match rate. The data contain no Social Security numbers and no information on living people, and will be deposited on Harvard Dataverse under an open license.

1 Why the Existing Data Are Not Enough

Deaths of key individuals are widely used as shocks in the economics of innovation and of organizations, from superstar scientists (Azoulay et al. 2010) and inventor teams (Jaravel et al. 2018) to founders (Becker and Hvide 2022) and executives (Bennedsen et al. 2020). The credibility of these designs rests on the underlying death data. For inventors, the available data fall short along three dimensions, which I discuss in turn: whether the person recorded as deceased is in fact the inventor, when the death occurred, and whether it could have been anticipated. The first of these has received the least attention in the literature and is where the dataset I propose makes its main contribution.

Verification: whether the person who died is the inventor. Any inventor mortality file is built by matching a name in the patent record to a name in a death register, and whether the file is usable depends on whether the two names belong to the same person. No public file addresses this question. The only large one (Kaltenberg et al. 2021) was built to measure inventor *ages*; its death fields are a by-product of that match, which uses first and last name, a state, and a birth year itself scraped from people-search sites, and scores the agreement on an ordinal scale the authors describe as a proxy for accuracy. It carries no hand-validated false-match rate, no requirement that a second source agree, and no test for signs of life after the recorded date, and its authors are careful to say so: "we identify a birth or death year for a person associated to a city or state, which does not mean we have identified the birth or death year for the true inventor."[1]

How much room this leaves is visible in the deposit itself. For only 34.6% of its 535,120 inventors does the file list a single candidate date of death; the remaining two-thirds carry several (a median of three and, because common names generate long candidate lists, a mean of 9.4).[2] The file lists these candidates but does not adjudicate among them. A researcher who needs one death per inventor must therefore choose, for two inventors in three, which of several deceased persons of the same name is the inventor, and nothing in the file indicates which choice is correct.

The cost of leaving this choice to the user can be measured directly. My own procedure generates 72,529 candidates by the same logic (name, state, and a plausible birth year) and then adjudicates each one. Of these candidates, 42.0% are confirmed, 29.5% cannot be confirmed after searching, and 557 are refuted outright: for 241 a document shows that the record belongs to someone else, for 187 a document places the person alive after the recorded date, and for 129 the date itself is contradicted. A behavioral bound sharpens this picture. Inventors my procedure confirms stop filing patents about 95% of the time, whereas inventors it refutes stop only 37% of the time. Taking these two classes as anchors, the stopping rate of an unadjudicated register match implies that roughly one candidate in three is not the inventor at all. The ranking holds within every bin of pre-death filing intensity, so it does not reflect mean reversion.[3]

[1] Kaltenberg et al. (2021, p. 21). The death fields are in the working paper's data deposit; the published article (Kaltenberg et al. 2023) reports the ages only.

[2] Computed from the public deposit, `inventor_deathdate.csv`, Harvard Dataverse `doi:10.7910/DVN/YRLSKU`.

[3] This is a mixture bound rather than an identification result, as it assumes the two anchor classes are pure. Figure 3 reports the exact rates for every evidence grade.

The one U.S. source in which an inventor death *is* verified carries no dates. The "deceased" designation on the patent front page, which the USPTO records only after a legal representative files proof of death, allows Balsmeier et al. (2023) to identify 1,621 deceased inventors over 1976–2005. It establishes that a death occurred between application and grant, but it provides neither the date nor the place nor the cause. As a result, no public dataset of U.S. inventor deaths is both dated and verified.

Timing: years, not dates. The Kaltenberg et al. (2021) file records death *years*, scraped from genealogy aggregators, and every claim its authors make about the data is at the year level. For an event-study design this is too coarse: patents are filed continuously throughout the year, so a death that cannot be placed within the year cannot be ordered relative to the filings it is meant to have disrupted. I recover an exact date for 24,708 deaths and flag the 774 further 1 January entries that registers use as placeholders, rather than treating them as actual dates.

Anticipation: expected deaths are not shocks. The existing data do not indicate whether a death was expected, a distinction that matters for identification. When a firm watches an inventor decline over two years, it hires, trains, and reassigns before the death occurs, so that the estimated effect of the death includes a response that began earlier; when it loses an inventor suddenly, it has done none of this. I classify the cause of death wherever an obituary states one and mark the death as anticipated or unanticipated. Because a cause is stated for few deaths, I also record, for every inventor I can trace, the ages at which the inventor's own mother and father died, which provide a family-specific benchmark for how early the death came.

