Research careers in Europe: New evidence from the Doc-Track database

Alberto Corsini [1], Johannes Koenig [2,3], Burcu Özgun [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
[4] Department of Human Geography and Spatial Planning, Utrecht University
[5] BSE UMR CNRS 6060 INRAE UMR 1441 – Université de Bordeaux ; [6] Department of Economics – Universita` dell'Insubria
[7] Department of Economic Policy – Universita` Cattolica del Sacro Cuore ; [8] AQR-IREA – University of Barcelona
[9] GREDEG UMR CNRS 7321 – Université Côte d'Azur

This draft: November 24th, 2025 (please do not quote nor cite)

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

Abstract

We present early evidence from Doc-Track, a new cross-country and longitudinal database on doctoral graduates' research careers in science, technology, engineering, mathematics and medicine (STEMM). Based on all dissertations discussed from 2000 to 2020 and archived in electronic form in Germany, France, Spain, the Netherlands, and Austria, we first identify all STEMM graduates in the same period and retrieve their publications during and after their doctoral spell. Based on this, we confirm results from a number of country-specific studies on the declining share of graduates undertaking a career in research – as measured by scientific authorship - but also find significant cross-country differences in both levels and trends. The declining trend as well as the observed cross-country differences become clearer after controlling for a number of observables, both at the individual and the country level, which suggests they are due to structural cross-country differences in the nature of doctoral training and/or the labour market. When distinguishing between academic and non-academic publications, we find that – conditional on publishing after graduating – the probability to do so outside academia is increasing, albeit being very low. Post-doctoral mobility is also increasing, as measured by the probability to publish with a different affiliation and in a different country than those of the PhD period. This is coherent with both the increasing integration of the European Research Area and the increasing share of international students. The Doc-Track data mining and record linkage methods are scalable and transferable to other countries, which will allow us to create a permanent observatory on doctoral graduates' careers.

JEL classification: I23, J24, O57

Keywords: STEMM; Doctoral training; EDT; Publications; Research careers; EU; Doc-Track

Acknowledgements: This work results from the 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), grant number 16WIK2101A. We also acknowledge funding from the Deutsche Forschungsgemeinschaft (Germany) under grant BU 2454/8-1 (FOR 5234), from the BMBF under grant 16RBM1011, and from the Spanish national research project INNDOC AEI PID2023-149135NB-IOO. We thank the Spanish Ministry of Science, Innovation and Universities, ABES, DNB, DANS and ONB for access to, respectively, the Spanish, French, German, Dutch and Austrian EDT data. Pauline Menu, Carla Fournier, Luca Codecasa, Francesca Ammerata, Bradley Butcher, Aman Araissi, Valentine Petzold, Baptiste Comby, Lucien Boucart, Samuel Desneulin and Clovis Sourisse provided 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.

1. Introduction

Doctoral training has become the worldwide dominant model for early career researchers who aspire to pursue academic careers. The PhD degree attests the doctoral candidates' ability to perform independent research after several years of doctoral education, during which they acquire knowledge at the frontier of science as well as research skills. Moreover, their scientific contributions have become essential to their laboratories' research production (et al., 2022; Larivière, 2012; Shibayama, 2022).Corsini et al., 2022; Larivière, 2012; Shibayama, 2022).

A growing body of literature investigates various aspects of doctoral education, with special emphasis on its private and public economic value. A dominant theme is the observation that the number of PhD degrees awarded has increased significantly over time in most countries with a strong university system, while the share of graduates securing an academic position has declined (OECD, 2021; Sarrico, 2022). These trends have given rise to various concerns, namely: that doctoral students and young graduates, especially foreign ones, are used as temporary workforce in place of permanent laboratory technicians or tenure-track faculty (Stephan and Ma, 2005; Stephan, 2012 and 2013); that imbalances between the increasing supply of graduates and a limited availability of faculty positions may cut too many research careers short (Milojevic et al., 2018; Kwon, 2025); and that the doctoral training and/or the years spent on postdoc contracts may not result in higher productivity and wages when ultimately moving to non-academic jobs (Koenig, 2022 and 2024; Marini and Henseke, 2023). More nuanced views of the career perspective of doctoral holders, especially in science, technology, engineering, mathematics and medicine (STEMM), point out that many of those who do not undertake an academic career decide so early on during their studies and are less attracted to it from the start (Sauermann and Roach, 2012; Balsmeier and Pellens, 2014; Roach and Sauermann, 2017); and that many STEMM graduates may profit of an increasing offer of research-oriented jobs in the business sector (Buenstorf et al., 2023; Martinez and Parlane, 2023). This, in turn, raises the question of the adequacy of doctoral training models for preparing to non-academic jobs and even non-research ones. Two related themes concern the individual determinants of success in establishing a research career, whether in academic or outside it, with the quality of mentorship and gender having been proven to play a significant role, alongside with the prestige and visibility of the degree-granting institution (Hancock 2023; Buenstorf et al., 2025).

The data mobilized by the literature vary considerably, each source presenting its own set of trade-offs. Survey data, collected through questionnaires administered to doctoral students during and/or after graduation, provide detailed information on individual biographical data, motivations, and subsequent employment (in or out the academia, research-related or not). However, they often lack a longitudinal dimension. With the exception of the Survey of Earned Doctorates (Okrent and Burke, 2021; Opsomer et al., 2021) they are generally occasional or highly irregular over time; moreover, they do not follow the same individuals over time and, if they do, they cover a limited time period. They are generally conducted on a national basis and do not track the graduates who go abroad, the main exception being the Careers of Doctorate Holders survey, conducted jointly by the OECD, UNESCO and Eurostat, but only in 2009 (Auriol, 2010; and Auriol et al., 2013). This prevents international comparisons and a detailed analysis of mobility patterns, a limitation whose importance increases with the rise in number of international graduate students (Ganguli and MacGarvie, 2025). Survey data also come - more often than not - in anonymized form, which prevents data linkage to other information sources.

The main alternative to surveys consists in combining archival records on doctoral graduates to other information sources, such a bibliographic databases or social security and other employer-employee databases as well as social media records. In the first case, the main goal is to find out which graduates keep publishing after their graduation and for how long, with the publication activity used a proxy measure for an employment in research and the author affiliation used for establishing the graduate's employer and location over time, conditional on publishing. This strategy is limited to the extent that it assumes that research lead inevitably to publishing (a more appropriate assumption in academia than outside it) and does not allow to find out what jobs the graduates who stop publishing take up and where (for a discussion, see Hanlon, 2019; and Milojevic et al., 2019). At the same time, it has the advantage of allowing for a long-term monitoring of active researchers, including of their mobility and of the quality of their research, especially within the academia. In the second case, information can be retrieved on non-academic careers but access to sensitive data such as employment records is very unequal across countries, which limits the potential for international comparisons. In both cases, the increasing availability of theses and dissertation in electronic form (EDTs) provides the possibility to go back in time and consider at once many cohorts of graduates as well as to update whatever data are collected for the graduation cohorts to come. So far, this possibility has been exploited for country-level study (Geuna and Shibayama, 2015, on Japan; Kahn and MacGarvie, 2016, on the US; Corsini et al., 2022, on France; Carriero et al., 2024, on Italy). The Doc-Track

project's main goal is to exploit it in an international perspective.

To do so, it relies on matching information extracted from EDTs to a variety of bibliographic resources. In this paper, we present early results for 21 cohorts of 784,727 graduates (from 2000 to –2020) in five European countries (Germany, France, Spain, Netherlands, and Austria), matched to more than 35 million scientific authors from the Elsevier Scopus database for the period 1996-2022.

Based on these data, we provide cross-country and longitudinal evidence on the graduates' research careers – as defined by their publication activity at different points in time after graduation. We first trace post-graduation publication activities, where ongoing publication activities, either in academic or non-academic settings, provide evidence that PhD graduates remain active researchers. And then estimate the likelihood of choosing academic and industrial research careers, conditional on publishing. We also explore the mobility patterns of those graduates who remain active in research, across both organizations and countries.

Building a large-scale dataset such a Doc-Track poses substantial conceptual and computational challenges, which we addressed by developing a five-step methodology. First, we cleared and harmonized names and surnames in both national EDT repositories and a widely used bibliometric database, i.e., Elsevier Scopus. Second, for each EDT 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 EDT 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.

We show that the likelihood of having a long-term research career has declined across cohorts and for all countries throughout the last 20 years, net of any cyclical effect due to more R&D funding and other macroeconomic conditions in specific years. Short-term research careers, such as in the years immediately following the graduation, are in decline, too. Moreover, we find that sustained publication engagement is increasingly more likely to be concentrated in selected academic institutions, while the

probability of contributing to scientific publishing from corporate settings is increasing but still low.

At the individual level, the two most important predictors for persistency in research are gender and publications during the PhD. We find that, all else equal and in all countries, female graduates are less likely to have a long-term research career and that the gender gap does not close across cohorts. Yet, important cross-country differences exist and persist. We also find that graduates who engaged in publishing early on during their studies are much more likely to keep publishing later on than those who did not, but that also for them the probability to have a long-term research career decreases across cohorts.

Conditional on having such a career, it is increasingly likely that researchers will change organization and also country to pursue it, with the geographic mobility being most likely explained by the increasing share of foreign graduates in all countries, both from the European Union and outside it.

The remainder of the paper is organized as follows. Section 2 describes the various national data sources to retrieve information about PhD recipients. Section 3 outlines the record linkage methodology used to match dissertations with Scopus author profiles. In Section 4 we describe our strategy to tracing doctorate recipients' publication activities during time and between countries. Section 4 reports the main regression results. Section 5 reports the main regression results and Section 6 concludes.

2. Data and methods

2.1 Data sources

We draw on two complementary sources of large-scale administrative and bibliographic data. Concerning the former, we combine information from five national EDT repositories, all of them originating from legal obligations for universities to either deposit their doctoral dissertations at national libraries or to make them otherwise accessible via official initiatives. They are respectively: *theses.fr*, maintained by ABES, the French national bibliographic agency of higher education; the collections and metadata maintained by DNB, the German National Library Germany; TESEO, maintained by the Spanish ministry of universities; Narcis, a digital platform maintained untile recently by Dutch National Center of Expertise and Repository for Research Data (DANS); and the catalogue maintained by ÖNB, the Austrian National Library.

The French repository is the most stable and complete and it can be regarded as the standard of reference. Each thesis it records identified through a unique code and it comes with title, language, graduate's name

and surname, supervisor and co-supervisors' names and surnames, defense date, discipline, and institution of defense. All other repositories contain similar information, albeit not always as complete and with important cross-country differences. In particular, disciplinary classification is missing for both Austria and the Netherlands, so that we had to infer it from either the dissertations' titles and keywords or the supervisors' publications and their journal classifications. Once solved this problem, we harmonize disciplines across countries by classifying each dissertations into one of five broad categories: i) *Engineering* (comprising General Engineering, Chemical Engineering, and Energy); ii) *Life Science*s (comprising Agricultural and Biological Sciences, Biochemistry, Genetics and Molecular Biology, Immunology and Microbiology, Neuroscience, Pharmacology, Toxicology and Pharmaceutics); iii) *Mathematics & Computer Sciences* (including Mathematics and Computer Science); iv) *Physics* (comprising General Physics and Astronomy, Material Science, Environmental Science, Earth and Planetary Sciences, and also Chemistry as a neighboring discipline); and v) *Medicine* (comprising Medical Science, Nursing, Veterinary, Dentistry, and Health Professions). It is important to stress a peculiarity concerning Medicine in Germany, whose education system equates all medical dissertations produced at the end of single-cycle degrees to doctoral ones, whereas in other countries they are considered as master theses. This makes data for German medical dissertation hardly comparable to those of other countries.

Table 1 reports summary information on the consistency of each repository, in total and for the STEMM subset.

Concerning bibliographic data, our main source of information is Scopus, avery large commercial database maintained by the academic publisher Elsevier. The main reason for choosing this source is the presence and reliability of unique identifiers for authors (Scopus ID), which provides a highly useful research tool when it comes to find the doctoral graduates' publication, based on name matching. To the extent that the same scientific author (with a unique ID) appears with alternative name representations in different publications, we can assign all of them to the doctoral graduates whose name match just one representation.

