Tracking knowledge production among European PhD graduates: publication and exit patterns over time

Early draft – please do not cite or quote

Alberto Corsini[1], Johannes Koenig[2,3], Burcu Özgün[2,4], Andriy Romanyuk[5,6], Guido Buenstorf[2], Francesco Lissoni[5,7], Ernest Miguelez[5,8], Michele Pezzoni[9] & Catalina Martinez[1]

[1]Institute of Public Goods and Policies, IPP-CSIC – Spanish National Research Council

[2]INCHER and Institute of Economics – University of Kassel

[3]IAB – Institute for Employment Research, Saarbruecken, Germany

[4]Department of Human Geography and Spatial Planning, Utrecht University

[5]BSE UMR CNRS 6060 – Université de Bordeaux

[6]Department of Economics – Università dell'Insubria

[7]Department of Economic Policy – Università Cattolica del Sacro Cuore

[8]AQR-IREA – University of Barcelona

[9]CNRS-GREDEG – Université Côte d'Azur

June 22, 2025

This work has been carried out in the framework of the international research project "DOC-TRACK: STEM Doctoral Graduates and Inventive Activities in European Countries" funded by the European Patent Office Academic Research Program (EPO-ARP 2021). Use of publication data from Scopus was enabled by the German Competence Network for Bibliometrics funded by Bundesministerium für Bildung und Forschung (BMBF) under grant number 16WIK2101A. We acknowledge funding from Deutsche Forschungsgemeinschaft (Germany) under grant BU 2454/8-1 (FOR 5234), from German Bundesministerium für Bildung und Forschung under grant 16RBM1011, and from the national research project INNDOC AEI PID2023-149135NB-IOO (Spain). We thank the Spanish Ministry of Science, Innovation and Universities for access to the Spanish data; ABES for access to the French data; DNB for access to the German data; DANS for access to the Dutch data; ONB for access to the Austrian data. We are thankful to Pauline, Carla, Luca, Francesca, Bradley, Araissi, Valentine, Baptiste, Lucien, and Samuel for their excellent research assistance. We thank participants at the STI2024 Conference held in Berlin for their comments. Research results and views included in the paper are only those of the authors and do not necessarily represent the views of funding institutions.

Abstract

Doctoral education plays a central role in the production of scientific knowledge and the training of future researchers, with PhD programs adopted as the dominant model of research training worldwide. Yet systematic cross-country analyses remain rare, even though national systems vary substantially in the structure of doctoral training and in the career trajectories of PhD graduates. We use newly linked large-scale data from five European countries to provide comparative evidence on publication activities during and after the PhD. We match national Electronic Theses and Dissertations (ETD) records from Germany, France, Spain, the Netherlands, and Austria with PhD defense dates from 2000 to 2020 (N = 784,727) to author profiles from the Scopus bibliometric database in the years 1996 to 2022. The resulting data allow us to trace publication trajectories across fields, countries, and cohorts. We find marked differences in early publication activity and show that publishing during the PhD is strongly associated with sustained research engagement after graduation. At the same time, the likelihood of long-term research activity has declined in recent cohorts, and has become increasingly concentrated in academic institutions. These developments may suggest a growing segmentation of science labour markets and a diminishing contribution of industry to open scientific knowledge production. The novel record linkage method developed in this study is scalable and transferable to other contexts, providing a foundation for future comparative research on doctoral education and scientific careers across additional countries.

JEL classification: I23, J24, O57

Keywords: STEMM PhD graduates; PhD theses; Publications; Careers; EU countries

0 Introduction

In various parts of the world, the PhD has become the dominant model for training early-career researchers who aspire to pursue academic careers. The PhD attests the doctoral candidate's ability to perform independent research. It is based on several years of doctoral education, during which PhD students acquire knowledge at the frontier of science as well as research skills. Moreover, students directly contribute to their lab's scientific production, with their scientific contributions becoming comparable to those of senior scientists in terms of both quantity and quality (Corsini et al., 2022; Larivière, 2012; Shibayama, 2022).

Despite the global diffusion of PhD training, considerable heterogeneity across national university systems has persisted in several dimensions, such as in the relative importance of structured coursework and on-the-job training during doctoral education, the time required to obtain a PhD, and the extent to which PhD research is published in scientific journals. In addition, national labour markets differ markedly in the extent to which non-academic employers reward holding a PhD degree.

A growing body of literature investigates various aspects of doctoral education, including PhD graduates' productivity and career trajectories (Heinisch et al., 2020; Recotillet, 2007; König, 2022; Kahn and Ginther, 2017). A general pattern emerging from this literature is that the number of PhD degrees awarded has increased significantly in several countries (Buenstorf et al., 2023; Cyranoski et al., 2011).

At the same time, the share of graduates holding faculty positions has declined (Buechele et al., 2025; Schillebeeckx et al., 2013), with the majority of science and engineering PhDs eventually taking positions outside of academia (König et al., 2025). These trends have given rise to concerns that imbalances between the increasing supply of graduates and the limited number of available faculty positions may force many PhDs away from careers in academia (Roach and Sauermann, 2017; Sauermann and Roach, 2012). Yet, there is increasing demand for PhD graduates from research-oriented start ups and established companies in sectors with a strong reliance on complex technologies and specialized knowledge workers (Martinez and Parlane, 2023).

Systematic comparisons of such developments across countries have been complicated, however, by a lack of unified datasets. For the same reason, researchers have remained limited in

their ability to assess the impacts of national reforms in doctoral education (such as the shift to an Anglo-Saxon graduate school model replacing individual supervision of doctoral students), or in the academic career models (such as variations in recruitment policies or the introduction of tenure-track models). Against this backdrop, this paper explores how new, internationally comparable data can be utilized to enhance our understanding of PhD graduates' knowledge production during and after the PhD training period. To do so, we have matched information about the population of STEMM (Science, Technology, Engineering, Mathematics, and Medicine) PhD graduates (N = 784,727) from national Electronic Theses and Dissertations (ETD) repositories in five European countries (Germany, France, Spain, Netherlands, and Austria) for the period 2000-2020 to information about 35,624,048 scientific authors from the Scopus worldwide publications database for the period 1996-2022.

Based on these data, we provide systematic evidence on PhD graduates' knowledge generation, in terms of publication productivity, during and after the PhD training period in the five countries, across fields and over time, for cohorts spanning 21 years (2000-2020). We begin by documenting how publication activities during the PhD program differ by country, field, and cohort. We then trace post-graduation publication activities, where ongoing publication activities provide evidence that PhD graduates remain active researchers in the early, middle, and late times after the PhD. Lastly, we compare how academic and industrial research careers are related to individual publication activities during doctoral education in the five countries, as well as by field and across cohorts.

Building the large-scale dataset on which these analyses are based posed substantial conceptual and computational challenges. A six-step methodology was developed to address these challenges. First, we cleared and harmonized names and surnames in both national ETD repositories and a widely used bibliometric database, i.e., Elsevier Scopus. Second, for each ETD repository, we produced all possible combinations of first names, initials, and surnames for each graduate and matched them to all the name variations associated with the Scopus author IDs to create a list of potential PhD-author matches. Third, for each ETD repository, we produced a manually curated dataset of PhD-author true and false matches from which we built training, test, and evaluation sets. Fourth, we applied the random forest machine learning algorithm, trained on the training sets, to classify positive and negative matches. The final linked dataset demonstrates high levels of

precision and recall rates.