2 What the Data Contain

The release has four linked files, all keyed to PatentsView inventor identifiers, with the counts in table 1. **Deaths** carries the date, age, state and county of death, the confirming channel and source document, the number of independent channels, the cause class and anticipation flag, the employer named in the obituary, and the inventor's patent record as of the death. **Candidates** carries every register match that entered the procedure with the verdict reached on it. **Parents** carries both parents' names and the inventor's birthplace from the Numident and the parents' own dates of death from a second register match, each with its match tier and candidate count. The **evidence ledger** carries, per death, the sources consulted, which fields agreed, the adjudication note, the disagreements recorded against it, and a crosswalk to Kaltenberg et al. (2021) identifiers, which 29,881 confirmed deaths carry.

Table 1. The data at a glance

Candidate deaths entering the procedure		72,529
confirmed, 1995–2015		**30,457**
with an exact date of death		24,708 (81%)
with a state of death		26,072 (86%)
confirmed by a source outside the SSA registers		14,656 (48%)
confirmed through two or more independent channels		4,398 (14%)
at or under age 60	2,807 (9.2%; 10.9% of the 25,738 with a recorded age)	
with an employer identified at the death		16,840 (55%)
at or under 60 *and* at an identified firm		1,988
refuted by a later sign of life or a contradiction		557
searched but not confirmed / not yet searched		21,386 / 19,737
under review, a merge the procedure has not settled		392
Inventors with both parents' names		21,306
with both parents' dates of death		**18,982**

3 How the Data Are Built

Figure 2 applies the procedure to two inventors whose deaths are publicly documented, one usable as a shock and one not. Christopher Hipp, who designed the first blade server, died at 47 of a cardiac arrhythmia while cycling, two years after his last patent filing; his death is confirmed, early, and unanticipated. Anita Borg died at 54 of cancer, fourteen years after her last filing; her death is equally well confirmed but cannot serve as a shock to any firm. For both inventors the release contains the confirmation together with the fields (date, cause, anticipation flag, last filing) that allow a user to tell the two cases apart.

The confirming channels for the 30,457 deaths are the SSA registers joined by SSN (15,773), the VA Nationwide Gravesite Locator (4,765), state death indexes (4,259), open-web obituary and memorial documents (2,770) and NewsBank local newspapers under a text-and-data-mining permission (1,754), with the remaining 1,136 from

Wikidata and Wikipedia, cemetery records, USPTO filings naming an estate or legal representative, and genealogy aggregators. Where an obituary states a cause it is classified as natural illness (391), cancer (287), sudden (278) or other (149). Anticipation is coded separately and only where the text supports a call: 453 deaths are marked anticipated and 309 unanticipated, leaving 343 with a stated cause but no call either way.

