Quantifying Attrition in Science: A Longitudinal Study of Scientists in 38 OECD Countries

University of Oxford, CGHE, January 30, 2024

Professor Marek Kwiek

Institute for Advanced Studies in Social Sciences and Humanities (IAS), Director

UNESCO Chair in Institutional Research and Higher Education Policy

University of Poznan, Poland

kwiekm@amu.edu.pl

Twitter: @Marek_Kwiek

INSTITUTE FOR ADVANCED STUDIES IN SOCIAL SCIENCES AND HUMANITIES

2. Introduction (1/3): Main Questions

- **How can we *quantify* the phenomenon of *leaving science?***
- How do members of the **global** scientific community actually **disappear** from science?
- How attrition differs between **men and women? Across disciplines? And over time**?
- How does a **global, cohort-based, longitudinal** approach work in practice (major limitations)?
- **How can we meaningfully *track* scientists over time (here: two decades, different cohorts)?**

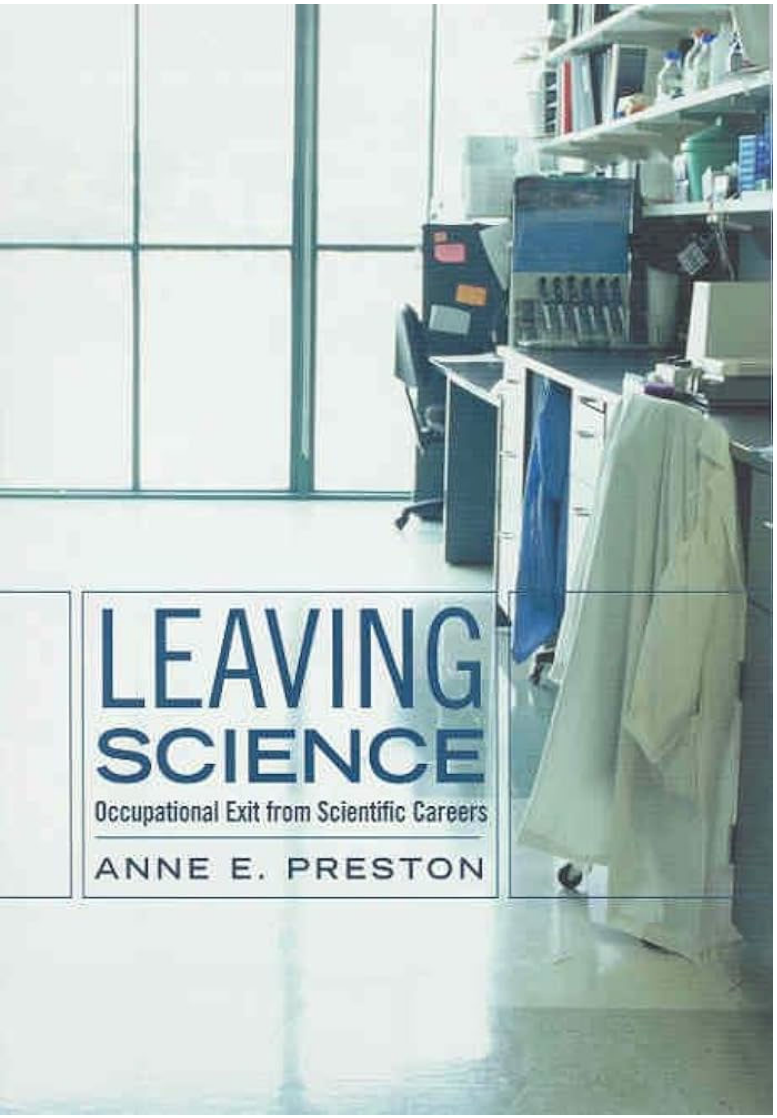
3. Introduction (2/3): Testing Traditional Narratives with a New Global Dataset

- **Testing validity of traditional narratives on attrition** (which have supported science policies for decades).
- **Traditional narratives**: **(1) Women tend to disappear from academia earlier than men**; and (2) women tend to disappear **in higher proportions than men** (Alper, 1993; Blickenstaff, 2005; Deutsch & Yao, 2014; Goulden et al., 2011; Preston, 2004; Shaw & Stanton, 2012).
- **Big numbers: tracking scientists who started publishing in 2000** (N=142,776) and **2010** (N=232,843) over time.
- **Comprehensive**: publication and citation metadata (Scopus raw dataset, Elsevier's ICSR Lab): **careers in 38 OECD countries, 16 STEMM disciplines**.
- **Testing new possibilities opened up by global bibliometric datasets for large-scale studies of scientific careers.**

4. Introduction (3/3): Global Datasets and Big Data Approaches to Scientists

- **Global & longitudinal approaches** to academic careers possible today (why: **increasing access to digital databases)**.
- Difficult, expensive, temporary… time-consuming, team work…
- The databases offer **comprehensive information** about scientists (research outputs, citation-based impact) – we can **build individual lifetime histories**.
- **New opportunities to test traditional conceptual frameworks about science and scientists (academia and the academic profession).**
- Systematic explorations of **career histories** of hundreds of thousands of individual scientists possible.
- **Men and women in science // Big Data** (e.g., King et al., 2017; Nielsen & Andersen, 2021; Wang & Barabási, 2021; Sugimoto & Larivière, 2023).

5. *Leaving Science* as a Scholarly Theme



- Not examined both **longitudinally** (year by year) and **globally** (many countries, with a focus on disciplines) so far.
- **Traditionally explored through small-scale, case study research, mostly survey- and interview-based.**
- Concepts of "**faculty departure intentions**", "**faculty turnover**", etc. (Zhou & Volkwein 2004; Rosser 2004).
- Focus on **single institutions**, limited to the USA.
- Explanations of quitting science (in surveys, interviews) (e.g. Cornelius et al., 1988; Goulden et al., 2011; Levine et al., 2011):
  - the problems of keeping **work–life balance**, parenthood,
  - low **job security** and low **salaries**,
  - **colleagues and workload** concerns,
  - **discrimination** in the workplace,
  - hostile workplace (chilly) **climates** (e.g. Spoon et al. 2023)
  - (internal) **push** vs. (outside) **pull** factors.

6. This Research vs. Previous Research

- A different geographical scale; moving away from a single-country research design - **toward *disciplines* and changes over *time* (cohort-based approach).**
- A different methodology (**survival analysis & logistic regression analysis**), cross-disciplinary and gender differences in attrition.
- Using **large cohorts of scientists**.
- **Longitudinal in the strict sense of the term: cohorts of exactly the same scientists tracked over time on a yearly basis (up to 22 years).**
- A wealth of **individual micro-level data** used.

- **Research with Lukasz Szymula from Poznan CPPS Team: "Quantifying Attrition in Science"** preprint at ArXiv.

