doctrack

2021 EPO Academic Research Program (EPO-ARP)

STEM Doctoral Graduates and Inventive Activities in European Countries (DOC-TRACK)

DOC-TRACK 2022-2025

Final Technical Report

France, Germany, Spain, Netherlands

Austria, UK, Italy

DISCLAIMER This is the final technical report of the DOC-TRACK project for the European Patent Office. Please contact the project coordinator (catalina.martinez@csic.es) before citing and/or quoting, to make sure you refer to the latest published version of the research findings. More information at https://doc-track.eu/

December 2025

CSIC Consejo Superior de Investigaciones Científicas | IPP Instituto de Políticas y Bienes Públicos

BxSE Bordeaux Sciences Economiques Bordeaux School of Economics

UNIKASSEL VERSITÄT | INCHER International Center for Higher Education Research

Universiteit Utrecht

Università Bocconi ICRIOS The Invernizzi Center for Research on Innovation, Organization, Strategy and Entrepreneurship

MANCHESTER 1824 The University of Manchester

EPO-ARP 2021

Thematic area title with the corresponding code SA-RA1-TA2: Value creation through university-industry technology transfer in Europe
Title of the research scheme DOC-TRACK: STEMM Doctoral Graduates and Inventive Activities in European Countries
Research theme Data linkage between PATSTAT and doctoral graduates' dissertations and publications
Abstract This report describes the DOC-TRACK methodology and the results obtained from its application to seven European countries (Austria, France, Germany, the Netherlands, Spain, the UK, and Italy). The methodology matches data on doctoral graduates obtained from Electronic Theses and Dissertations (ETD) repositories to patent and publication data, based on supervised machine learning algorithms. We use it to build a new dataset with information on the publishing and patenting activity of STEMM doctoral graduates in the selected countries, for the period 2000-2020. For the publications, we provide citation-based metrics to their distance from patents, whether signed by the graduates or (most often) other inventors.

Project team

Institute of Public Goods and Policies – Spanish National Research Council (IPP-CSIC, Spain): Catalina Martínez (coordinator), Alberto Corsini, Luis Sanz-Menéndez, Laura Cruz-Castro.

Bordeaux School of Economics – University of Bordeaux (BxSE, France): Ernest Miguélez (co-researcher 1), Francesco Lissoni, Andriy Romanyuk; associated researcher: Michele Pezzoni (University of Nice).

International Center for Higher Education Research – University of Kassel (INCHER, Germany): Guido Buenstorf (co-researcher 2), Johannes Koenig, Burcu Ozgun.

Dept. of Human Geography and Planning - Utrecht University (The Netherlands): Andrea Morrison (co-researcher 3); associated researcher: Fabiana Visentin (University of Maastricht).

Dept. of Management and Technology - Bocconi University (Italy): Stefano Breschi (co-researcher 4)

Manchester Institute of Innovation Research - University of Manchester (UK): Cornelia Lawson (co-researcher 5), Xin Deng, An Yu Chen, Liangping Ding.

Acknowledgements

We are grateful for the financial support from the European Patent Office Academic Research Program (EPO-ARP 2021) and contract extensions to include Austria, Italy and United Kingdom in the database. This work has also benefited from additional funding to individual teams. The BSE team thanks the French National Research Agency TKC project (ANR-17-CE26-0016) as well as the French State in the framework of the Investments for the Future programme IdEx université de Bordeaux / GPR HOPE. The University of Kassel team thanks Deutsche Forschungsgemeinschaft (DFG 447967785 - FOR 5234) and German Bundesministerium für Forschung, Technologie und Raumfahrt (BMFTR 16RBM1011 and 16WIK2101A/German Competence Network for Bibliometrics). The CSIC team thanks the Spanish National Research Agency INNDOC project (AEI PID2023-149135NB-IOO) and the Juan de la Cierva postdoctoral grant (JDC2023-052972-I). The BSE and CSIC team thank the CNRS-CSIC Laboratoire International Associé ALLIES-IRP (LIA2019FR1). The Bocconi team gratefully acknowledges financial support from the projects Multilayered Urban Sustainability Action (project code ECS 000037) and Growing Resilient, Inclusive, and Sustainable (project code PE00000018), funded by the European Union's NextGenerationEU initiative under the National Recovery and Resilience Plan. The Manchester team acknowledges financial support from the Innovation and Research Caucus (IRC/UKRI FFOpen001). We also thank the support of the institutions responsible of the Electronic Doctoral Theses repositories in France (ABES), Spain (TESEO, Ministry of Science, Innovation and Universities), the Netherlands (DANS), Germany (DNB) and Austria (ONB); and for their participation at the kick-off meeting of the DOC-TRACK project in September 2022:Tomás Mayoral and José María Gallego from the Spanish Ministry; Chris Baars from DANS in the Netherlands; Mäité Roux, Olivier Cian and Mathias Eon from ABES in France; and Carmen Simon (then head of the Postgraduate and Specialisation department at CSIC). We are grateful to Pauline Menu, Carla Fournier, Luca Codecasa, Francesca Ammerata, Bradley Butcher, Aman Araissi, Valentine Petzold, Baptiste Comby, Lucien Boucart, and Samuel Desneulin for their excellent research assistance and help in building the training and validation publication datasets, as well as to Eugenia Espinosa, Estefanía Herrán, and Melanie Ortíz for their outstanding work with the manual annotation for the patent datasets. We are grateful to participants at the DOC-TRACK kick-off meeting in Madrid in September 2022, the STI2024 Conference held in Berlin in September 2024, and the EmELHE workshop in Kassel in June 2025 for their comments. Lastly, we are very grateful to Cornelia Lawson for her comments on the interim report, to Alexander Klenner-Bajaja and Marco del Rey for their advice on how to implement the methodology; and to Yann Ménière and Victor Arribas for their support and guidance.

Research results and views included in this report are only those of the authors and do not necessarily represent the views of funding institutions. All remaining errors are our own.

Table of contents

1. Introduction

Human capital with advanced education in Science, Technology, Engineering, Mathematics, and Medicine (STEMM) is a key input to the invention processes of firms, universities, and public laboratories. Doctoral graduates, in particular, contribute to the advancement of science and technology, often beginning to do so during their doctoral studies (Buenstorf and Heinisch, 2020; Corsini et al., 2022). Being trained as researchers, they are expected to acquire and produce knowledge at the science frontier and to contribute directly to their laboratories' scientific and inventive production. With increasing "publish-or-perish" pressure (which the prevailing science funding systems impose on both them and their supervisors), their scientific production in the early career stages is increasing in terms of quantity and quality (Larivière, 2012; Shibayama, 2019) and reflects changing patterns of collaboration, evaluation and career prospects. How much of this pressure translates into more technological innovation remains, however, an open question, whose answer requires more and better data than those collected so far by both the national and international statistical offices and individual science and innovation scholars, especially in Europe. The DOC-TRACK project explicitly addresses this data gap.

In order to do so, DOC-TRACK exploits the increasing time depth, cross-country coverage, and information accuracy of data on doctoral dissertations, which is made possible by the diffusion of Electronic Doctoral Theses (ETD) repositories. In many countries, these repositories are created and maintained by national libraries, enabled by universities' legal obligation to provide access to the dissertation contents or, at least, their metadata (such as the names of authors and supervisors, the title and abstract, and the university and year of defence). In other countries, universities manage their own repositories individually, but are equally engaged in providing access to them. The DOC-TRACK project focuses, in particular, on a number of countries with well-developed ETD repositories, namely France, Germany, Spain, the Netherlands, Austria, Italy, and the United Kingdom, and it links the dissertations and their authors to the authors of scientific publications and the inventors named in patent applications. Its ultimate objectives consist of identifying the doctoral graduates who contribute to innovation either directly or indirectly, both during their doctoral studies and afterwards, and in collecting and testing the usefulness of data that may help explore what factors explain these contributions.

We measure direct contributions to invention by counting the patents signed by the doctoral graduates, whether alone or with co-inventors, before and after their thesis defence.[1] As for indirect contributions, we build upon Ahmadpoor and Jones (2017) and

[1] For simplicity, even if we refer to 'patents' or 'patented inventions' throughout the text, our focus is mostly on patent applications. More specifically, when we use patent data, we rely on information included in two types of patent documents: 1) published EPO patent applications from PATSTAT (for the

consider any scientific publication authored or co-authored by the doctoral graduates that is either cited directly or indirectly by one or more patent documents, whether signed by the graduate or not. A publication is directly cited by a patent whenever we find it listed either among the patent's front-page citations or among the citations included in the invention description; and it is indirectly cited when it originates a citation chain that ultimately leads to a patent document (the doctoral graduate's publication is cited by other publications, which are themselves cited by a patent or by other publications similarly leading to a patent citation). A simple metric (the number of citations composing the citation chain between a patent and each publication, whether directly or indirectly cited) allows us to assess the distance between a doctoral graduate's scientific production and its use in the invention realm.

Throughout the report, when analysing the graduates' productivity, we refer to periods during and after doctoral studies. As the ETD repositories do not generally provide information on the start date of the doctoral degree programs, we define the period of doctoral study as ranging from three years before to one year after the defence year, as shown in Figure 1.[2]

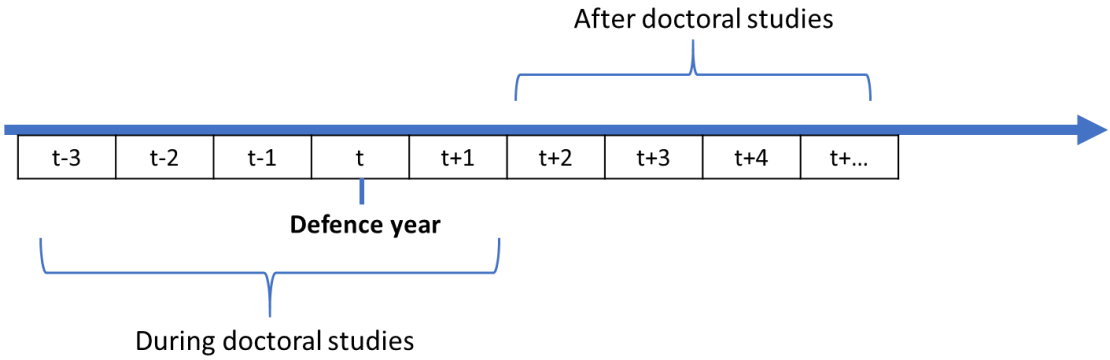
For cross-country comparability purposes, we also classify the doctoral dissertation in five large disciplines, namely: i) *Engineering* (comprising General Engineering, Chemical Engineering, and Energy); ii) *Life Sciences* (comprising Agricultural and Biological Sciences, Biochemistry, Genetics and Molecular Biology, Immunology and Microbiology, Neuroscience, Pharmacology, Toxicology and Pharmaceutics); iii) *Mathematics & Computer Science* (including Mathematics and Computer Science); iv) *Physical Sciences* (comprising General Physics and Astronomy, Material Sciences, 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).[3] Note that, in all figures, we use the term *Physics* to refer to *Physical Sciences.*

graduate-inventors matching) and 2) EPO and USPTO published patent documents citing non patent literature as included in the Reliance on Science (RoS) database by Marx and Fuegi (2020a, 2020b) (for the publication-to-patent metric).

[2] Adding one year after the thesis defence accounts for potential delays in the publication of outputs related to doctoral studies. This approach is commonly used in the literature (see, for instance, Corsini et al., 2022, and Koenig, 2025).

[3] This classification follows the "All Science Journal Classifications" (ASJC) system used by Scopus to categorise scientific journals. Specifically, we refer to the three Scopus Subject Areas in STEMM disciplines, except that we have divided the Physical Sciences into three subcategories. https://service.elsevier.com/app/answers/detail/a_id/14882/supporthub/scopus/~/what-are-the-most-frequent-subject-area-categories-and-classifications-used-in/

Figure 1. Definition of doctoral study period



Based on this methodology, we produce data that, according to our explorative analyses, reveal a substantial contribution of doctoral graduates to inventive activities, which has largely escaped –so far– the attention of both scholars and policy-makers.

Our main results, which consider the graduation cohorts from 2000 to 2020, can be summarised as follows:

- *Doctoral graduates' publication propensity:* France has the highest share of publishing graduates over the period considered (around 80% already publish during the PhD), whereas Italy stands out in the most recent cohorts (almost 90% publish during the PhD). Austria, the Netherlands, Spain, Germany (excluding medical doctors), and the UK exhibit broadly similar trends, with 50–80% of graduates publishing already during the PhD and with upward trends. Physical Sciences (broadly defined) shows the highest publication propensity across all fields. France consistently records the largest share of publishing graduates across all disciplines, followed closely by Italy and the UK. The Netherlands, Austria, and Spain follow with broadly similar levels, while Germany lags in Medicine, reflecting the structure of its medical doctorates. Overall, publication propensities are high and comparable across countries, but France remains at the top, and Italy shows a marked increase in the most recent cohorts.

- *Doctoral graduates' scientific productivity:* The cross-country variations of productivity levels are more marked than those of publication propensity. During doctoral studies, Italian and Dutch graduates are the most productive across nearly all disciplines, closely followed by Spanish and UK graduates, while French graduates generally show moderate output despite their high propensity to publish. After graduation, productivity rises in all countries, with Italy emerging as the strongest performer in several fields, particularly Physical Sciences and Mathematics & Computer Science. Germany and Austria record the lowest productivity both during and after the PhD, reflecting the larger share of graduates who –according to the literature– do not pursue an academic career.

Regarding gender disparities, men consistently publish more than women, both during and after studies. The gender gap is narrowing in the recent cohorts when considering publications after doctoral studies. The smallest gender gap during doctoral studies is in Life Sciences, but it widens post-graduation. Overall, productivity differences appear driven by structural features of national research and higher education systems rather than disciplinary composition alone.

- *Doctoral graduates' publications as knowledge inputs for patents:* An increasing share of graduates' publications is linked to patents over time, with the publications of doctoral graduates from the Netherlands standing out as the most connected. Direct citations from patents to publications remain stable over time. Austria and the Netherlands show higher shares, suggesting stronger science-industry links, also due to their scientific specialization. We observe that starting in 2013, graduates' publications seem to move farther from the technological frontier (they become more distant from patents), with France and Spain being the farthest. Concerning gender differences, men are more likely to have a publication directly cited by a patent; whereas, until 2013, female graduates were slightly more likely to be connected to the frontier (at any distance) than their male counterparts. From 2013, both genders' outputs are more distant from the technological frontier, with women's output standing farther. We should note, however, that these results are likely to be affected by some truncation issues for the last cohorts, despite our focus on a rolling time window to ensure that publications from different points in time have the same probability of being cited by patents.

- *Doctoral graduates as inventors of EPO patent applications:* France has the highest overall share of graduate-inventors (between 20 and 23% in early 2000s cohorts), followed by Germany (without considering medicine); but differences across countries narrow over time (also due, in part, to truncation issues). During doctoral studies, less than 5% of graduates in all countries become inventors (with some exceptions in 2011 in France and 2019 in Austria), though this share is rising. By discipline, Engineering and Physical Sciences (broadly defined) dominate patenting activity, while Medicine has the lowest rate. By country and discipline, France, Germany, and Austria exhibit very high inventive activity in Engineering and Physical Sciences, especially in early cohorts (up to almost 25% in France across both disciplines). France has the highest female inventor share (between 12 and 15% in early cohorts). The UK has the lowest share across all cohorts (less than 3%). Gender gaps are partly explained by women's lower representation in patent-intensive disciplines across countries (Engineering and Physical Sciences).

The DOC-TRACK methodology is replicable across countries whose universities make – or can be invited to make– their doctoral graduates' dissertations available via either

centralised or individual ETD repositories. Equally important, the methodology is also replicable over time, as long as the ETD repositories are regularly maintained. This suggests the possibility of using it for setting up a permanent observatory at the European level, for both the observation of general trends and the production of in-depth, topic-specific studies.[4]

The remainder of the report is organised as follows. Section 2 provides an overview of the DOC-TRACK project's activities as well as the data sources used for each of them. The following sections provide details on each activity, namely: the matching between doctoral graduates and authors of scientific publications (Section 3); the calculation of direct and indirect citations from patents to publications (Section 4); and the matching between doctoral graduates and inventors (Section 5). Each of these sections also includes a description of the methodological results. Section 6 employs the newly constructed DOC-TRACK database to analyze graduates' productivity, providing descriptive evidence across graduation cohorts and cross-country comparisons that highlight graduates' relevance for innovation studies. Section 7 concludes.

2. Project overview and data sources

Figure 2 provides a first overview of the data sources and main activities of the DOC-TRACK project. The former are:

i) The national ETD repositories of France (*Theses.fr*), Germany (*DNB catalogue and DissOnline*), Spain (*TESEO*), the Netherlands (*Narcis*), Austria (*ONB catalogue*), the UK (*EThOS*), and Italy (*BCNF OPAC and OAI-PMH*), from which we extract information on the STEMM doctorate recipients and their theses, from 2000 to 2020;
ii) The OpenAlex publication and citation database (release 2024-09-27);
iii) The Scopus publication database, published by Elsevier, for the years between 1996 and 2021;
iv) The PATSTAT database, release Autumn 2024;
v) The Reliance on Science (RoS) database, as of 2024, which contains information on patent citations to scientific literature, for both USPTO and EPO patent documents (Marx and Fuegi, 2020a, 2020b).