This systematic record linkage procedure is transparent and applicable to various national contexts, and the data generated can be applied to a variety of research questions and international comparative studies. Our study aims to address the gap in the lack of comprehensive studies that cover different national contexts.

Our findings reveal pronounced differences in early publication activity across countries and fields. We show that publishing during the PhD has become more common over time and that early publication output is a strong predictor of sustained research engagement after graduation. However, the likelihood of long-term publication activity has declined across cohorts since the early 2010s, suggesting increasing difficulty in maintaining research careers. Moreover, we find that sustained publication engagement is increasingly more likely to be concentrated in academic institutions, while the probability of PhD graduates contributing to scientific publishing from corporate settings is shrinking. These patterns point to a growing segmentation of science labor markets and raise questions about the long-term role of industry in open scientific knowledge production.

The remainder of the paper is organized as follows. Section 1 describes the various national data sources to retrieve information about PhD recipients. Section 2 outlines the record linkage methodology used to match dissertations with Scopus author profiles. In Section 3 we describe stratgey to tracing doctorate recipients' publication activities during time and between countries. Section 4 reports the main regression results. Section 5 concludes.

1 Data: ETD and publication repositories

To investigate the publication patterns of 21 graduation cohorts of STEMM PhDs (2000-2020 defense years), during and after the PhD, we draw on two complementary sources of large-scale administrative and bibliometric data.

1. ETD (Electronic Doctoral Theses) repositories of France, Germany, Spain, Austria, and the Netherlands, maintained, respectively, by ABES, DNB, the Spanish Ministry of Science, Innovation and Universities, ONB, and DANS.

2. The Scopus publication database, maintained by the academic publisher Elsevier.

The national ETD repositories contain information on 166,607; 420,927; 113,772; 29,496; and 53,925 STEMM theses for the period 2000–2020 for France, Germany, Spain, Austria, and the Netherlands, respectively. Scopus contains a total number of 35,624,048 authors who have published at least one article between 1996 and 2022, which we use as the pool of authors to match with the ETD graduate students.

The methodology, based on the linkage between ETD data and publication data, offers several advantages: it provides retrospective, relatively unbiased information on the full population of doctoral graduates; it is not subject to the attrition or selective participation typical of survey-based studies; and it allows for consistent cross-national comparisons over time.

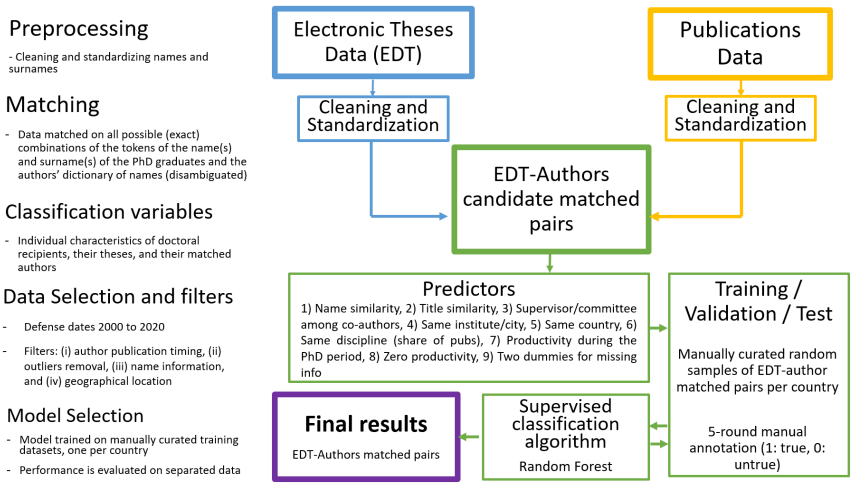
The next section provides a detailed explanation of how we match ETD graduates with authors of Scopus publications.

2 Linking doctorate recipients from five countries with their publication outcome

The objective of our linkage approach is first to identify as many potential matches as possible between ETD graduates and publication authors (i.e., maximize recall), and then to minimize the number of incorrect linkages (i.e., improve precision). This problem can be framed as a classification task, which can be addressed using machine learning methods (see, e.g., Heinisch et al., 2020; Gareth et al., 2013). Various machine learning algorithms are available for classification tasks, including supervised and unsupervised methods (Christen, 2012; Bishop, 2006; Heinisch et al., 2020; Rehs, 2021; Donner, 2022). We opt for a Random Forest supervised classifier, and frame our classification process into six key phases: parsing, merging, creation of training-validation-test sets, feature engineering, training and evaluation of the algorithm, and classification. See Figure 1 for an overview.

- **Parsing:** In this initial phase, the ETD and Scopus data are converted into a structured format suitable for linkage. This involves extracting and standardising relevant information from the ETD repositories and transforming name information into a consistent format for matching.

Figure 1: ETD-Scopus record-linkage approach



- **Merging:** The second phase focuses on merging entries from the ETD and publication repositories. Since an exhaustive n-to-m comparison of all entities in the ETD repositories with all authors in the Scopus database would be computationally prohibitive, we merge datasets based on name information that serves as an effective pre-selection criterion for linkage (Schnell et al., 2004, cf.).

- **Creation of Training, Validation, and Test datasets:** For our supervised learning approach, we constructed training, validation, and test sets based on manually curated gold standard datasets of merged ETD authors - publication authors, one for each country of analysis. These sets are used to train the model and evaluate its classification performance.

- **Feature Engineering:** This step involves creating predicting variables from raw data, such as name similarity scores and other relevant attributes, that the model can use to classify matches accurately. These features help identify true-positive and true-negative matches, ensuring that the machine learning model can distinguish between correct and incorrect linkages.

- **Training and evaluation of the algorithm:** In this step, a Random Forest algorithm is trained and evaluated using the manually constructed, curated training, validation, and test sets mentioned above. This step allows the model to learn and adapt to the structured data. It involves iterative processes for adjusting the model parameters to optimise the classification

performance. The algorithm with the highest F1 score is selected for the final classification step.

- **Classification:** For the final classification task, we apply the Random Forest classifier algorithm to the full set of graduate-author matches obtained from the Merging step, and identify the true positive matches.

The resulting matched samples include 248,962 STEMM PhD graduates for Germany, 140,408 for France, 80,982 for Spain, 20,016 for Austria, and 35,418 for the Netherlands. Further details on our methodological approach linking EDT author with their publication outcome can be found in the Appendix A.

3 Tracing doctorate recipients' publication activities

Building on the dataset created through the record linkage approach described above, we examine publication trajectories before and after graduation to gain insight into the research activity and career paths of PhD holders in five European countries.

1. First, we describe how publication activities during the PhD differ across countries, disciplines, and graduation cohorts.

2. Second, we examine whether PhD graduates continue to publish after graduation and, if so, for how long.

3. Third, we analyze whether publication-active graduates are affiliated with academia or industry research, and how these outcomes relate to early publication behavior.

To address these questions, we utilize the previously presented large-scale, harmonized dataset that links STEMM PhD graduates from Austria, France, Germany, the Netherlands, and Spain to their publication profiles in Scopus. We decided to select only STEM graduates that we can correctly assign to four sub-fields, namely engineering, life sciences, mathematics and computer sciences, and physics. In other words, we remove from our sample graduates in medicine and those who have a thesis in an interdisciplinary field. Medicine is excluded due to the high proportion

of medical doctors who work as practitioners, whose publication patterns differ substantially from PhD students in other fields, and because of the considerable cross-country heterogeneity in the educational requirements for practicing the medical profession. We remain with 18,589 graduates from Austria, 202,995 graduates from Germany, 86,445 graduates from Spain, 143,524 graduates from France, and 22,823 graduates from the Netherlands.