Figure 1. How a death is found and confirmed

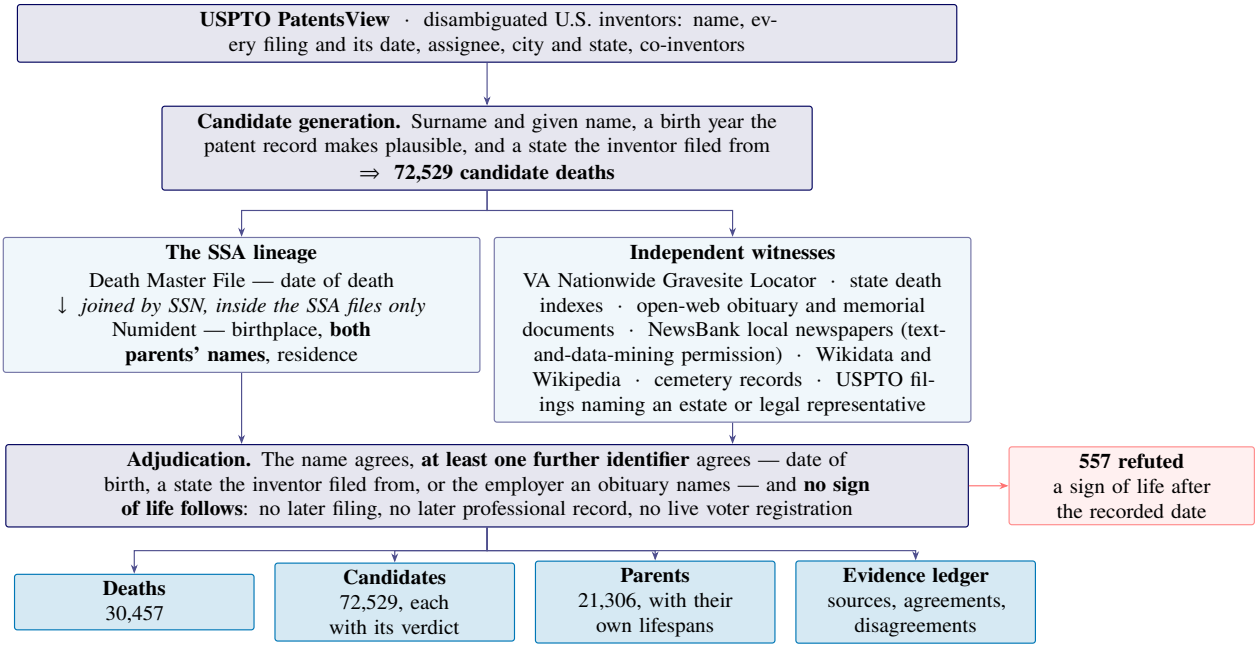


Figure 2 applies the procedure to two inventors whose deaths are publicly documented: one whose death can serve as a shock, and one whose death cannot. Christopher Hipp, who designed the first blade server, died at 47 of a cardiac arrhythmia while cycling, two years after his last filing — confirmed, early, unanticipated, and usable as a shock. Anita Borg died at 54 of cancer, fourteen years after her last filing: equally well confirmed, and not usable. For both inventors I release the confirmation and the fields a user needs to tell the two cases apart.

Figure 2. The procedure applied to two inventors

	Christopher G. Hipp designed the first blade server	**Anita Borg** founded the Institute that bears her name
Patent record	six utility patents, 2000–2007; last filing 2007 at Teradici, a public company; six co-inventors; San Mateo County	two utility patents, 1982–1989; last filing 1989 at Digital Equipment Corporation; four co-inventors; San Mateo County
Register match	one genealogical register; **14 July 2009**; an alternative record says 19 July	two register families agree on the year; **6 April 2003** in one, **7 April** in the other. The disagreement is recorded, not resolved
Witness	Wikidata `Q5112545` → Wikipedia → the obituary it cites; name, age and place agree	Social Security Death Index and a second, independent genealogical family
Cause	cardiac arrhythmia, coded from the obituary, which places the death during a bicycle ride ⇒ **sudden, unanticipated**	cancer ⇒ **anticipated**; the employer had time to adjust
Age, place	47, Menlo Park, California	54, San Mateo County, California
Verdict	no sign of life after the date; **confirmed**, two channels, independent of the SSA	no sign of life after the date; **confirmed**, two channels, independent of the SSA
Usable as a shock?	**yes**: early, unanticipated, and still filing two years before he died	**no**: the death was anticipated and her last filing is fourteen years earlier