As concerns the activities, we first identify and extract from each ETD repository the doctoral dissertations in STEMM disciplines. Since ETD repositories are not harmonized across countries, with dissertations catalogued according to different national classification systems, we identify the STEMM theses making use of different criteria for each repository (see below). This suggests caution when comparing absolute numbers of STEMM theses (and, consequently, the numbers of STEMM graduates and their

[4] A list of DOC-TRACK work-in-progress, working papers and published articles will be updated regularly in the project website, https://doc-track.eu/

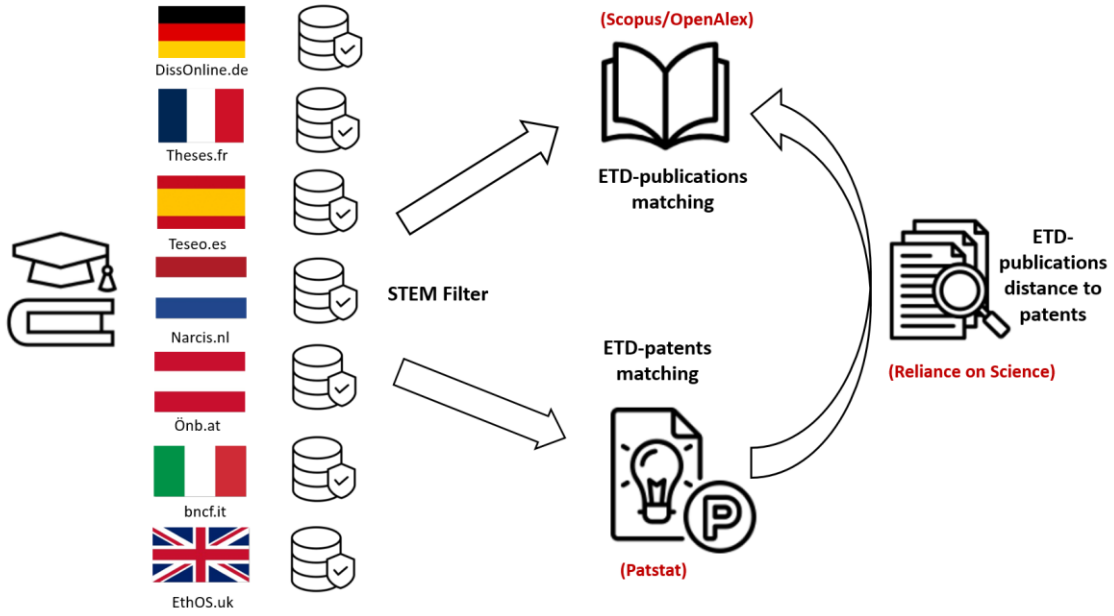
publications and patents) across countries, and it may also generate compositional effects that are equally responsible for differences in graduates' propensity to publish and/or patent.

Second, for each country, we match the doctoral graduates to both the authors of scientific publications, making use of the OpenAlex and Scopus databases, and the inventors of patent applications at the EPO. The matching methodology first pairs the names of the graduates to those of the authors and inventors, then it applies a Random Forest algorithm (Gareth et al., 2013) to remove the false positives, based on a training dataset with numerous filter variables (such as the presence –among the presumed co-authors and co-inventors– of the name of the graduate's supervisor; or the similarity between the title of the graduate's dissertation and the title of the paired publications and patents). Although the methodology is the same across countries, the weights assigned by the Random Forest algorithm to the various filter variables vary across countries. This is a second reason for exercising caution when undertaking cross-country comparisons.

Third, we look at citations to scientific literature in patent documents from both the EPO and the USPTO, relying on Marx and Fuegi's (2020a, 2020b) database 'Reliance on Science' (RoS). We use OpenAlex to link the graduates' publications to the publications included in the RoS database and identify the graduates' publications that are directly or indirectly cited by a patent, where indirectly means that the graduate's publication is at the root of a citation chain linking several publications and –ultimately– a patent.

In the remainder of this section, we provide further details on the contents and limitations of the various ETD repositories, as well as on the methodology for extracting STEMM dissertations. For details on the methodology of the other actions, see the following sections.

Figure 2. DOC-TRACK project overview: data sources and activities



2.1 ETD repositories: contents and limitations

All the ETD repositories used by the DOC-TRACK project consist of nationwide, open-access data collections managed by dedicated public agencies. The latter encourages both the use and the development of their repositories, including for data linkage. The selected ETD repositories contain detailed personal and bibliographic information – albeit with cross-country variations in terms of completeness and coverage– such as: the doctoral graduates' surname and name (or initials); the title, abstract and defence date of the theses; the surname and name (or initials) of supervisors and possibly jury members; as well as information on the discipline and degree-granting institutions. In what follows, we examine each ETD in detail.

2.1.1 France

The doctoral theses of French graduates are collected in the National French repository of Electronic Doctoral Theses, whose metadata are recorded in the database *Theses.fr* and go back in time to 1981.[5] The repository is managed by ABES (Agence Bibliographique de l'Enseignement Supérieur), the French national bibliographic agency of higher education.

The dissertation data in *Theses.fr* come from the "Système Universitaire de Documentation" (*Sudoc*), the collective catalogue of the French higher education and research libraries and documentation centres. This includes the metadata of the theses

[5] See: https://theses.fr/ (last accessed November 23, 2023).

archived in printed format between 1981 and 2016, as well as of the theses directly archived in electronic format, starting 2006, via two centralized systems called *STAR* (for the theses that have been successfully defended) and *STEP* (for those still underway). Each French higher education establishment authorized to award doctoral degrees is responsible for producing the dissertation data. Hence, depending on the university and year, data may not be complete.

Each thesis recorded in *Theses.*fr is identified through a unique code, *Numéro National de Thèse* (NNT)[6]. Information on the thesis title, language, graduate's name and surname, supervisor and co-supervisors' names and surnames, defence date, discipline, and degree-granting institution are available for almost 100% of the theses. Information on thesis summary, doctoral school, committee members, and keywords is also present but with lower coverage, which ranges from 25% to 89%, depending on the university and year. Titles and summaries are reported both in French and English, although the English version is sometimes missing. The disciplines of the theses are recorded in two variables, respectively based on a fine-grained classification that comes from the *STEP* source and on the Open Access Initiative (OAI) classification. The OAI follows a Dewey decimal classification system also used for German theses (see below), making them comparable.[7]

Starting in 2006, all graduates, supervisors, committee members, and doctoral degree-granting institutions are also provided with a unique identifier, called IdRef, which links to an authority record to be used for identifying the authors of all documents included in *Sudoc*. The authority records contain both other identifiers for the same person, to be found in other bibliographic databases (for example, Orcid and Scopus ID), plus some biographical information, such as the year or country of birth, but not in a systematic way, and with some errors (especially for the country of birth). For the graduates who defended their theses before 2006, IdRef identifiers are generally created retrospectively if and when the graduate produces a scientific publication or other document included in the *Sudoc* catalogue. For the years of our interest, the IdRef coverage of doctoral graduates is almost 100%, with the missing values all concentrated in the early 2000s.[8]

[6] See: https://documentation.abes.fr/sudoc/formats/unmb/zones/029.htm (last accessed November 23, 2023).

[7] See: https://www.theses.fr/schemas/tef/recommandation/oai_sets.html (last accessed November 23, 2023).

[8] For more information in IdRef, see:ohttps://www.idref.fr/ (last accessed November 23, 2023). More generally, the IdRef records allow linking a person's records to other open access databases managed by the French government (see https://www.data.gouv.fr/fr/, last accessed November 23, 2023). But these do not include the year or country of birth, nor the gender, to which we would be interested and can only found in the IdRef authority records, if input manually by a librarian and for the only purpose of better identifying an author.

For the DOC-TRACK project, we retrieved the *Theses.fr* data via a dump made available by data.gouv.fr (gouv.fr, version downloaded in January 2024, JSON format), which contains all the metadata available for defence years from 1985 to 2023, although data after 2021 suffer from truncation problems.

2.1.2 Germany

The dissertation data for Germany used in DOC-TRACK result from a combination of different repositories and databases maintained by the German National Library (Deutsche Nationalbibliothek or DNB), as part of its institutional mandate to collect all printed or digital work published in Germany since 1913.[9] The extension of this mandate to doctoral dissertations dates back to the late 1960s, when the Conference of Ministers of Education and Cultural Affairs established that publishing the dissertation was an integral part of the academic qualifications of doctoral candidates, absent which no German university can confer the doctoral title. The DNB's full catalogue, including the bibliographic information on about a million dissertations, is available for research under a Creative Commons Zero License (CC0 1.0) and includes details on the year of submission, subject group, and the corresponding university name. In addition to its full catalogue, the DNB houses DissOnline, which is the largest ETD repository in Europe. Opened in 1998, the DissOnline portal was integrated into DNB's bibliographic system in 2012.[10] Metadata from DNB repositories are available online.[11]

The DOC-TRACK project has made use of two additional repositories, namely the "Gesamtauszüge Titeldaten" and the "Gesamtauszüge Normdatei (GND)". The first one contains meta information on all publications, including doctoral dissertations, while the latter contains information on their authors. The *Gemeinsame Normdatei* and *Gesamtauszüge Titeldaten* were obtained from the German National Library as of November 2021 in MARC21 and processed in R to create a relational database. Full texts of online dissertations were downloaded between August and September 2022 and read out using text mining methods. Online available tables of contents were downloaded in September 2022.

Altogether, these resources provide relevant information such as each dissertation's title, subject groups, author, university of affiliation, defence and publication dates, as well as, for some doctoral candidates, details about supervisors and jury members. Standard data often also contains additional personal information about individuals, such as birth date and birthplace.

[9] DNB - Über uns, https://www.dnb.de/DE/Ueber-uns/ueberUns_node.html (last accessed November 17, 2013).

[10] www.dissonline.de (last accessed November 25, 2023).

[11] https://www.dnb.de/DE/Professionell/Metadatendienste/Datenbezug/Gesamtabzuege/gesamtabzuege_node.html (last accessed November 25, 2023).

2.1.3 Spain

TESEO, which stands for *TESis Españolas Ordenadas* (Organized Spanish Theses), is the national database providing information on the doctoral dissertations defended at Spanish universities and doctoral centres since 1976. It is maintained by the Spanish Ministry of Science, Innovation and Universities, and includes information from more than 80 higher education institutions in all disciplines. The regulation that first established that doctoral theses must be deposited in open access in Spain is Royal Decree 99/2011, of January 28.

The bulk of TESEO data comes from the dedicated forms that doctoral graduates are required to fill in with information about themselves as well as their thesis, doctoral degree committee, and supervisors, under the supervision of their universities' administrations, which are in charge of entering the information in the system. For each dissertation, TESEO reports the title and a summary (mostly in Spanish), the name of the doctoral program, university, and department where the thesis was carried out, the graduate's and supervisors' names and surnames, as well as the defence date and names and surnames of the defence committee members. It also includes a list of keywords, in Spanish, chosen from the UNESCO Thesaurus nomenclature for fields of science and technology.[12]

Data for the DOC-TRACK project was provided directly by the Ministry of Universities, General Secretariat of Universities, Subsecretariat of Academic Research Activity, which provided a data dump (CSV files) in March 2023 to the IPP-CSIC team.

2.1.4 The Netherlands

The main data source for Dutch data is NARCIS, a digital platform for cataloguing the outputs of scientific research from the Netherlands established in 2004 by the Royal Netherlands Academy of Arts and Sciences (KNAW) and the Netherlands Organisation for Scientific Research (NWO). It gathered and integrated a large quantity of information from Dutch university libraries, including their ETD repositories, dating back to the 1990s. In July 2023, NARCIS was decommissioned, and a new portal, managed by the UKB (Dutch University Libraries and The Royal Library partnership) and SURF, has now succeeded it.[13]

NARCIS data come with some more limitations than the other ETD repositories. Most notably, they do not include disciplinary classifications for the theses and quite often do not provide the first names of graduates and supervisors, but only their initials (plus surnames). To address this issue, we enriched the original NARCIS data by matching them to external information sources (web-scraping of names based on other available information), at the cost of some data loss. Regarding the disciplinary classification, we

[12] https://vocabularies.unesco.org/browser/thesaurus/es/?clang=en (accessed November 23, 2023).
[13] See https://www.surf.nl/en (last accessed November 25, 2023).

assigned the theses into STEMM and non-STEMM categories based on text analysis of their titles, with an original machine learning model trained on data from *Theses.fr* (see Annex 2). This resulted in 53,925 theses classified as STEMM, with values of accuracy, recall, and F1 score all around 95%. This approach allowed us to maintain a comprehensive dataset and conduct our analysis exclusively on STEMM theses.

2.1.5 Austria

The Austrian ETD repository is based on the bibliographic catalogue maintained by the Austrian National Library (Österreichische Nationalbibliothek, ÖNB). This is similar to the German DNB catalogue, particularly regarding the legal requirements for the dissertation publication and the role of the national library in maintaining a comprehensive catalogue. The legal framework for dissertation publication in Austria is established by Section 86(2) of the Universities Act (Universitätsgesetz) 2002, which mandates that one copy of each successfully defended dissertation must be submitted to the ÖNB. Furthermore, universities regulate the obligation to publish dissertations within their respective study and examination regulations. As a result, the ÖNB catalogue contains a comprehensive collection of nearly all doctoral dissertations completed at Austrian universities. The ÖNB catalogue provides detailed metadata on dissertations, including: i) personal information of the graduate (full name); ii) university affiliation (name of the awarding university); iii) defence and examination details (year and month of defence, as well as, in some cases, details on the examination board); iv) dissertation metadata (title, department, keywords); v) supervisor information. Unlike the German ones, the Austrian dissertations are not systematically classified by subject. To identify the STEMM ones, we used the available classification codes and, where missing, the titles of doctoral degrees, the names of the university departments associated with the thesis, and the search keywords.

The ÖNB catalogue data were extracted in January 2024 and processed in Marc XML format. The dataset was then transformed into a relational database, ensuring structured access to dissertation records. Information about 29,496 STEMM theses was identified, accounting for 59.22% of all dissertations listed in ONB. Data are highly complete regarding bibliographic information. Information about candidates is less complete, such as birthdate and examination board details, which are missing for some entries. Supervisor details are recorded in 81.75% of all theses and 74.26% of STEMM theses.

2.1.6 The United Kingdom

The UK doctoral thesis data were obtained from EThOS (Electronic Theses Online Service), the British Library's national database for UK doctoral theses.[14] EThOS functions as a centralised digital repository designed to ensure the long-term preservation and accessibility of the UK's doctoral research output. EThOS covers

[14] https://www.bl.uk/collection/ethos (last accessed December 10, 2025)

approximately 98% of all PhDs awarded in the UK since 1787, making it one of the most comprehensive national thesis repositories in the world. EThOS is coordinated through institutional agreements between the British Library and UK universities. Each participating institution deposits doctoral theses, either directly or through institutional repositories, ensuring broad coverage across the UK higher education landscape. The metadata is available via Creative Commons licence.[15] As of November 2023[16], the collection contains more than 600,000 theses, spanning all major academic disciplines, of which 214,036 are PhD dissertations published in STEMM fields between 2000 and 2020.

Bibliographic metadata covers the name of the graduate, awarding institution, year of publication, discipline, title, and abstract (for 72% of entries) and thus provides a comprehensive and structured dataset suitable for large-scale quantitative analysis of doctoral education in the UK. Supervisor information is available but shows gaps, particularly in older records, with less than 10% of pre-2010 records and less than 50% of post-2009 records containing supervisor details. To address this, a data recovery strategy was implemented, involving direct verification through universities' library portals for theses awarded since 2010, which combined web-crawling and Named Entity Recognition (NER) to extract supervisor names from thesis PDF acknowledgements. The procedure achieved a validated accuracy of 95.44% in confirming at least one supervisor for theses from 2010 onwards, allowing us to achieve a completion rate of 81%.

2.1.7 Italy

The construction of a comprehensive database of Italian doctoral theses relied on two complementary sources. The first is the Online Public Access Catalog (OPAC) of the *Biblioteca Nazionale Centrale di Firenze* (BNCF), which has been the legal repository for doctoral theses since the 1980s. The second consists of the institutional research repositories of Italian universities, accessible via the OAI-PMH protocol, from which metadata and, in some cases, full texts have been harvested.

The OPAC of the BNCF represents the main national collection of doctoral theses.[17] It includes both the early publication-based deposits and the later electronic submissions, whether transmitted on CD/DVD, uploaded to *Magazzini Digitali* (the BCNF's digital archives), or harvested from university repositories.[18] As a result, it is the most extensive source available, covering over 210,000 doctoral theses (as of the time of last access). However, its primary function has been to guarantee legal preservation rather than to

[15] See https://doi.org/10.23636/rcm4-zk44 (last accessed December 10, 2025).

[16] November 2023 is the last available version of EThOS, which was suspended the same year following a severe cyber attack.

[17] See https://opac.bncf.firenze.sbn.it (last accessed December 31, 2023).

[18] See Storti, Chiara (2019). "Il deposito, la valorizzazione e la conservazione delle tesi di dottorato nell'esperienza di Magazzini digitali: un contributo per la ricerca e l'accesso". *JLIS.it: Italian Journal of Library and Information Science* 10(4): 114–124.

facilitate discovery and research use. As a consequence, the catalog often lacks essential metadata (such as university, year, or disciplinary field), contains several duplicate records, and includes occasional misclassified or even non-doctoral documents. Moreover, the theses themselves remain accessible only on-site at the national libraries, limiting the usability of the collection for researchers and policymakers.

Since the late 2000s, most universities have created institutional repositories that collect and disseminate doctoral theses in open access. These repositories, generally OAI-PMH compliant, provide more structured metadata, usually following the Dublin Core standard. They also tend to offer online access to the full text of theses.[19]

We leveraged these two sources by cross-referencing the theses collected from the BNCF's OPAC with those harvested from the OAI repositories of individual universities. In total, the final dataset comprises 202,142 distinct PhD theses published between 1986 and 2023, of which 159,899 were published between 2000 and 2020. To identify STEMM theses, we used the disciplinary field reported in the metadata. For the theses lacking this information, we applied machine learning algorithms based on the thesis titles.[20] Overall, we identified 106,500 theses in STEMM fields published in the period 2000-2020 (around 67% of the total).