We estimate linear regression models to analyze two types of outcomes: (1) the probability of being publication-active and (2) the number of publications among those who publish. We define "PhD publication activity" as any Scopus-indexed article published from five years before to one year after the defense. For post-PhD activity, we track publications in three subsequent windows: 2–3 years (short term), 4–5 years (medium term), and 6–10 years (long term) after graduation. We estimate the following linear regression model:

$$Y_{ict} = \alpha + \sum_{j=1}^{J} \beta_j \left(\text{Country}_i \times \text{Cohort}_t\right)_j + \gamma X_{ict} + \varepsilon_{ict} \tag{1}$$

where:

- Y_{ict} denotes the outcome for individual i in country c and cohort t. Outcomes include (1) a binary indicator for post-PhD publication activity and (2) the number of publications among those who publish.
- $\left(\text{Country}_i \times \text{Cohort}_t\right)_j$ is a set of interaction terms between country of graduation and PhD cohort (graduation year grouped into intervals or single years).
- X_{ict} is a vector of control variables including field of study, gender (as inferred from first names), and base effects of Country_i and Cohort_t. For regressions explaining long-term research activities, we also include a measure of PhD cohort size (standardized relative to 2000) to capture changes in the scale of doctoral training and the resulting competition for research jobs.
- ε_{ict} is the error term.

For interpretability, we estimate linear probability models for binary outcomes. All models are descriptive and do not aim to identify causal effects.

While the models control for observable differences, we do not aim to identify causal effects. Linear probability models are chosen for interpretability; logit models yield similar results.[1]

4 Cross Country Patterns in Doctoral Education

4.1 Expansion of doctoral education

The past two decades have seen a notable expansion of doctoral education in STEM fields across all five countries in our sample. While the overall trend is similar, with more PhDs trained year by year, there are pronounced differences in scale and growth rates. Germany, for example, has maintained a consistently high number of PhD graduates, producing close to 9,400 STEM doctorates in 2020. The Netherlands, by contrast, started from a much lower level but saw a remarkable increase of over 200% (see Table 1).

Table 1: Growth in STEM PhD Graduates by Country (2000–2020)

Year	Austria	Germany	Spain	France	Netherlands
2000	496	8,854	2,939	5,524	439
2020	934	9,394	3,745	6,521	1,396
% Change	88%	6%	27%	18%	218%

This dynamic has important implications for academic labor markets. While the absolute number of STEM doctorates is growing, it remains unclear whether this expansion has been matched by corresponding increases in research positions—an issue we revisit when examining long-term research trajectories.

4.2 Field Composition

The disciplinary makeup of STEM PhDs differs slightly across countries. Table 2 shows the relative shares of Engineering, Life Sciences, Mathematics, and Physics among doctoral graduates. France and the Netherlands stand out for their emphasis on Physics, with nearly 40% of PhD graduates

[1] Two technical caveats deserve mention. First, although the record linkage procedure performs well across all countries, small differences in recall of the Random Forest algorithm across countries may affect cross-country comparisons. Second, institutional differences in dissertation requirements (e.g., mandatory publication vs. monograph format) shape incentives to publish and should be kept in mind when interpreting differences in early publication rates.

in this field. Germany and Spain also have strong Physics representation, whereas Austria shows a more balanced distribution across disciplines. Life Sciences account for around one-third of all STEM PhDs in Austria and Spain, suggesting different research infrastructures and traditions. These differences are not merely academic: field composition influences publication patterns, labor market trajectories, and international mobility, all career indicators we analyze in later sections, which need to be taken into account when interpreting stay rates in academia.

Table 2: STEM Field Distribution by Country

Country	Field	Nbr	Share
Austria	Engineering	4,725	25.4%
	Life Sciences	6,084	32.7%
	Mathematics and Computer Science	2,394	12.9%
	Physics	5,386	29.0%
Germany	Engineering	59,259	29.2%
	Life Sciences	55,938	27.6%
	Mathematics and Computer Science	22,052	10.9%
	Physics	65,746	32.4%
Spain	Engineering	17,419	20.2%
	Life Sciences	26,899	31.1%
	Mathematics and Computer Science	12,148	14.1%
	Physics	29,979	34.7%
France	Engineering	28,154	19.6%
	Life Sciences	31,865	22.2%
	Mathematics and Computer Science	27,286	19.0%
	Physics	56,219	39.2%
Netherlands	Engineering	4,580	20.1%
	Life Sciences	6,341	27.8%
	Mathematics and Computer Science	2,948	12.9%
	Physics	8,954	39.2%

4.3 Gender Disparities

Table 3 shows the gender composition of PhD graduates by country. Spain comes closest to gender parity with 45% women among STEM PhDs. In Germany, by contrast, only 30.5% of STEM doctorates were awarded to women between 2000 and 2020. Austria, France, and the Netherlands fall somewhere in between. These gaps are relevant not only from an equality perspective but also because they shape aggregate productivity and influence retention in academic careers, as we will show.

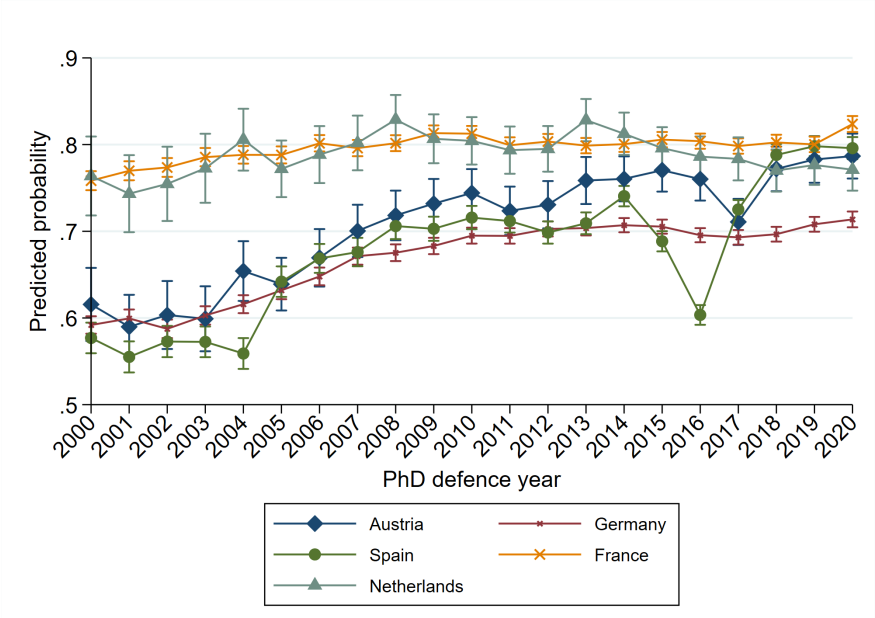
Table 3: Gender Composition of STEM PhD Graduates (2000–2020)

Country	Male	Female
Austria	65.8%	34.3%
Germany	69.5%	30.5%
Spain	55.1%	44.9%
France	64.0%	36.0%
Netherlands	67.8%	32.3%

4.4 Publishing during the PhD

As shown in Figure 2, in some countries, there is an increasing likelihood to publish during the PhD. In France and the Netherlands, PhD students have had a consistent high probability to publish at least one article during the PhD training period (around 0.8) in the past two decades, controlling for differences across disciplines. In contrast, Germany, Austria, and Spain started from lower baselines (around 0.6 in the early 2000s), but have caught up substantially since then (between 0.7 and 0.8).