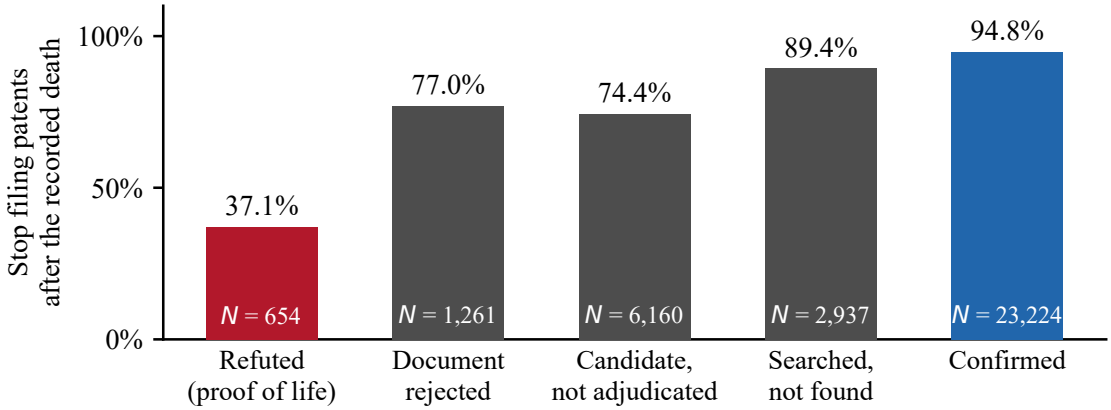
4 Validation

Section 1 argues that an unchecked name match is wrong often enough to matter. The two tests below establish the complementary claim, namely that the evidence grades in this file separate real deaths from spurious ones.

An identity test with a known answer. For 6,592 inventors the correct Numident record is known by construction: their Social Security number comes from the Death Master File match, and that number appears in the Numident window. The test is whether the name-and-date rule, which never sees the number, recovers the same record. It returns at least one candidate for 89.3% of them, 2.15 candidates on average; the correct record is among the candidates for 85.1%; and of the records graded CONFIRMED, **94.8% are the correct person** (92.5% where the confirming field is geographic, 96.8% where it is age). The test population is inventors who do have an SSA record, which is the easiest case, so this is an upper bound for the rest of the file.

A behavioral test across evidence grades. If the evidence grade measures whether a death is real, then inventors with stronger evidence should stop filing patents more reliably after the recorded date. Figure 3 shows that they do. Among inventors graded CONFIRMED who were filing in the three years before the recorded date, 94.8% stop; among those searched but not found, 89.4%; among candidates never adjudicated, 74.4%; and among those refuted by a proof of life, 37.1%. Because inventors stop filing eventually in any case, the evidence lies in the gradient across grades rather than in the level of any one rate.

Figure 3. Stopping rates by evidence grade



Share of people who stop filing patents after the recorded death, conditional on having filed in the three years before it, by the verdict assigned.

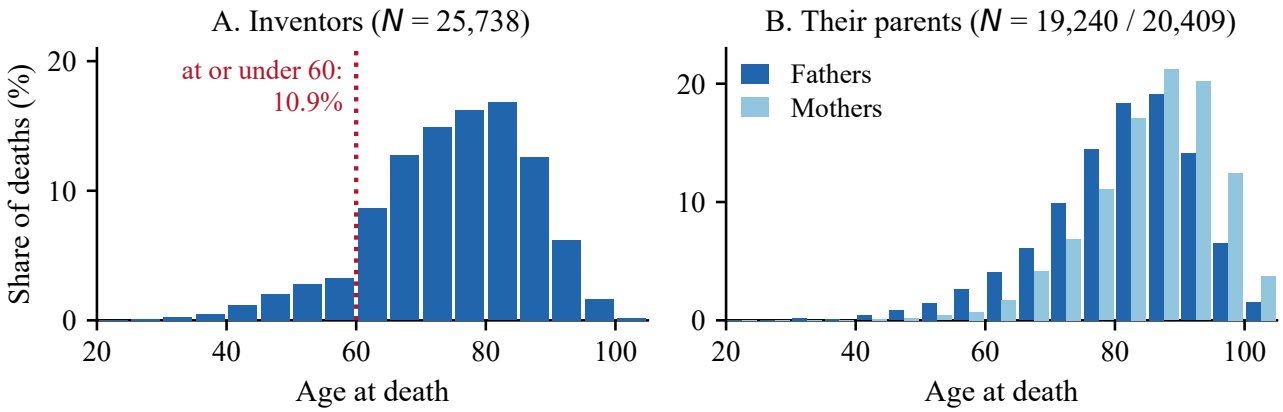
Limitations. The Death Master File ends in November 2011 and the Numident in 2007, so later deaths rest on state indexes, veterans' files and obituaries and are under-covered; year-by-year coverage is released with the data. Generating candidates by name misses the most common names and deaths abroad, and 19,737 candidates are not yet searched. A cause is stated for only 3.6% of confirmed deaths, which is why the parents' lifespans are included. Disagreements between a document and the record are recorded rather than resolved silently: 4,904 differ on the day, 2,130 on the year. Women and foreign-born inventors are harder to trace, and every candidate's search status is released so that coverage gaps are visible.