Table 1. EDT sources and DOC-TRACK coverage

Country	Sources	Organization	Coverage	Number of theses, of which STEMM
France	*theses.fr*	Agence Bibliographique de l'Enseignement Supérieur (ABES)	All universities, 2000-2020	257,739 theses, of which 166,607 STEMM (65%)
Germany	DNB catalogue; DissOnline	Deutsche Nationalbibliothek (DNB)	All universities, 2000-2020	544,237 theses, of which 420,927 STEMM (77%)
Spain	TESEO	Spanish Ministry of Science, Innovation and Universities	All universities, 2000-2020	204,506 theses, of which 113,772 STEMM (56%)
Netherlands	Narcis	Dutch National Center of Expertise and Repository for Research Data (DANS)	All universities, 2000-2020	81,571 theses, of which 53,925 STEMM (66 %)
Austria	ONB catalogue	Österreichische Nationalbibliothek (ONB)	All universities, 2000-2020	49,808 theses, of which 29,496 STEMM (59%)

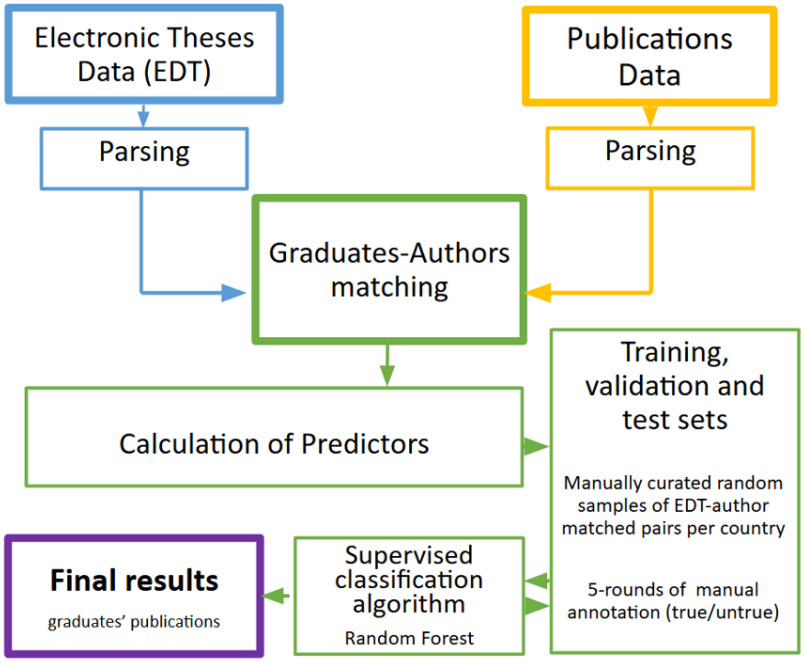
Before proceeding to the match, the author data also required substantial parsing. First, we keep all publication types. Second, we exclude publications with authors lacking all information about either names or surnames but keep cases where only the initial of the first name is included. Third, we standardize both the graduates' and authors' names or the remaining articles by applying the same alphabetical standardization rules, thus ensuring comparability. Last, we generate different name variants for each unique Scopus to account for potential variations in spelling and structure across publications. Once again, full details can be found on the DocTrack technical report.

2.2 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 EDT 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 five key phases: parsing, name matching, creation of training-validation-test sets, training and evaluation of the algorithm, and final classification. See Figure 1 for an overview.

Figure 1: EDT-Scopus record-linkage approach



- Parsing: We first converted the EDT and publications data into a structured format. This included extracting and standardizing relevant information from the EDT repositories and transforming name data into a consistent format across databases for matching.

- Doctoral graduates-authors name matching: Given the computational infeasibility of an exhaustive N-to-M comparison between all the N doctoral graduates in the EDT repositories and all M authors of scientific publications, we applied a pre-selection step based on EDT and publication authors, based on their name similarity (Schnell et al., 2004; Heinisch et al., 2020) and the temporality of the publications with respect to the defence years. We also excluded outliers (very common names

resulting in a disproportionate number of matches) to efficiently reduce the number of comparisons.

- Creation of training, validation, and test datasets: In view of filtering the results of the graduates-authors matching by means of a supervised machine learning method, we constructed a number of manually curated gold standard datasets for training, validation, and testing.

- Training of the classification algorithm: We set a maximum of twelve predictors, per country, as input to the classification algorithms that help predict true positives and true negatives in the original name-based matches. We assign weights to the predictors by means of the Random Forest classification algorithm, making use of the training and validation datasets.

- Final classification and sample construction: We select the best-performing weighing model (the one with the highest F1 score) for undertaking the final classification task on the entire set of graduate-author 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. One technical caveats deserves mention: 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. Further details on our methodological approach linking EDT author with their publication outcome can be found in the Appendix A.

3. Study sample and methodology for tracking doctorate recipients' publication activities

The final study sample is obtained after applying several data refinements. First, we remove outliers in the denominator, which reduces the sample size from 784,727 to 749,530 observations. Second, the usable cohorts are constrained by truncation: for regressions where the outcome corresponds to the short-term period after the PhD, cohorts are included only up to 2017, whereas for regressions examining longer-term outcomes, cohorts are restricted to those graduating up to 2013. Finally, because of missing information on gender, the regressions are estimated on 596,075 graduates for analyses of short-term post-PhD outcomes and on 436,635 graduates for analyses of long-term

outcomes.

Building on the resulting dataset, we examine publication trajectories to gain insight into the research activity and career paths of PhD holders in five European countries:

1. First, we examine whether PhD graduates continue to publish after graduation and, if so, for how long, by estimating (1) the probability of being publication-active after having obtained the PhD, in two subsequent windows: 2–5 years (short term) and 6–9 years (long term) after graduation;

2. Second, we analyze whether in the long term PhD graduates remain affiliated with academia or industry research, and whether they stay in their country of graduation or not, by estimating different affiliation outcomes of those who remain publication-active.

To do so, we estimate the following linear regression models:

$$PubShort_i = \alpha + \beta X_i' + FE(i) + \varepsilon_i \qquad (1)$$

$$PubLong_i = \alpha + \beta X_i' + FE(i) + \varepsilon_i \qquad (2)$$

$$AffiliationLongi = \alpha + \beta X_i' + FE(i) + \varepsilon_i \qquad (3)$$

Where the dependent variables are defined as follows:

- $PubShort_i$ dummy equal to one if the PhD graduate *i* publishes between +2 and +5 years after graduation.

- $PubLongi$ dummy equal to one if the PhD graduate *i* publishes between +6 and + 9 years after graduation.

- $AffiliationLongi$ set of different dummies equal to one if the PhD graduate *i* publishes with a given type of academic affiliation between +6 and +9 years after graduation, conditional on having at least one publication in those years. These include: corporate, national, same as PhD granting institution and foreign.

And the independent variables X_i' include:

First, variables related to the productivity and network of the graduate, calculated for i) the period *during the PhD* (between -3 and +1 years from the PhD graduation) in regressions estimating the probability of publishing in the short term; and ii) the period *shortly after the PhD* (between +2 and +5 years from the PhD graduation) added to the regressions estimating the probability of publishing in the long term:

- $Pubs$ number of publications.
- $AvgCit$ average field-normalized citations (received in the first 3 years) per publication (all types).
- $Coauth$ number of distinct co-authors in publications.
- $Foreign$ dummy equal to one if the PhD graduate *i* had at least one foreign (different from the country of graduation) in a publication.
- $Corporate$ dummy equal to one if the PhD graduate *i* had at least one corporate affiliation in a publication.

Second, variables related to the PhD graduate *i*'s university of defence and field of research:

- $TopUni_i$ dummy equal to one if the PhD graduate *i*'s university of defence is among the top 5 universities in the country according to the Leiden ranking.
- $SizeField_i$ number of graduates in the same defence year, country, and field as the focal PhD graduate *i*.

Third, macroeconomic variables for the country in the corresponding time period, that is, during the PhD for regression (1) and in the period shortly after the PhD for regression (2), sourced from OECD Main Science and Technology Indicators:

- GDP average GDP per capita (purchasing power parity) adjusted for inflation, in thousand euros, in the PhD graduate's country.

- $HERD$ average Higher Education R&D expenditure, as a percentage of GDP, in the PhD graduate's country during the PhD.

- $BERD$ average Business R&D expenditure, as a percentage of GDP, in the PhD graduate's country during the PhD.

- $GOVERD$ average Government R&D expenditure, as a percentage of GDP, in the PhD graduate's country during the PhD.

And fourth, a number of fixed effects:

- $PhDcohort$

- $PhDcountry$

- $PhDfield$

- $PhDfemale$

As mentioned earlier, in order to alleviate truncation issues, regression (1) only considers PhD graduates with graduation years 2000-2017 and all other regressions consider only PhD graduates with graduation years 2000-2013.

While the models control for observable differences, we do not aim to identify causal effects.

4. Cross Country Patterns in Doctoral Education

4.1 Expansion of doctoral education

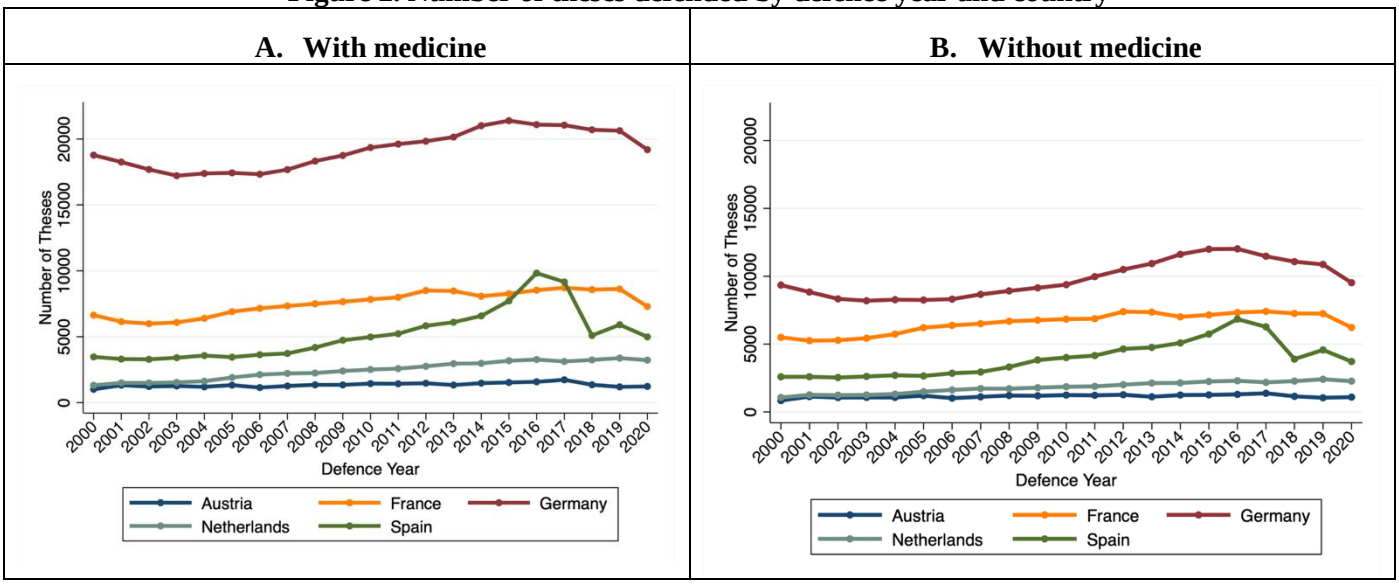
Data in table 2 confirm the general increase in the number of doctoral degrees granted each year in all countries, as well as some important cross-country differences, not all of them highlighted by the literature. The table reports the overall number of degrees issued in 2000 and 2020, with and without Medicine (for which Germany, as explained above, does not distinguish between master and doctoral dissertations). In both cases Germany is the only country that exhibit no meaningful increase and stands at the opposite end of the Netherlands, which doubles its numbers. France, Spain and Austria stand in between, with increases ranging from around 10% to more than 40%, regardless of the inclusion of Medicine.

Table 2: Growth in STEMM PhD Graduates by Country (2000–2020)

Year	Austria	Germany	Spain	France	Netherlands
Including Medicine					
2000	1,009	18,780	3,472	6,635	1,317
2020	1,229	19,193	4,989	7,294	3,225
% Change	21.8%	2.2%	43.7%	9.9%	144.8%
Excluding Medicine					
2000	841	9,348	2,587	5,496	1,065
2020	1,090	9,526	3,714	6,223	2,268
% Change	29.6%	1.9%	43.5%	13.2%	113.0%

Figure 2 provides year-per–year information. This reveals that in Spain, the number of dissertations – besides being generally increasing - spiked up dramatically in 2015 and 2016. One explanation for this pattern is the application, in academic year 2014/15, of a reform approved in 2011 (Royal Decree 99/2011) that, among others, limited the duration of doctoral studies, hitherto unregulated, to three years (five for motivated exceptions). This produced a "graduation run" and a rapid exhaustion of the theses backlog (see Corsini et al., 2025, for an overview on the effect of the Royal Decree 99/2011 on Spanish doctoral graduates). A comparison of the two graphs in figure 2 also make clear the extent of Germany's anomaly concerning Medicine, which has to be kept in mind in the rest of the analysis.

Figure 2. Number of theses defended by defence year and country



4.2 Field Composition

The disciplinary makeup of STEM PhDs differs slightly across countries. Table 3 shows the relative shares of Engineering, Life Sciences, Mathematics and Computer Science, Medicine, and Physics among doctoral graduates.[1] France stands out for their emphasis on Physics, with 34% of PhD graduates in this field, and for the importance of theses in Mathematics and Computer sciences relative to other countries (16.6%). Life Sciences account for around one-fourth of all STEM PhDs in Austria, the Netherlands, and Spain, suggesting different research infrastructures and traditions. Austria has also a relative proportion of theses in Engineering, compared to other countries (21.7%). Medicine is overrepresented in Germany, accounting for half of the theses defendend. This is due to the institutional organization of Germany, which requires medical doctors to have a PhD degree. 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

[1] For cross-country comparability purposes, DOC-TRACK classifies the doctoral dissertation in five large disciplines, namely: i) *Engineering* (comprising General Engineering, Chemical Engineering, and Energy); ii) *Life Sciences* (comprising Agricultural and Biological Sciences, Biochemistry, Genetics and Molecular Biology, Immunology and Microbiology, Neuroscience, Pharmacology, Toxicology and Pharmaceutics); iii) *Mathematics & Computer Sciences* (including Mathematics & Computer Sciences); iv) *Physics* (comprising General Physics and Astronomy, Material Science, Environmental Science, Earth and Planetary Sciences, and also Chemistry as a neighboring discipline); and v) *Medicine* (comprising Medical Science, Nursing, Veterinary, Dentistry, and Health Professions).

to be taken into account when interpreting stay rates in academia.