7. Scholarly Publishing *Events* and *Survival Analysis*

- **Leaving science conceptualized as an event: analyzed within survival analysis** (Allison, 2014; Mills, 2011).
- **Scientific life conceptualized as a sequence of scholarly publishing events** (from the first publication event - onwards).
- **In event analysis, we compute probabilities of occurrence of an event (here: stopping publishing)** at a certain point in time.
- The last publication ever: when **scientists stop publishing** (uncensored observations only, 2019 vs. 2022; about 90% publish every year).
- No studies combining 3 perspectives: **longitudinal**, **global, and quantitative**!

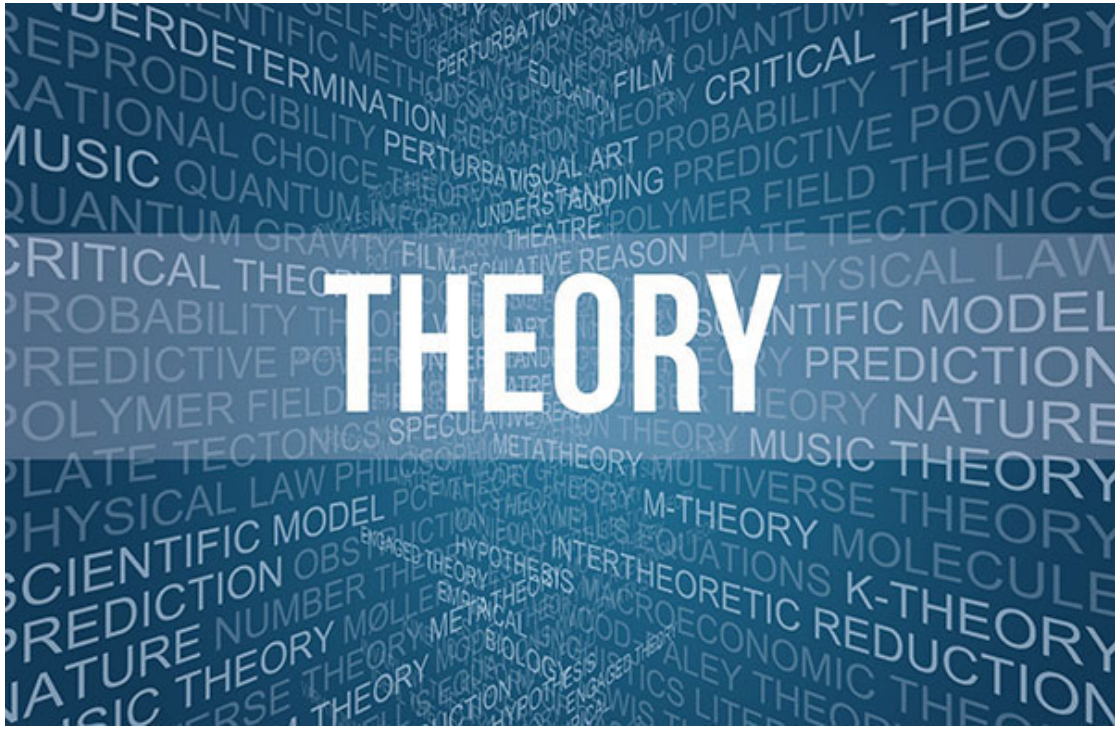
8. Theories of Women Leaving Science – Revisited?

The chilly climate theory: a **hostile or unwelcoming** work environment in STEM fields can **discourage** women from pursuing and persisting in these fields The "**perception of exclusion**" and **a sense of "not belonging".**

(Maranto & Griffin 2011, Wolfinger et al. 2008).

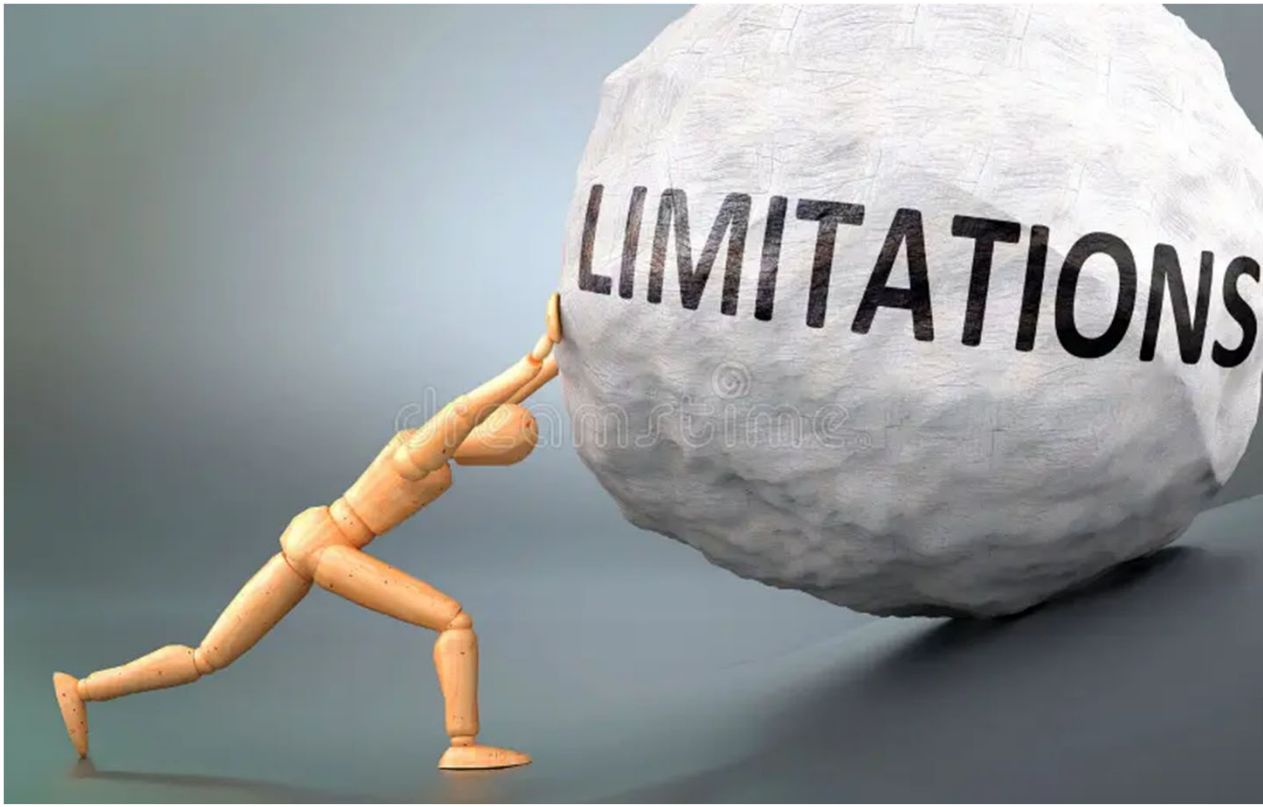
The leaky pipeline theory: a significant **loss of talent at every stage** of the academic career pipeline due to **systemic barriers** (such as bias and discrimination). Women either **progress through a series of academic stages** or **leave** academia altogether.