5 Value to the Research Community

The immediate use is the death-as-shock design, which a researcher can run from these files instead of rebuilding the mortality data, as each paper in this literature has so far had to do. Exact dates place an event relative to filing dates and fiscal years; the cause class and the parents' lifespans separate unanticipated deaths from expected ones; the ledger lets a referee check any individual case. As a measure of scale, my job-market paper uses 989 of the 1,988 early deaths at an identified firm reported in table 1 (those in which the inventor was still filing), each matched to living inventors at comparable firms. A year-of-death file cannot support a sample of this kind.

The data also support demographic work. Figure 4 plots age at death for the inventors and their parents; 10.9% of confirmed deaths with a recorded age are at or under 60. The parents' distributions sit roughly a decade to the right and separate by sex as period life tables imply, which is a check on the parent match and a starting point for work on inventor longevity and on the transmission of health and of inventiveness across generations. The data link directly to PatentsView, to Kaltenberg et al. (2023), and through the patent record to firm-level sources.

Figure 4. Age at death: inventors and their parents



Panel A: all confirmed inventor deaths with a recorded age. Panel B: the fathers and mothers of the inventors in the parents file, from Numident names matched to their own death records.

6 Dissemination

Version 1, at the time of the meeting: the deaths file, the candidate file, the evidence ledger, a codebook and a source catalog, on Harvard Dataverse under CC BY 4.0 with a DOI and a stated versioning policy. Personal information is limited to what is already public: no Social Security numbers, no living people, and no obituary text, only the extracted fields and a citation to the source. **Version 2**, with the companion paper in 2028: the parents file, the build code (GitHub, MIT license), the remaining candidates adjudicated, and an extension past 2015. Both carry a public issue tracker, where a wrongly confirmed or a missed death can be reported with evidence and re-adjudicated, and a template for contributing a state death index or a regional obituary archive under the same fields and grades.

7 What I Would Present

The presentation would focus on the data and on how to use them rather than on research findings, and would proceed in three parts. The first part would walk through the construction of the dataset end to end — the source catalog, the generation of candidates from the registers, the adjudication of a single death from candidate to verdict, and the two validation tests — so that the audience can assess each step and the evidence grades that result from it. The second part would show how a researcher works with the files: how to run a death-as-shock design directly from the deaths file; how to use the cause class and the parents' lifespans to restrict attention to unanticipated deaths; how to link a death to PatentsView and, through the patent record, to firm-level sources; how to use the candidate file to characterize coverage in a given sample; and how to use the evidence ledger to document any individual case for a referee. The third part would cover the dissemination plan and the versioning and issue-tracking mechanics.

Because the effort is in progress, I would also seek the group's feedback on four open questions: how to handle common names when the register carries no occupation field; whether there is a defensible way past the 2011 Death Master File cut-off; whether cause-of-death classification can be standardized across papers, so that "unanticipated" means the same thing in each; and what the appropriate disclosure standard is for a public file on deceased individuals and their parents.

References

Azoulay, P., J. Graff Zivin and J. Wang (2010). Superstar extinction. *Quarterly Journal of Economics* 125(2), 549–589.

Becker, S. O. and H. K. Hvide (2022). Entrepreneur death and startup performance. *Review of Finance* 26(1), 163–185.

Bennedsen, M., F. Pérez-González and D. Wolfenzon (2020). Do CEOs matter? Evidence from hospitalization events. *Journal of Finance* 75(4), 1877–1911.

Balsmeier, B., L. Fleming and S. Lück (2023). Isolating personal knowledge spillovers: Coinventor deaths and spatial citation differentials. *American Economic Review: Insights* 5(1), 126–140.

Jaravel, X., N. Petkova and A. Bell (2018). Team-specific capital and innovation. *American Economic Review* 108(4–5), 1034–1073.

Kaltenberg, M., A. B. Jaffe and M. E. Lachman (2021). The age of invention: Matching inventor ages to patents based on web-scraped sources. NBER Working Paper 28768.

Kaltenberg, M., A. B. Jaffe and M. E. Lachman (2023). Invention and the life course: Age differences in patenting. *Research Policy* 52(1), 104629.