Although the use of metadata from the institutional repositories of individual universities helped fill some of the gaps present in the BNCF's OPAC, information remained incomplete for certain fields. Notably, for about 18% of the STEMM theses in the sample, we were unable to identify the university where the thesis was defended. Similarly, the names of supervisors are available for only about 78% of all STEMM theses.

2.2 Cross-country ETD comparison

Table 1 lists the ETDs used in the DOC-TRACK project, together with the name of the organization in charge of their maintenance, the years covered in our analysis, and the number of theses (with an estimate of the share we classified as STEMM). The German repository is the largest one, followed closely by the UK, France, Spain, Italy, and, more distant, the Netherlands and Austria. This ranking reflects that of the countries' population size, but it also depends on the structure of their higher education system. In particular, in Germany all medical dissertations completed at the end of single-cycle degrees in Medicine are equated to doctoral ones, whereas in other countries they are considered as master's theses; second, the German industry regards doctoral degrees more favourably than industry elsewhere in Europe, thereby increasing the number of doctoral graduates, who have more career options outside academia (Diez et al., 2000;

[19] However, practices differ significantly across institutions: some repositories are complete and well-curated, while others contain only partial collections or display inconsistent metadata. In a number of cases, advisors were mistakenly listed as thesis authors, or disciplinary classifications were missing or idiosyncratic.

[20] All thesis titles in Italian were translated into English using the DeepL API.

Enders, 2002; Buenstorf et al., 2023). In the UK, in addition to PhD dissertations, there are traditionally a large share of professional doctorates, e.g., Doctorate in Medicine (MD). We were able to identify and exclude these professional doctorates from the UK sample of regular research theses; our data do not allow us to do so for Germany. We collected a total of 1,645,598 theses from the 7 countries under analysis, of which 1,105,263 (67.2%) are in STEMM disciplines.

Appendix A summarizes all the data features available in the seven repositories, allowing readers to appreciate each repository's strengths and weaknesses as well as their comparability. Information in *Theses.fr* and TESEO is the most complete, followed by DissOnline and ONB, with Narcis and BNCF.IT lagging behind.

From Figure 3 onward, all statistics we provide are based exclusively on STEMM disciplines, as these make up the DOC-TRACK dataset. Figure 3 provides an overview of the temporal distribution of the dissertations by defence year. The trend is generally increasing, but with some divergences across countries. In Germany, there is a high volume of theses due largely —though not exclusively— to medical dissertations, as shown by the significantly lower levels when excluding this discipline. For this reason, from this point onward, we include in all figures –whenever relevant– a separate line for "Germany without medicine". As for Spain, the number of dissertations generally increased from 2008 to 2014, but spiked up dramatically in the two following years, during which Spain recorded more dissertations than France (a country with double its population), and then slumped back to the 2008 level afterwards. 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 – until then with no limited time for delivering the dissertation– to three years (up to 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). Italy displays a distinct pattern: after steady growth through the mid-2010s, the number of dissertations declines noticeably from 2017 onward. This decline is largely attributable to a series of reforms, most notably Law 240/2010 and Ministerial Decree 45/2013, which introduced stricter accreditation requirements for doctoral programmes, including minimum faculty sizes and mandatory teaching and research standards. Combined with sustained cuts in public funding for higher education, these changes reduced the number of accredited programmes and the intake capacity of existing ones. The decline in dissertations after 2017 is therefore likely to reflect a structural contraction of doctoral provision, rather than temporary reporting issues.[21]

[21] The temporary drop in the number of Italian theses in 2006–2007 is instead fully explained by the transition from paper-based to electronic submission systems, during which a number of dissertations were not transferred correctly to the BNCF.IT repository and were effectively "lost in translation."

Table 1. ETD 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 ⬇ 166,607 STEMM theses (65%)
Germany	DNB catalogue; DissOnline	Deutsche Nationalbibliothek (DNB)	All universities, 2000-2020	544,237 theses ⬇ 420,927 STEMM theses (77%)
Spain	TESEO	Spanish Ministry of Science, Innovation and Universities	All universities, 2000-2020	204,506 theses ⬇ 113,772 STEMM theses (56%)
Netherlands	Narcis	Dutch National Center of Expertise and Repository for Research Data (DANS)	All universities, 2000-2020	81,571 theses ⬇ 53,925 STEMM theses (66 %)
Austria	ONB catalogue	Österreichische Nationalbibliothek (ONB)	All universities, 2000-2020	49,808 theses ⬇ 29,496 STEMM theses (59%)
UK	EThOS (Electronic Theses Online Service)	British Library (BL)	All universities, 2000-2020	347,838 theses ⬇ 214,036 STEMM theses (61.5%)
Italy	BCNF OPAC, plus OAI-PMH repositories of individual universities	Biblioteca Centrale Nazionale di Firenze (BNCF)	All universities, 2000-2020	159,899 theses ⬇ 106,500 STEMM theses (67%)

Figure 3. Dissertations collected by defence year and country (all disciplines)

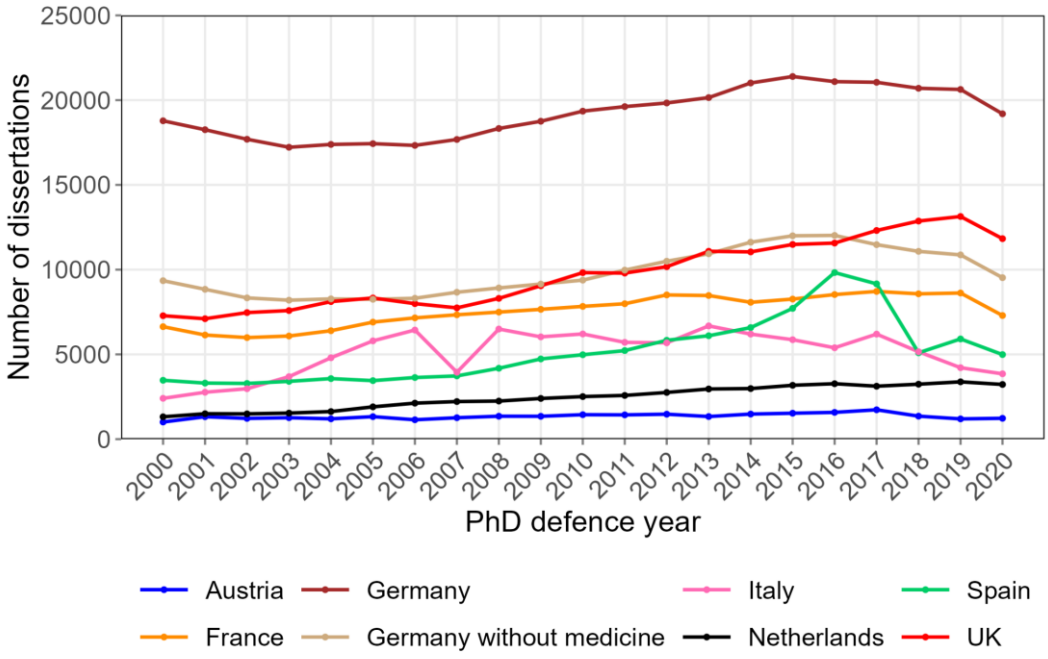


Figure 4 presents the disciplinary composition of theses by country, both including and excluding theses in Medicine. Disciplinary classification was relatively straightforward for Spain, Germany, France, Italy, and the UK, thanks to the presence of field classification in their metadata (UNESCO codes, DDC classes, SSD classes, and the REF Units of Assessment[22]). For Austria, we had to infer it based on the awarded degree names, faculty and department names, and thesis keywords. For the Netherlands, where no disciplinary information was available, and for a relatively small number of cases in other countries with missing metadata, we assigned the most likely STEMM field using a machine-learning model trained on thesis titles for which a discipline could be identified.[23] From panel (a), we observe, once again, the disproportionate number of medical dissertations in Germany (49%). France shows the highest share of theses in

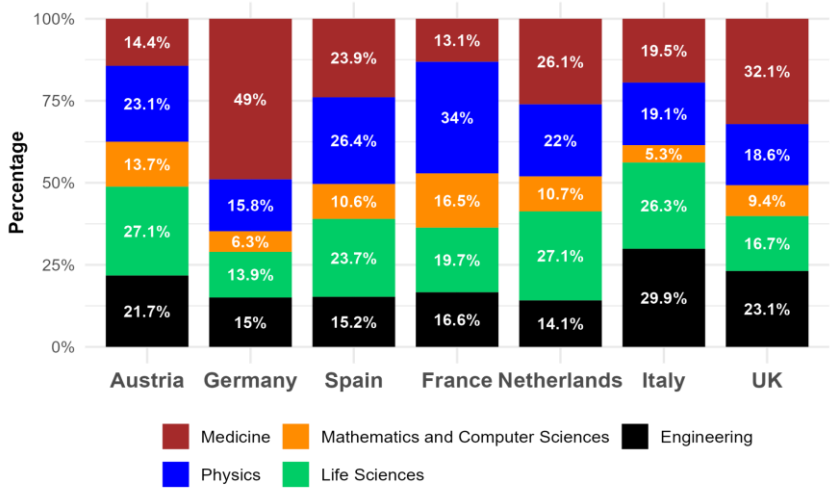
[22] UNESCO codes: https://skos.um.es/unesco6/view.php?l=en&fmt=1
DDC classes (Germany): https://www.dnb.de/EN/Professionell/DDC-Deutsch/ddc-deutsch_node.html
DDC classes (France): https://theses.fr/schemas/tef/recommandation/index.html
SSD classes: *settori scientifico-disciplinari* (scientific disciplinary codes as defined by the Ministry of University and Research): https://www.mur.gov.it/it/aree-tematiche/universita/docenti-universitari-e-carriera-accademica/settori-concorsuali-e-settori
REF Units of Assessment: https://2029.ref.ac.uk/panels/units-of-assessment/

[23] We use a neural-network text-analysis algorithm to embed thesis titles into vectors that capture the semantic meaning of the words in each title, and we assign each thesis to the discipline whose vector is most similar (Mikolov et al., 2013). Disciplinary information obtained through this procedure covers 100% of Dutch theses, 24.0% of Austrian theses, 4.3% of German theses, 1.2% of French theses, and 1.6% of UK theses.

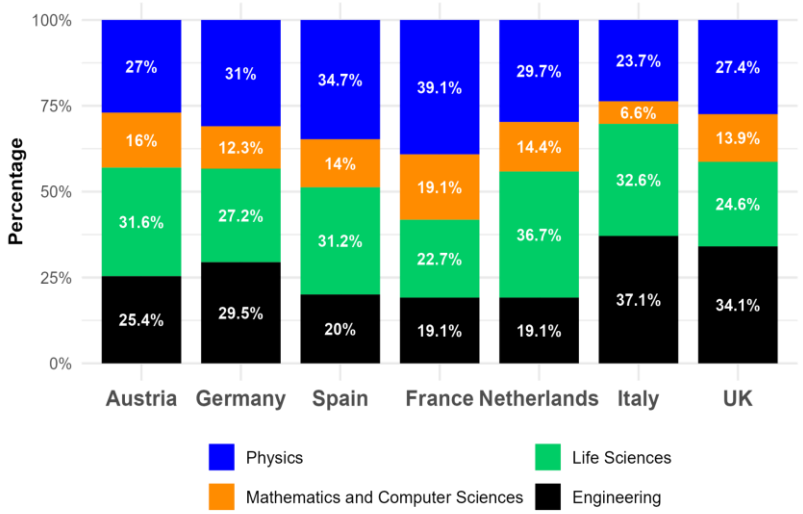
Physical Sciences and in Mathematics & Computer Science (34% and 16.5%, respectively), while Austria and the Netherlands have the largest concentration in the Life Sciences (27.1%). When excluding medical theses, panel (b) shows that the concentration of theses in life sciences in the Netherlands rises further relative to other countries (36.7%), followed by Italy (32.6%). Moreover, in Italy and the UK, the concentration of theses in Engineering exceeds that of other countries (37.1% and 34.1%, respectively). Spain exhibits a relatively balanced distribution across fields, with the largest share of theses in Physical Sciences (34.7%).

Figure 4. Disciplinary composition of dissertations by country (all defence years)

(a) Including Medicine



(b) Excluding Medicine



NOTE. Doctoral dissertation disciplines are harmonised into five broad categories. Note that Physics refers to the macro discipline Physical Sciences, which includes General Physics and Astronomy, Material Sciences, Environmental Science, Earth and Planetary Sciences, and also Chemistry.

Figures 5 and 6 summarize the gender composition of dissertations over time between 2000 and 2020 and across the seven countries. Since gender information is not included in the ETD data, we infer it using two name gender databases, one included in the Global Name Recognition by IBM, the other produced by WIPO for the purpose of classifying inventors (Lax-Martinez et al., 2016, 2021; Miguelez et al., 2019; Di Iasio et al., 2022). We were unable to assign gender to only 7.6% of graduates overall: 2.4% in Austria, 2.3% in Germany, 1.3% in Spain, 3.0% in France, 65.7% in the Netherlands, and 14.8% in the UK.[24] All analyses disaggregated by gender exclude dissertations for which gender could not be attributed.

Figure 5 shows that female graduates are underrepresented in all countries for the entire time window, with the exception of some older cohorts in Italy and some more recent cohorts in Spain. Overall, the share of women has increased over time. In contrast, Italy shows a decreasing trend, and Austria shows a relatively flat trend over the same period. The increasing trend in Germany appears to have stopped in 2011 and reversed when excluding medicine. Figure 6 shows that, across all countries, Mathematics & Computer Science and Engineering are the most male-dominated disciplines. On the contrary, female graduates in Medicine and Life Sciences exceed male graduates in all countries, except for the Netherlands. Spain stands out as the country with the most balanced gender composition across disciplines. Italy also displays relatively high female shares across disciplines, higher than those observed in Austria, France, Germany, the Netherlands, and the UK. Germany shows a particularly low representation of female graduates in Physical Sciences, Mathematics & Computer Science, and Engineering.

[24] The particularly high share of missing gender data in the Netherlands is due to the limited availability of full names of graduates in the Narcis ETD database. For the UK, we relied on WIPO name-gender database and applied a confidence threshold of 80%. The higher share of missing gender for the UK is driven by non-European names, with e.g. gender missing for more than 50% of Chinese names. For Italy, gender was successfully assigned to all thesis authors. In the few cases where the IBM GNR and WIPO name-gender databases did not provide an unequivocal classification, we completed the Italian name assignment using a manually curated list of names and the Ethnea gender predictor (http://abel.lis.illinois.edu/cgi-bin/ethnea/search.py).

Figure 5. Share of female graduates by defence year and country (all disciplines)

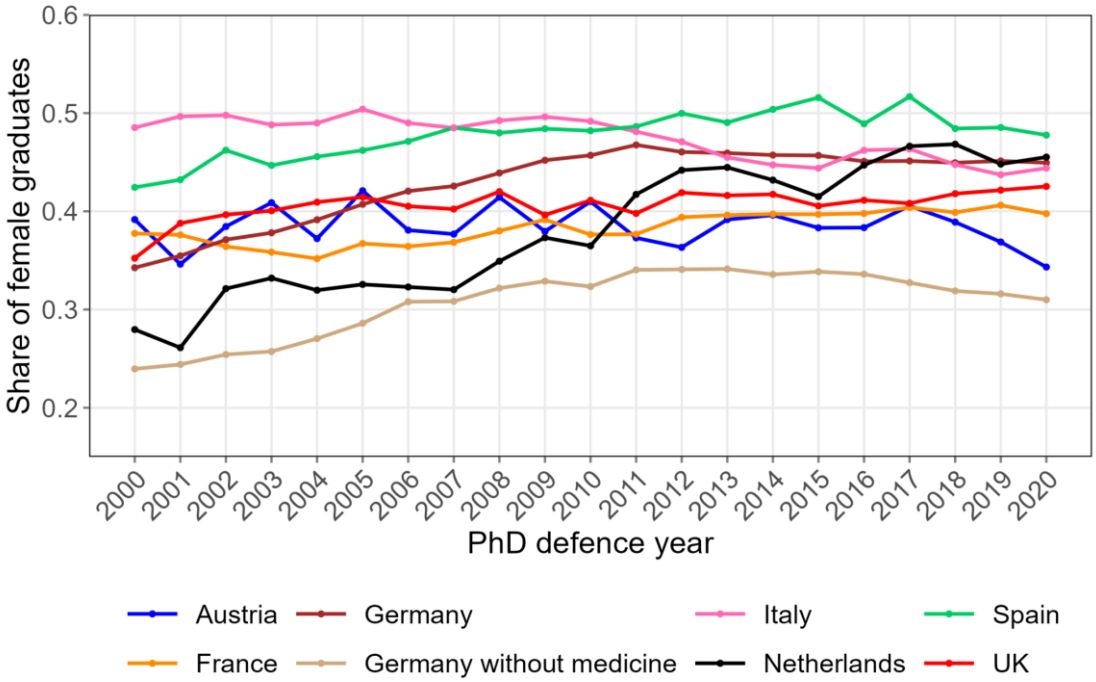
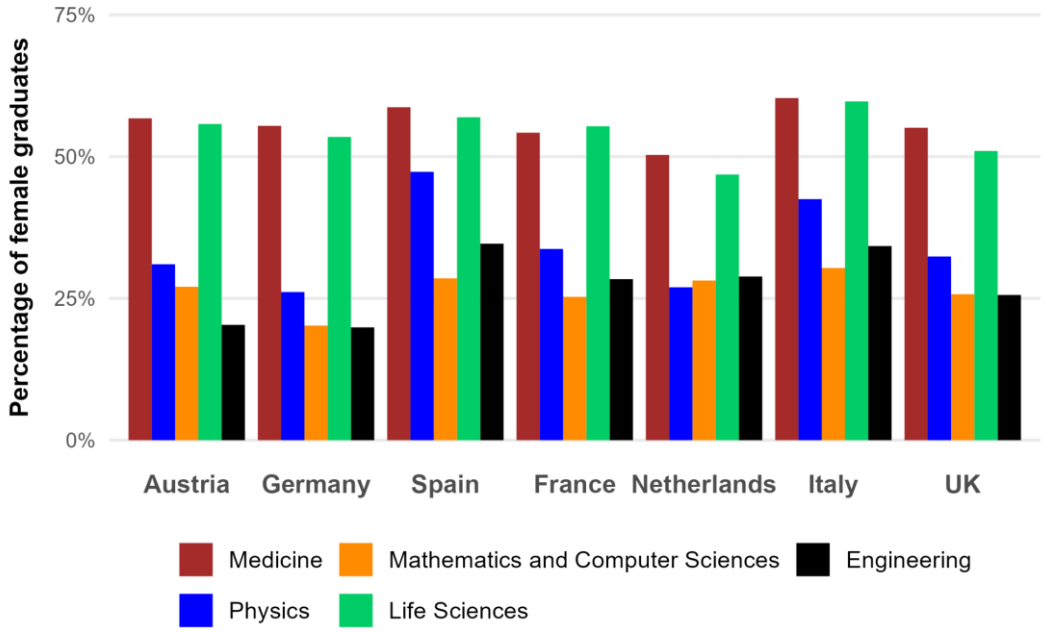