(Blickenstaff 2005; Goulden et al. 2011; Shaw & Stanton 2012).



9. Global Datasets and Their Limitations

- **Only bibliometric-type sources** provide access to micro-level data **longitudinally?**
- Useful **to treat global bibliometric datasets as 'structured' Big Data** (requiring new algorithmic techniques) for **useful information extraction?**
- **Old limitations** of bibliometric datasets: **language and STEMM focus**, Anglo-Saxon bias, and article-only content. Etc.
- Discussed for years!
- **New limitations, on top of previous bibliometric-type limitations.**

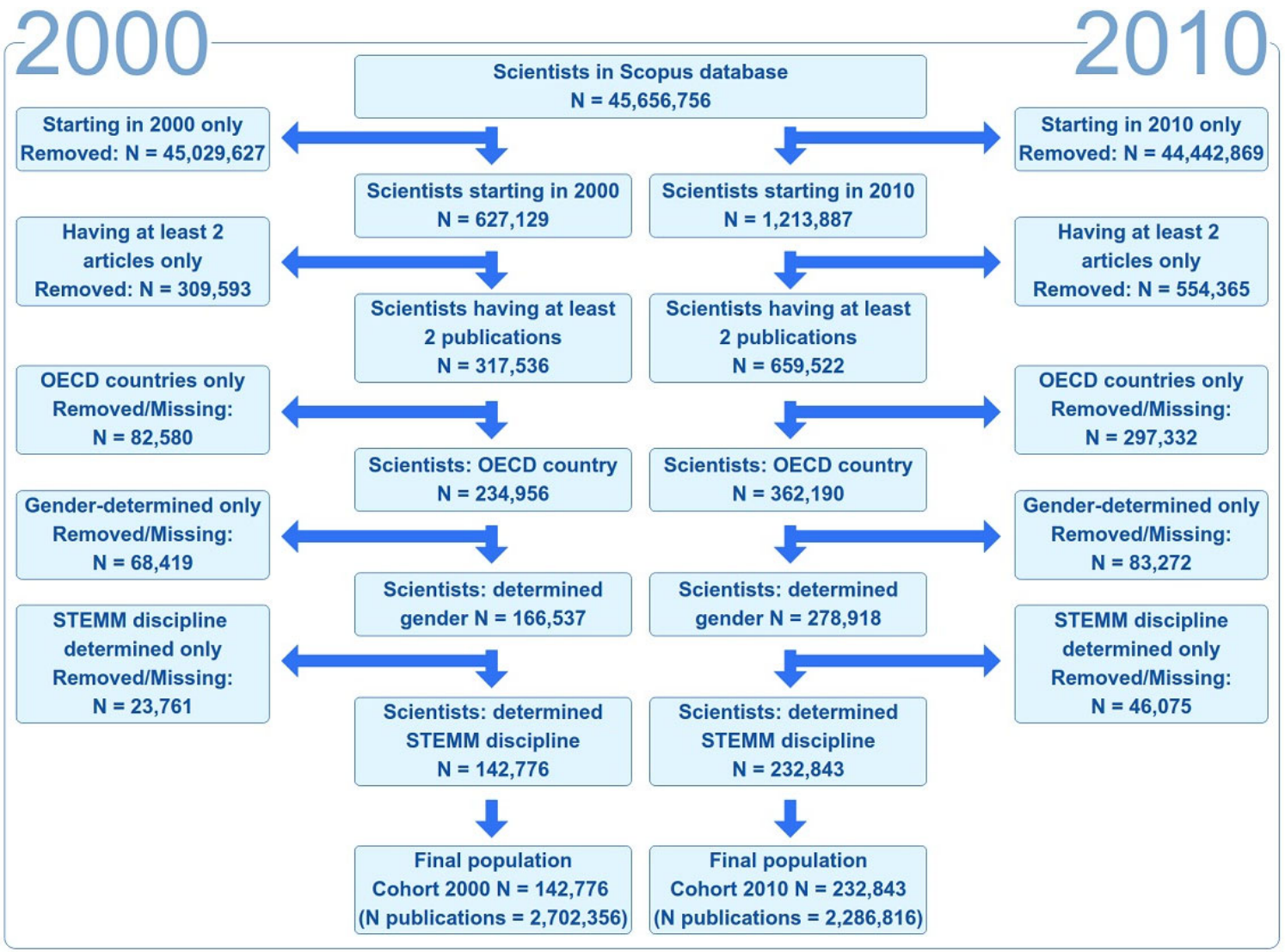


10. New Limitations in This Research

- Leaving science as **'stopping publishing'**: other academic roles **dismissed** (teaching, service, administration).
- **'In science'** and **'out of science':** slippery concepts (Preston, 2004).
- Doing science is **more than publishing**: **simplified,** much wider cognitive & social processes (Sugimoto & Lariviere, 2023). Various dimensions omitted.
- **'Not publishing anymore'** as '**not doing science any more'** (as opposed to 'leaving academia'): mentoring, reviewing grant proposals, editing journals).
- Not possible to verify **intra-sectorial or extra-sectorial employment** of **'non-publishers'** at the global level (our datasets).
- **Active participation** – only through publishing.
- **Non-English, non-indexed** publications not counted (but: STEMM).
- **Testing the power of structured, reliable, and curated Big Data (of the Scopus type).**

11. Dataset

- Two cohorts.
- Also: all 11 cohorts 2000-2010.
- The steps taken:
  - **non-occasional** scientists with at least two journal articles;
  - **country** affiliation as an OECD country;
  - **gender** (binary: male or female);
  - **discipline** as STEMM .



12. For every scientist, we have micro-level data (mostly computed by us): demographic, institutional, publishing & collaboration patterns. Examples: Cohort 2000 & Cohort 2010, N=375,619.