Table 3: STEM Field Distribution by Country

Country	Field	Nbr	Share
Austria	Engineering	6,135	21.7%
	Life Sciences	7,641	27.1%
	Mathematics and Computer Science	3,872	13.7%
	Medicine	4,054	14.4%
	Physics	6,524	23.1%
	TOT	28,226	100%
Germany	Engineering	60,574	15.0%
	Life Sciences	56,023	13.9%
	Mathematics and Computer Science	25,396	6.3%
	Medicine	197,231	49.0%
	Physics	63,662	15.8%
	TOT	402,886	100%
Spain	Engineering	16,491	15.2%
	Life Sciences	25,683	23.7%
	Mathematics and Computer Science	11,515	10.6%
	Medicine	25,909	24.0%
	Physics	28,587	26.4%
	TOT	108,185	100%
France	Engineering	26,377	16.6%
	Life Sciences	31,248	19.7%
	Mathematics and Computer Science	26,259	16.6%
	Medicine	20,843	13.1%
	Physics	53,941	34.0%
	TOT	158,668	100%
Netherlands	Engineering	7,296	14.1%
	Life Sciences	13,994	27.1%
	Mathematics and Computer Science	5,497	10.7%
	Medicine	13,452	26.1%
	Physics	11,326	22.0%
	TOT	51,565	100%

4.3 Gender Disparities

Table 4 shows the gender composition of PhD graduates by country. Spain comes closest to gender parity with a women/men ratio of about 0.94. France and Austria, by contrast, have the lowest ratio, corresponding to 0.62 and 0.63, respectively. Germany and the Netherlands fall somewhere in between. Note that, for the Netherlands, 64.3% of graduates lack information about gender. This is due to the peculiarity of Dutch EDT data, which lack of full graduate name information, making impossible to attribute a gender for most of graduates. These gaps are relevant not only from a gender balance perspective but also because they may shape aggregate productivity and influence retention in academic careers.

Table 4: Gender Composition of STEM PhD Graduates (2000–2020)

Country	Male	Female	Missing info	Ratio F / M
Austria	60.2%	37.8%	2.0%	0.63
Germany	56.0%	42.3%	1.7%	0.75
Spain	51.0%	47.8%	1.2%	0.94
France	60.1%	37.5%	2.4%	0.62
Netherlands	21.1%	14.6%	64.3%	0.69

4.4 Post-PhD publishing: Patterns of persistence and exit

Figure 3 plots the observed and predicted probabilities of publishing after graduation in the short and long term (respectively, 2 to 5 and 6 to 9 years after graduation). The predicted probabilities come from the regressions described in section 3 and are obtained by fixing all explanatory variables, but the cohort fixed effects, at their average value. They should be interpreted as baseline estimates of the dependent variable for each cohort, after taking into account composition effects (by country and discipline) as well as the changing conditions of the labour market (size of graduation cohorts and macroeconomic conditions). The full regression tables with estimated predicted probabilities and additional figures with observed and predicted probabilities by country and field are available in the Annex. All figures exhibit right truncation, proportional to the time lag necessary for producing the observations (at the 2017 cohort for the short-term graphs and 2013 for the long-term ones).

We first notice that until around 2010, the observed probability of short-term publishing keeps

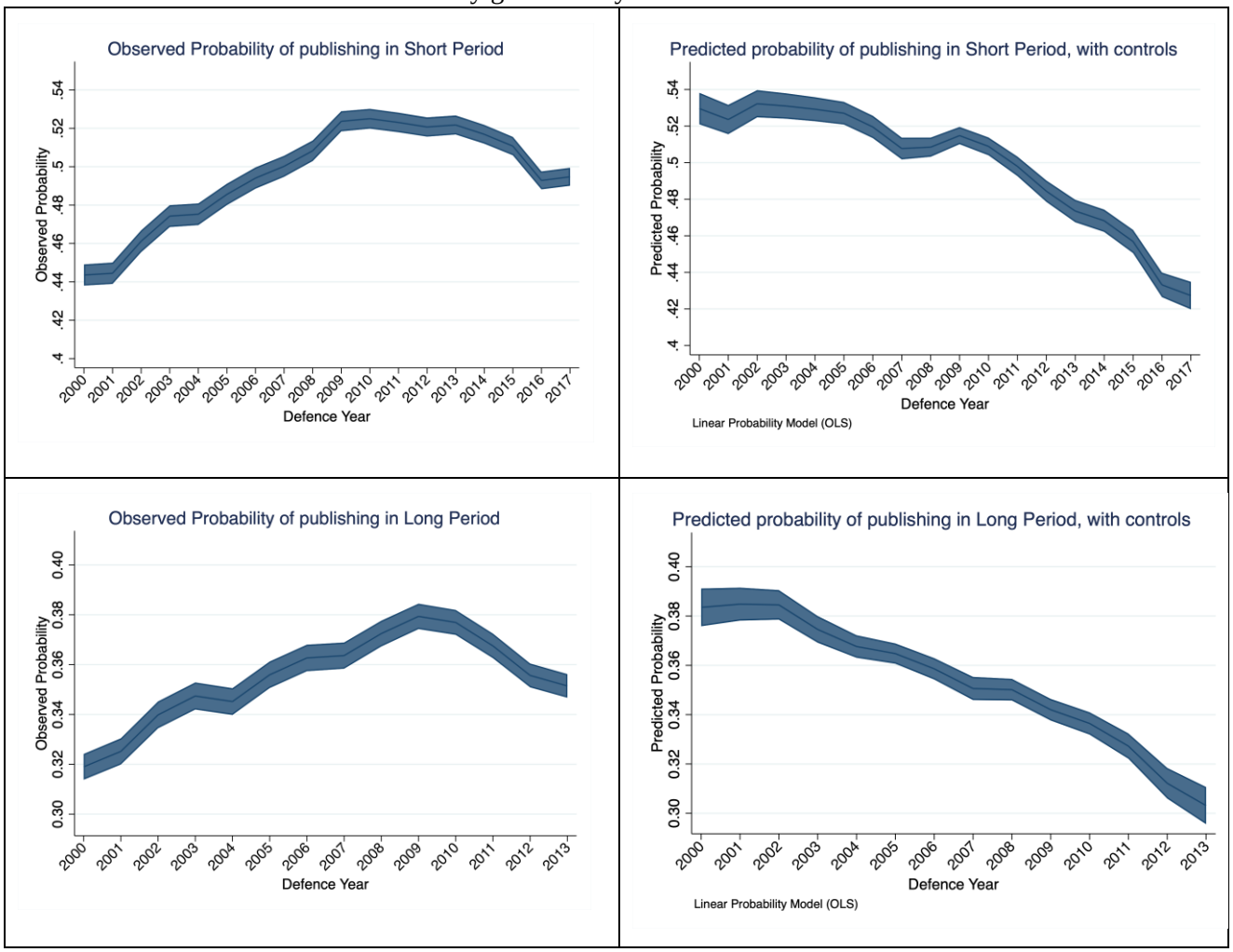
increasing, just to invert its trend in the following decade. The observed probability of long-term publishing follows a similar pattern, but with the inversion point shifted occurring about two year later. The predicted probability, instead, keeps decreasing across all cohorts, which we interpret as indicative of a structural trend that sees fewer and fewer doctoral graduates, after controlling for the economic context at the time of their graduation and soon afterwards. Figures C1 and C2 in the Appendix show that this result holds across countries (with the exception of Germany) and especially disciplines. The same figures in the Appendix also show remarkable differences across countries and disciplines differences exist, not so much with respect to trends, but to levels. Cross-disciplinary differences are intuitively related to the value of a doctoral degree outside the academia, from which most of the graduates' postdoc publications come from (see below).

Setting aside Medicine (for the representativeness problems discussed above), it is Engineering graduates who exhibit the lowest estimated probability to publish, whether in the short or long postdoc term, at great distance from the graduates in Mathematics and Computer Science and, further above, those in the Life Sciences and Physics. (It is also interesting to notice that the line for Mathematics and Computer Science is especially low, relative to those for Life Sciences and Physics for the more recent cohorts). As for cross-country differences, we know from the literature that Germany stands out for the value attached by industry employers to doctoral training, which explains its line's position at the bottom of the graph; and we presume that something similar may hold, in a more attenuated way, for Austria and the Netherlands. It may also be that our fixed effects for disciplines do not capture entirely the composition effects coming from the specific fields of research in different countries, whose exposure to industrial collaboration and research may vary.

Another important observation concerning figure 3 and its analogues in the Appendix relates to the temporal distribution of attrition points, that is of the moments in which the doctoral graduates appear to stop publishing. For the average graduate (see figure 3) the most important attrition point is graduation. Consider, for example, the 2000 cohort: the estimated probability to keep publishing in the short-term it is around 0.54, which implies a 0.46 probability not to start research career; in the long-run, the probability to quit research increases only of 16 percentage point (to around 0.62). Similarly, for the 2013 cohort, the probability to quit research is around 0.52 right after graduation and 0.70 in the longer term.

Figure 3: Observed vs predicted probability of publishing shortly and long after the PhD

By graduation year



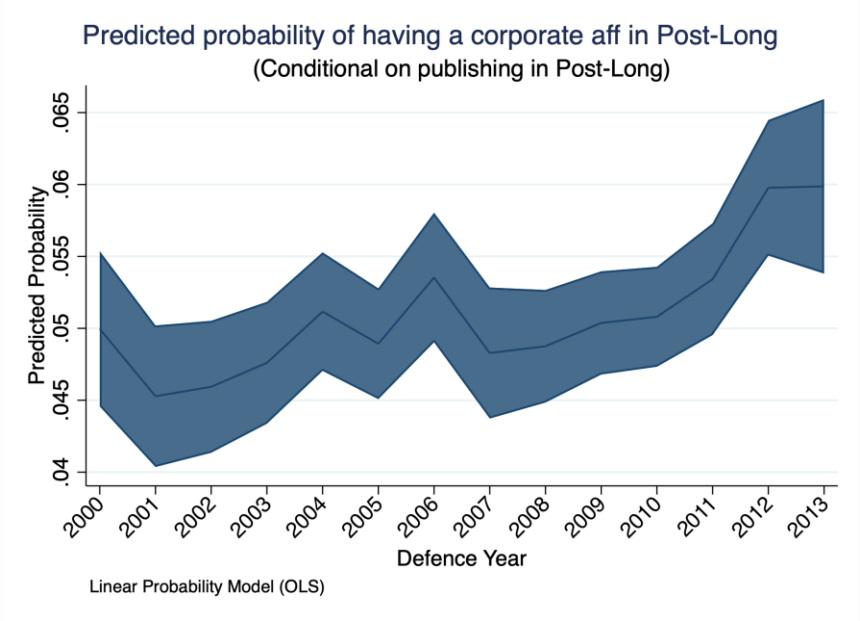
4.5 Who Stays—and Where?

Conditional on publishing long after the PhD, where do PhD graduates work? We now present findings from our regression exploring, for those who continue to publish in the long-run, the likelihood of publishing with a national or foreign affiliation, with a corporate affiliation, and with the same affiliation as the PhD. For every scenario, we also present the heterogeneity analyses by country and field in the Annex. Figures 4 and 5 plot the predicted probabilities of publishing with different types of affiliations long after the PhD defence (6-9 years after the PhD). Regression tables with estimated predicted probabilities and additional figures with observed and predicted probabilities by country and field are available in the Annex. Figure 4 shows an upward trend in the predicted

probability of pursuing a research career in industry across successive graduation cohorts. Although this probability increases steadily over time, it remains small, and thus substantially lower than the likelihood of following an academic research path.

Figure 4: Predicted probability of a corporate affiliation in publications long after the PhD

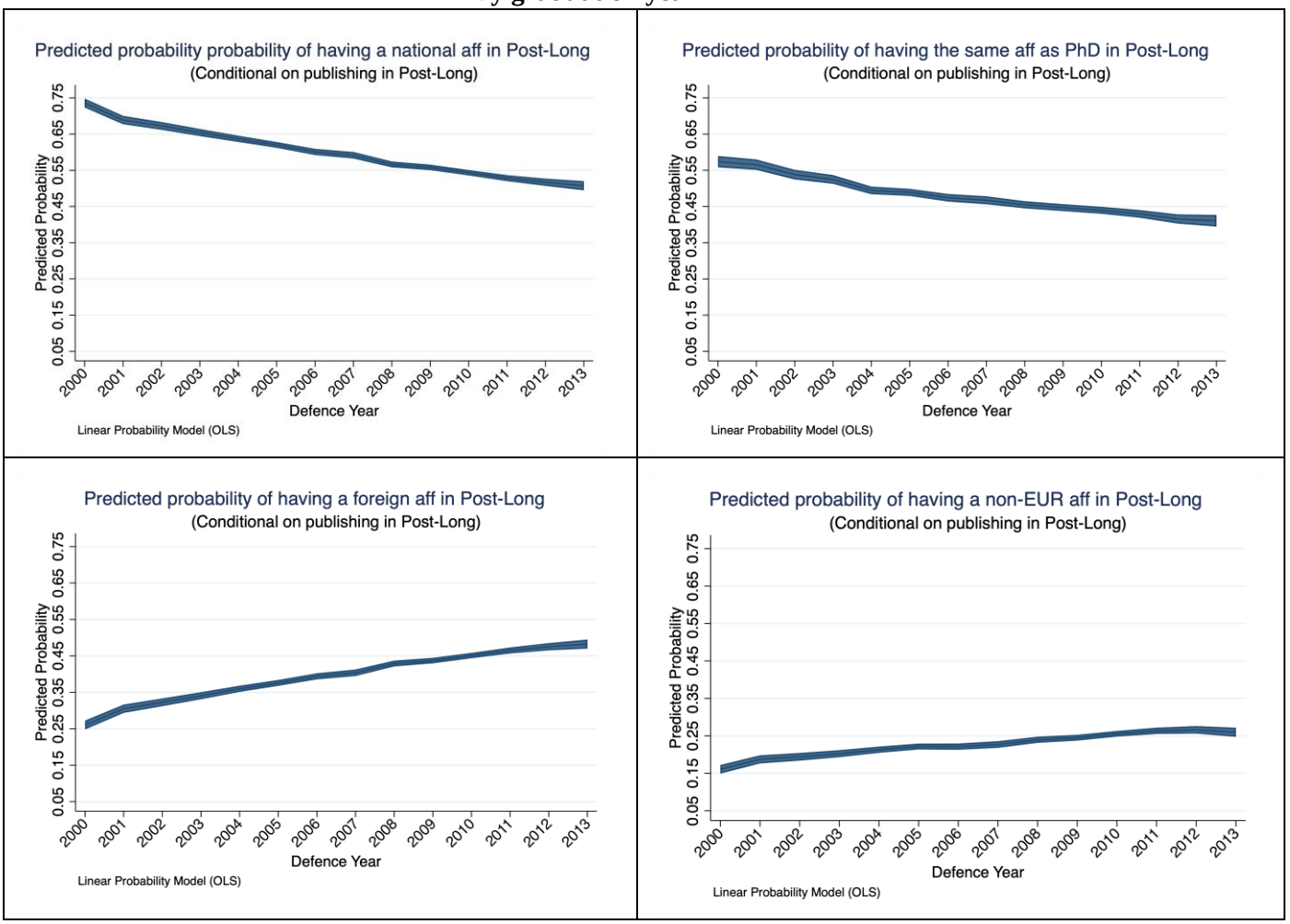
By graduation year



Without distinguishing between academic or corporate, Figure 5 shows that the probability of holding a national affiliation in publications in the long term is declining, as it is the likelihood of retaining the same affiliation as at the time of the PhD. These patterns suggest increasing post-doctoral mobility, consistent with the higher probability of having a foreign affiliation over time. This is also reflected in the gradual upward trend of non-European affiliations.