Figure 6. Percentage of female graduates by discipline and country (all defence years)

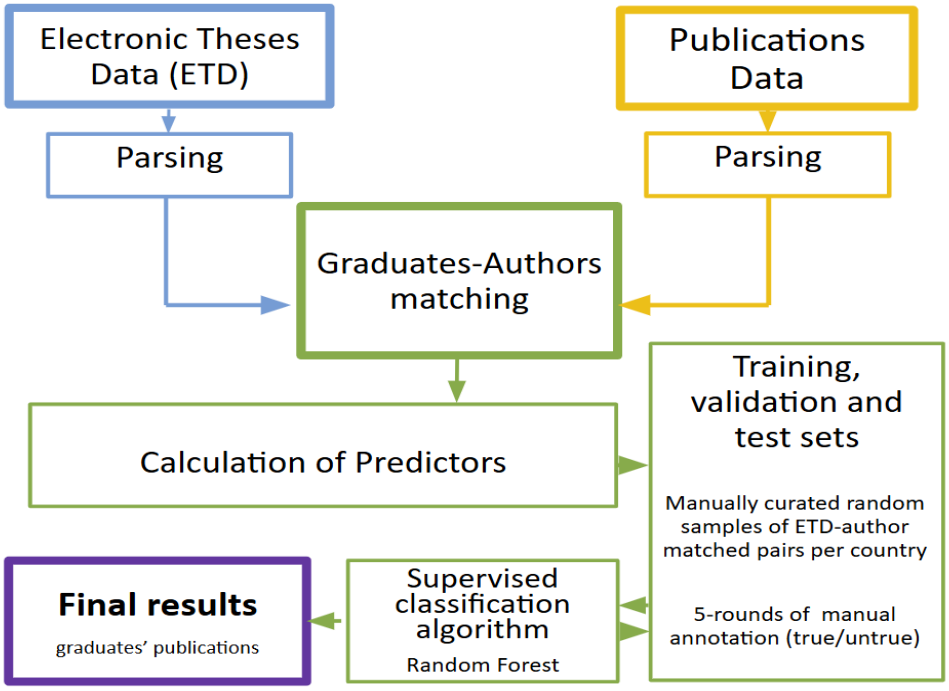


3. Doctoral graduates' publications: methodology

Figure 7 summarizes the methodology we followed for linking the doctoral graduates (authors of dissertations in the ETD repositories) to the authors of the scientific publications included in OpenAlex and Scopus. This is based on five steps, as follows:

- **Parsing**: We first converted the ETD and publication data into a structured format. This included extracting and standardizing relevant information from the ETD 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 ETD repositories and all *m* authors of scientific publications, we applied a pre-selection step based on ETD and publication authors, based on their name similarity (Schnell et al., 2004; Heinisch et al., 2020) and the timing 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 identify true positives and true negatives out of the initial graduate-author matches, based on a maximum of twelve predictors per country. 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 weighting model (the one with the highest F1 score) for undertaking the final classification task on the entire set of graduate-author matches.

Figure 7. Doctoral graduate-publications matching methodology



Step 1: Parsing

ETD data

We compiled ETD records from the national repositories described above. Due to differences in structure, metadata availability, and language conventions, this required extensive preprocessing. First, we normalized the discipline metadata and structured the repository data into a unified relational database schema. Where available, we enriched the data with additional information. For example, for Germany, we extracted committee member names from the title pages of the dissertations by applying text mining algorithms, for the UK, we extracted supervisor names from dissertation acknowledgements and through web scraping, and for the Netherlands, we retrieved as many missing authors' names as possible through web scraping.

Second, in order to prepare the names and surnames for matching, we converted them to lowercase, stripped all non-alphabetic characters (excluding dots and apostrophes), and removed titles and honorifics.

Third, we applied transliteration rules to account for language-specific characters. For instance, characters such as "ä", "ö", "ü", or "ß" were systematically replaced with "ae" "oe", "ue", and "ss", respectively. In addition, alternative forms common in Spanish names (e.g., "mª" to "Maria", "fco" to "Francisco") were mapped accordingly.[25]

[25] For more specificities about Spanish names, see Maraut and Martinez (2014).

Third, we generated several alternative name representations for each individual, including using initials instead of names and common permutations of names and surnames. These combinations reflect common inconsistencies in the way names may be entered across systems, whether in the ETD repositories, scientific publications, or patents.

Publication Data

We extracted publications from both the Scopus and OpenAlex databases. Scientific publications in both databases come either with the full names of authors or just the initials of the names and the surnames. Very importantly, in Scopus, they also come with unique identifiers for authors (Scopus ID), produced and maintained by the database publisher (Elsevier). This provides a highly useful research tool for matching. An author appearing with alternative name representations in different publications is still identifiable as the same person, as long as it is assigned the same Scopus ID. This implies that whenever we match a doctoral graduate to a publication author with a Scopus ID, we can assign to the former all publications of the latter, even those in which the names escape the match.

However, before proceeding to the match, the author data also required substantial parsing. First, we kept all publication types. Second, we excluded publications with authors lacking all information about either names or surnames. Therefore, we keep cases where only the initial of the first name is included. Third, we processed the authors' names or the remaining articles by applying the same standardization rules described above for the ETD records, thus ensuring comparability.

Fourth, we generated different name variants for each unique author Scopus ID to account for potential variations in spelling and structure across publications. These were obtained by combining two alternative transliteration rules with as many variations of name and surname combinations as possible. For transliteration, we applied both a standard conversion of accented and special characters into basic Latin letters and a procedure replicating the transliteration of language-specific characters applied to ETD data.

As for the name variations, in one case we preserved single-character elements (e.g., initials) that appear at the boundaries between names and surnames, as we did for ETD data, particularly in cases where the segmentation between name and surname is ambiguous. In another case, we combined the name's initials with the surnames, reflecting common author name formats used in academic publications.

Step 2: Doctoral graduates-authors matching

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.[26] 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 satisfactory match rates, which vary by country, with Spain achieving nearly full coverage (99.6%) and Italy the lowest (86.9%; France: 98.2%; Germany: 93.1%; the Netherlands: 94.6%; Austria: 93.0%; UK: 93.5%).[27]

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 (see next steps). 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).[28]

The rules we used to deal with the low precision issue are as follows:

- **Temporal filter**: We retained only the matches where publication authors have published at least once within a plausible time window centred on the doctoral graduate's dissertation defence year (namely, no more than three years before and up to five years after the defence year).
- **Outlier filter**: We discarded all *1-to-m* matches where *m* is unusually large (namely, it falls in the top-5% percentile of the frequency distribution of *m*).
- **Name consistency filter**: If a given Scopus ID includes the exact full name of the doctoral graduate (among its associated author's name variants), we discarded other Scopus IDs that only included name variants with initials.

Additionally, in the case of the Netherlands and the UK, where a disproportionate amount of candidate authors was matched to graduates[29], we further applied:

- **Geographic filter**: We dropped all matches in which neither the matched publication authors nor their coauthors had ever published with an affiliation from the same country as the doctoral graduate's *alma mater*.

[26] Italy represents an exception to this pattern. For this country, we made use of the Scopus API. Using the Python package *pybliometrics* (Rose and Kitchen, 2019), we searched for each author by given name and surname.

[27] The lower percentage of Italian PhD students with at least one matched author in Scopus may be related to the different matching approach (see footnote above). To further verify this, we randomly selected 1,500 PhD students and manually matched them with Scopus. We found only seven cases (less than 0.5%) in which a PhD student had a corresponding Scopus author that our automated retrieval through the Scopus API had missed. Most of these mismatches were due to misspellings in the students' names.

[28] 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.

[29] In the Netherlands, this is due to a lack of full information on graduates' names. In the UK, it is due to a disproportionate number of common and Asian names.

Step 3: Creation of the training, validation, and test datasets

After imposing the step-2 filters described above, and in preparation for a supervised classification model (see step 4), we 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-author pairs, more precisely: 4,058 matches for France; 7,159 for Spain; 5,101 for Germany; 5,722 for Austria; 4,776 for the Netherlands; 11,284 for the UK; and 6,610 for Italy.[30] We then proceeded to manually annotate all the sampled matches, based on a three-stage procedure aiming at ensuring consistency and reducing classification errors. This can be summarized as follows:

- **Stage 1:** We assigned each random sample to at least two different independent annotators, who were tasked with attributing one of three possible labels: 'true match', 'false match', or 'not sure'. Annotators were presented with URL links to the ETD and Scopus publication author pages, and were suggested to use all additional available resources on the internet needed to reach a decision, at their discretion (for example: ResearchGate, LinkedIn, Google Scholar, ORCID, institutional and personal websites).
- **Stage 2:** We first classified as true or false each sampled match with no disagreement between the manual annotators, then reviewed personally all uncertain cases and cases with disagreements by means of further inspections. Second, we paid special attention to the *1-to-m* and *n-to-1* matches in order to detect and correct issues related to name ambiguity or inconsistencies in the metadata sources (where *n-to-1* matches consist of multiple graduates matched to a single author). This led to a single, consolidated classification of sampled matches per country, where no unresolved cases were left, after exhausting all available resources. Annotation error rates varied by country: estimated discrepancies attributable to human error ranged from 0.7 to 0.8 percentage points in Spain to approximately 2.8 to 4 percentage points in Germany. The iterative design of the annotation workflow was intended to minimize such errors and improve the robustness of subsequent classification steps.
- **Stage 3:** A number of final quality checks were performed on a random subset of annotated cases to identify any remaining errors due to the annotators' consensual agreement on incorrect classification or systematic biases. One of these checks consisted of further revising the authors with the most common names and performing additional validation searches for the most prolific authors included in the training, validation, and test sets.

This multi-stage annotation process was designed to produce a reliable gold standard dataset for training and testing. The resulting datasets remain unbalanced, with the

[30] 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. For the UK we sampled 1750 authors.

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, 78–22% for the Netherlands, 86-14% for the UK, and 79-21% for Italy. 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 an appropriate machine learning algorithm, i.e., the Random Forest (see Step 4).[31]

Step 4: Model training, validation, and testing

We trained a Random Forest classification algorithm with the manually annotated data produced at step 3. This specific type of machine learning algorithm has been widely adopted in similar classification contexts (e.g., Donner, 2022; Shin et al., 2014) and compares favourably to alternative approaches such as logistic regression 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 (Gareth et al., 2013).

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.[32] As for the former, we chose up to 12 features of the doctoral graduate-author matches, grouped into five categories, as follows (see the Appendix for details of the features used for each country):

- **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).
- **Geographic alignment** (such as the presence, among the matched author's affiliations, of either the graduate's *alma mater* or its city or country).
- **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).
- **Name similarity** (such as a low Levenshtein distance between the text strings containing the standardized name forms of the graduate and the author of the publication).

[31] We also experimented with the Synthetic Minority Over-sampling Technique (SMOTE) to increase the representation of the positive class (Chawla et al., 2002). 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.

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

- **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 a Random Forest model separately for each DOC-TRACK 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.[33]

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 strongly. Part of this performance reflects the fact that these are 'engine' metrics computed on the test sample after applying the filters in step 2, which are likely to remove false matches. By restricting the sample in this way, the Random Forest is tested on a dataset with a higher prevalence of true positives, which can lead to higher precision and recall. However, this filtering step also excludes some graduate–author pairs that could in fact be true positives, reflecting the inherent trade-off between stricter filtering and capturing all valid matches. As follows, the performance by country:

- **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
- **Italy**: 97.2% precision, 94.6% recall
- **UK**: 96.7% precision, 92.3% recall
- **Netherlands:** 99.0% precision, 95.3% recall

Step 5: Final classification and sample construction

We applied the trained models to the full set of candidate matches produced at Step 2. In cases where the algorithm validates a *1-to-m* match, we retained the two graduate-author pairs with the highest predicted probability, based on the assumption of a possible error in Elsevier's disambiguation of Scopus records (see again footnote 28). For *n-to-1* matches, we retained only the highest-probability pair, while also accounting for possible cross-country matches. This conservative strategy minimized false positives while preserving the most plausible links.

[33] 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 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.

The matches we finally retained as true positives are set out in Table 2 below. They include 248,068 German doctoral graduates (58.9% of the original STEMM German population), 140,017 French graduates (84.0%), 80,667 Spanish graduates (70.9%), 19,949 Austrian graduates (67.6%), 35,263 Dutch graduates (65.4%), 85,928 Italian graduates (80.7%), and 158,499 UK graduates (74.1%). Differences in match coverage reflect national differences in doctoral graduates' trajectories, such as the higher share of industry employment or medical doctors among German graduates. Overall, 768,391 DOC-TRACK graduates are matched with publication authors (69.5%).

Table 2 - Random Forest predicted true positives: graduates with at least one publication

	N. of matched STEMM graduates	% of total STEMM graduates
Germany	248,068	58.9
France	140,017	84.0
Spain	80,667	70.9
Austria	19,949	67.6
The Netherlands	35,263	65.4
Italy	85,928	80.7
UK	158,499	74.1

For all these true matches, we consider as authored by the doctoral graduates not only the publications made in the time interval used for the graduate-author matching of Step 2, but also all the articles assigned to the same author by Scopus, based on the Scopus author ID, irrespective of the distance from the defence year. As for the graduates for whom no matches were found at Step 2 or no matches were validated as true in this final Step 5, we consider them as not having ever published a scientific article, during or after their dissertation.

To prepare the DOC-TRACK datasets, we matched this sample with the OpenAlex publication database based on the publications' DOIs, and included all OpenAlex entries with matched DOIs.[34] Moreover, we discarded the graduates classified as outliers in step 2 of the doctoral graduates-authors matching, since we cannot determine whether they are matchable in Scopus. Descriptive analyses of graduates' productivity are reported in Section 6 below. The results may differ slightly from those of analogous analyses based

[34] Note that, for some publications, the publication year differs slightly between Scopus and OpenAlex, and some OpenAlex publication years exceed 2021 or precede 1996, the coverage period of our Scopus data.

on the DOC-TRACK datasets, as DOIs could be retrieved for 85.2% of the Scopus publications matched to our doctoral graduates.

4. Doctoral graduates' publication-to-patent distance: methodology

Research by doctoral graduates can contribute to technological advancement in many ways. In some cases, it may include the production of a prototype or proof of concept for either a new product or process, which may result in a patent directly field by the graduate (see the following Section 5). More often, and very much like academic science in general, it contributes to a cumulative process of discovery and understanding of either natural phenomena or technologies, which may ultimately result in a patent, not necessarily filed by the graduate. In this case, we may trace the graduates' contribution by examining non-patent literature citations of any patent filed after the graduate's publications, in particular those we deemed most closely related to their doctoral research. This type of contribution is as important as the direct production of patents and possibly more frequent. In order to appreciate it, however, we need to go beyond the dissertations and publications directly cited by patents. For example, Ahmadpoor and Jones (2017) find that 80% of all articles indexed in the Web of Science and published in 1945-2013 are linked forward to future patents issued by USPTO from 1976-2015, not by a direct citation, but *via* a chain of publication-to-publication citations ultimately leading to a patent citing a publication. In the authors' own words, there exists a "majority connectivity between the corpus of patented inventions and the corpus of scientific papers [but] these connections are typically indirect, and both scientific fields and patenting technology classes vary enormously in their connectivity and proximity to the other domain" (Amahdpoor and Jones 2017, p.3).