Scientist ID, the two-cohort database	Gender	Discipline	Country affiliation	Institutional type	Year entering science (year of first publication)	Year leaving science (year of last publication plus 1)	International collaboration rate, lifetime (%)	Average publication journal percentile, lifetime (1-99)	Median team size, lifetime	FWCI 4y - Field-Weighted Citation Impact, 4 years	Scholarly output, lifetime
Panel 1: Scientists – Cohort 2000 (N=142,776)											
ID 1	Female	MED	Spain	Rest	2000	2020	60.26	31.24	6.5	0.81	78
ID 2	Male	COMP	United States	TOP200	2000	2004	40.00	99.00	4	4.95	10
ID 3	Female	AGRI	France	Rest	2000	2008	21.43	68.15	4	0.88	14
ID 4	Male	PHYS	Japan	TOP200	2000	2013	0.00	90.00	5	1.37	3
ID 5	Female	CHEM	Denmark	Rest	2000	2001	75.00	1.00	3	1.19	4
…											
ID 142776	Male	MED	Germany	Rest	2000	2017	26.67	72.60	3	2.05	30
Panel 2: Scientists – Cohort 2010 (N=232,843)											
ID 142777	Male	ENER	United Kingdom	TOP200	2010	2012	33.33	98.00	5	1.15	6
ID 142778	Female	IMMU	Switzerland	TOP200	2010	2020	27.27	82.10	5	0.78	11
ID 142779	Female	BIO	Belgium	Rest	2010	2017	100.00	29.50	4	0.10	2
ID 142780	Male	ENG	Canada	Rest	2010	2014	14.29	31.43	2.5	2.04	7
ID 142781	Male	MED	Italy	Rest	2010	2012	100.00	14.00	10	0.13	3
…											
ID 375619	Female	AGRI	Australia	TOP200	2010	2015	0.00	91.08	5	1.93	9

13. Kaplan–Meier estimate for the 2000 cohort population, by gender (all disciplines combined).

The Kaplan–Meier probability of staying: lower than 50% for women in year 10, for men in year 12.

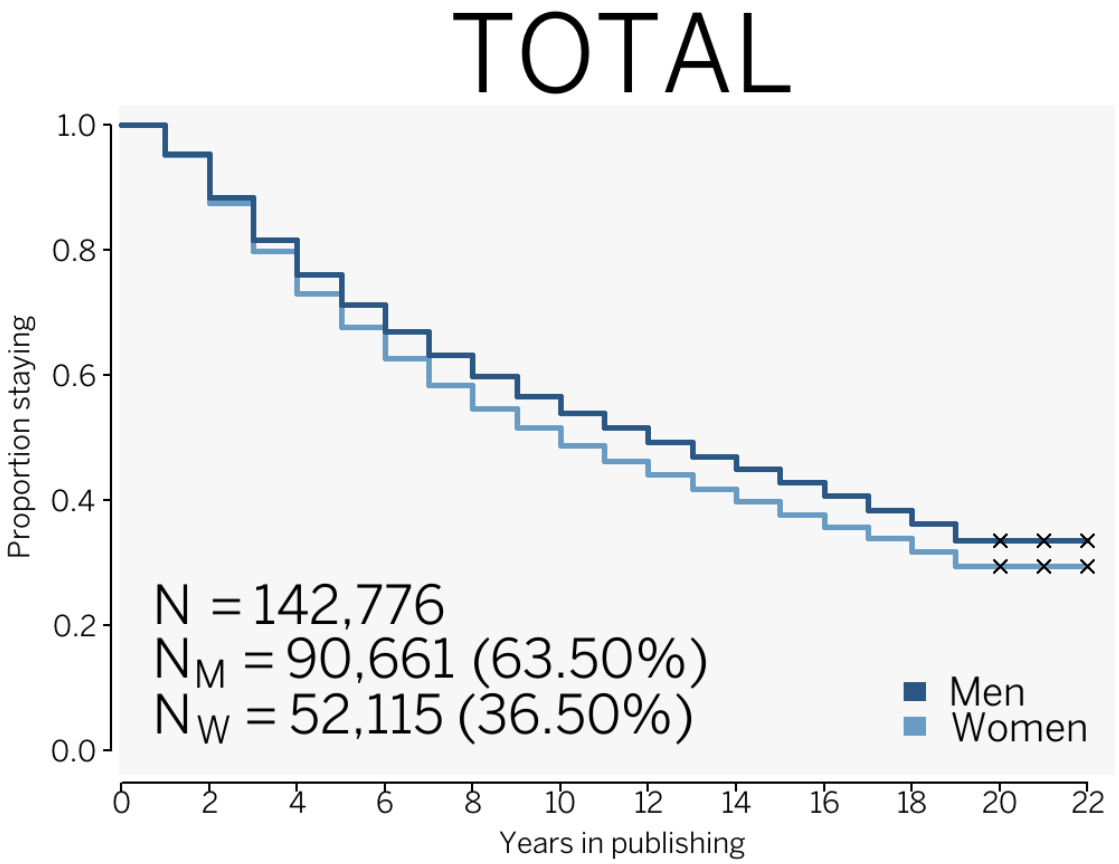
	Women			Men		
Time (years)	n	n leaving science	KM probability (staying) with 95% CI and SE	n	n leaving science	KM probability (staying) with 95% CI and SE
1	52,115	2,530	0.951 (0.950–0.953)[1]	90,661	4,151	0.954 (0.953–0.956)[1]
2	49,585	3,985	0.875 (0.872–0.878)[1]	86,510	6,302	0.885 (0.883–0.887)[1]
3	45,600	3,948	0.799 (0.796–0.803)[2]	80,208	6,114	0.817 (0.815–0.820)[1]
4	41,652	3,553	0.731 (0.727–0.735)[2]	74,094	5,062	0.761 (0.759–0.764)[1]
5	38,099	2,838	0.677 (0.673–0.681)[2]	69,032	4,356	0.713 (0.710–0.716)[2]
6	35,261	2,602	0.627 (0.623–0.631)[2]	64,676	3,934	0.670 (0.667–0.673)[2]
7	32,659	2,183	0.585 (0.581–0.589)[2]	60,742	3,458	0.632 (0.629–0.635)[2]
8	30,476	1,961	0.547 (0.543–0.551)[2]	57,284	3,110	0.598 (0.594–0.601)[2]
9	28,515	1,665	0.515 (0.511–0.520)[2]	54,174	2,774	0.567 (0.564–0.570)[2]
10	26,850	1,472	0.487 (0.483–0.491)[2]	51,400	2,465	0.540 (0.537–0.543)[2]
11	25,378	1,264	0.463 (0.458–0.467)[2]	48,935	2,225	0.515 (0.512–0.518)[2]
12	24,114	1,158	0.440 (0.436–0.445)[2]	46,710	2,055	0.493 (0.489–0.496)[2]
13	22,956	1,151	0.418 (0.414–0.423)[2]	44,655	2,032	0.470 (0.467–0.473)[2]
14	21,805	1,089	0.398 (0.393–0.402)[2]	42,623	1,889	0.449 (0.446–0.453)[2]
15	20,716	1,048	0.377 (0.373–0.382)[2]	40,734	1,884	0.429 (0.425–0.432)[2]
16	19,668	1,033	0.358 (0.353–0.362)[2]	38,850	1,959	0.407 (0.404–0.410)[2]
17	18,635	1,002	0.338 (0.334–0.342)[2]	36,891	2,020	0.385 (0.381–0.388)[2]
18	17,633	1,064	0.318 (0.314–0.322)[2]	34,871	2,070	0.362 (0.359–0.365)[2]
19	16,569	1,228	0.294 (0.290–0.298)[2]	32,801	2,350	0.336 (0.333–0.339)[2]