Figure 2: Predicted probability of publishing during the PhD by country and graduation cohort

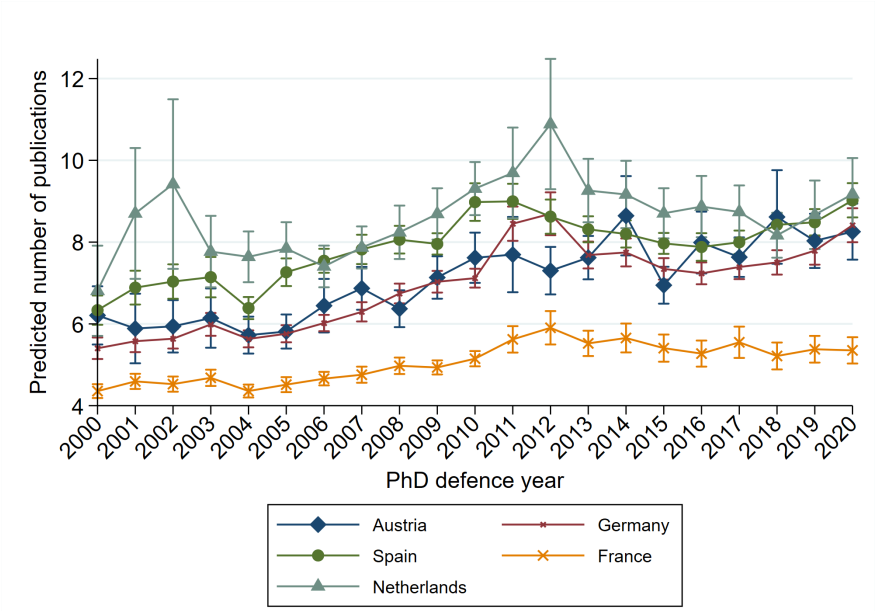


These trends reflect changes in the requirement for receiving a PhD, where at some universities and faculties, the publication of at least one article became a requirement for the academic article,

and broader shifts in academic expectations regarding requirements for a tenured position after the PhD.

Among those who publish, the predicted average number of publications differs substantially, as shown in Figure 3: in Spain, the Netherlands, and Austria PhD students are likely to publish on average around six articles per person (controlling for differences across disciplines); for France, despite the high predicted probability of publication participation, we find the lowest productivity likelihood (under four publications). Our regressions indicate that field-specific patterns are strong. Physics leads in publication intensity, followed by Engineering. Life Sciences and Mathematics show lower output. Gender differences are consistent and substantial: across all five countries, women publish less frequently during the PhD.

Figure 3: Predicted number of publications during the PhD by country and graduation cohort



4.5 Post-PhD publishing: Patterns of persistence and exit

Not all publication-active graduates during the PhD training period remain engaged with publishing at a later stage. A high share of students may stop doing research after graduation and pursue new career paths, for example, in the administration or management of companies, or may join

Figure 4: Predicted probability of publishing 2 to 3 years after the PhD by country and graduation cohort

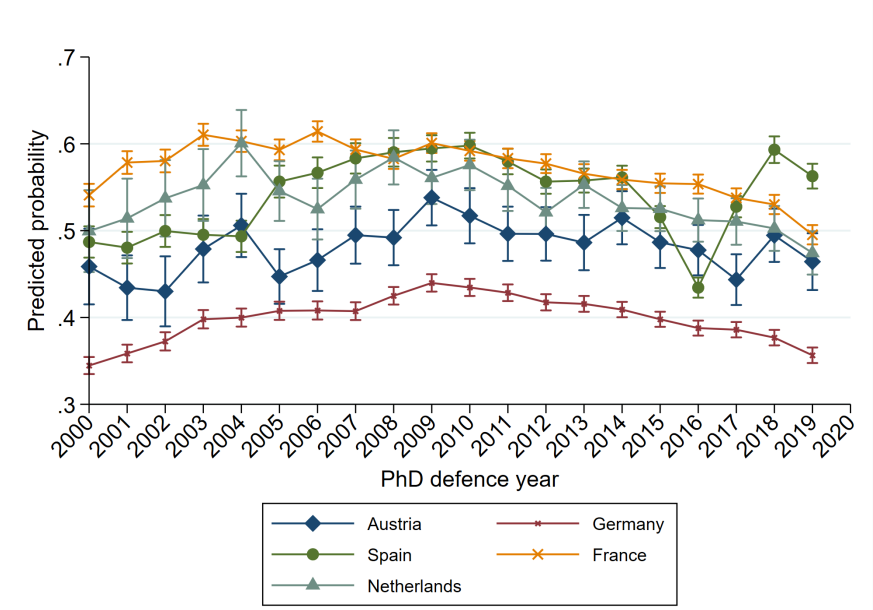
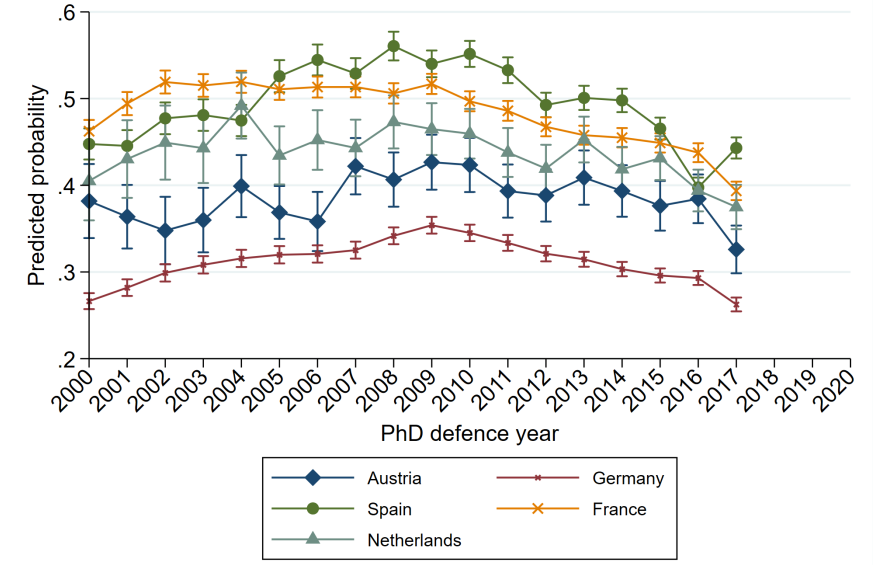


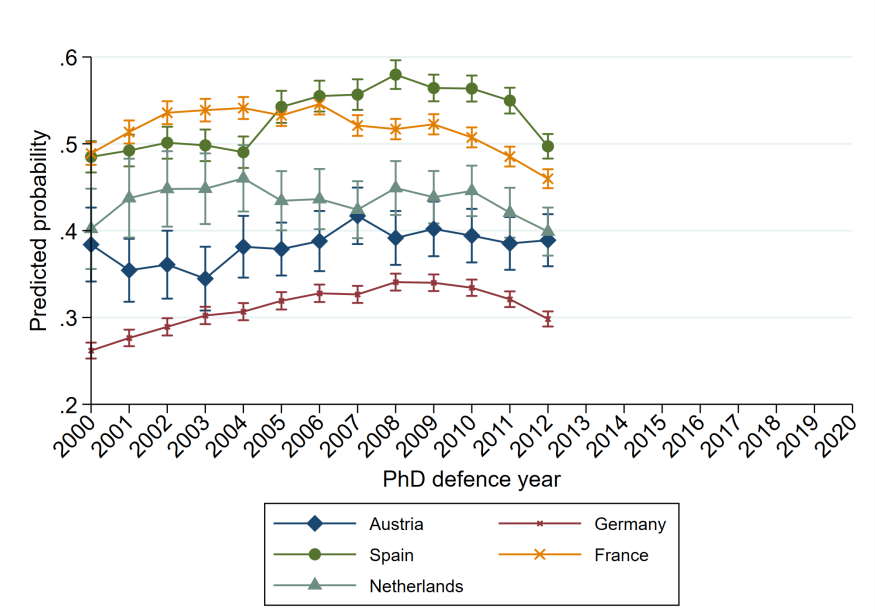
Figure 5: Predicted probability of publishing 4 to 5 years after the PhD by country and graduation cohort



private research labs where the norm is to protect research results with secrecy. The largest drop in publishing occurs immediately after graduation (as comparison of Figure 2 and Figure 4 reveals).