Figure 5: Predicted probability of different types of affiliations in publications long after the PhD

By graduation year



4.6 Temporal evolution of structural determinants of long-term publication activity

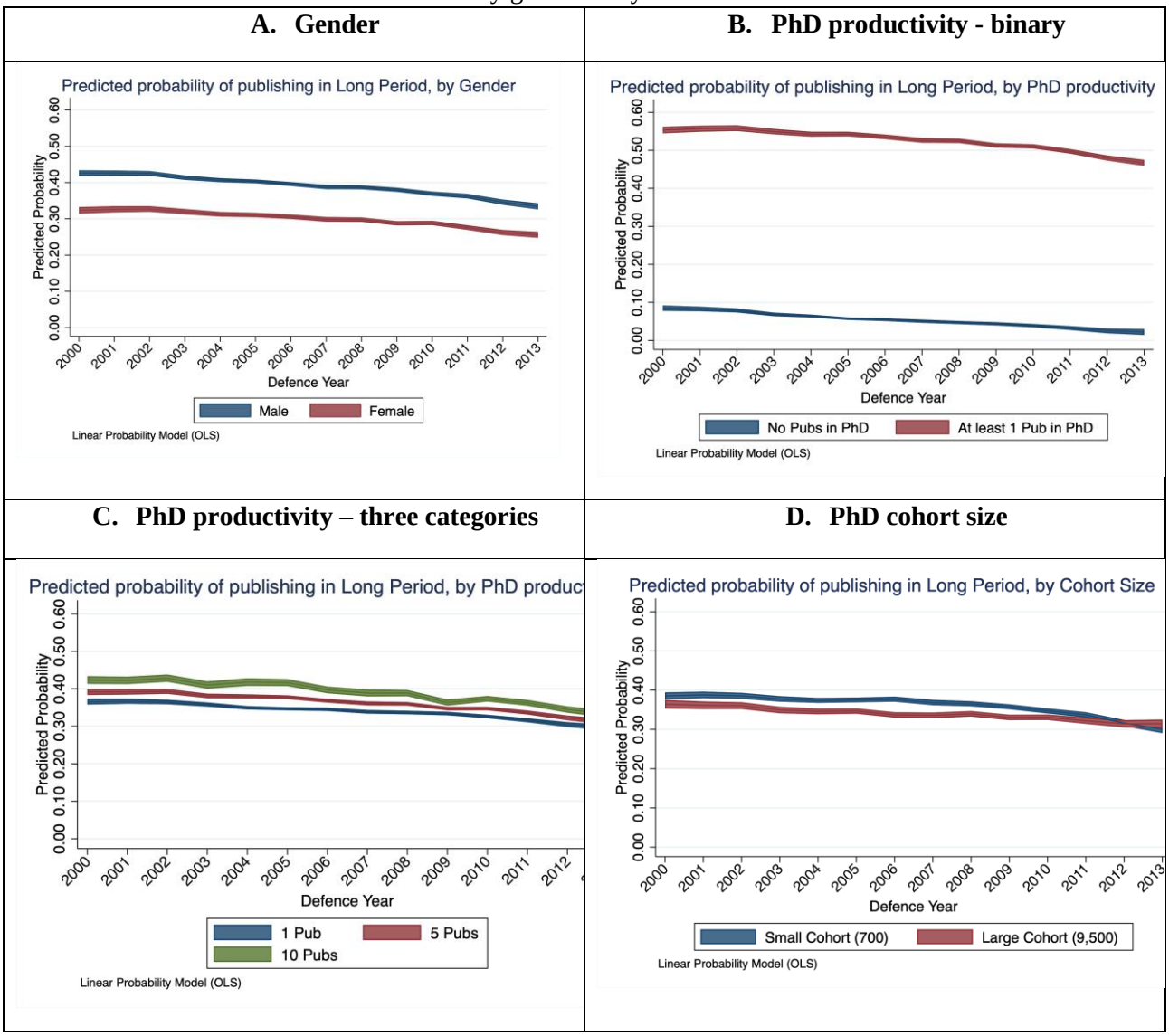
We now build on the earlier results by examining how key structural determinants evolve over time. For these regressions (available in the Annex).[2] Figure 6 presents the results on the relative importance of different determinants of long-term productivity in the long term, where we distinguish between the two sets of graphs: one set relies on a binary indicator of "at least one publication," whereas the other uses the

[2] For the two specifications focusing on PhD productivity, a small number of extreme outliers were removed. These cases correspond to graduates with more than 50 publications during the PhD and represent fewer than 1% of the sample and exert disproportionate influence on the fitted values. In addition, the controls for citations and coauthors accumulated during the PhD were excluded from these models due to their high collinearity with the main explanatory variable.

three productivity categories, namely zero publications (approximately the 10th percentile), five publications (between the median and the 75th percentile), and ten publications (around the 90th percentile). As shown in the first graph (A. Gender) of Figure 6, male PhD graduates consistently display higher probabilities than female graduates, with a gap of roughly 5-7 percentage points for all cohorts, and with predicted probabilities declining for both in parallel. The second graph (B. PhD productivity - binary) shows that PhD graduates who published during the PhD have predicted probabilities around 0.50-0.55 at the start of the period, declining modestly over time but remaining well above 0.40. In contrast, graduates with no publications during their PhD start below 0.15 and decline toward nearly 0.05 by 2013. The large and persistent gap illustrates the strong predictive value of early productivity in the careers of PhD graduates. The third graph (C. PhD productivity – three categories) introduces three levels of PhD productivity: 1 publication, 5 publications, and 10 publications. The three trajectories are vertically ordered as expected, with higher PhD productivity associated with higher long-period publication probabilities and all three groups show downward trends over time. Lastly, the fourth graph (D. PhD cohort size) compares predicted probabilities for graduates from small versus large cohorts and somewhat reflects the impact of competition for finding permanent jobs in academia. Belonging to small PhD cohorts (roughly 700 PhD graduates per year, country and discipline) is related to a higher probability of continuing publishing in the long term, than belonging to large cohorts (roughly 9,500 graduates), but this order reverses in the most recent cohorts.

Figure 6: Determinants of long-term publication activity

By graduation year



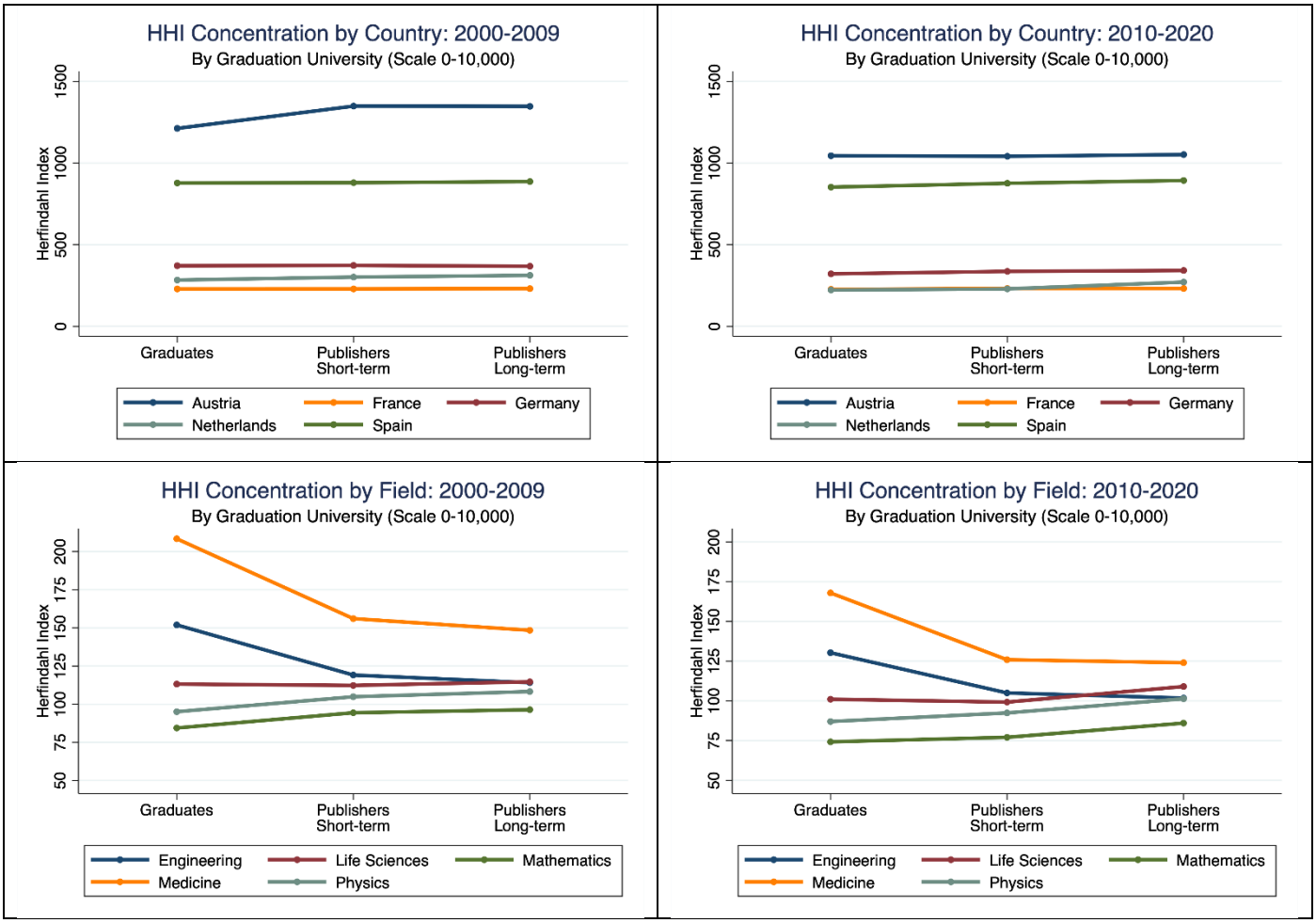
4.7 Concentration in top institutions

Figure 7 displays the evolution of institutional concentration of PhD graduates by granting institution, measured through the Herfindahl–Hirschman Index (HHI), across countries and disciplinary fields for the graduation periods 2000–2009 and 2010–2020, for three categories of PhDs: all graduates, graduates who continue publishing in the short term and graduates who continue publishing in the long term. Austria and

Spain show higher institutional concentration than Germany, the Netherlands, and France, but the gap has diminished over graduation cohorts. At the discipline level, Medicine has the highest HHI values, reflecting a stronger centralization of doctoral training in a small number of universities, whereas other fields present more dispersed institutional patterns. It is worth noting that in the most recent period (2010-2020) the HHI is higher for PhD graduates who publish in the long term in Life Sciences, Physics and Mathematics and Computer Science, than for PhD graduates who publish in the short term. This may indicate a trend towards increasing concentration of permanent academic positions for PhD graduates from a smaller subset of institutions in those fields.

Figure 7. Institutional concentration by PhD granting institution of different categories of graduates by graduation year

Herfindahl Index. Country (upper graphs) and Field (bottom graphs) Trends



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.

[TO BE COMPLETED]

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.

Auriol, L., 2010. *Careers of Doctorate Holders: Employment and Mobility Patterns*, OECD Science, Technology and Industry Working Papers No. 2010/4. Organisation for Economic Co-operation and Development: Paris

Auriol, L., Misu, M. and Freeman, R.A., 2013. *Careers of Doctorate Holders: Analysis of Labour Market and Mobility Indicators.* OECD Science, Technology and Industry Working Papers No. 2013/4. Organisation for Economic Co-operation and Development: Paris

Balsmeier B, Pellens M. Who makes, who breaks: Which scientists stay in academe? Economics Letters 2014; 122(2):229–32

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Information Science and Statis-tics. 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.

Carriero, R., Coda Zabetta, M., Geuna, A. and Tomatis, F., 2024. Investigating PhDs' early career occupational outcomes in Italy: individual motivations, role of supervisor and gender differences. *Higher Education*, *87*(5), pp.1375-1392.