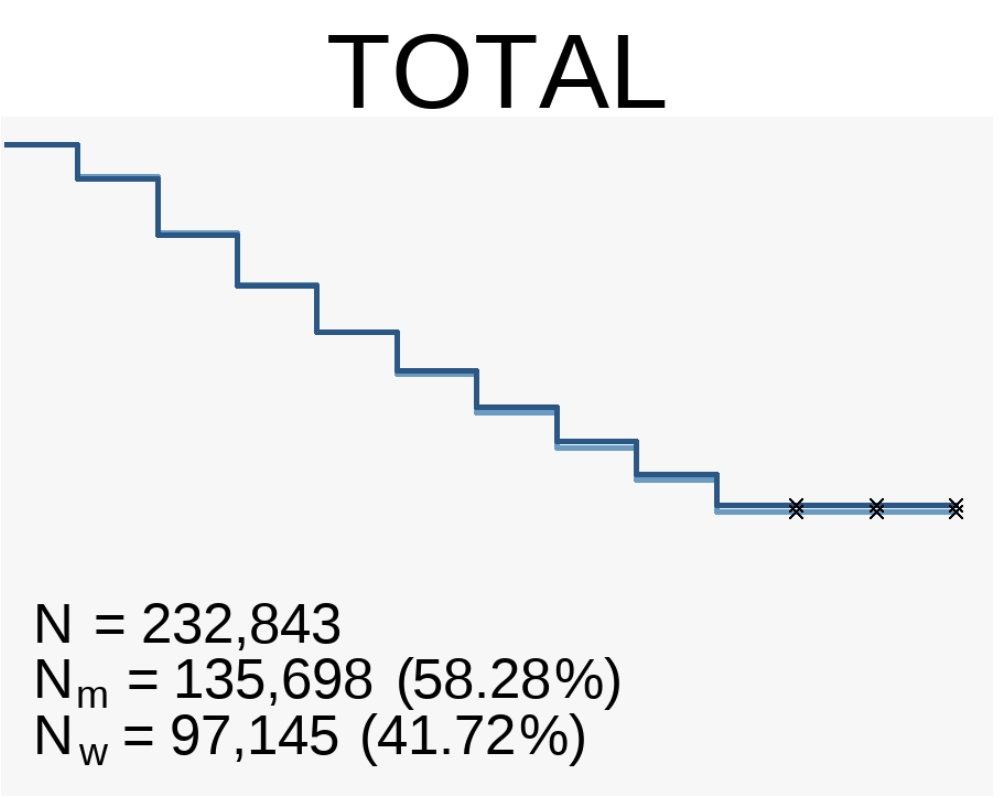
Note: (1) Standard Error 0.001, (2) Standard Error 0.002.

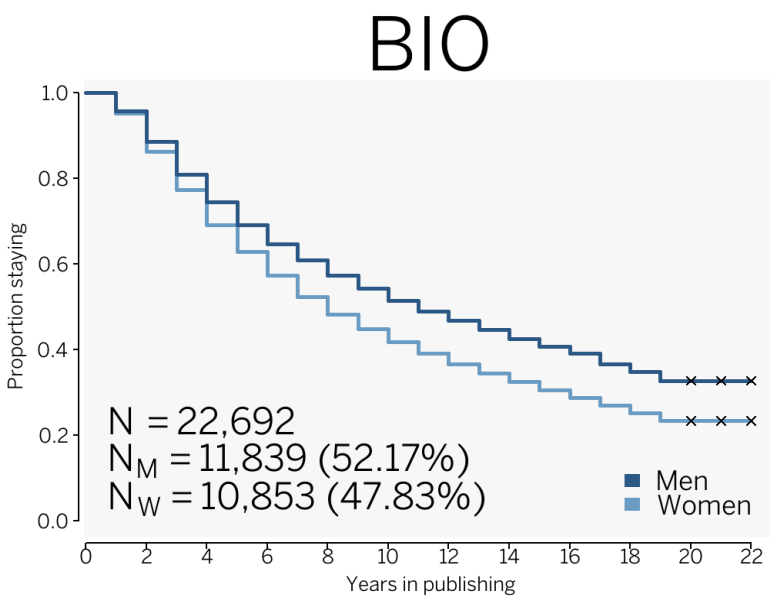
- **Multiplying all the probabilities of survival** across all time intervals preceding given point in time.
- The estimated probability that a woman will survive in science 10 years is 48.7% (54% for men).
- **The cumulative probability of staying** at the end of the study period (19 years): **women 29.4%, men 33.6%**
- **Significantly higher probabilities of staying for men for each year studied**!

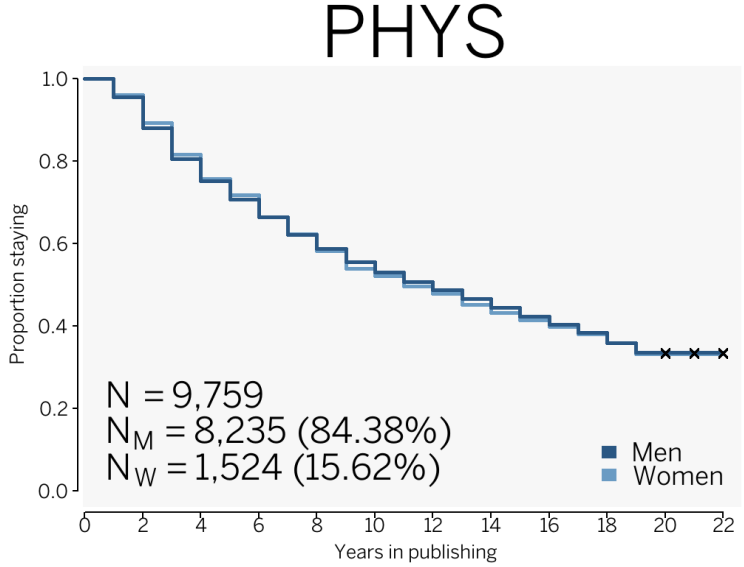
14. Kaplan–Meier survival curve by gender, all disciplines combined.

KM estimates: computing probabilities of occurrence of an event (= leaving science) at a certain point in time. Tick-marks: observations whose survival times have been right-censored.
The 2000 cohort (left) vs. the 2010 cohort (right). Uncensored observations only.









15. Disciplinary variations: Kaplan-Meier survival curve by gender, BIO Biochemistry, Genetics & Molecular Biology vs. PHYS Physics and Astronomy, the 2000 cohort

- **In BIO, women are one-fifth more likely to drop out of science after both 5 and 10 years** (20.78% and 19.96%).
- **In contrast, PHYS is a perfect example of the lack of gender differences in attrition.**
- **Strikingly, in the three math-intensive disciplines, MATH, COMP, and PHYS – which have very low numbers and percentages of women – the survival curves for men and women are nearly identical (overlapping survival curves).**
- For scientists starting publishing in 2000, **gender differences in attrition in PHYS do not exist**.