In contrast, graduates who, after graduation, remain in academia or publish while working in

Figure 6: Predicted probability of publishing 6 to 10 years after PhD by country and graduation cohort



the private sector are likely to remain research-active in the long run (see Figure 5 and Figure 6):

- In France, the predicted probability of publishing graduates drops from 0.75 during the PhD to below 0.50 ten years later.
- Other countries show similar trajectories, with different baseline levels.

An interesting insight is that until the late 2000s, the likelihood of researchers being publication active in nearly all countries in the medium and long term after graduation was increasing. However, since the early 2010s, the likelihood of long-term publishing PhDs has started to decline across all five countries. This pattern may be linked to the increasing size of PhD cohorts or the emergence of more temporary or non-research academic positions. It may also signal a realignment of research careers away from publishing roles. Indeed, in recent years, the share of Higher Education Professionals (HEPROs) staff expanded at universities (cf. Schneijderberg, 2020). This consists of a group of individuals who often hold a PhD and work in research support areas, such as technology transfer offices or third-party funding offices. Importantly, a higher likelihood of PhDs exiting from publishing in the long run is not necessarily a negative outcome. Many PhD graduates find fulfilling, high-level employment in non-academic sectors—often applying their skills in indus-

try, management, or policy (cf. König et al., 2025; Agarwal and Ohyama, 2013), where especially early exit out of the academic system can have long-lasting positive effects on earnings (Recotillet, 2007; König, 2022; Kahn and Ginther, 2017)

4.6 Who Stays—and Where?

Who is more likely to continue publishing long after the PhD, and where do they work? We now present findings from our regression exploring the probability of staying active in academia 6 to 10 years after graduation. We first estimate the general probability of continuing publishing for those who already had a publication during the PhD period. For those who continue to publish in the long run, we investigate the explanatory factors of publishing with a corporate affiliation or a non-corporate affiliation.

First, publication experience during the PhD training period appears to be a strong predictor of long-term research activity. PhD graduates who published more during their doctorate are significantly more likely to remain active in research a decade later. This suggests that early integration into the scientific publication system creates durable advantages, perhaps by signaling competence or facilitating access to postdoctoral opportunities. In contrast, holding a patent during the PhD does not predict continued publication activity.

Second, we find persistent cross-country differences. Graduates from Spain are most likely to remain publication-active, while German PhDs are least likely to do so. This pattern may reflect structural differences in labor market opportunities for PhD outside Research and postdoctoral employment systems. In Germany, for example, attractive and relatively research-distant non-academic career paths may draw early-career researchers out of publication-intensive roles (Koenig, 2024). In Austria, by contrast, we observe a stronger transition likelihood into corporate research, suggesting that industry-based research careers play a more prominent role in absorbing trained scientists.

Third, early career indicators also predict sectoral affiliation among long-term publishing researchers. Those who had higher publication output and patent activity during their PhD are more likely to publish from corporate institutions later on. Conversely, long-term academic researchers—those affiliated with universities or public research institutes— tend to have a lower publication probability during their PhD. This result may indicate different selection mechanisms

Table 4: Regression results on long-term publication activity and sectoral affiliation

VARIABLES	(1) Publication 6-10 years later	(2) Corporate affil.	(3) Non-corporate affil.
Nbr. pub PhD	0.003***	0.001***	-0.001***
Patent PhD = 1	0.001	0.114***	-0.114***
aff_foreign_early_reg = 1	0.344***	-0.029***	0.029***
aff_corporate_early_reg = 1	0.103***	0.417***	-0.417***
Germany	-0.099***	-0.025***	0.025***
Spain	0.188***	-0.104***	0.104***
France	0.050***	-0.109***	0.109***
Netherlands	0.041***	-0.052***	0.052***
2001	0.016***	-0.004	0.004
2002	0.027***	-0.001	0.001
2003	0.026***	-0.010*	0.010*
2004	0.030***	-0.003	0.003
2005	0.033***	-0.016***	0.016***
2006	0.033***	-0.009	0.009
2007	0.020***	-0.010*	0.010*
2008	0.027***	-0.013**	0.013**
2009	0.019***	-0.014**	0.014**
2010	0.008	-0.010*	0.010*
2011	-0.010*	-0.004	0.004
2012	-0.040***	-0.014**	0.014**
Female	-0.050***	-0.037***	0.037***
Missing gender	-0.006	-0.021***	0.021***
Change in PhDs since 2000 (%)	-0.001***	0.000	-0.000
Constant	0.462***	0.204***	0.796***
Observations	185,562	104,947	104,947
R-squared	0.153	0.135	0.135

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

across sectors. The industry may prefer highly productive early-career researchers with demonstrated applied research output.

In terms of demographic patterns, we find that male PhD graduates are slightly more likely to end up publishing from corporate affiliations. The gender gap does not seem to be large.

Cohort size also plays a role. Larger PhD cohorts are associated with a lower probability of remaining publication-active in the long run. This suggests intensified competition for a limited number of stable research positions.

Finally, we document a subtle yet significant trend: the likelihood of having long-term publishing researchers affiliated with the corporate sector has declined over time. In other words, the locus of

sustained scientific publishing is increasingly likely to be shifting toward academic institutions. This pattern may reflect a retreat from open science by industry, stricter corporate disclosure policies, or a strategic shift toward proprietary research outputs.

5 Conclusion

This study demonstrates the potential of large-scale record linkage for advancing our understanding of doctoral education and research careers. By linking national ETD repositories to individual publication records, we constructed a harmonized dataset covering STEMM PhD graduates from five European countries over a 21-year period. The linkage approach proved to be transferable across different national contexts.

The linkage of ETD data with publication data offers several advantages: it provides retrospective, relatively unbiased information on the full population of doctoral graduates; it is not subject to attrition or selective participation typical of survey-based studies; and it allows for consistent cross-national comparisons over time.

Building on this dataset, we explored the likelihood of observing continuous engagement or exit from publication activities by PhD graduates across different countries and over time, by focusing on STEM PhD graduates over time (excluding medicine). Our findings document differences across countries in the volume of doctoral education. While Germany remains the largest producer of STEM PhDs, relative growth has been particularly strong in Austria and the Netherlands. Field distributions vary markedly, with physics dominating in France and Germany, and a more balanced subject composition in Austria and Spain.

Most importantly, we identify systematic relationships between early publication activity and the likelihood of long-term publishing research careers. Publishing during the PhD is a strong predictor of research activity six to ten years after graduation, across all national contexts. However, the nature of this research activity varies: graduates who published more—and especially those who produced patents—are more likely to continue publishing in corporate settings.