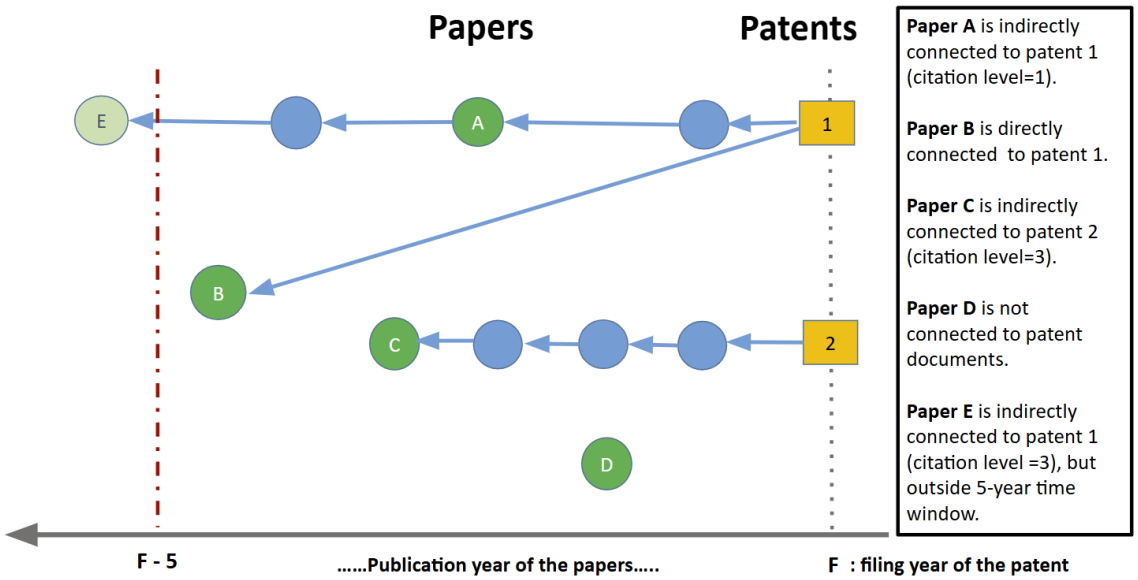
Based on Ahmadpoor's and Jones' (2017) metrics, we first define as "technological frontier" the set of all patents directly citing a scientific publication (we consider both the front-page citations and the citations included in the invention description, in both EPO and USPTO patents). We then calculate the distance from this frontier for all the scientific publications authored by the doctoral graduates during the doctoral period, namely those dated from *t-3* to *t+1*, where *t* is the thesis defence year. Any publication's distance from the technological frontier equals zero when the publication is directly cited by a patent, and it is larger than zero when it is not directly cited by a patent, but it is cited by other publications that generate a citation chain ultimately leading to a citation by a patent. In this latter case, the distance is equal to the number of steps in the shortest citation chain reaching a patent. Finally, a publication that receives no citation or standing on no citation chain reaching a patent is considered disconnected from the technological frontier.

One issue to consider in the calculation of the distance from the technological frontier is that older publications have more time to get cited than recent ones. Therefore, the older publications are more likely to be connected and closer to the technological frontier. In order to allow for a comparison between the publication of doctoral graduates from different graduation years, we therefore considered only the citation chains leading to patents filed no later than 5 years after the appearance of the graduates' publications.

Figure 8 illustrates our methodology. Yellow nodes represent USPTO and EPO patent documents, green nodes represent publications authored by doctoral graduates during the doctoral period, and blue nodes represent publications not authored by doctoral graduates but present in the citation chain. Edges (ties) in the citation network represent citations either from a publication to other publications or from a patent to a publication. The 5-year window imposed to collect citations from patents is indicated by the vertical red dashed line. For each doctoral graduate's publication, we calculate its distance from the technological frontier as the minimum number of steps in the citation network needed to connect the graduate's publication to any patent. For instance, publication A stands at distance one from the frontier because one is the number of citing publications on the citation path to the closest patent (patent 1 in the figure). Publication B stands at distance zero from the frontier because it is directly cited by a patent (once again, patent 1). Following the same logic, publication C stands at a distance of three (three publications leading to patent 2). Finally, publications D and E are not connected to the technological frontier for two different reasons. Publication D is not cited at all, so there is no citation chain that leads to a citation by a patent. Publication E is connected to a patent, but beyond the 5-year limit imposed between the publication's year and the patent filing year, and therefore escapes our measurement.

For all these calculations, we rely on two data sources. The first one is citation data from OpenAlex, which we use to reconstruct publication-to-publication citation chains for each graduate's publication identified in the "ETD-publication matching" section (the green nodes). The second source is the "Reliance on Science" dataset, which reports information on patent-to-publication citations (Marx and Fuegi, 2020a, 2020b).

Figure 8. Calculation of the DOC-TRACK distance-to-patent metric



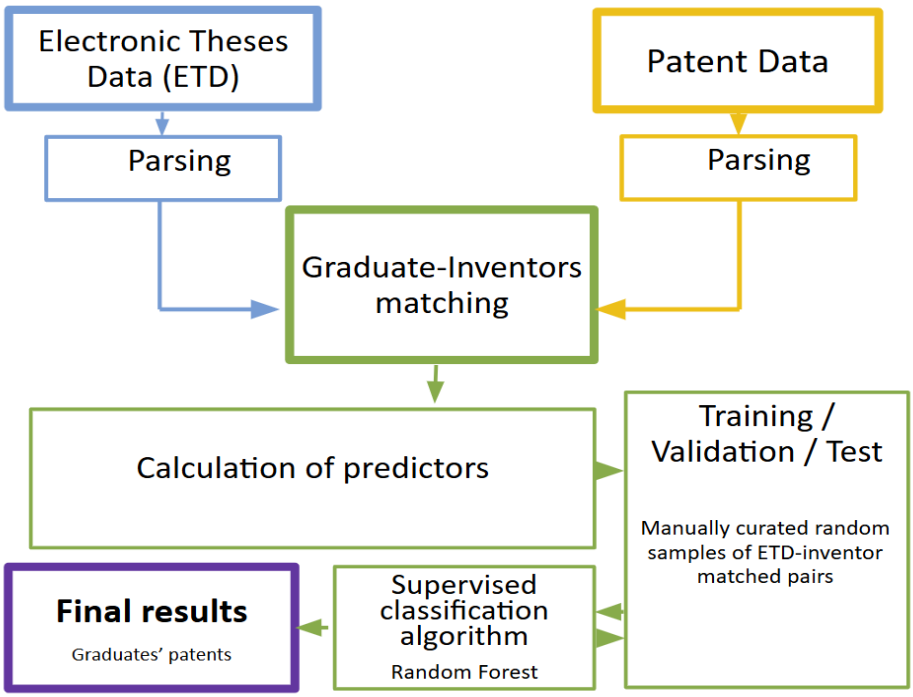
NOTE: Own elaboration metric, based on the methodology of Ahmadpoor and Jones (2017).

5. Doctoral graduates' patents: methodology

Figure 9 summarizes the methodology we followed for linking the doctoral graduates (authors of the dissertations in the ETD repositories) to the inventors named on EPO patent applications. This is based on five steps, similar to those followed for the attribution of publications (see section 3 above), but with some important modifications due to differences in contents and structure between PATSTAT (our patent data source) and the publication data. Most noticeably, inventor data in PATSTAT are not disambiguated so that inventors are not uniquely identified; they are also less curated, with more frequent cases of misspelling or inversions of names and surnames, which affects our matching strategy. It also occurs that the name text strings include alien components such as portions of the inventor's employers and/or addresses, or her academic titles, which need to be removed at the parsing stage.[35]

[35] More in detail, PATSTAT assigns unique identifiers ("person_id") to identical combinations of inventors' full names and addresses. If the same inventor files multiple patents from different addresses or under different name spellings, these are not identified as coming from the same person (multiple person_id are produced). To mitigate this problem, we rely on the PSN_ID identifier in the KUL EEE-PPAT database (also based on PATSTAT; see Magerman et al 2006), which provides a first rough grouping of name variants referring to the same individual. Although useful for reducing the number of graduate-inventor matches to be produced at Step 2, the recall rate of the underlying disambiguation algorithm is too low for PSN_ID to be considered a reliable unique identifier.

Figure 9. Doctoral graduate-patent matching methodology



Step 1 - Parsing

As a first step, we converted all name strings in PATSTAT to lowercase, standardized them to Latin ASCII format, and removed all academic and honorific titles (such as "Dr.", "Prof.", or "Dipl."). Second, we identified and removed from the name strings the 3,000 most common spurious tokens[36]. As for the less common spurious tokens, also referred to as the inventors' employers, we identified and removed them based on a list of text patterns to which they are most commonly associated (for example, "c/o" in the name string frequently precedes an organization's name or address). To complete this process, we finally compiled and used iterative lists of frequently occurring academic titles, institutional and company suffixes, legal business forms, address terms, country names, and other generic non-name words.[37]

[36] This includes the 1000 most common names of universities, companies, and other inventors' employers.

[37] We took care to avoid removing elements that could also be legitimate parts of personal names. For example, academic titles usually appear at the beginning or end of a name string, in which case we removed it directly. When they appeared elsewhere, we treated them as delimiters indicating the end of the actual name, and deleted both them and all the following characters in the string. We applied a similar rule to legal business suffixes (such as "AG" or "Ltd"), address-related tokens (such as "street", "room", or country names) and concept tokens such "networks", "solar", "planning", or "IP". In addition to the 3,000 most common inventors' employers, we also removed common company name components such as "apple" or "samsung". Last, given that personal names typically consist of a name and surname (sometimes with a middle name or two or more surnames), and based on manual inspection, we removed

For symmetry reasons, we applied this parsing procedure also to doctoral graduates' names in our ETD repositories. However, as the graduates' names are much less error-prone, this did not affect them in any noticeable way.

Step 2: Doctoral graduates-inventors name matching

Two problems with inventor data that our parsing strategy could not address are: *i)* the frequent misspellings of names, inconsistent use of name initials, and token inversions (e.g., between name and surname); and *ii)* the absence of unique identifiers. This forced us to adopt, for the graduate-inventor matching, a different strategy than that used for the matching with publication authors. In particular, we opted for name-matching with a fuzzy string-matching algorithm based on 3-grams and a Jaccard similarity metric. We then retained for the subsequent steps all name pairs with a score of at least 0.75.[38]

To reduce computational load, especially due to the large Cartesian product of all doctoral graduates and inventors, we also restricted the string distance calculation to name pairs sharing the same initial letter of the last name. This limits the comparison space without significantly affecting match quality.

Step 3: Creation of the training, validation, and test datasets

We created a single manually annotated dataset by randomly selecting a number of doctoral graduates per country and their candidate matched patents (753 records for France, 840 for Spain, 1,476 for Germany, 439 for Austria, 905 for the Netherlands, 465 for Italy, and 1,918 for the UK), where the size of each country subset depends both on the relative size of each country in terms of doctoral graduates and inventors, and the recall of the name matching. The lack of a unique inventor identifier for EPO inventors in PATSTAT required us to focus on doctoral graduate-patent pairs, rather than doctoral graduate-inventor pairs, and this increased the effort required for manual annotation. This is the main reason why we decided to have a lower number of records to annotate, compared to the publications matching described in Section 3, and to pool the records of all DOC-TRACK countries into a single dataset to ensure a sufficiently large dataset as input for the classification algorithm.

The pooled dataset was reviewed by three professional annotators, who iteratively labeled all pairs as either true or false matches.[39] Following the same approach as in the

all words after the fourth word in any name string. This truncation effectively eliminates residual non-name components in most remaining cases.

[38] We chose a 3-grams algorithm because it offers a better balance between precision and tolerance than a 2-grams one. With 2-grams, we observed frequent partial overlaps between unrelated names, which increases the number of false positives, while 3-grams provide instead more reliable differentiation between similar and dissimilar names.

[39] An excessively large temporal distance between the doctoral defence and the patent application typically does not represent true positive cases. We therefore excluded patents for which the application had been filed five or more years prior to the defence date. In addition, we limited the manually annotated dataset to the up to two patents closest to the graduation date for each person_id, as these provided the most reliable basis for classification.

publication-matching process, we carefully revised the *1-to-m* and *n-to-1* relationships in the labeled dataset. Finally, since our focus is on identifying doctoral graduates who become inventors, we constructed a final dataset at the doctoral graduate–inventor pair level, inferring the true and false pairs from the manually annotated graduate–patent pair dataset. The graduate–inventor pair dataset includes 5,560 records.

The resulting labelled dataset remained unbalanced, with the proportion of negative (0) and positive (1) labels as follows: 72–28%. As we did in the publications matching, to mitigate imbalance issues, we applied stratified cross-validation to preserve class proportions during training and validation and used the Random Forest machine learning algorithm.[40]

Step 4: Model training and testing

The Random Forest algorithm was applied to classify the merged candidate pairs as true or false matches. As indicated in Section 3, training the algorithm requires selecting a number of features to which the algorithm assigns weights to maximise its performance (based on F1). In this case, we chose 20 features of the doctoral graduate-inventor matches, grouped into seven categories, as follows (see Appendix C for details of the features used):

- **Social proximity** (such as the presence, estimated with a low Jaccard distance between names, of the graduate's supervisor or of a member of the doctoral committee among the co-inventors of the patent applications).
- **Geographic alignment** (such as the presence, among the addresses in the inventor's patent applications, of the country or city of the graduate's *alma mater*).
- **Topical similarity** (such as a significant overlap of the text strings containing the titles of the graduate's dissertation and the titles of the inventor's patent applications, or the overlap between the theses' disciplines and the patents' technological sectors).
- **Name similarity** (such as a low Jaccard distance between the text strings containing the standardized name forms of the graduate and the inventor).
- **Patenting activity** (such as the number of patents filed by the inventor in relation to the time available from the graduate's defence year, or the presence of patenting activity from the inventor before the graduate's defence year, or the time elapsed between the graduate's defence year and the inventor's first patent application).
- **Patent academic features** (such as the presence of an academic title in the inventor or co-inventor names reported in any of their patent applications, or the

[40] Also, in this case we experimented with the Synthetic Minority Over-sampling Technique (SMOTE), but ultimately decided against implementing it as it offered no significant advantages.

presence of a university or another academic institution, broadly defined, among the applicants of any of the inventor's patents).

- **Thesis country** (such as the country where the doctoral graduate defended her thesis).

The algorithm for training and validation was implemented on 90% of the manually annotated dataset, extracted randomly, but with a stratification to preserve the class distribution between true and false matches to deal with the imbalance in the data.[41]

The test was performed against the remaining 10% of the annotated sample and yielded the following results: 87.6% precision and 82.5% recall.

Step 5: Final classification and sample construction

We applied the trained Random Forest model to the full set of candidate graduate-inventor pairs obtained at Step 2. For doctoral graduates matched to multiple inventors (*1-to-m*), we kept up to three most likely inventors, based on the Random Forest predicted probabilities.[42] In *n-to-1* matches, we retained only the graduate-inventor pairs with the highest probability of a match. In addition, we checked the *n-to-1* matches at the graduate-patent level, and again retained the pairs with the highest probability of match[43]. Finally, we correct for extreme outliers: graduates with more than 55 patents (the top 0.1% of the distribution of patents per graduate) are considered as false positives, while graduates flagged as outliers in the previous publication matching due to unusually high productivity or common names are excluded from the sample.[44] In line with the publication matching, this strategy is used to maximize the reliability of the dataset while preserving the most plausible links.

After these adjustments, the final dataset of true positives includes 86,499 graduates (8.16% of the full ETD population) and 142,417 graduate-inventor pairs. Match coverage varies by country, as set out in Table 3, showing rates of 8.0% for German graduates, 15.5% for French graduates, 7.2% for Spanish graduates, 11.2% for Austrian graduates, 4.9% for Dutch graduates, 9.0% for Italian graduates, and 3.2% for UK graduates.

[41] As for the publication matching, we conducted a stratified 5-fold cross-validation procedure combined with an exhaustive grid search to tune 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, as in the publication matching. However, mtry was tested across a higher number of values, i.e., the values 1, 2, 3, 4, 5, 6 and 20, the maximum value of features used. This is because the features used in the patent matching exceed those used in the publication matching.

[42] We kept a maximum of 3 patent inventors per graduate, rather than 2, as was done for publication authors, since there is barely no disambiguation of inventors in PATSTAT person_id. Furthermore, the Random Forest linked, on average, 2.7 inventors per graduate.

[43] *n-to-1* matches at the graduate-patent level occurs due to the poor disambiguation of inventors in PATSTAT person_id, and the presence of co-inventors with very similar names according to our fuzzy string-matching algorithm used in step 2.

[44] We decided to remove publication outliers to maintain coherence throughout the entire procedure.

Mirroring the approach adopted for the publication matching, we consider all patents assigned by PATSTAT to the IDs of the inventors matched to doctoral graduates, as filed by the graduates. Graduates without validated matches are considered to have no patents.

Table 3 - Random Forest predicted true positives: graduates with at least one patent

	N. of matched STEMM graduates	% of total STEMM graduates
Germany	32,283	8.0
France	24,570	15.5
Spain	7,783	7.2
Austria	3,160	11.2
The Netherlands	2,549	4.9
Italy	9,551	9.0
UK	6,603	3.2

Note: graduates flagged as outliers in the publication matching of Section 3 are excluded from the sample, consistently with the entire procedure.