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 PhD 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), 5863–58

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.

Ganguli, I. and MacGarvie, M. (2025) "International students' migration: policy impacts and implications for innovation." In: LIssoni F. and Morrison A. (eds.) *A Research Agenda for Migration and Innovation*, pp. 129-150. Edward Elgar Publishing

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

112. Springer.

Geuna, Aldo, and Sotaro Shibayama. "Moving out of academic research: Why do scientists stop doing research?." In Geuna, A. (ed.) *Global mobility of research scientists: The economics of who goes where and why*. Academic Press

Hancock, S., 2023. Knowledge or science-based economy? The employment of UK PhD graduates in research roles beyond academia. *Studies in Higher Education*, *48*(10), pp.1523-1537.

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 MacGarvie, M.J., 2016. How important is US location for research in science?. *Review of Economics and Statistics*, *98*(2), pp.397-414.

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

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

Kwon, D., 2025a. How many PhDs does the world need? Doctoral graduates vastly outnumber jobs in academia. *Nat*ure, 643(8070), pp.16-17.

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

Marini, G. and Henseke, G., 2023. Is a PhD worth more than a Master's in the UK labour market? The role of specialisation and managerial position. Studies in Higher Education, 48(10), pp.1538-1550

Martínez, C. and S. Parlane, 2023, Academic scientists in corporate R&D: a theoretical model, Research Policy, 52, 5, 104744

Milojević, S., Radicchi, F. and Walsh, J.P., 2018. Changing demographics of scientific careers: The rise of the temporary workforce. *Proceedings of the National Academy of Sciences*, *115*(50), pp.12616-23

OECD. 2021. Reducing the precarity of academic research careers. OECD Science, Technology and Industry Policy Papers 113. OECD. Paris.

Okrent, A. and Burke, A., 2021. *Where are they now? Most early career US-trained S&E doctorate recipients with temporary visas at graduation stay and work in the United States after graduation*. NSF 21-336. Alexandria, VA: National Science Foundation. Available at https://ncses.nsf.gov/pubs/nsf21336/

Opsomer, J., Chen, A., Chang, W.Y., Foley, D 2021. *US employment higher in the private sector than in the education sector for US-trained doctoral scientists and engineers: Findings from the 2019 survey of doctorate recipients*. NSF 21-319. Alexandria, VA: National Science Foundation. Available at https://ncses.nsf.gov/pubs/nsf21319/

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). The declining interest in an academic career. *PLOS ONE 12* (9), e0184130. .

Sarrico, C.S., 2022. The expansion of doctoral education and the changing nature and purpose of the doctorate. *Higher Education*, *84*(6), pp.1299-1315.

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

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 Interna-tional 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

Stephan P (2012) Research efficiency: Perverse incentives. *Nature* 484:29

Stephan, P., 2013. How to exploit postdocs. *BioScience*, *63*(4), pp.245-246.

Stephan, P. and Ma, J., 2005. The increased frequency and duration of the postdoctorate career stage. *American Economic Review*, *95*(2), pp.71-75.

Appendix A Matching EDT repositories to Publication Data

To identify potential matches between doctoral graduates (authors of dissertations) and unique authors in Scopus (authors of publications), we conducted a many-to-many exact match using both the standardized and the name variants. We opted for exact rather than fuzzy string matching due to both the high quality of name metadata in our sources and the computational infeasibility of string-distance calculations at the scale of our operations. This procedure yielded at least one author match for over 95.2% of the doctoral graduates. Match rates vary by country, with Spain achieving nearly full coverage, 99.6%, and Austria the lowest, 93.0% (France: 98.2%; Germany: 93.1%; the Netherlands: 94.6%). While these figures are indicative of a high recall, they also suggest very low precision, which we need to improve both with the application of a few rules and then by means of our machine learning algorithm. Notice that low precision may arise from both 1-to-1 false positives (one graduate matched, wrongly, to one author) and 1-to-m' false positives (one graduate matched to m authors, m'≤m of which are wrong).[3]

We then produced, for each country, a random sample of graduate-author pairs to be used for training, validating, and testing the model. All countries' samples corresponded to around 5000 graduate-authors, more precisely: 4,058 matches for France; 7,159 for Spain; 5,101 for Germany; 5,722 for Austria; and 4,776 for the Netherlands.[4] Independent annotators manually annotated all the sampled matches, based on a three-stage procedure aiming at ensuring consistency and reducing classification errors. The multi-stage annotation process was designed to produce a reliable gold standard dataset for training and testing. The resulting datasets remain unbalanced, with the proportion of negative (0) and positive (1) labels as follows: 61–39% for France, 80–20% for Spain, 77–23% for Germany, 76–24% for Austria, and 78–22% for the Netherlands. Class imbalance of this magnitude is common in binary classification tasks involving real-world data (e.g., He and Garcia, 2009; Fernández et al., 2018). To mitigate imbalance issues, we applied stratified cross-validation to preserve class proportions during training and validation, and used a proper machine learning algorithm, i.e., the

[3] As long as the disambiguation used by Elsevier to produce the Scopus IDs may include some false negatives (the same author receives different Scopus author IDs), it could be the case that the number of mistakes in *1-to-m* matches may be inferior to n-to-1. Absent reliable performance statistics for the Elsevier's disambiguation, we therefore decided not to drop 1-*to-m* matches, but to treat them with our machine learning algorithm as separate entities.

[4] The number of graduate-authors is based on a random sample of 1500 doctoral graduates from each country who have at least one matched author.

Random Forest.[5]

We trained a Random Forest classification algorithm with the manually annotated data produced. This specific type of machine learning algorithm has been widely adopted in similar classification contexts (e.g., Fan et al., 2011; Shin et al., 2014) and compares favourably to alternative approaches such as logistic regressions and AdaBoost (cf. Heinisch et al., 2020; Rehs, 2021). In particular, Random Forest algorithms are robust to class imbalances, handle high-dimensional feature spaces well, are less prone to overfitting, and perform efficiently with large datasets. Training a Random Forest algorithm requires, besides the data, the choice of a number of features to which the algorithm assigns weights to maximise its performance, and the choice of one or more performance indicators. Concerning the latter, we opted for using the average F1 score.[6] As for the former, we chose up to 12 features of the doctoral graduate-author matches, grouped into five categories: i) Social proximity (such as the presence, estimated with a low Levenshtein distance between names, of the graduate's supervisor or of a member of the doctoral committee among the co-authors of matched publications); ii) Geographic alignment (such as the presence, among the matched author's affiliations, of the country, city, or of the graduate's institutional alma mater); iii) Topical similarity (such as a significant overlap of the text strings containing the titles of the graduate's dissertation and the author's publication titles, or the overlap between disciplines or keywords); iv) Name similarity (such as a low Levenshtein distance between the text strings containing the standardized name forms of the graduate and the publication author); and v) publication productivity (such as the author being more active in publishing around the period of the graduate's doctoral studies and defence). We trained and tested separately one Random Forest algorithm for each country. In each case, we conducted stratified cross-validation. The algorithm for training and validation was implemented on 90% of the manually annotated dataset of each country, extracted randomly, but with a stratification to preserve the class distribution between true and false matches to deal with the imbalance in the data.[7]

[5] We also experimented with the Synthetic Minority Over-sampling Technique (SMOTE) to increase the representation of the positive class. This approach slightly improved the recall of our models but at the cost of reducing precision. Since our primary concern is maintaining high precision, we chose not to implement SMOTE in the final models.

[6] The F1 score is the harmonic mean of precision and recall, balancing the trade-off between false positives and false negatives.

[7] Specifically, we conducted a stratified 5-fold cross-validation procedure combined with an exhaustive grid search to tune two hyperparameters: the number of trees (*ntree*) and the number of variables considered at each split (*mtry*). The number of trees was varied from 500 to 2,000 in increments of 50, while *mtry* was tested across the values 1, 2, 3, 4, and the maximum value of features used for a given country. The data were divided into five equally sized subsets, which were

The results of the trained country-specific Random Forest models were tested against the 10% records from the manually annotated datasets that were not used for training. In terms of precision and recall, all models perform rather strongly, as follows:

- Germany: 93.5% precision, 87.8% recall
- France: 96.1% precision, 94.2% recall
- Spain: 96.2% precision, 89.5% recall
- Austria: 95.2% precision, 88.2% recall
- Netherlands: 99.0% precision, 95.3% recall[8]

extracted randomly and stratified. There were five iterations, in each of which we used four out of five subsets for model training and held the fifth one out for validation. The validation subset was rotated in each iteration.

[8] The 99% precision observed for the Netherlands is partly due to the additional geographical filter applied in step 2, aimed at reducing false positives. By restricting the sample in this way, the Random Forest is applied to a sample where true positives are more prevalent, leading to higher measured precision and recall. However, this filter also excludes some graduate-author pairs that could include true positives. In this case, 2,570 graduates were removed, representing 4.8% of the total Dutch graduates.

Appendix B Regression results

Table B1. Determinants of publication activity, short and long after the PhD. OLS estimates.

	(1)	(2)
	Publishing Short Period (+2,+5)	Publishing Long Period (+6,+9)
2001	-0.0060*	0.0013
	(0.0035)	(0.0027)
2002	0.0026	0.0010
	(0.0036)	(0.0033)
2003	0.0014	-0.0089**
	(0.0037)	(0.0038)
2004	-0.0004	-0.0159***
	(0.0038)	(0.0038)
2005	-0.0025	-0.0187***
	(0.0039)	(0.0038)
2006	-0.0101**	-0.0249***
	(0.0042)	(0.0041)
2007	-0.0219***	-0.0329***
	(0.0045)	(0.0043)
2008	-0.0211***	-0.0334***
	(0.0047)	(0.0047)
2009	-0.0148***	-0.0415***
	(0.0050)	(0.0051)
2010	-0.0207***	-0.0471***
	(0.0054)	(0.0054)
2011	-0.0316***	-0.0563***
	(0.0058)	(0.0059)
2012	-0.0452***	-0.0713***
	(0.0062)	(0.0065)
2013	-0.0561***	-0.0804***
	(0.0065)	(0.0072)
2014	-0.0614***	
	(0.0066)	
2015	-0.0727***	
	(0.0068)	
2016	-0.0964***	
	(0.0071)	
2017	-0.1023***	
	(0.0075)	
Germany	-0.0100	-0.0439***
	(0.0141)	(0.0142)
Spain	0.2716***	0.1777***
	(0.0160)	(0.0158)
France	0.2467***	0.0724***
	(0.0129)	(0.0124)
Netherlands	-0.0071	0.0562***

	(0.0094)	(0.0108)
Life Sciences	0.1398***	0.0448***
	(0.0022)	(0.0019)
Mathematics	0.0378***	0.0159***
	(0.0024)	(0.0022)
Medicine	0.0873***	0.0414***
	(0.0026)	(0.0025)
Physics	0.1123***	-0.0077***
	(0.0022)	(0.0019)
Female	-0.0328***	-0.0367***
	(0.0012)	(0.0011)
TopUni (d)	-0.0007	-0.0014
	(0.0014)	(0.0012)
SizeField (d)	-0.0000***	-0.0000***
	(0.0000)	(0.0000)
PubsDuringPhD (n)	0.0150***	0.0028***
	(0.0003)	(0.0002)
AvgCitDuringPhD (n)	0.0321***	0.0056***
	(0.0018)	(0.0004)
CoauthDuringPhD (n)	-0.0006***	-0.0003***
	(0.0000)	(0.0000)
ForeignDuringPhD (d)	0.2177***	0.0289***
	(0.0030)	(0.0023)
CompanyDuringPhD (d)	0.0368***	0.0136**
	(0.0057)	(0.0061)
GDPDuringPhD	0.0060***	
	(0.0004)	
HERDDuringPhD	-0.0845	
	(0.0623)	
BERDDuringPhD	0.0787***	
	(0.0125)	
GOVERDDuringPhD	0.0169	
	(0.0226)	
PubsShort (d)		0.4435***
		(0.0023)
PubsShort (n)		0.0097***
		(0.0002)
AvgCitsShort (n)		0.0043***
		(0.0005)
CoauthShort (n)		-0.0002***
		(0.0000)
AcademicShort (n)		0.1991***
		(0.0021)
CompanyShort (n)		0.0047
		(0.0052)
GDPshort		0.0024***
		(0.0006)
HERDshort		0.0155

		(0.0605)
BERDShort		0.0447***
		(0.0132)
GOVERDShort		0.1870***
		(0.0264)
Constant	-0.0359	-0.1834***
	(0.0270)	(0.0256)
Adjusted R_sq	0.242	0.508
Observations	596,075	436,635