Our results also shed light on broader structural dynamics. We observe a steady decline in the predicted probability of long-term corporate-affiliated researchers among those who remain publication-active, suggesting a possible narrowing of industry's role in open science.

Several limitations of our study merit discussion. First, our focus on Scopus-indexed publications excludes non-indexed and non-English language outputs, which may be more prevalent in certain countries or fields. Second, we are unable to fully observe non-publication-based career success—such as teaching-focused or managerial roles—limiting our ability to interpret publication inactivity as an academic exit.

References

Agarwal, R. and A. Ohyama (2013). Industry or academia, basic or applied? career choices and earnings trajectories of scientists. *Management Science 59*(4), 950–970.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Information Science and Statistics. New York, NY: Springer.

Buechele, S., G. Buenstorf, M. Huegel, J. Koenig, and M. Theissen (2025). Is the dominance of graduates from top-tier universities among tenured faculty driven by prestige or output? evidence from 50 years of university appointments in germany. Discussion Paper 16-2025, MAGKS Joint Discussion Paper Series in Economics.

Buenstorf, G., J. Koenig, and A. Otto (2023, August). Expansion of doctoral training and doctorate recipients' labour market outcomes: evidence from German register data. *Studies in Higher Education 48*(8), 1216–1242. Publisher: Routledge _eprint: https://doi.org/10.1080/03075079.2023.2188397.

Buenstorf, G., J. Koenig, and A. Otto (2024). Keeping up with the max plancks? germany's quest for university excellence and the role of public research institutes in doctoral education. *Scientometrics*, 1–42.

Christen, P. (2012). *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Data-Centric Systems and Applications. Berlin, Heidelberg: Springer.

Corsini, A., M. Pezzoni, and F. Visentin (2022). What makes a productive ph. d. student? *Research Policy 51*(10), 104561.

Cyranoski, D., N. Gilbert, H. Ledford, A. Nayar, and M. Yahia (2011). Education: the phd factory. *Nature 472*(7343), 276–279.

Donner, P. (2022). Algorithmic identification of ph. d. thesis-related publications: a proof-of-concept study. *Scientometrics 127*(10), 5863–5877.

Fan, X., J. Wang, X. Pu, L. Zhou, and B. Lv (2011). On graph-based name disambiguation. *Journal of Data and Information Quality 2*(2), 1–23.

Gareth, J., D. Witten, T. Hastie, R. Tibshirani, et al. (2013). *An introduction to statistical learning*, Volume 112. Springer.

Heinisch, D. P., J. Koenig, and A. Otto (2020). A supervised machine learning approach to trace doctorate recipients' employment trajectories. *Quantitative Science Studies 1*(1), 94–116.

Kahn, S. and D. K. Ginther (2017). The impact of postdoctoral training on early careers in biomedicine. *Nature biotechnology 35*(1), 90–94.

Klie, J.-C., R. E. d. Castilho, and I. Gurevych (2024). Analyzing dataset annotation quality management in the wild. *Computational Linguistics 50*(3), 817–866.

Koenig, J. (2024). Costs and benefits of a formal academic qualification beyond the phd. *Higher Education*, 1–35.

König, J. (2022). Postdoctoral employment and future non-academic career prospects. *Plos one 17*(12), e0278091.

König, J., A. Otto, G. Bünstorf, K. Briedis, F. Cordua, and H. Schirmer (2021). Karriereentscheidungen und karriereverläufe promovierter–zur multifunktionalität der promotion. *Studien im Rahmen des Bundesberichts Wissenschaftlicher Nachwuchs (BuWiN)*.

König, J., A. Otto, K. Briedis, and S. Gäckle (2025). Karriereentscheidungen und -verläufe promovierter unter besonderer berücksichtigung der befristung. Online under https://buwik.de/.

Larivière, V. (2012). On the shoulders of students? the contribution of phd students to the advancement of knowledge. *Scientometrics 90*(2), 463–481.

Recotillet, I. (2007). Phd graduates with post-doctoral qualification in the private sector: Does it pay off? *Labour 21*(3), 473–502.

Rehs, A. (2021). A supervised machine learning approach to author disambiguation in the web of science. *Journal of Informetrics 15*(3), 101166.

Roach, M. and H. Sauermann (2017, September). The declining interest in an academic career. *PLOS ONE 12*(9), e0184130. Publisher: Public Library of Science.

Sauermann, H. and M. Roach (2012, May). Science PhD Career Preferences: Levels, Changes, and Advisor Encouragement. *PLOS ONE 7*(5), e36307. Publisher: Public Library of Science.

Schillebeeckx, M., B. Maricque, and C. Lewis (2013). The missing piece to changing the university culture. *Nature biotechnology 31*(10), 938–941.

Schneijderberg, C. (2020). Higher education professionals, a growing profession. In *The International Encyclopedia of Higher Education Systems and Institutions*, pp. 731–736. Springer.

Schnell, R., T. Bachteler, and S. Bender (2004). A toolbox for record linkage. *Austrian Journal of Statistics 33*(1&2), 125–133.

Shibayama, S. (2022). Development of originality under inbreeding: A case of life science labs in japan. *Higher Education Quarterly 76*(1), 63–75.

Shin, D., T. Kim, J. Choi, and J. Kim (2014). Author name disambiguation using a graph model with node splitting and merging based on bibliographic information. *Scientometrics 100*(1), 15–50.

Appendix A Parsing Publication and ETD Data

The following sections explain each of the six steps applied to merge ETD data and PhDs publication outcomes in detail.

Appendix A.1 Parsing ETD Data

In the first step, we extracted the ETD data from various sources outlined in Section 1. Since these data were collected from different national contexts and, thus, were differently structured, we processed and standardized them to create a common database that is as homogeneous as possible. Specific national information has been taken into account. For Germany, e.g., additional information on reviewers was extracted from the full texts of the theses' title pages. For the Netherlands, name information was supplemented through web searches. The ETD repositories were then transferred into a relational database.

We then processed name information in the ETD databases. We first normalize names across languages and formats. As part of this process, names were transliterated into basic Latin characters (ASCII transliteration) to ensure consistency.[2] Then, we tokenized first and last names and generated "name combinations" based on likely spelling variations across different databases. Additionally, we created combinations using only initials, as some authors list only their first-name initials in publications. The creation of these name combinations follows consistent criteria across countries, with country-specific adjustments to account for national naming conventions.

Appendix A.2 Parsing Publication Author Name Information

Similarly, we preprocessed the authors' given and family names from Scopus publications to match them with the ETD names. First, we removed records where either the given or family name was

[2]We converted all string characters to lowercase, removed text in brackets (e.g., affiliations), and eliminated all non-alphabetical elements (numbers, brackets, hyphens, etc.) except for periods and apostrophes. Additionally, we excluded non-name words (e.g., titles such as "Mr.," "Prof.," or "Madame") and removed single letters at word boundaries between given and family names. Finally, for the first name alias, we apply a generic character transliteration into basic Latin letters. For the second name alias, we perform an additional country-specific transliteration on top of the basic Latin conversion, taking into account the linguistic characteristics of the analysed countries. Specifically, in the latter case, before converting to basic Latin, we replace "ö" with "oe," "ü" with "ue," "ä" with "ae," and "ß" with "ss." Additionally, we convert "m^{a}" to "Maria," "fco" to "Francisco," and "fca" to "Francisca." While the first four transformations are common in German and Turkish names (as Turkish migrants constitute a significant proportion of Germany's population), the latter three are typical for Spanish names.

missing. Then, we applied the same name-cleaning procedures to publication author names as we did for ETD graduate names.