16. Explanation?

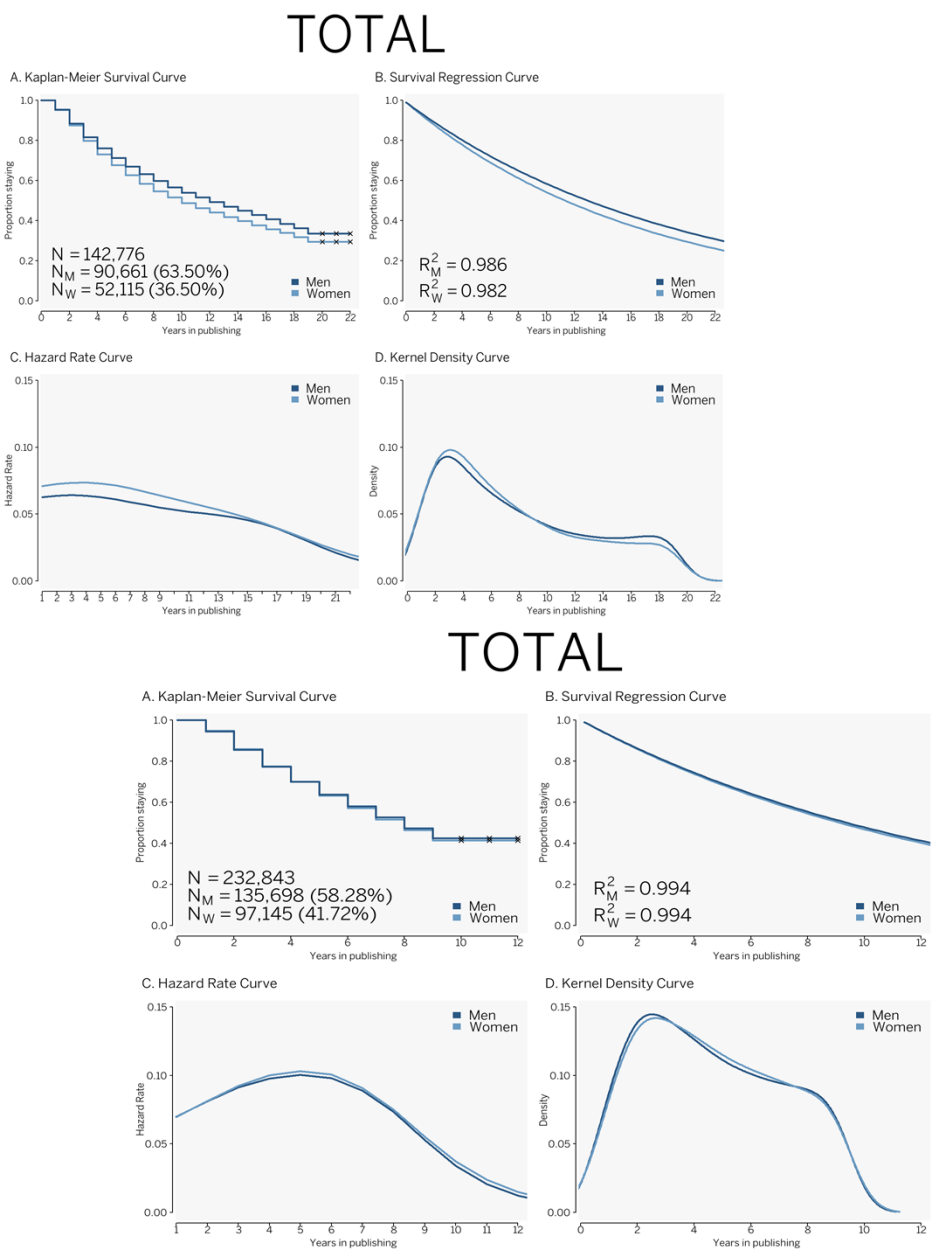
- In disciplines with very **low representation** of women (PHYS, COMP, MATH, ENG), **the newcoming and surviving women are extremely talented and hardworking?**
- Women as very visible minorities (10%-20%) may act as **exemplary figures, representatives of women** in university departments (as in companies, Kanter 1977). All their actions are public.
- **Small numbers & percentages of women** (alone or nearly alone in a peer group of men scientists)**: highly competitive from the very beginning?**
- Despite any discrimination women might meet in **heavily male-dominated environments**, they stay in the system of science as powerfully as men do.

17. Inter-Cohort Differences: Cohort 2000 vs. Cohort 2010 (Approaches).

(1) Kaplan-Meier Curve
(2) Survival Regression Curve
(3) Hazard Rate Curve, and
(4) Kernel Density Curve, all disciplines combined.

A dramatic lack of difference for cohort 2010 - compared with the 2000 cohort, where the results were substantially gender-sensitive.