Table B2. Determinants of different types of affiliations in publications long after the PhD. OLS estimates.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Corporate affiliation	National affiliation	Same affiliation as PhD	Foreign affiliation	German affiliation	Non-EUR affiliation	USA affiliation
2001	-0.0047	-0.0469***	-0.0083	0.0448***	0.0051	0.0266***	0.0094**
	(0.0030)	(0.0058)	(0.0070)	(0.0057)	(0.0036)	(0.0053)	(0.0044)
2002	-0.0040	-0.0631***	-0.0359***	0.0624***	-0.0007	0.0335***	0.0140***
	(0.0030)	(0.0062)	(0.0076)	(0.0061)	(0.0049)	(0.0058)	(0.0047)
2003	-0.0024	-0.0814***	-0.0492***	0.0804***	0.0070	0.0418***	0.0075
	(0.0031)	(0.0066)	(0.0082)	(0.0066)	(0.0066)	(0.0062)	(0.0050)
2004	0.0012	-0.0988***	-0.0794***	0.0994***	0.0017	0.0523***	0.0080
	(0.0032)	(0.0066)	(0.0082)	(0.0065)	(0.0075)	(0.0062)	(0.0050)
2005	-0.0010	-0.1157***	-0.0849***	0.1157***	0.0104	0.0609***	0.0108**
	(0.0032)	(0.0066)	(0.0083)	(0.0066)	(0.0078)	(0.0063)	(0.0051)
2006	0.0036	-0.1351***	-0.0998***	0.1337***	0.0038	0.0605***	0.0049
	(0.0034)	(0.0070)	(0.0089)	(0.0069)	(0.0081)	(0.0067)	(0.0054)
2007	-0.0017	-0.1442***	-0.1067***	0.1434***	0.0095	0.0662***	-0.0005
	(0.0035)	(0.0074)	(0.0093)	(0.0073)	(0.0081)	(0.0071)	(0.0056)
2008	-0.0012	-0.1695***	-0.1191***	0.1688***	0.0112	0.0789***	0.0036
	(0.0034)	(0.0077)	(0.0097)	(0.0076)	(0.0083)	(0.0073)	(0.0058)
2009	0.0004	-0.1777***	-0.1270***	0.1769***	0.0171*	0.0844***	0.0138**
	(0.0034)	(0.0080)	(0.0103)	(0.0079)	(0.0092)	(0.0077)	(0.0061)
2010	0.0008	-0.1925***	-0.1344***	0.1907***	0.0148	0.0949***	0.0109*
	(0.0035)	(0.0083)	(0.0107)	(0.0082)	(0.0102)	(0.0079)	(0.0063)
2011	0.0035	-0.2075***	-0.1442***	0.2047***	0.0154	0.1029***	0.0100
	(0.0039)	(0.0091)	(0.0116)	(0.0089)	(0.0123)	(0.0087)	(0.0069)
2012	0.0098**	-0.2187***	-0.1585***	0.2149***	0.0100	0.1055***	0.0112
	(0.0043)	(0.0100)	(0.0129)	(0.0099)	(0.0148)	(0.0096)	(0.0076)
2013	0.0099**	-0.2284***	-0.1633***	0.2222***	0.0085	0.0989***	0.0081
	(0.0049)	(0.0113)	(0.0144)	(0.0111)	(0.0175)	(0.0107)	(0.0085)
Germany	0.0100	0.0123	-0.0409	-0.0088		0.1853***	0.0922***
	(0.0112)	(0.0230)	(0.0284)	(0.0227)		(0.0229)	(0.0177)
Spain	-0.0551***	0.3019***	0.4845***	-0.2908***	-0.1244***	0.0172	0.0136
	(0.0116)	(0.0254)	(0.0323)	(0.0251)	(0.0233)	(0.0244)	(0.0191)
France	-0.0273***	0.1554***	-0.0131	-0.1456***	-0.1080***	0.1287***	0.0300**
	(0.0083)	(0.0190)	(0.0239)	(0.0187)	(0.0133)	(0.0186)	(0.0143)
Netherlands	-0.0148	0.0125	0.1897***	-0.0148	-0.0896***	0.0427**	0.0281*
	(0.0105)	(0.0196)	(0.0246)	(0.0193)	(0.0191)	(0.0195)	(0.0154)
Life Sciences	-0.0217***	-0.0516***	0.0155***	0.0522***	0.0034	0.0242***	0.0645***
	(0.0020)	(0.0033)	(0.0043)	(0.0033)	(0.0024)	(0.0033)	(0.0024)
Mathematics	-0.0200***	-0.0160***	-0.0417***	0.0149***	-0.0013	0.0029	0.0103***
	(0.0023)	(0.0039)	(0.0048)	(0.0039)	(0.0023)	(0.0039)	(0.0026)
Medicine	-0.0249***	-0.0135***	0.0729***	0.0092**	-0.0074***	0.0119***	0.0635***
	(0.0021)	(0.0041)	(0.0051)	(0.0040)	(0.0021)	(0.0040)	(0.0030)
Physics	-0.0097***	-0.0085***	-0.0103**	0.0083***	0.0267***	-0.0263***	0.0168***
	(0.0021)	(0.0032)	(0.0041)	(0.0032)	(0.0037)	(0.0032)	(0.0022)
Female	-0.0106***	0.0149***	-0.0139***	-0.0161***	-0.0040***	-0.0289***	-0.0085***
	(0.0010)	(0.0021)	(0.0026)	(0.0021)	(0.0014)	(0.0021)	(0.0016)

TopUni(d)	0.0025**	-0.0125***	-0.0273***	0.0132***	0.0045***	0.0037	0.0248***
	(0.0011)	(0.0023)	(0.0028)	(0.0023)	(0.0015)	(0.0023)	(0.0018)
FieldSize (d)	-0.0000***	0.0000***	-0.0000**	-0.0000***	-0.0000***	-0.0000***	-0.0000***
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
PubsDuringPhD (n)	0.0003***	0.0015***	0.0065***	-0.0015***	0.0000	-0.0013***	0.0003**
	(0.0001)	(0.0001)	(0.0002)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
AvgCitsDuringPhD(n)	0.0009***	0.0005**	0.0057***	-0.0005*	0.0012***	-0.0018***	0.0018***
	(0.0002)	(0.0003)	(0.0006)	(0.0003)	(0.0003)	(0.0003)	(0.0004)
CoauthDuringPhD(n)	0.0000	0.0000	-0.0002***	-0.0000	-0.0000***	-0.0000	-0.0000***
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
ForeignDuringPhD(d)	-0.0052***	-0.0946***	0.1780***	0.0951***	0.0104***	0.0782***	0.0008
	(0.0013)	(0.0028)	(0.0033)	(0.0028)	(0.0021)	(0.0031)	(0.0024)
CompanyDuringPhD (d)	0.1321***	0.0208***	0.1379***	-0.0221***	0.0013	-0.0213***	0.0136**
	(0.0094)	(0.0078)	(0.0105)	(0.0077)	(0.0063)	(0.0073)	(0.0063)
PubsShort(d)	-0.0409***	0.1251***	0.0830***	-0.1136***	-0.0133***	-0.0837***	-0.0316***
	(0.0028)	(0.0047)	(0.0055)	(0.0046)	(0.0023)	(0.0039)	(0.0027)
PubsShort(n)	-0.0004***	-0.0003***	0.0054***	0.0005***	0.0005***	0.0004***	0.0011***
	(0.0000)	(0.0001)	(0.0002)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
AvgCitsShort(n)	0.0007***	-0.0018***	0.0015***	0.0018***	0.0008***	0.0002	0.0048***
	(0.0002)	(0.0003)	(0.0003)	(0.0003)	(0.0003)	(0.0003)	(0.0006)
CoauthShort(n)	-0.0000	-0.0000***	-0.0001***	0.0000***	0.0000***	0.0000***	-0.0000
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
AcademicShort(d)	-0.0045***	-0.5999***	-0.1610***	0.6027***	0.0605***	0.3484***	0.1770***
	(0.0011)	(0.0024)	(0.0027)	(0.0024)	(0.0016)	(0.0025)	(0.0020)
CompanyShort(d)	0.5384***	-0.0317***	-0.1434***	0.0303***	0.0191***	-0.0045	0.0435***
	(0.0070)	(0.0058)	(0.0067)	(0.0058)	(0.0052)	(0.0053)	(0.0047)
GDPShort	-0.0016***	0.0122***	0.0062***	-0.0119***	0.0006	-0.0060***	-0.0039***
	(0.0005)	(0.0011)	(0.0014)	(0.0011)	(0.0022)	(0.0010)	(0.0008)
HERDShort	0.0858	0.0197	0.1971	0.0141	0.0049	0.3409***	0.2052**
	(0.0542)	(0.1064)	(0.1336)	(0.1048)	(0.0781)	(0.1074)	(0.0843)
BERDShort	-0.0161	0.0570**	0.2177***	-0.0615***	-0.0328*	-0.0438*	-0.0089
	(0.0132)	(0.0239)	(0.0297)	(0.0236)	(0.0190)	(0.0241)	(0.0190)
GOVERDShort	0.0047	0.2208***	0.3795***	-0.2153***	0.0371	-0.3892***	-0.0885***
	(0.0181)	(0.0398)	(0.0480)	(0.0395)	(0.0579)	(0.0389)	(0.0298)
Constant	0.1591***	0.0874**	-0.4395***	0.8644***	0.1168	0.2737***	0.0581*
	(0.0186)	(0.0416)	(0.0537)	(0.0410)	(0.0879)	(0.0400)	(0.0316)
Adjusted R_sq	0.251	0.412	0.131	0.421	0.053	0.204	0.105
Observations	153,162	153,162	153,162	153,162	92,326	153,162	153,162

Table B3. Temporal Evolution of Structural Determinants. OLS estimates.

	(1)	(2)	(3)	(4)
	Publishing Long Period	Publishing Long Period	Publishing Long Period	Publishing Long Period
2001	0.0003	-0.0020	0.0014	0.0029
	(0.0034)	(0.0022)	(0.0029)	(0.0043)
2002	-0.0007	-0.0060**	-0.0015	0.0005
	(0.0039)	(0.0030)	(0.0035)	(0.0047)
2003	-0.0124***	-0.0164***	-0.0072*	-0.0071
	(0.0043)	(0.0035)	(0.0040)	(0.0050)
2004	-0.0192***	-0.0207***	-0.0177***	-0.0110**
	(0.0043)	(0.0036)	(0.0040)	(0.0050)
2005	-0.0226***	-0.0277***	-0.0206***	-0.0097*
	(0.0043)	(0.0036)	(0.0040)	(0.0051)
2006	-0.0299***	-0.0303***	-0.0201***	-0.0068
	(0.0046)	(0.0039)	(0.0043)	(0.0054)
2007	-0.0384***	-0.0343***	-0.0263***	-0.0160***
	(0.0048)	(0.0041)	(0.0046)	(0.0056)
2008	-0.0390***	-0.0379***	-0.0283***	-0.0198***
	(0.0052)	(0.0046)	(0.0051)	(0.0058)
2009	-0.0458***	-0.0411***	-0.0286***	-0.0274***
	(0.0056)	(0.0051)	(0.0056)	(0.0060)
2010	-0.0566***	-0.0459***	-0.0385***	-0.0389***
	(0.0058)	(0.0053)	(0.0057)	(0.0062)
2011	-0.0633***	-0.0522***	-0.0484***	-0.0489***
	(0.0063)	(0.0059)	(0.0062)	(0.0066)
2012	-0.0799***	-0.0597***	-0.0592***	-0.0708***
	(0.0069)	(0.0065)	(0.0069)	(0.0073)
2013	-0.0915***	-0.0628***	-0.0668***	-0.0881***
	(0.0076)	(0.0072)	(0.0076)	(0.0081)
Female	-0.0491***	-0.0334***	-0.0322***	-0.0368***
	(0.0038)	(0.0011)	(0.0011)	(0.0011)
2001 # Female	0.0028			
	(0.0054)			
2002 # Female	0.0047			
	(0.0054)			
2003 # Female	0.0088			
	(0.0055)			
2004 # Female	0.0085			
	(0.0054)			
2005 # Female	0.0099*			
	(0.0054)			
2006 # Female	0.0124**			
	(0.0054)			
2007 # Female	0.0136**			
	(0.0054)			
2008 # Female	0.0135**			
	(0.0054)			
2009 # Female	0.0102*			
	(0.0054)			
2010 # Female	0.0219***			

	(0.0054)			
2011 # Female	0.0159***			
	(0.0053)			
2012 # Female	0.0190***			
	(0.0053)			
2013 # Female	0.0242***			
	(0.0053)			
Germany	-0.0426***	-0.0479***	-0.0419***	-0.0774***
	(0.0142)	(0.0141)	(0.0142)	(0.0151)
Spain	0.1806***	0.1472***	0.1593***	0.1381***
	(0.0158)	(0.0155)	(0.0158)	(0.0166)
France	0.0747***	0.0520***	0.0728***	0.0424***
	(0.0124)	(0.0122)	(0.0124)	(0.0130)
Netherlands	0.0567***	0.0479***	0.0476***	0.0575***
	(0.0108)	(0.0107)	(0.0108)	(0.0113)
Life Sciences	0.0447***	0.0458***	0.0493***	0.0446***
	(0.0019)	(0.0019)	(0.0019)	(0.0019)
Mathematics	0.0160***	0.0165***	0.0161***	0.0161***
	(0.0022)	(0.0022)	(0.0022)	(0.0022)
Medicine	0.0412***	0.0411***	0.0342***	0.0412***
	(0.0025)	(0.0025)	(0.0025)	(0.0025)
Physics	-0.0079***	-0.0142***	-0.0112***	-0.0079***
	(0.0019)	(0.0019)	(0.0019)	(0.0019)
TopUni (d)	-0.0014	-0.0026**	-0.0018	-0.0013
	(0.0012)	(0.0012)	(0.0012)	(0.0012)
FieldSize (d)	-0.0000***	-0.0000***	-0.0000***	-0.0000***
	(0.0000)	(0.0000)	(0.0000)	(0.0000)
PubsDuringPhD (n)	0.0028***		0.0064***	0.0028***
	(0.0002)		(0.0005)	(0.0002)
AvgCitsDuringPhD (n)	0.0056***			0.0056***
	(0.0004)			(0.0004)
CoauthDuringPhD (n)	-0.0003***			-0.0003***
	(0.0000)			(0.0000)
ForeignDuringPhD (d)	0.0289***	0.0108***	0.0158***	0.0289***
	(0.0023)	(0.0022)	(0.0022)	(0.0023)
CompanyDuringPhD (d)	0.0137**	-0.0036	0.0117*	0.0139**
	(0.0061)	(0.0060)	(0.0061)	(0.0061)
PubsShort (d)	0.4435***	0.3783***	0.4247***	0.4434***
	(0.0023)	(0.0024)	(0.0022)	(0.0023)
PubsShort (n)	0.0097***	0.0142***	0.0131***	0.0097***
	(0.0002)	(0.0003)	(0.0003)	(0.0002)
AvgCitsShort (n)	0.0043***	0.0040***	0.0040***	0.0043***
	(0.0005)	(0.0005)	(0.0005)	(0.0005)
CoauthShort (n)	-0.0002***	-0.0003***	-0.0003***	-0.0002***
	(0.0000)	(0.0000)	(0.0000)	(0.0000)
AcademicForeignShort (d)	0.1992***	0.1912***	0.1906***	0.1991***