6. DOC-TRACK graduates' productivity results

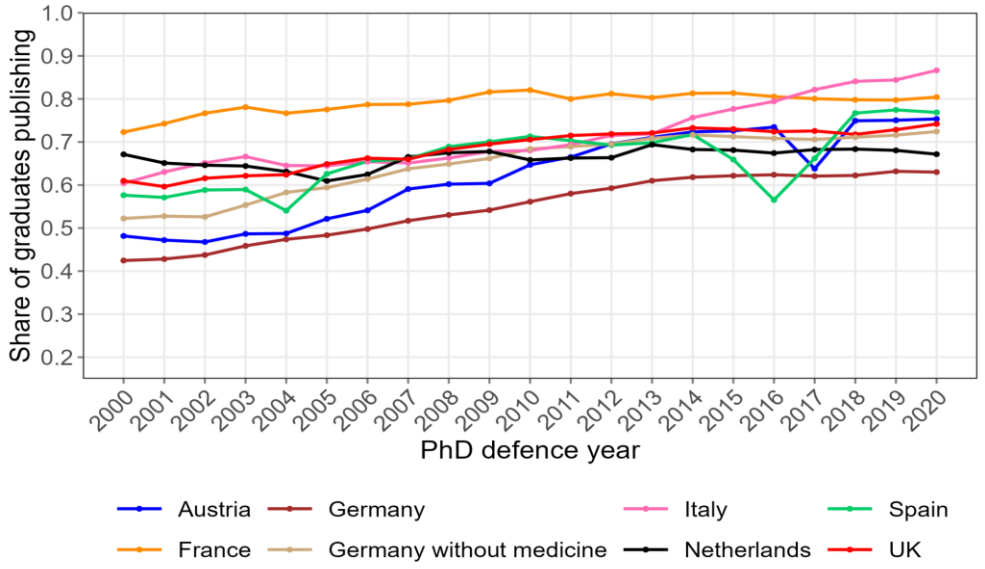
6.1 Doctoral graduates' publication propensity and productivity

Figure 10 plots the share of STEMM doctoral graduates with at least one publication, by defence year and country, during their doctoral studies (*t-3, t+1*) in Figure 10a, and afterwards (from *t+2* until 2021, which is the last date for which we have publication information) in Figure 10b.[45] We observe that levels in Figure 10b are lower than in Figure 10a. This is coherent with two well-established findings in the literature, namely that a substantial share of doctoral graduates does not undertake an academic career (which weighs negatively on the probability to publish once graduated); and that the probability to engage in such a career is strictly correlated to the probability to publish before graduating (that is, very likely to publish one's own doctoral findings) (Black and Stephan, 2010; Brischoux and Angelier, 2015; Horta and Santos, 2016).

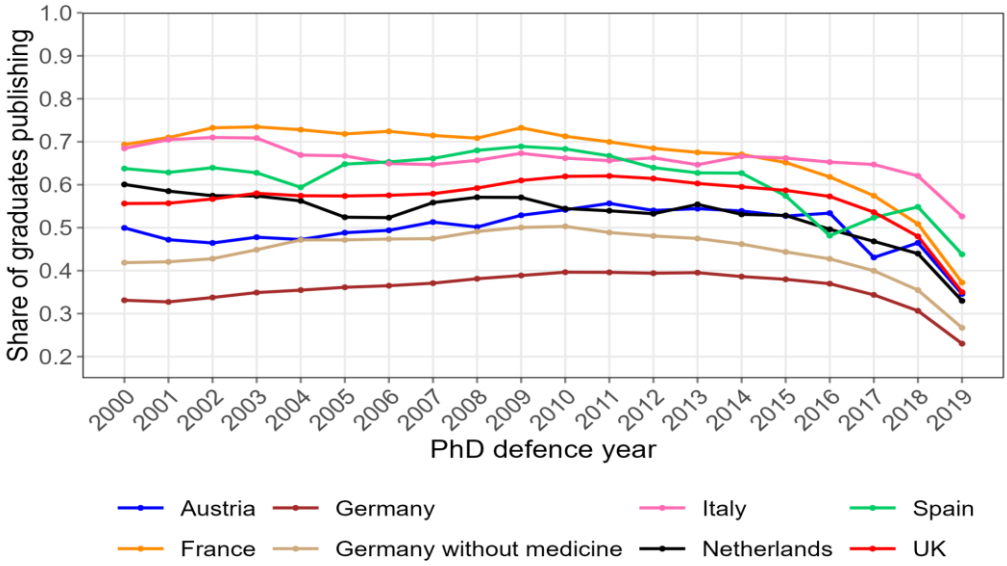
[45] Publication data are available through 2021, and post-phd output is measured from two years after the defence onward. Consequently, the 2019 cohort has one observable year, while later cohorts have no post-PhD observations.

Figure 10. Share of graduates with at least one publication by defence year and country (all disciplines)

(a) During Doctoral studies



(b) After doctoral studies



NOTE: The period during doctoral studies spans from three years before to one year after the PhD defence, while the period after doctoral studies starts two years after the defence. For the latter, we truncate the sample at the 2019 cohort, as publication data are available through 2021; consequently, the 2019 cohort is the last cohort with at least one observable post-PhD year.

Looking at the individual countries in Figure 10a, we see that four out of seven (namely: Austria, the Netherlands, the UK, and Spain) exhibit broadly similar levels and trends: shares generally range between 50% and 80%, with a moderate upward trend until the mid-2010s. Once excluding medical doctors, Germany follows the same pattern.

Italy, by contrast, shows a distinct pattern. Italian graduates publish at relatively high levels throughout the period, with a steady and marked increase in the propensity to publish during the doctoral studies from the mid-2000s onward, surpassing 80% in the most recent cohorts. This upward trend is consistent with Italy's research environment, where research evaluation has become increasingly central to academic careers (Baccini et al., 2019). Since the early 2000s, the implementation of national research assessment exercises (VQR) and, from 2010, the introduction of the National Scientific Habilitation (ASN) have tied access to academic positions and a significant share of university funding to bibliometric performance. These mechanisms, which require international journal publications and citation indicators as core components of evaluation, have fostered strong incentives for doctoral programmes and early-career researchers to prioritise publishing. As a result, Italian doctoral graduates face more pronounced pressures to build competitive publication portfolios early in their careers, which likely contributes to the comparatively high and rising propensity to publish observable in Figure 10a.

France is another exception, with a share of graduates with publications always around 80% and a flat trend. This could be due to the higher quality of the French ETD repository, which enables a higher recall rate, or to the specificity of French names, which are easier to match with Scopus due to a low number of homonyms. However, so far, we cannot exclude substantive reasons, such as the existence of different publication requirements for doctoral candidates across countries or compositional effects (different disciplinary mixes across countries, with varying publication propensities across disciplines). We also notice, upon comparing Figure 10b to Figure 10a, that the gap separating France from Spain appears to be reducing, which suggests that the French exception is primarily due to a higher propensity to publish during doctoral studies (especially until around 2010), rather than different career prospects.

Both Spain and Austria exhibit short-term deviations from the common trend. For Spain, we observe a sharp drop in 2015-2016 and a successive rebound in 2017-2019. Most likely, this is the result of the Spanish Royal Decree 99/2011, upon which we already commented in section 2.2 (Corsini et al., 2025). This shortened to three years the *maximum* time for completing doctoral dissertations (with exceptions under certain conditions), starting 2014/2015, and forced many senior doctoral students who had not yet defended their dissertations to rush to defend their thesis by 2016-2017, possibly by delaying their publication efforts. For Austria, a one-year deviation from the general upward trend can be observed in 2017, when the Austrian University Law introduced a *minimum* duration of three years and additional regulations for doctoral studies. This may have induced students enrolled in shorter-term doctoral programs and without

pronounced academic ambitions to graduate before the minimum duration became binding.

In Figures 11 and 12, we explore composition effects. Figure 11 reports the share of graduates with at least one publication, over time and by discipline, considering all countries together. Graduates in the Physical Sciences have the highest propensity to publish, and those in Medicine have the lowest, although when excluding Germany, the trend and levels of Medicine become very similar to those of the other disciplines. Engineering, which starts at lower levels, follows a generally increasing trend. Figure 12 reports the share of graduates with at least one publication, by discipline and country, considering all defence years. This allows us to separate disciplinary composition effects from within-discipline productivity differences. France consistently records the highest shares of publishing graduates across all five disciplines, indicating a systemic country effect beyond discipline mix. Italy also performs strongly across the disciplinary spectrum, with publication propensities close to the French levels in several fields, most notably in Medicine, Mathematics & Computer Science, Life Sciences, and Engineering. The UK generally ranks third, but shows a higher propensity than Italy in Engineering. Germany, the Netherlands, and Austria lag behind in most disciplines, notably in Engineering and Mathematics & Computer Science, suggesting that graduates in these disciplines in these countries are less likely to pursue an academic career, and possibly more likely to work in industry. Spain shows more balanced levels. Persistent within-discipline gaps across countries indicate that productivity differences are driven not only by the different distribution of graduates across high-publishing disciplines, but also by structural country-specific factors.

Figure 11. Share of graduates with at least one publication by defence year and discipline (all countries)

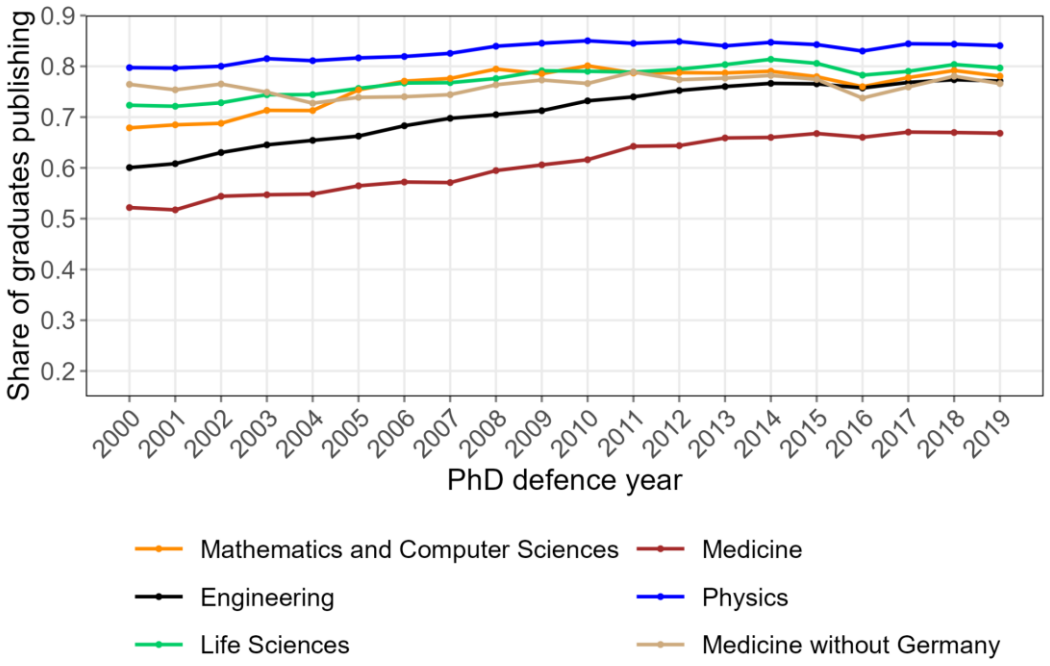
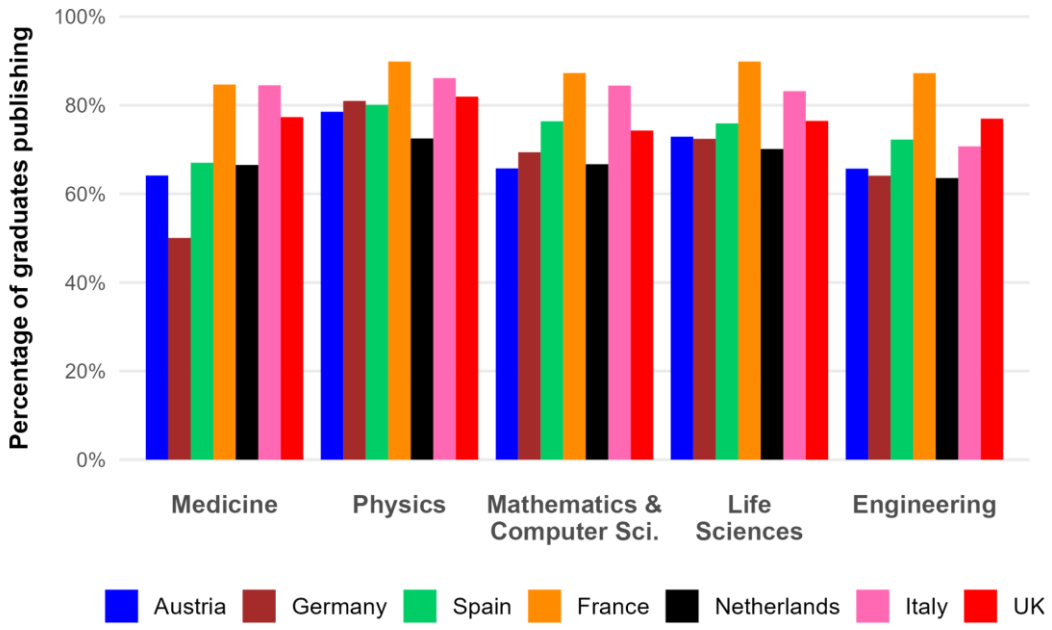


Figure 12. Percentage of graduates with at least one publication by discipline and country (all defence years)



In Figures 13 and 14, we examine the average number of publications produced by doctoral graduates (both before and after their studies), also by discipline.[46]

A first observation from Figure 13a is that Italy clearly stands out as the most productive country during the doctoral period, with Italian graduates surpassing all others from the mid-2010s onward and reaching the highest average outputs in the most recent cohorts. The Netherlands and Spain follow, displaying consistently strong productivity levels, while France shows only moderate performance during the PhD despite its high publication propensity (see Figure 10).

These differences persist after graduation (Figure 13b). Post-PhD, Italy again records the highest productivity across cohorts, followed by the Netherlands and Spain, and the UK in the most recent cohorts; whereas France no longer emerges as a top performer. Germany continues to lag behind, reinforcing the view that a substantial share of its doctoral graduates transition into non-academic careers with limited incentives to publish.[47]

Figure 14 examines productivity differences across disciplines, allowing us to assess whether cross-country gaps are driven by disciplinary composition. During the doctoral training period (Figure 14a), Italian graduates stand out as the most productive in almost every field, especially in Physical Sciences, where their average output is by far the highest. The Netherlands and Spain follow with strong performance across most disciplines, while Germany and the UK stand out particularly in Physical Sciences. This confirms that Italy's leading position during the PhD is not simply a compositional effect but reflects consistently higher output within fields. As discussed above, the strong pressures to publish created by the introduction of bibliometric criteria for academic careers in Italy are likely to explain, at least in part, this distinctive pattern.

The picture remains similar after the defence (Figure 14b). Italy continues to dominate scientific productivity across disciplines, with exceptionally high averages in Physical Sciences and above-average performance in Life Sciences, Mathematics & Computer Science, and Engineering.[48] Spain also performs strongly after graduation but remains

[46] For the main analyses, we use full counting, where each graduate receives full credit for a publication. We also replicated the analyses using fractional counting, assigning each graduate a share equal to the inverse of the number of co-authors. Qualitatively, results remain consistent.

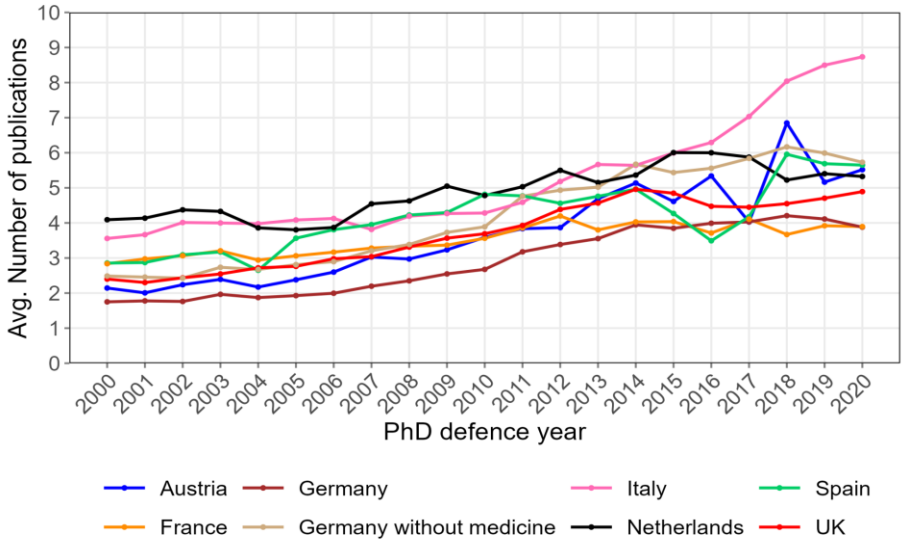
[47] This assumption is consistent with previous research showing that Germany has a long tradition of employing scientists in industry (Murmann, 2025). In Germany, a large proportion of doctoral graduates leave the academic sector immediately after completing their doctorates and often take up well-paid positions in the private sector (Koenig, 2025; Konsortium BuWiK, 2025).

[48] In Appendix D, we provide a more detailed examination of cross-country differences in scientific productivity. This analysis confirms that these differences are systemic and not driven solely by a small group of exceptionally productive individuals. Countries such as Italy, Spain and

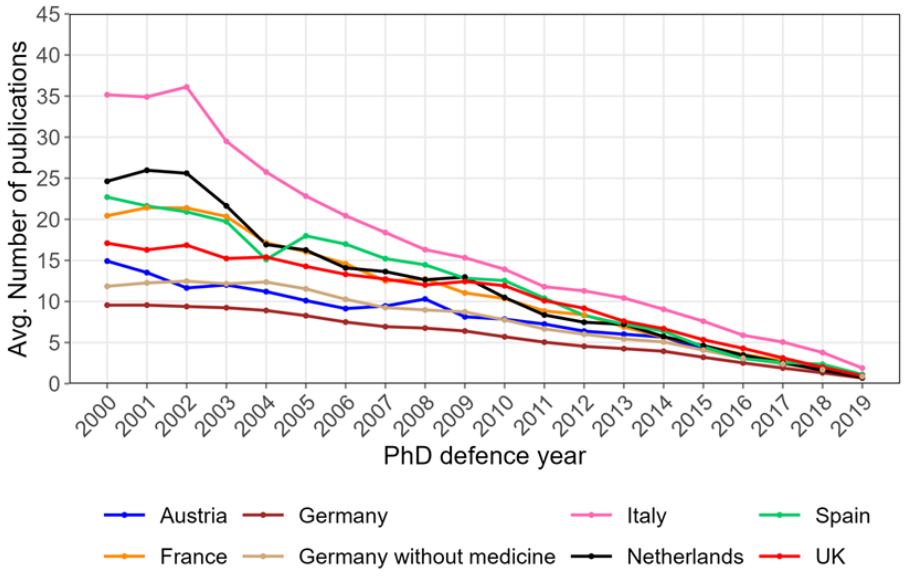
below Italy in all fields. France emerges as a good performer in Medicine and Physical Sciences after the PhD; while the UK in Physical Sciences. Germany and Austria, by contrast, show the lowest productivity within nearly every field after the doctorate.