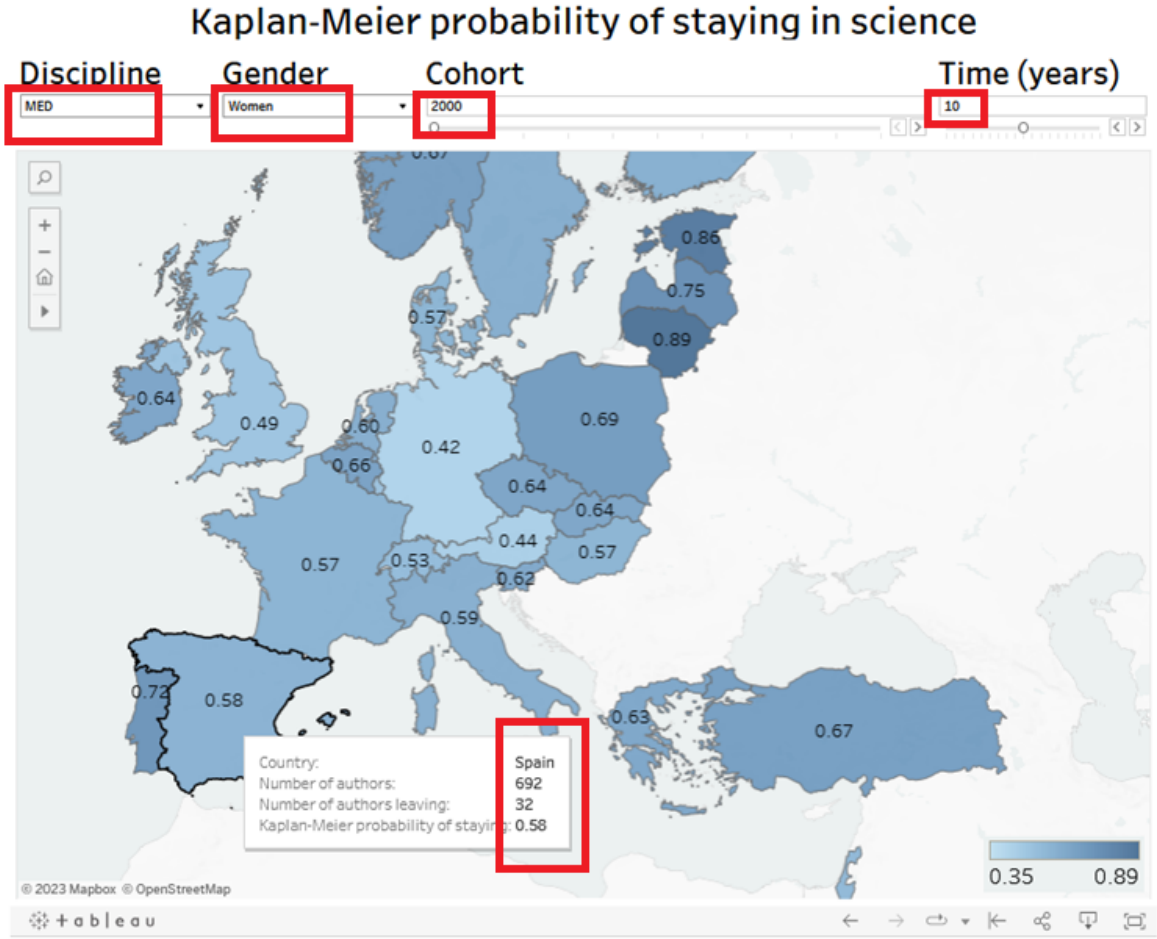
However, big story (Total: all disciplines combined) hides smaller-scale disciplinary stories… not discussed today (separate figures and analyses for each discipline)!



18. Attrition Patterns – Less Gendered Over Time…

- **The differences between men and women – so starkly visible for the 2000 cohort – almost disappear for the 2010 cohort.**
- **Different attrition patterns for different cohorts of scientists!**
- **The findings valid for older cohorts of scientists** (here: 2000 cohort) **may not be applicable to younger cohorts** (here: 2010 cohort).
- **Time in science matters!** Science enviroment is different – cohorts experience attrition differently!

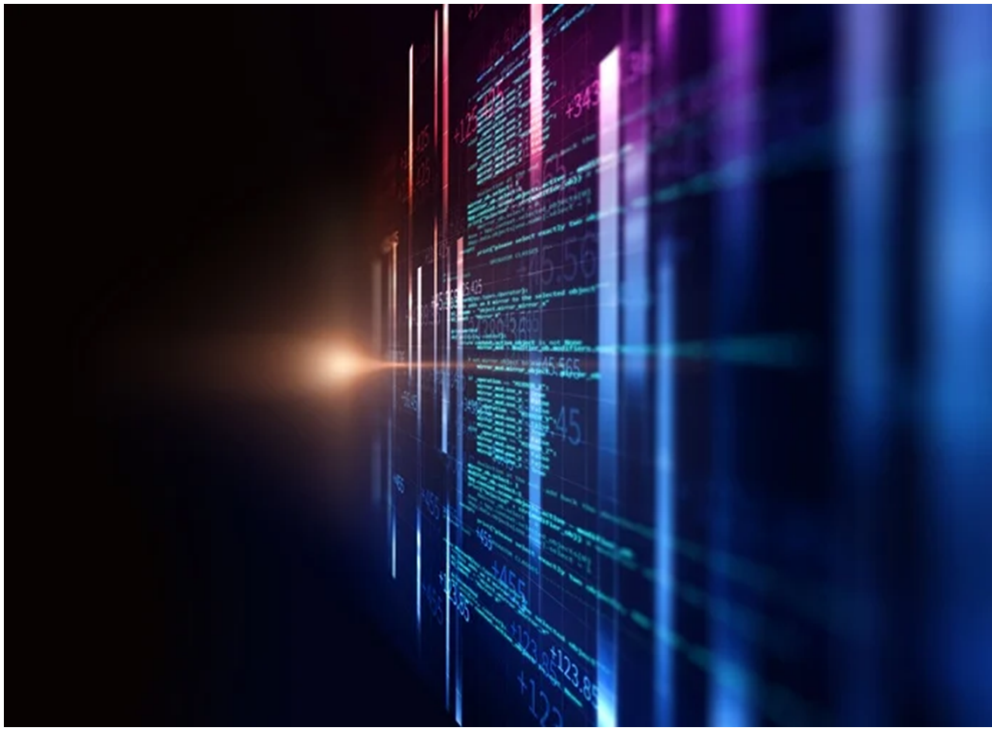
19. A snapshot of an interactive dashboard: Kaplan-Meier probabilities of staying in science by country, discipline, gender and cohort (11 cohorts 2000-2010) (N=2,127,803). Who stays in science, who leaves, where?



- Select: country, discipline, gender.
- For 11 cohorts, 2000-2010
- N=2,127,803 scientists.
- **Example: the probability of staying in science, women, 2000 cohort, in MED (Medicine), after 10 years, Europe: only about 40% for women in Germany and Austria, as opposed to about 90% in Estonia and Lithuania.**
- Different career prospects!
- Highlighted country data in red: Spain.
- **Address**: https://public.tableau.com/app/profile/marek.kwiek/viz/Attrition-in-science-OECD/Dashboard