	(0.0021)	(0.0022)	(0.0022)	(0.0021)
CompanyShort (d)	0.0048	0.0113**	0.0055	0.0048
	(0.0052)	(0.0052)	(0.0052)	(0.0052)
GDPShort	0.0025***	0.0025***	0.0023***	0.0028***
	(0.0006)	(0.0006)	(0.0006)	(0.0008)
HERDShort	0.0157	-0.0136	-0.0030	-0.1873***
	(0.0605)	(0.0601)	(0.0608)	(0.0672)
BERDShort	0.0456***	0.0278**	0.0377***	0.0596***
	(0.0132)	(0.0130)	(0.0131)	(0.0135)
GOVERDShort	0.1819***	0.1339***	0.1500***	0.1624***
	(0.0264)	(0.0264)	(0.0263)	(0.0273)
PubsDuringPhD(d)		0.1233***		
		(0.0035)		
2001 # PubsDuringPhD(d)		0.0054		
		(0.0047)		
2002 # PubsDuringPhD(d)		0.0109**		
		(0.0048)		
2003 # PubsDuringPhD(d)		0.0123**		
		(0.0048)		
2004 # PubsDuringPhD(d)		0.0098**		
		(0.0048)		
2005 # PubsDuringPhD(d)		0.0172***		
		(0.0047)		
2006 # PubsDuringPhD(d)		0.0125***		
		(0.0047)		
2007 # PubsDuringPhD(d)		0.0071		
		(0.0047)		
2008 # PubsDuringPhD(d)		0.0097**		
		(0.0047)		
2009 # PubsDuringPhD(d)		0.0007		
		(0.0047)		
2010 # PubsDuringPhD(d)		0.0030		

		(0.0046)		
2011 # PubsDuringPhD(d)		-0.0038		
		(0.0046)		
2012 # PubsDuringPhD(d)		-0.0138***		
		(0.0046)		
2013 # PubsDuringPhD(d)		-0.0235***		
		(0.0045)		
2001 # PubsDuringPhD(n)			-0.0002	
			(0.0006)	
2002 # PubsDuringPhD(n)			0.0006	
			(0.0006)	
2003 # PubsDuringPhD(n)			-0.0006	
			(0.0007)	
2004 # PubsDuringPhD(n)			0.0012*	
			(0.0007)	
2005 # PubsDuringPhD(n)			0.0014**	
			(0.0007)	
2006 # PubsDuringPhD(n)			-0.0006	
			(0.0006)	
2007 # PubsDuringPhD(n)			-0.0008	
			(0.0007)	
2008 # PubsDuringPhD(n)			-0.0007	
			(0.0006)	
2009 # PubsDuringPhD(n)			-0.0032***	
			(0.0007)	
2010 # PubsDuringPhD(n)			-0.0012*	
			(0.0006)	
2011 # PubsDuringPhD(n)			-0.0012**	
			(0.0006)	
2012 # PubsDuringPhD(n)			-0.0020***	
			(0.0006)	
2013 # PubsDuringPhD(n)			-0.0024***	
			(0.0006)	

2001 # FieldSize (d)				-0.0000
				(0.0000)
2002 # FieldSize (d)				-0.0000
				(0.0000)
2003 # FieldSize (d)				-0.0000
				(0.0000)
2004 # FieldSize (d)				-0.0000
				(0.0000)
2005 # FieldSize (d)				-0.0000
				(0.0000)
2006 # FieldSize (d)				-0.0000***
				(0.0000)
2007 # FieldSize (d)				-0.0000*
				(0.0000)
2008 # FieldSize (d)				-0.0000
				(0.0000)
2009 # FieldSize (d)				-0.0000
				(0.0000)
2010 # FieldSize (d)				0.0000
				(0.0000)
2011 # FieldSize (d)				0.0000
				(0.0000)
2012 # FieldSize (d)				0.0000***
				(0.0000)
2013 # FieldSize (d)				0.0000***
				(0.0000)
Constant	-0.1836***	-0.1629***	-0.1537***	-0.0968***
	(0.0256)	(0.0251)	(0.0257)	(0.0292)
Adjusted R_sq	0.508	0.524	0.516	0.508
Observations	436,635	435,481	435,481	436,635

Appendix C Additional Figures

Figure C1: Observed vs predicted probability of publishing shortly and long after the PhD

By graduation year & field

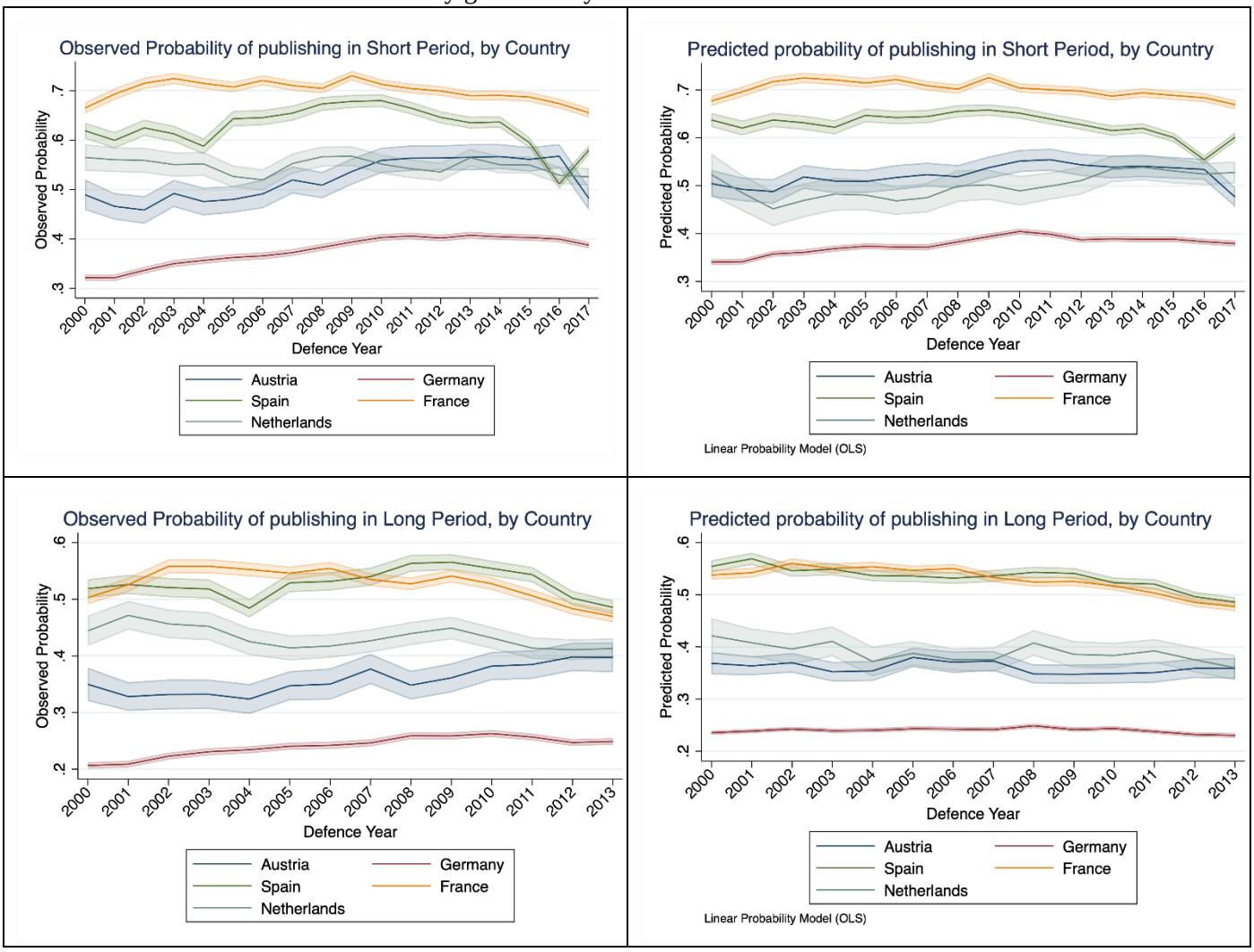


Figure C2: Observed vs predicted probability of publishing shortly and long after the PhD