Figure 13. Graduates' average number of publications by defence year and country

(a) During doctoral studies



(b) After doctoral studies



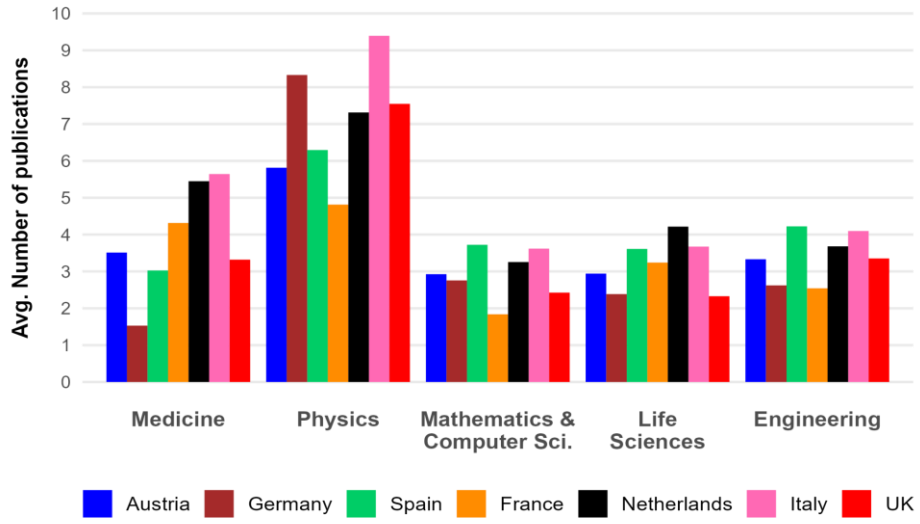
NOTE: The period during doctoral studies spans from three years before to one year after the PhD defence, while the period after doctoral studies starts two years after the defence. For the latter, we

the Netherlands do not merely produce more top performers; they also exhibit stronger middle distributions, with a substantially larger share of graduates sustaining moderate to high levels of publication activity. By contrast, Germany and France display more polarized patterns, with a large proportion of graduates publishing only one or very few papers and relatively few progressing into the 10–50 publication range. Countries that support strong publication activity during the PhD continue to do so in the early postdoctoral years.

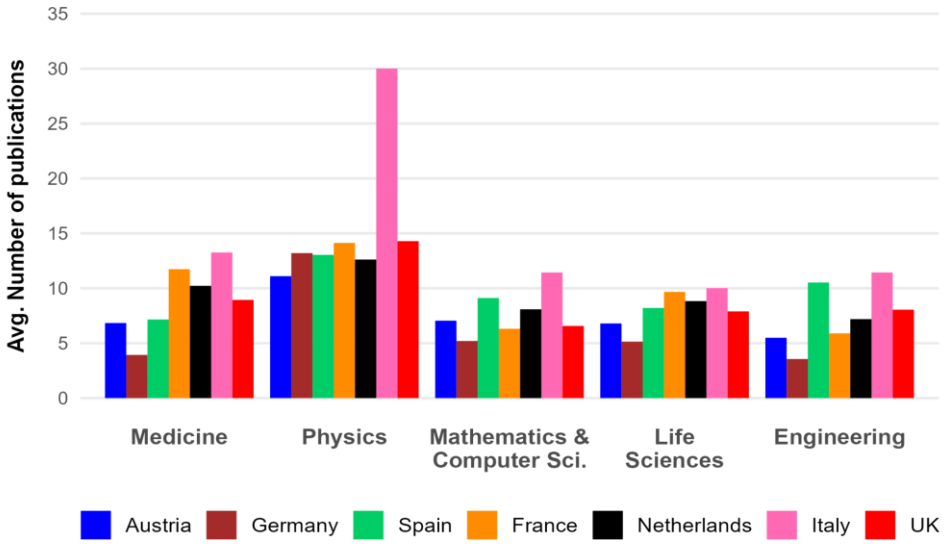
truncate the sample at the 2019 cohort, as publication data are available through 2021; consequently, the 2019 cohort is the last cohort with at least one observable post-PhD year.

Figure 14. Graduates' average number of publications by discipline and country

(a) During doctoral studies



(b) After doctoral studies



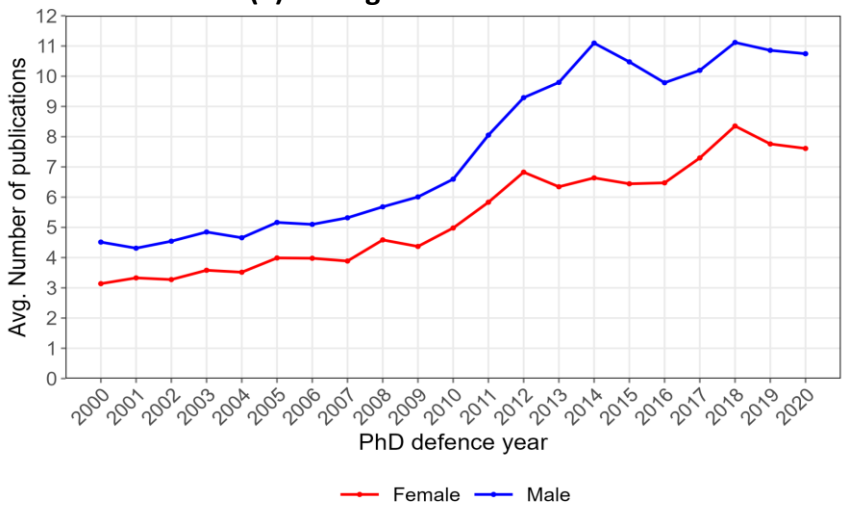
NOTE: The period during doctoral studies spans from three years before to one year after the PhD defence, while the period after doctoral studies starts two years after the defence. For the latter, we truncate the sample at the 2019 cohort, as publication data are available through 2021; consequently, the 2019 cohort is the last cohort with at least one observable post-PhD year.

In Figures 15 to 19, we examine gender effects by comparing the average number of publications of female and male graduates, both during and after their doctoral studies, by defence year and discipline. Recognizing the strong influence of disciplinary conventions, the analysis is broken down by discipline. We find that a persistent gender

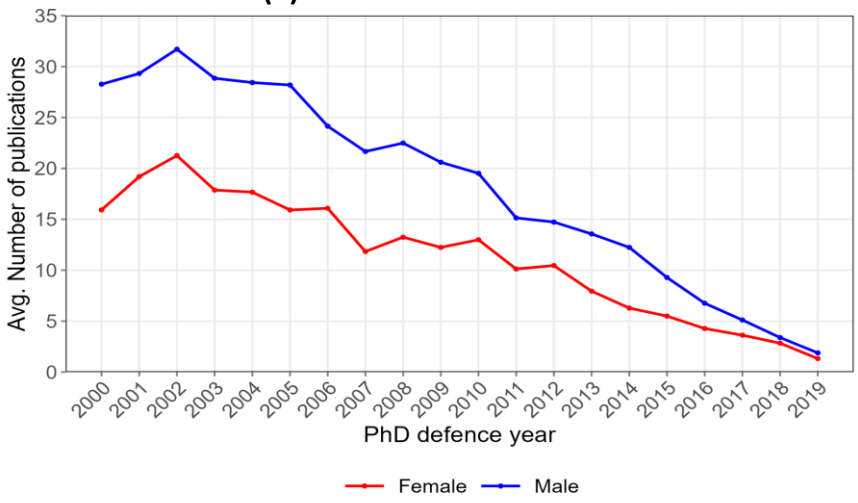
gap in productivity is evident across all disciplines and over time, favoring male graduates. However, this gap appears to be narrowing in recent years for the period after the doctoral studies. Notably, Life Sciences consistently exhibits the smallest productivity gender gap among the five disciplines during the doctoral period; however, this gap increases after graduation.

Figure 15. Graduates' average number of publications by defence year and gender: Physical Sciences

(a) During doctoral studies



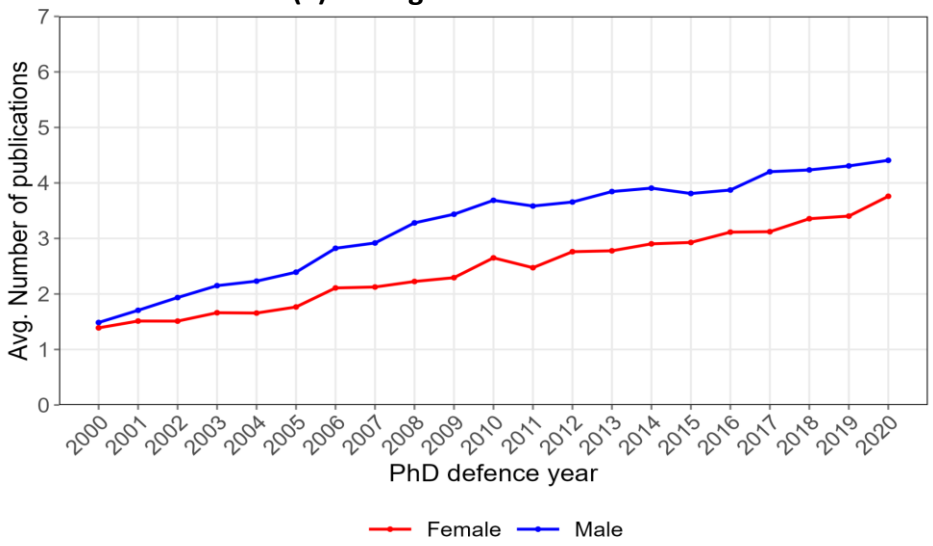
(b) After doctoral studies



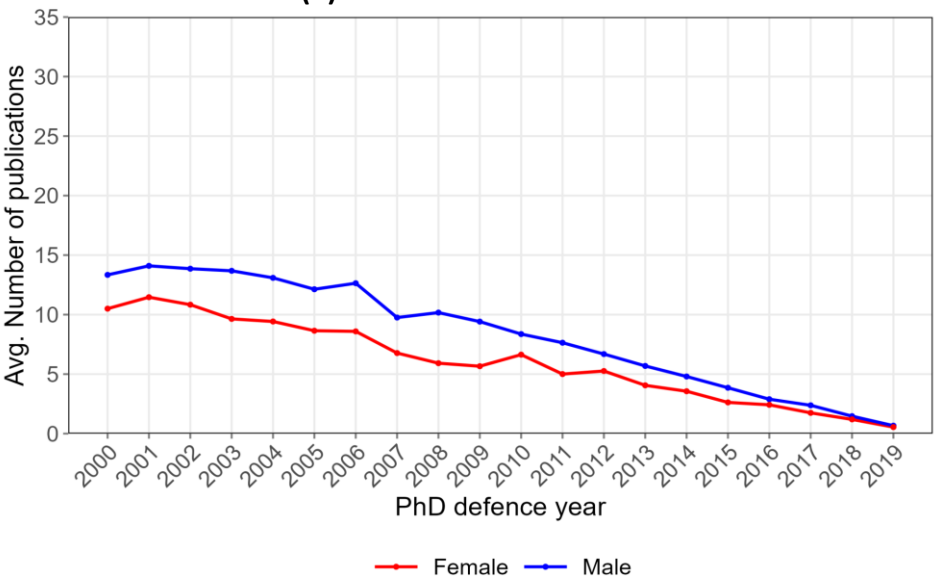
NOTE: The period during doctoral studies spans from three years before to one year after the PhD defence, while the period after doctoral studies starts two years after the defence. For the latter, we truncate the sample at the 2019 cohort, as publication data are available through 2021; consequently, the 2019 cohort is the last cohort with at least one observable post-PhD year.

Figure 16. Graduates' average number of publications by defence year and gender: Engineering

(a) During doctoral studies



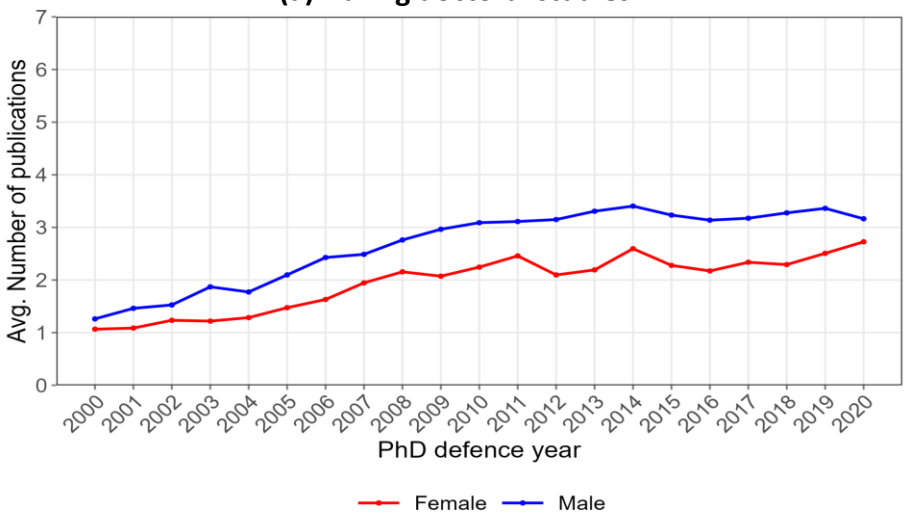
(b) After doctoral studies



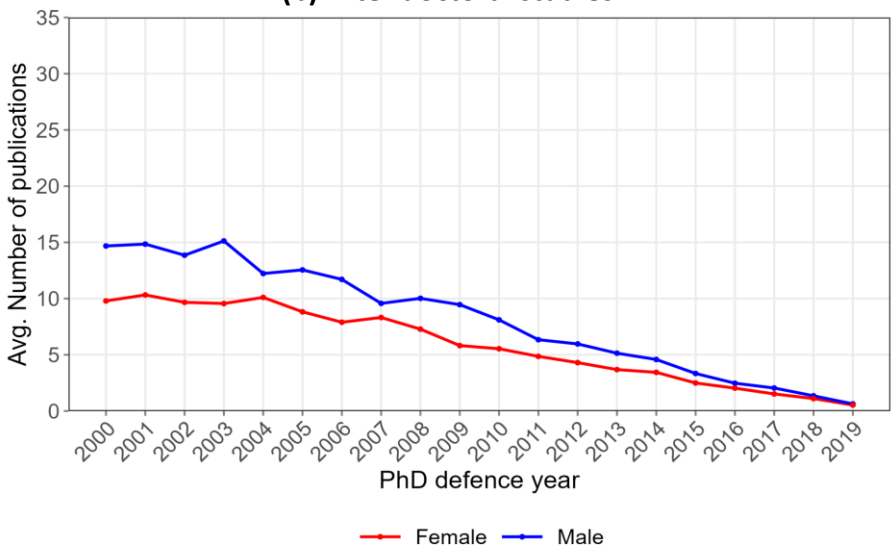
NOTE: The period during doctoral studies spans from three years before to one year after the PhD defence, while the period after doctoral studies starts two years after the defence. For the latter, we truncate the sample at the 2019 cohort, as publication data are available through 2021; consequently, the 2019 cohort is the last cohort with at least one observable post-PhD year.