Next, we generated four versions of each author's name. In the first one, we applied a generic character transliteration into basic Latin letters. In the second one, we performed country-specific transliteration—following the same approach as for ETD graduate names—before converting to basic Latin. Additionally, we created two name combinations that account for the structure of the names. The first retains single letters at word boundaries between names. This ensures consistency with the ETD graduate names, where single-letter tokens were preserved in cases where the distinction between given and family names was unclear. The second consists of the initial letter of the given name concatenated with the full family name.

Appendix A.3 Merging Author–PhD Name Combinations

We conducted an m-to-m exact merge between our processed ETD graduate name combinations and Scopus publication author names. Our decision to rely on exact matching is based on two key considerations.

First, both publication author names and ETD graduate names maintain a comparably high quality in terms of name spelling. We therefore assume that names are generally recorded correctly, while the processing steps outlined in Section **??** account for country-specific variations in the usage of name components between dissertation records and publications.

Second, employing fuzzy string matching would require computing string distances for all possible combinations of graduate names in the ETD and publication databases. Given the scale of our dataset—covering graduates over 21 years from five different countries—this approach would be computationally prohibitive. Maintaining computational efficiency is essential to ensure the feasibility of our linkage process.

We merge the name combinations of 784,727 ETD STEMM graduates from 2000 to 2020 from Spain, France, Germany, Austria, and the Netherlands with publication authors' names drawn from a pool of 35,624,048 Scopus authors between 1996 and 2022. Overall, 95.2% of ETD graduates (747,419) are merged to publication authors via name combination matching. Spain has the highest match rate, with 113,356 (99.6%) graduates merged, while Austria has the lowest match rate, with

27,429 (93.0%) graduates merged to publication authors. In France, 163,590 (98.2%) graduates are linked to publication authors, in Germany the figure accounts for 392,034 (93.1%) graduates matched, and in the Netherlands 51,010 (94.6%) graduates are matched.

Matching combinations of ETD graduate names with academic author names leads to high recall while at the same time generates many unlikely matches, inflating the set of candidate true matches. This poses three challenges: computational feasibility, imbalance in the training sample, and the need for manual curation of an inflated training data containing various highly unlikely combinations. To address these issues, we apply an expert-rated rule-based filtering approach to minimize unlikely matches while maintaining true positives. The filters, applied sequentially, exploit information on author publication timing, outliers, name information, and geographical location.

- **Author publication timing:** This filter restricts authors' publications, from which publication author names are retrieved, to a relevant time frame, ensuring that only publications produced around the PhD period are retained.
- **Outliers:** We define as outliers those ETD students merged with too many publication authors. We apply a 5% percentile cutoff for outliers.
- **Name information filter:** In cases where full name information is available in both databases and full names differ between ETD and publication authors, we exclude matches obtained with the combination that uses name initials, ensuring that we do not match "Maria Garcia" with "Mercedes Garcia", "Manuel Garcia", or "Marco Garcia".
- **Geographical location:** The geographical filter ensures that only authors with a link to the specific countries are considered.

Appendix A.4 Creation of Training, Validation, and Test sets

To train our supervised machine learning algorithm, we create a manually curated gold standard dataset for training, testing, and validation, where merges received from Section **??** were manually classified as true or false by our research assistants.

To create the dataset to curate manually, we apply in a first step a random sampling to select a subsample from our merged publication author - ETD graduates data, preserving the distribution of

true and false matches for effective model training. The random samples consist of 4,058 graduate-author pairs for France; 7,159 for Spain; 5,101 for Germany; 5,722 for Austria; and 4,776 for the Netherlands.

The annotation process, which follows in the second step, is performed in five rounds:

- **Round 1 & 2:** Two independent research assistants annotate the gold standard dataset for each country, classifying each publication author - ETD graduate pair as a true, false, or "not sure" match.
- **Round 3:** We then manually resolve disagreements between the two annotators, ensuring consistency in classifications.
- **Round 4:** We check cases of 1-to-m or m-to-1 matches, addressing annotation errors or disambiguation issues in both the ETD data and publication records.
- **Round 5:** A random sample of matches is checked to identify any annotation errors that remain unresolved and assess potential biases in the dataset.

This extensive manual process ensures that the identified true and false matches accurately reflect the underlying true-positive and true-negative cases. Multiple rounds of annotation were necessary to account for human errors in the process, which, according to the literature, can result in error rates of up to 30% (Klie et al., 2024). In our approach, the effect of human error varied across countries, ranging from 0.7 to 0.8 percentage points in Spain to a maximum of 2.8 to 4 percentage points in Germany. The iterative annotation process aimed to minimize these discrepancies and improve the overall reliability of the matching procedure.

Appendix A.5 Feature Engineering

We construct fourteen classification variables, which are listed and detailed in Table 1, which can be grouped into five categories that serve as a proxy for the distance between an ETD graduate and her corresponding matched publication authors, plus an additional category that proxies for the scientific productivity of the publication authors. The first category refers to social distance and uses information on the graduate's supervisors and jury members, as well as the publication author's co-authors. The second category serves as a proxy for geographical distance, utilizing

information on the graduate's defense institute, the city and country of that institute, and the author's affiliations as reported in corresponding publications. The third category refers to content distance. In particular, it considers the titles and disciplines of the graduate's thesis and the author's publications. The fourth group regards the name distance, which compares the similarity between the graduate's name and the Scopus author's name. Finally, the fifth category calculates the publication productivity of an author during the graduate's PhD period.

Table 1. Classification variables used by the machine learning algorithm.

Category	Variable	Description
Social distance	Supervisor is a co-author	Continuous. Maximum value among the Levenshtein similarities between the names of the PhD graduate's supervisor and co-supervisor and the names of the Scopus author's co-authors, considering all the name versions of a Scopus co-author associated with the same Scopus author ID. The Levenshtein similarity is calculated as: 1 – (Levenshtein distance / max length between the two full names)
	Committee member is a co-author	Continuous. Maximum value among the Levenshtein similarities between the names of the PhD graduate's jury committee members and the names of the Scopus author's co-authors, considering all the name versions of a Scopus co-author associated with the same Scopus author ID. The Levenshtein similarity is calculated as: 1 – (Levenshtein distance / max length between the two full names)
	No info on supervisor / committee	Dummy. Value 1 if no information on PhD graduate's supervisor/committee jury members is available.
Geographical distance	Affiliation: city/institute	Dummy. Value 1 if the institute or city where the PhD graduate defended the thesis appears in the affiliation of at least one Scopus author's article that is published between three years before and two years after the graduate's defence year.
	Affiliation: country	Dummy. Value 1 if the country where the PhD graduate defended the thesis appears in the affiliation of at least one Scopus author's article that is published between three years before and two years after the graduate's defence year.
Content distance	Title similarity	Continuous. Maximum similarity among the TF-IDF (Term Frequency Inverse Document Frequency) cosine similarities between the English translation of the thesis title and the English translation of the titles of all the Scopus author's articles published between three years before and five years after the graduate's defence year.
	Same discipline (share)	Continuous. Share of Scopus author's articles published in journals having a Scopus ASJC (All Science Journal Classifications) field that coincides with the graduate's thesis field.
	No info discipline	Dummy. Value 1 if no information on the thesis field is available.
Name distance	Name similarity	Continuous. Maximum similarity among the Levenshtein similarities between the PhD graduate's name and the Scopus author's names, considering all the Scopus author's names associated with the same Scopus author ID. The Levenshtein similarity is calculated as: 1 – (Levenshtein distance / max length between the two full names)
Publication productivity	Normalised productivity	Continuous. Number of normalised articles published by the Scopus author between three years before and two years after the graduate's defence year. The productivity is normalised by dividing it by the average researchers' productivity in the same country, field and PhD period as the focal graduate.
	Zero productivity	Dummy. Value 1 if no articles were published by the Scopus author between three years before and two years after the graduate's defence year.