By graduation year & country

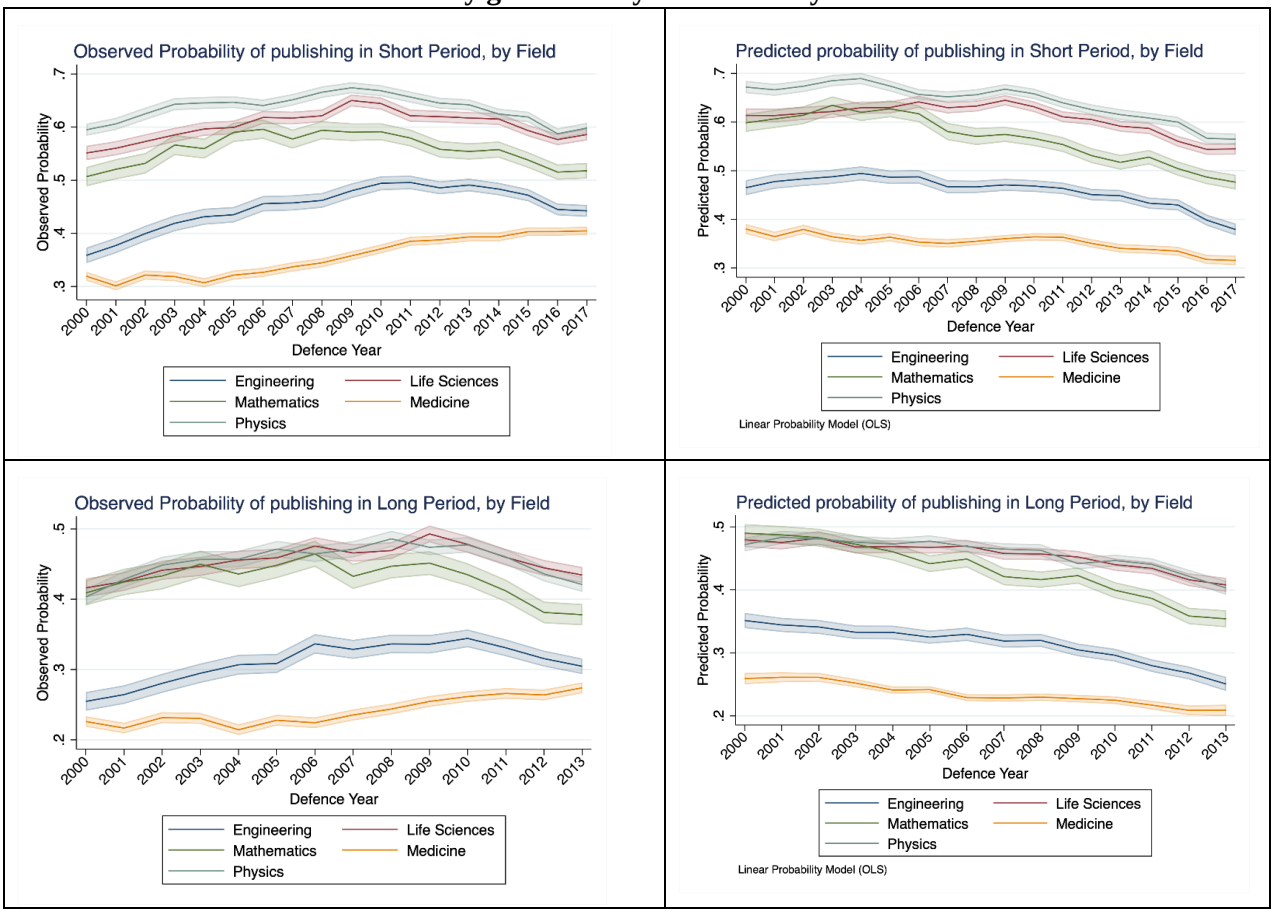
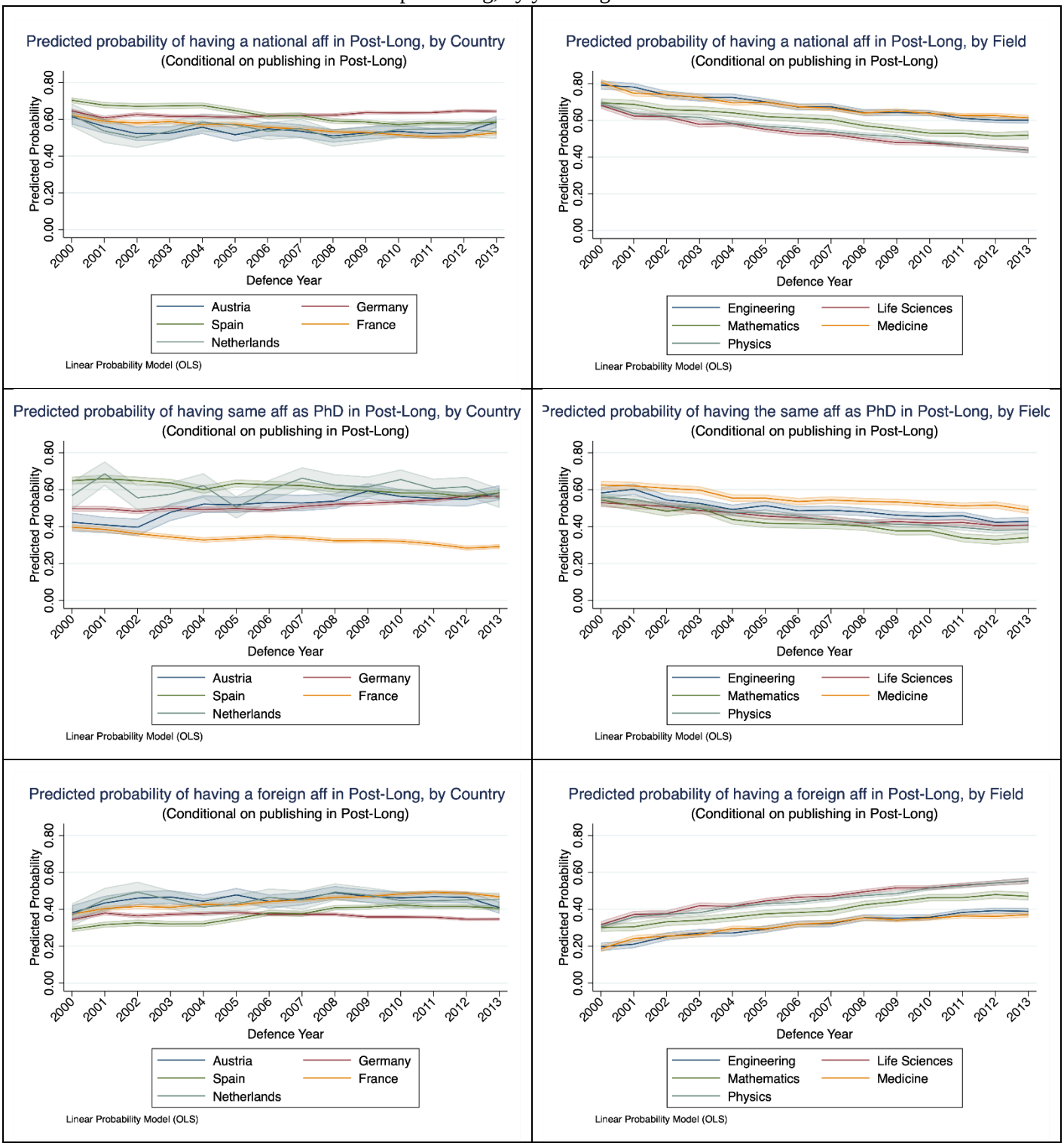
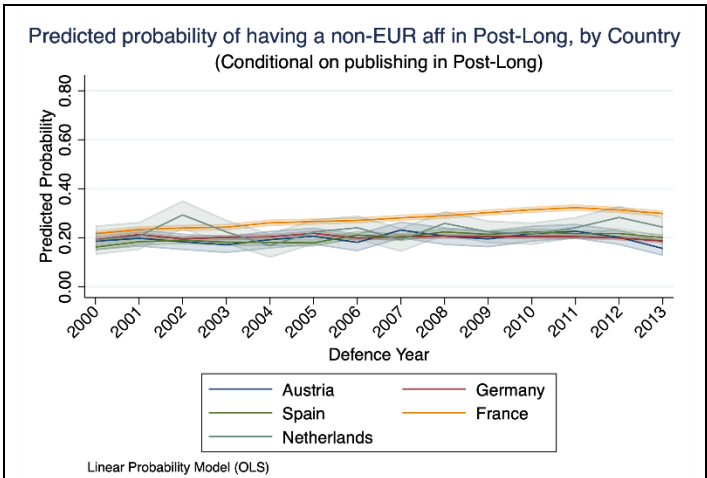
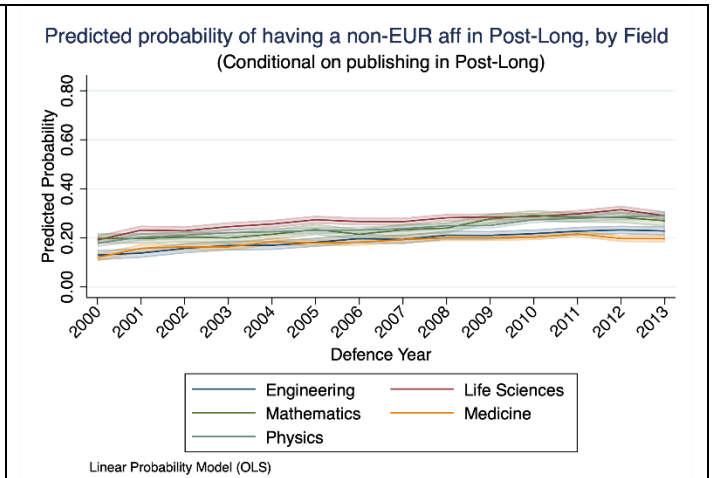


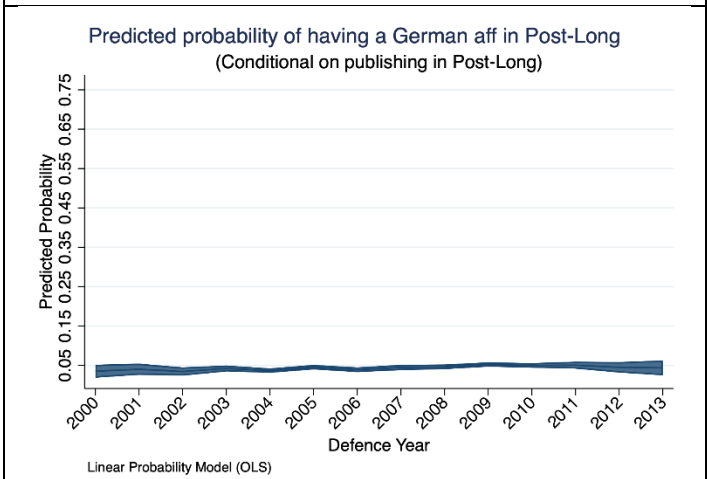
Figure C3: Predicted probability of having national vs foreign affiliations long after the PhD

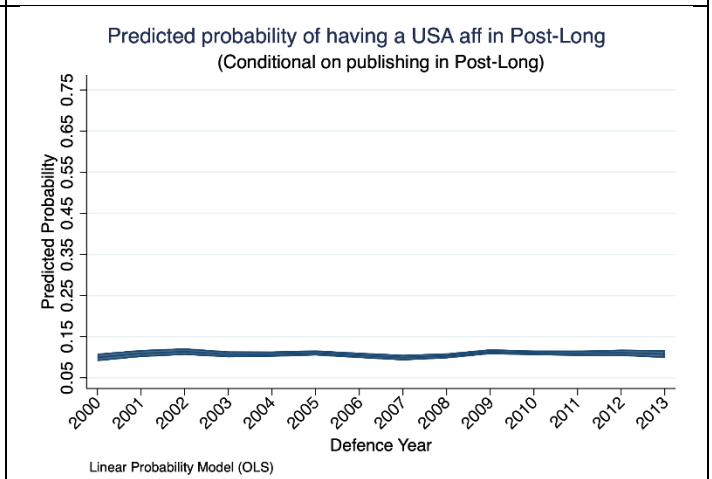
Conditional on publishing, by year of graduation

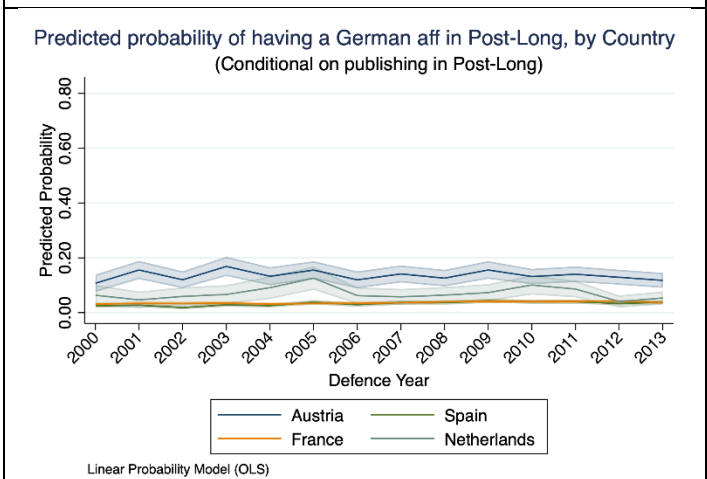


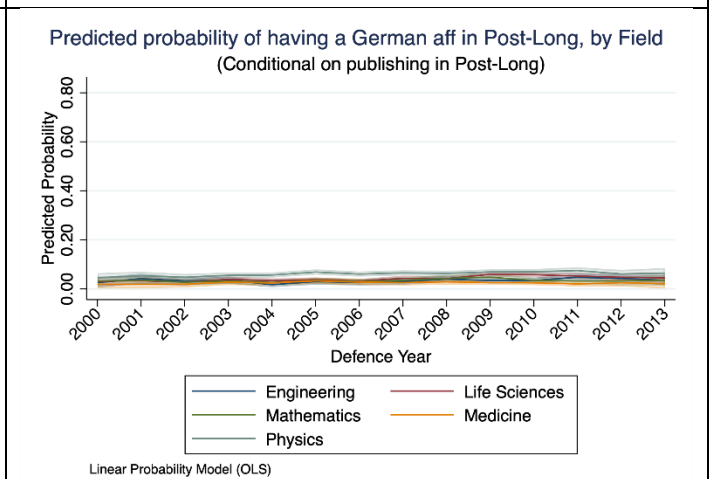

Predicted probability of having a non-EUR aff in Post-Long, by Country
(Conditional on publishing in Post-Long)
Predicted Probability
0.00 0.20 0.40 0.60 0.80
2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013
Defence Year
Austria
Germany
Spain
France
Netherlands
Linear Probability Model (OLS)


Predicted probability of having a non-EUR aff in Post-Long, by Field
(Conditional on publishing in Post-Long)
Predicted Probability
0.00 0.20 0.40 0.60 0.80
2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013
Defence Year
Engineering
Life Sciences
Mathematics
Medicine
Physics
Linear Probability Model (OLS)


Predicted probability of having a German aff in Post-Long
(Conditional on publishing in Post-Long)
Predicted Probability
0.05 0.15 0.25 0.35 0.45 0.55 0.65 0.75
2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013
Defence Year
Linear Probability Model (OLS)


Predicted probability of having a USA aff in Post-Long
(Conditional on publishing in Post-Long)
Predicted Probability
0.05 0.15 0.25 0.35 0.45 0.55 0.65 0.75
2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013
Defence Year
Linear Probability Model (OLS)


Predicted probability of having a German aff in Post-Long, by Country
(Conditional on publishing in Post-Long)
Predicted Probability
0.00 0.20 0.40 0.60 0.80
2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013
Defence Year
Austria
Spain
France
Netherlands
Linear Probability Model (OLS)


Predicted probability of having a German aff in Post-Long, by Field
(Conditional on publishing in Post-Long)
Predicted Probability
0.00 0.20 0.40 0.60 0.80
2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013
Defence Year
Engineering
Life Sciences
Mathematics
Medicine
Physics
Linear Probability Model (OLS)

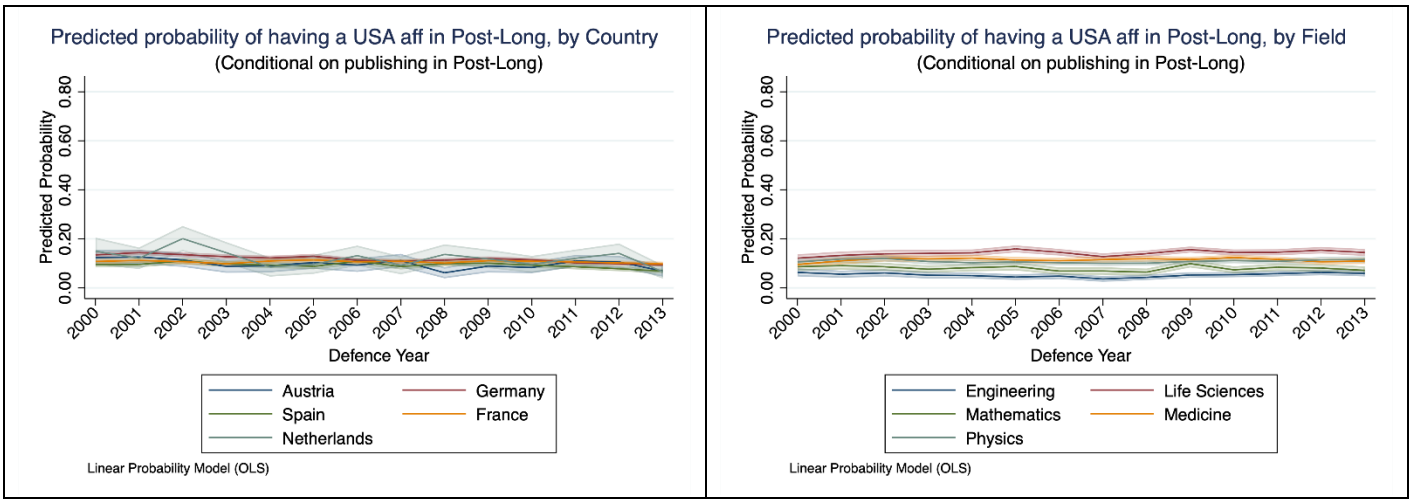


Figure C4: Predicted probability of having a corporate affiliation long after the PhD

Conditional on publishing, by year of graduation