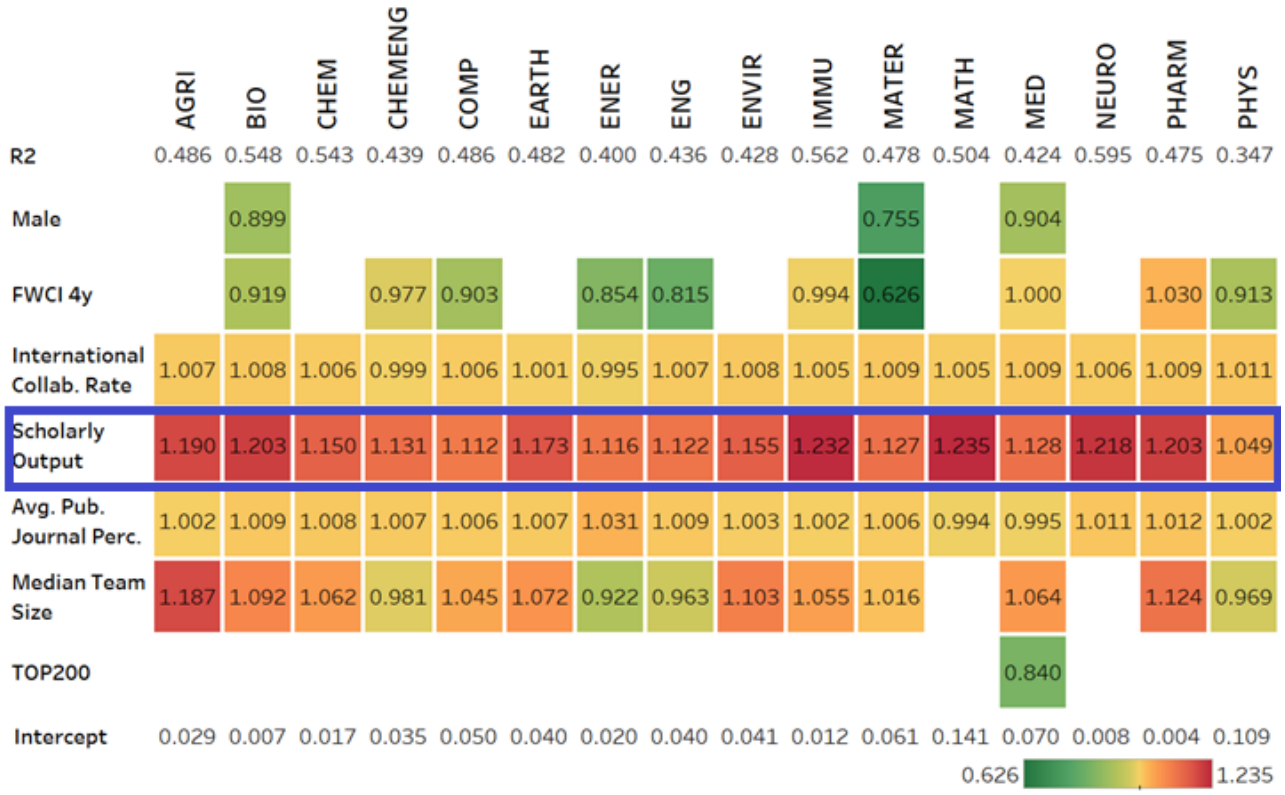
20. Logistic Regression Analysis

- **To estimate the odds ratios of staying in science** from a **multidimensional** perspective.
- Success = entering the 30% of scientists from the 2000 cohort who stayed in science after 19 years.
- **Is scientific career relatively predictable, based on mostly research-related requirements?**
- We hypothesized important roles of: **(1) high research productivity, (2) publishing in high-impact journals, and (3) gender.**



21. Logistic regression analysis, odds ratio estimates for staying in science. Cohort 2000 (N=142,776), publication period 2000–2019 (20 years)

- **A powerful message: publication number (variable: Scholarly Output) is the single most influential predictor of publishing in the following year!**
- **Lifetime (cumulative) number of publications** statistically significant for every discipline.
- **An increase in the total number of publications by one - increases the odds of** publishing in the following year on average **by about 10%-20%**.
- **Quality of publications** matters less (two variables): citations marginally, journal prestge in some disciplines only.
- 38 countries combined (diferent systems!)



	AGRI	BIO	CHEM	CHEMENG	COMP	EARTH	ENER	ENG	ENVIR	IMMU	MATER	MATH	MED	NEURO	PHARM	PHYS
R2	0.486	0.548	0.543	0.439	0.486	0.482	0.400	0.436	0.428	0.562	0.478	0.504	0.424	0.595	0.475	0.347
Male		0.899									0.755		0.904			
FWCI 4y		0.919		0.977	0.903		0.854	0.815		0.994	0.626		1.000		1.030	0.913
International Collab. Rate	1.007	1.008	1.006	0.999	1.006	1.001	0.995	1.007	1.008	1.005	1.009	1.005	1.009	1.006	1.009	1.011
Scholarly Output	1.190	1.203	1.150	1.131	1.112	1.173	1.116	1.122	1.155	1.232	1.127	1.235	1.128	1.218	1.203	1.049
Avg. Pub. Journal Perc.	1.002	1.009	1.008	1.007	1.006	1.007	1.031	1.009	1.003	1.002	1.006	0.994	0.995	1.011	1.012	1.002
Median Team Size	1.187	1.092	1.062	0.981	1.045	1.072	0.922	0.963	1.103	1.055	1.016		1.064		1.124	0.969
TOP200													0.840			
Intercept	0.029	0.007	0.017	0.035	0.050	0.040	0.020	0.040	0.041	0.012	0.061	0.141	0.070	0.008	0.004	0.109

0.626 — 1.235

22. Conclusions from Logistic Regression Analysis

- **Somehow surprising results!** A new large-scale data context - for extant (national) literatures.
- **The role of gender much smaller than expected**: *in the presence of other variables*, being male not a statistically significant predictor!
- **Quantity of publications *more* important** as a predictor than their **quality** (all other things being equal).
- **Quality-related independent variables related to publishing *less* important than expected**: **journal prestige** matters more than (field-normalized) **citations**.

23. Methodological (& Ontological) Takeaway Messages: Global Studies

- **Nationally: bibliometric** data can be **merged with administrative** and biographical data. But **datasets for a few countries only** (e.g., USA, Norway, Poland).
- **Globally: biographical information** (gender, year of birth, national discipline, employment history, promotions) is not a**vailable!**
- **All scientists registered nationally** = replaced with **publishing-only scientists indexed by Scopus** (or WoS).
- **Real scientists** with national IDs = replaced with **Scopus Author IDs**.
- **Perfect national** admin & biographical data (registries) = replaced with **inferred data or proxies**.
- Global studies are useful for **moving beyond national analytical containers** and toward **disciplines globally –** global academic profession.
- **Trade-offs needed to test new ideas – and to use new data to test old ideas!**

24. Final Words on Attrition in Science (1/2)

- **Attrition in science requires longitudinal, global datasets to study -** if we want to move **beyond single countries and examine it over time**.
- "Leaving science" is **undergoing significant transformations** as new cohorts enter science. New (working, professional, other) conditions!
- **Attrition in science means different things for men & different things for women in different disciplines**
- **Attrition in science means different things for scientists from different cohorts (entering the scientific workforce).**

25. Final Words (2/2)

- **Gender differences in attrition in science** are ***smaller*** **with each successive cohort** (results only for attrition, not participation!)
- **Traditional assumptions** about how scientists disappear from science – may need **careful revisions**.
- **New conditions – new data – new analyses – new policy implications?**
- Generally, attrition today is **very high, on the rise** (60% of scientists from cohort 2010 disappear within 9 yrs).
- **So, first: attrition is an issue (for both M & W)!**
- **So, second: job attractiveness, working conditions, career opportunities for scientists (M & W) - increasingly matter!**
- **So, third: more explanations needed: why** do we massively leave science, academia, research? (qualitative studies, accompanying Big Data analyses)… **pull** vs. **push** factors…

- Thank you! Contact? marek.kwiek@amu.edu.pl, Twitter: @Marek_Kwiek