Appendix A.6 Training and evaluation of the algorithm

To identify true-positive and true-negative links in our merged ETD graduates - Publication authors dataset, we adopt a Random Forest machine learning technique, commonly used for similar

classification exercises (see, e.g., Heinisch et al., 2020; Rehs, 2021). Previous literature shows that graph-based classification algorithms successfully solve similar classification problems (see, e.g., Fan et al., 2011; Shin et al., 2014; Rehs, 2021), while Random Forest often outperforms other algorithms like regularized logistic regressions or AdaBoost (Heinisch et al., 2020). Moreover, the Random Forest performs well in the case of unbalanced datasets and with long vectors, is good at avoiding overfitting, and is fast in terms of computational time.

We build a Random Forest model separately for each country, and train and validate the algorithm using stratified 5-folds cross-validation, with a hyperparameter tuning function of 155 iterations based on the combinations of two hyperparameters of the Random Forest, namely the number of trees (ntree) and the number of variables used at each split (mtry). Specifically, to fine-tune the hyperparameters, we repeat the cross-validation for each combination of ntree and mtry, where ntree ranges from 500 to 2000, increasing in steps of 50, and ntree can take the values of 1, 2, 3, 4, or 10.

Cross-validation is performed on the 90% part of each country's gold standard dataset, which is used together for training and validation. This allows us to maximize the amount of data available for training while still leveraging the validation set for robust tuning. This training-validation set is divided into 5 folds of equal size, in which a stratification procedure preserves the proportion of true positives and true negatives. Then, the training occurs sequentially on four folds, and the validation on the remaining one. In each sequence, the validation fold changes. The final validation is based on the average F1 score across the 5 folds used for validation.

The Random Forest classification algorithm is validated on a validation set and tested on a test sample. The test sample corresponds to the 10% of the gold standard dataset that is not used in the training and validation phases and acts as unseen data, allowing us to test the performance of the model in a real-world scenario. This procedure prevents the problem of overfitting the data, which occurs when the model provides accurate predictions for training and validation data but not for new data on which it needs to be applied. The best model specification can be chosen based on different metrics: recall, precision, or F1 statistics.

The performance of our Random Forest models reaches a precision of 93.5% and a recall of 87.8% for Germany, a precision of 96.1% and a recall of 94.2% for France, a precision of 96.2% and a recall of 89.5% for Spain, a precision of 95.2% and a recall of 88.2% for Austria, a precision of

99.0% and a recall of 95.3% for the Netherlands (See table 2).[3]

Table 2. Performance of the Random Forest, by country.

	Germany	**France**	**Spain**
Best parameter settings	mtry = 2, ntree = 950	mtry = 3, ntree = 1150	mtry = 4, ntree = 750
Cross-Val. Performance			
Precision	94.3%	98.2%	95.3%
Recall	89.5%	93.8%	91.8%
F1	91.8%	95.9%	93.5%
Test Performance			
Precision	93.5%	96.1%	96.2%
Recall	87.8%	94.2%	89.5%
F1	90.6%	95.2%	92.8%

	Austria	**Ntherlands**
Best parameter settings	mtry = 3, ntree = 800	mtry = 4, ntree = 750
Cross-Val. Performance		
Precision	95.5%	97.3%
Recall	91.4%	96.4%
F1	93.4%	96.8%
Test Performance		
Precision	95.2%	99.0%
Recall	88.2%	95.3%
F1	91.6%	97.1%

Appendix A.7 Classification

The selected Random Forest algorithms for each country are then used to classify true-positive matches in the full sample of ETD graduate - Scopus publication author datasets resulting from the name combination matching.[4]

[3]To address class imbalances in the gold standard dataset, we also apply SMOTE (Synthetic Minority Oversampling Technique) to augment the minority class (true positive matches), improving the balance of the classification algorithm. This multi-stage annotation process ensures the accuracy and reliability of the dataset, leading to a more effective and generalizable model. We combined the 5-fold cross-validation with the SMOTE technique by augmenting the true positive class in the training fold of each cross-validation sequence without contaminating the validation folds and the test set. The same procedure described in the main text to evaluate the Random Forest performance is then repeated with the augmented training sample. Applying SMOTE increases the recall of the model for every country, as expected due to augmenting the minority class, but it decreases the precision. Since the F1 performance is only slightly improved when applying the SMOTE, and we consider precision more important than recall, we decided to keep the model without SMOTE as the preferred model.

[4]We end up with cases where a single ETD graduate is matched with multiple publication authors (1-to-m matches) and a single Scopus author is matched with multiple graduates (m-to-1 matches). While the first case can

The true positive PhD-author pairs resulting from the Random Forest classification make up the final sample of matched graduates for the countries analyzed. The final sample for Germany is composed of 248,962 PhD graduates and 257,834 PhD-author pairs; the final sample for France is composed of 140,408 PhD graduates and 150,600 PhD-author pairs; the final sample for Spain is composed of 80,982 PhD graduates and 96,138 PhD-author pairs; the final sample for Austria is composed of 20,016 PhD graduates and 21,013 PhD-author pairs; and the final sample for the Netherlands is composed of 35,418 PhD graduates and 37,033 PhD-author pairs. The corresponding matching quota of PhD graduates relative to the initial population of graduates is the following: 59.1% for Germany, 84.3% for France, 71.2% for Spain, 67.9% for Austria, and 65.7% for the Netherlands.[5]

be justified by the publication data in-house author disambiguation, which is imperfect and assigns different author IDs to the same entity, the latter is a matching error. For both cases, we decided to adopt a conservative approach, which maximises the precision of the matching procedure. Specifically, when a graduate is matched to more than one publication author (1-to-m matches), we keep the two authors matched with the highest probability according to the Random Forest model. In so doing, we may underestimate the productivity of some graduates if their corresponding true publication authors are more than two. However, the cases where graduates are matched to more than two authors in our sample are minimal (less than 1%). When, instead, there is a publication author matched to multiple graduates (m-to-1 matches), we retain the graduate-author pair that is matched with the highest probability. This strategy eliminates the problem of m-to-1 matches but leads to a loss of graduates if they are matched to the same Scopus authors as other graduates, but with a smaller probability. Nevertheless, the loss of graduates is negligible, being less than 1% for all the countries.

[5]Note that Germany's lower percentage of graduates matched with Scopus depends on different role doctorate has in Germany compared to other countries. In Germany, a doctorate can be conducted in a broad range of employment contexts, including university-industry cooperation (Buenstorf et al., 2024). After completing their doctorates, the majority of graduates leave academia (König et al., 2021). This is especially true for medical professions (König, 2022). Therefore, it is not surprising that many German PhD graduates do not publish any scientific papers as they enter a PhD program with the aim of working in the private sector. When excluding PhDs graduated in medicine from the German sample, the quota of graduates matched with Scopus increases to 68.2%.