Figure 17. Graduates' average number of publications by defence year and gender: Mathematics & Computer Science

(a) During doctoral studies



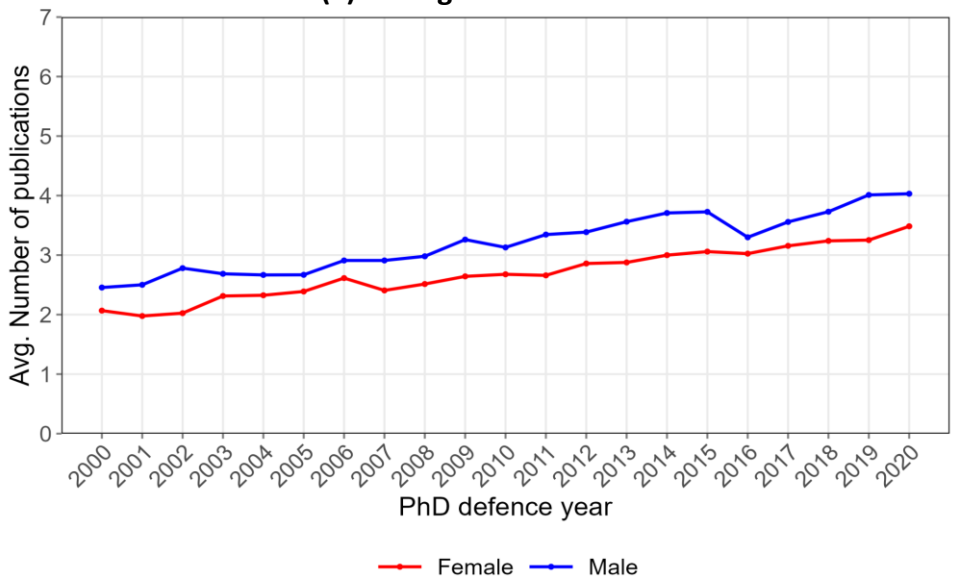
(b) After doctoral studies



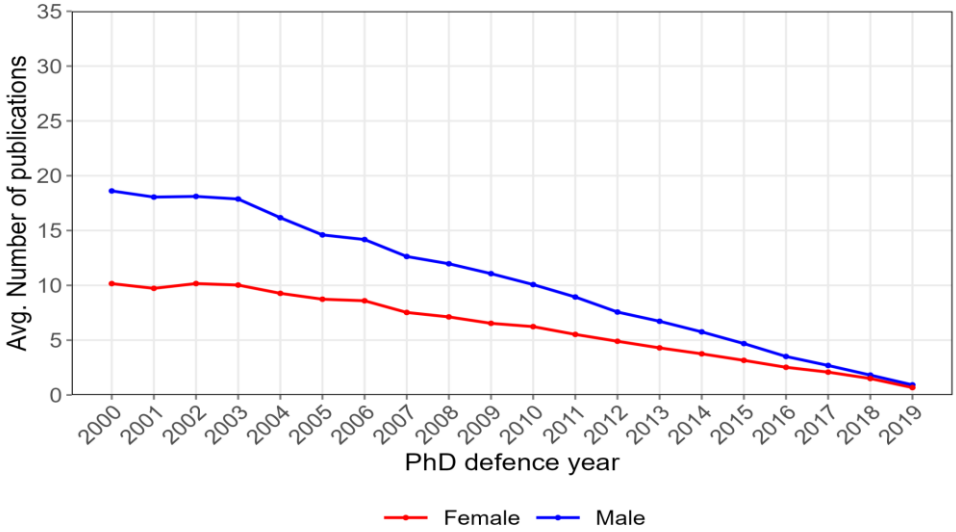
NOTE: The period during doctoral studies spans from three years before to one year after the PhD defence, while the period after doctoral studies starts two years after the defence. For the latter, we truncate the sample at the 2019 cohort, as publication data are available through 2021; consequently, the 2019 cohort is the last cohort with at least one observable post-PhD year.

Figure 18. Graduates' average number of publications by defence year and gender: Life Sciences

(a) During doctoral studies



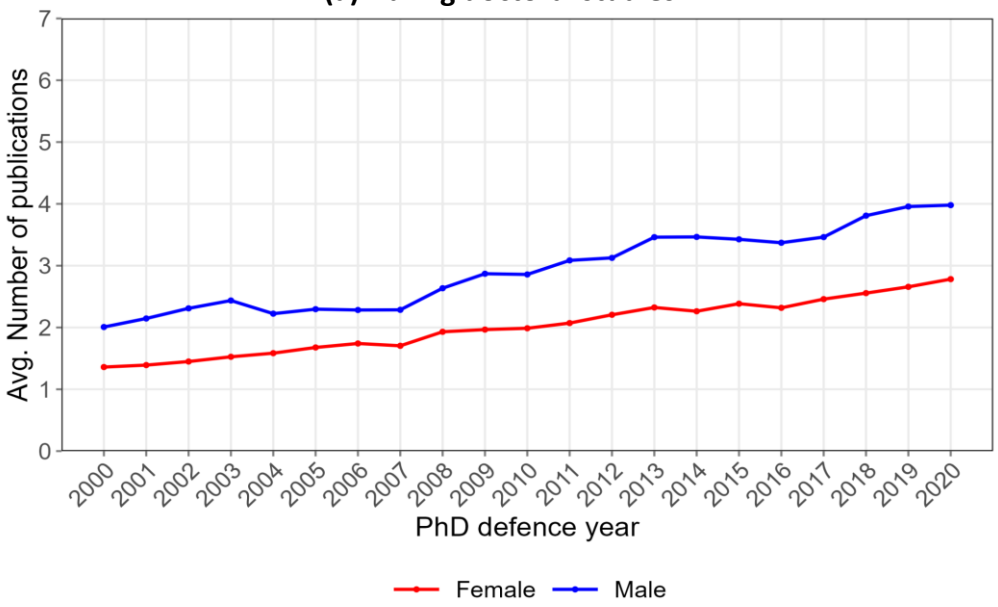
(b) After doctoral studies



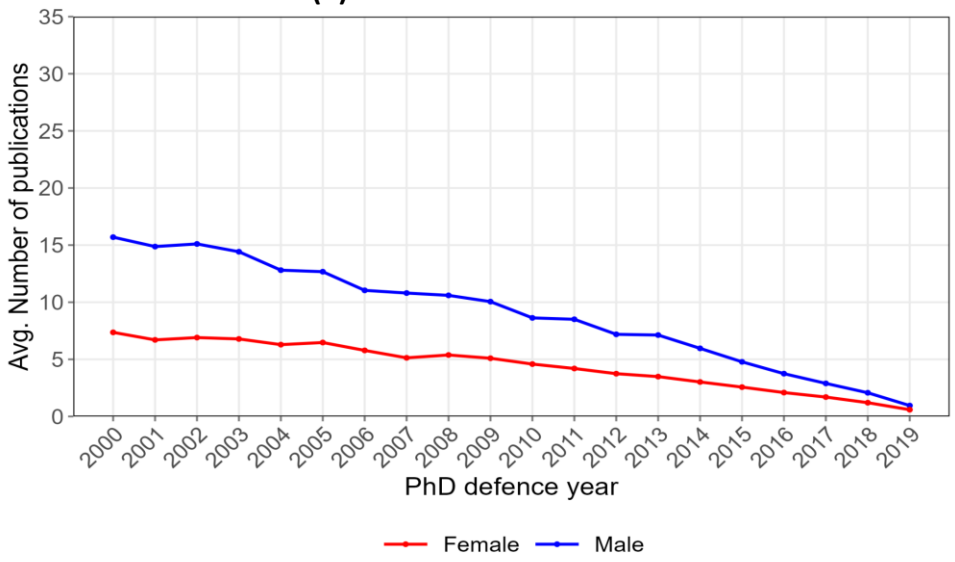
NOTE: The period during doctoral studies spans from three years before to one year after the PhD defence, while the period after doctoral studies starts two years after the defence. For the latter, we truncate the sample at the 2019 cohort, as publication data are available through 2021; consequently, the 2019 cohort is the last cohort with at least one observable post-PhD year.

Figure 19. Graduates' average number of publications by defence year and gender: Medicine

(a) During doctoral studies



(b) After doctoral studies



NOTE: The period during doctoral studies spans from three years before to one year after the PhD defence, while the period after doctoral studies starts two years after the defence. For the latter, we truncate the sample at the 2019 cohort, as publication data are available through 2021; consequently, the 2019 cohort is the last cohort with at least one observable post-PhD year.

6.2 Doctoral graduates' publications as knowledge inputs for patents

In this section, we examine the distance of doctoral graduates' publications from the technological frontier, based on the metrics described in Section 4.

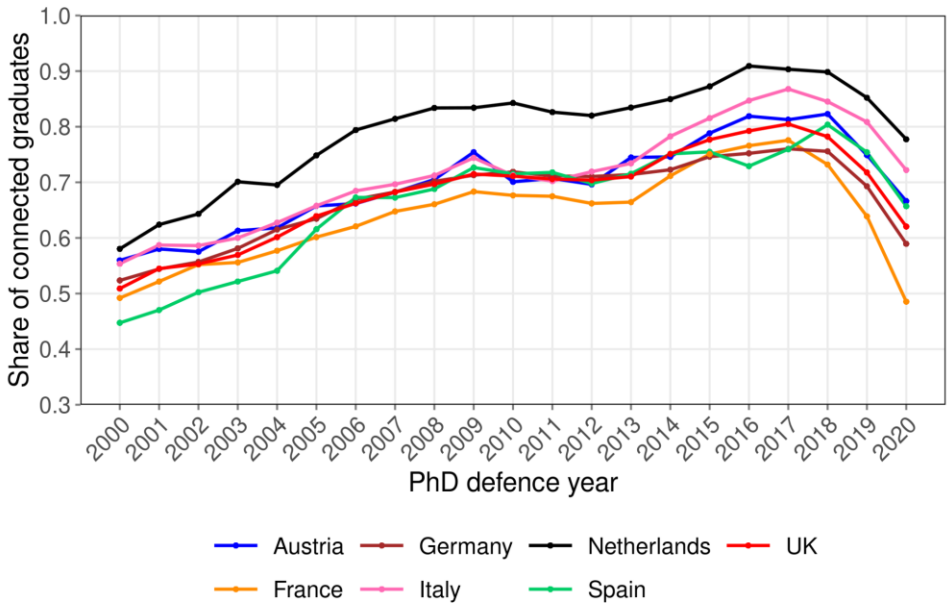
Note that, given the delay in the publication of patent documents, including references to non-patent literature and the additional time needed to prepare the database on reliance on science (Marx and Fuegi 2020a, 2020b), the latest doctoral defence cohorts in our database suffer from severe truncation issues. As a result, publications by doctoral graduates with a defence year in 2018-2020 are less likely to be cited in patents, as shown in some of the graphs below.

Figure 20 shows the shares of graduates with at least one publication (from the doctoral dissertation) connected to a patent ("connected graduates", in short), by country and defence year. The graduates considered are those with at least one publication during the doctoral training period, and the publications considered are those published from *t-3* to *t+1*, where *t* is the defence year. For all countries, we observe over time an increasing share of graduates connected to the technological frontier. The Netherlands shows a remarkably higher share of connected graduates compared to the other countries, with Italy approaching similar levels in the most recent cohorts.

Figure 21 shows the shares of graduates with at least one publication directly cited by a patent, by country and defence year. The share is calculated considering only graduates having at least one publication connected to a patent at any distance. We observe that the share of graduates with at least one publication cited by a patent is rather constant over the years for all countries. Interestingly, Austria and the Netherlands show a higher share of graduates with at least one publication directly cited by a patent if compared to other countries, reflecting a stronger orientation towards scientific areas more likely to be related to technological patent subject matter as well as science-industry links.

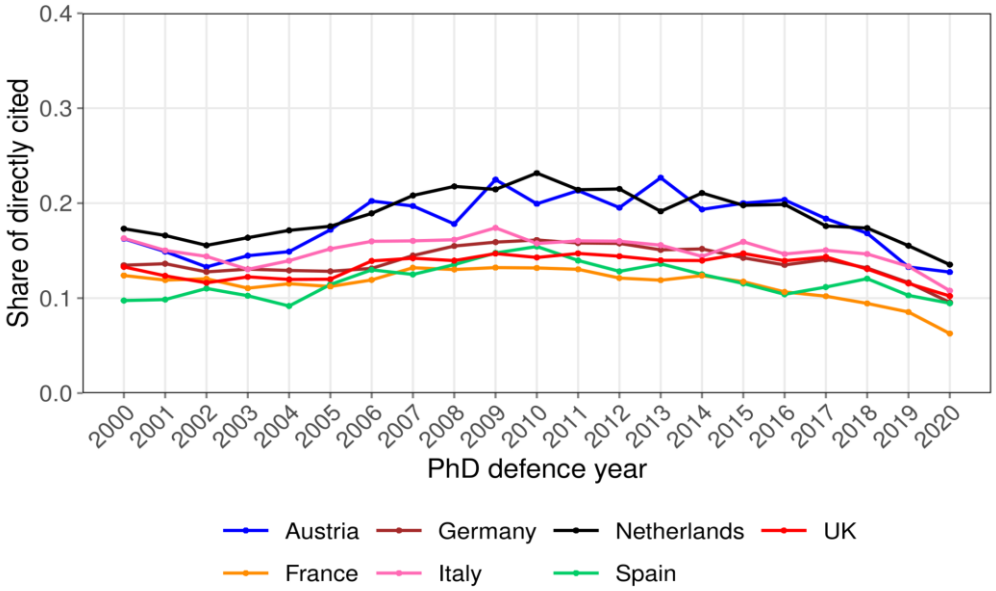
Figure 22 shows the average distance from the technological frontier of the graduates' closest publication to a patent, by country and defence year. The share is calculated considering only graduates having at least one publication connected to a patent at any distance. Interestingly, we observe that the average distance of the closest publication from the technological frontier is almost constant for all countries until 2013. From 2013, it seems to have increased for all countries, denoting a graduates' research output that is farther from the technological frontier in more recent years. French and Spanish graduates are the ones farthest from the technological frontier.

Figure 20. Share of doctoral graduates with at least one publication connected to the technological frontier by defence year and country (all disciplines)



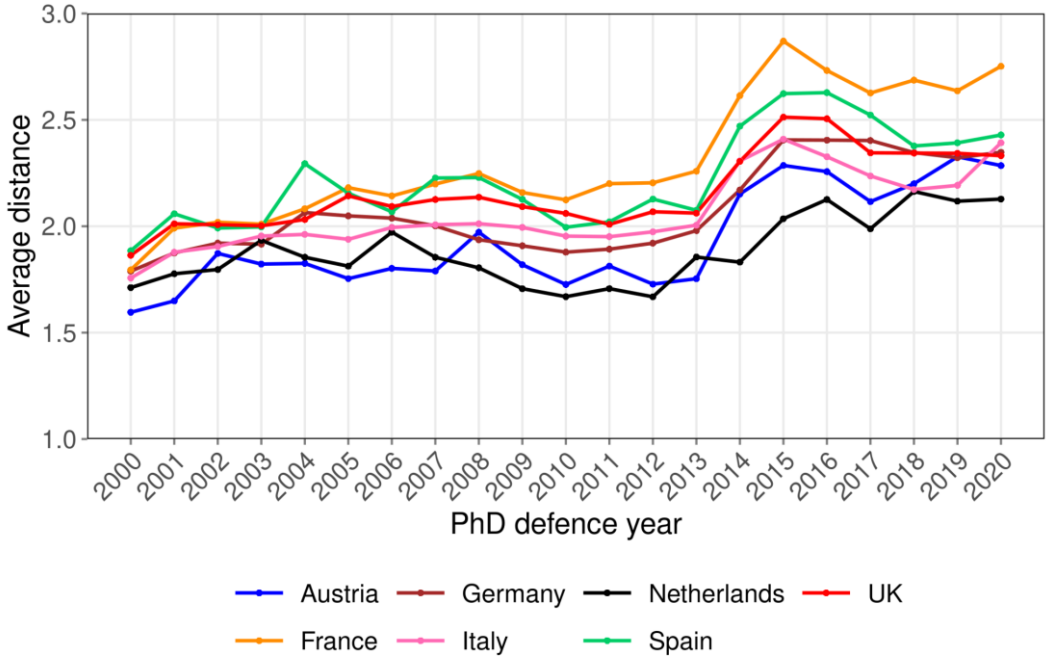
NOTE: This graph is calculated for graduates with at least one publication during the doctoral training period, from t-3 to t+1, where t is the defence year. Data for the most recent years may be incomplete, as our patent and patent citation data are available only through 2024.

Figure 21. Share of graduates with at least one publication directly cited by a patent by defence year and country (all disciplines)



NOTE: This graph is calculated for graduates with at least one publication connected to a patent at any distance during the doctoral training period, from t-3 to t+1, where t is the defence year. Data for the most recent years may be incomplete, as our patent and patent citation data are available only through 2024.

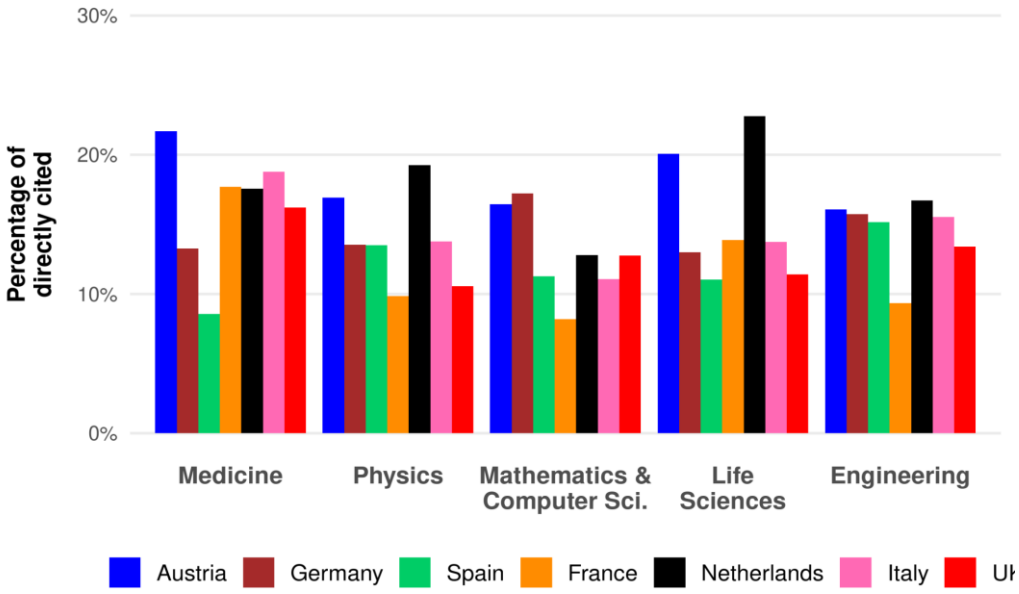
Figure 22. Average distance from the technological frontier of the graduates' closest publication to a patent by defence year and country (all disciplines)



NOTE: This graph is calculated for graduates with at least one publication connected to a patent at any distance during the doctoral training period, from t-3 to t+1, where t is the defence year. Data for the most recent years may be incomplete, as our patent and patent citation data are available only through 2024.

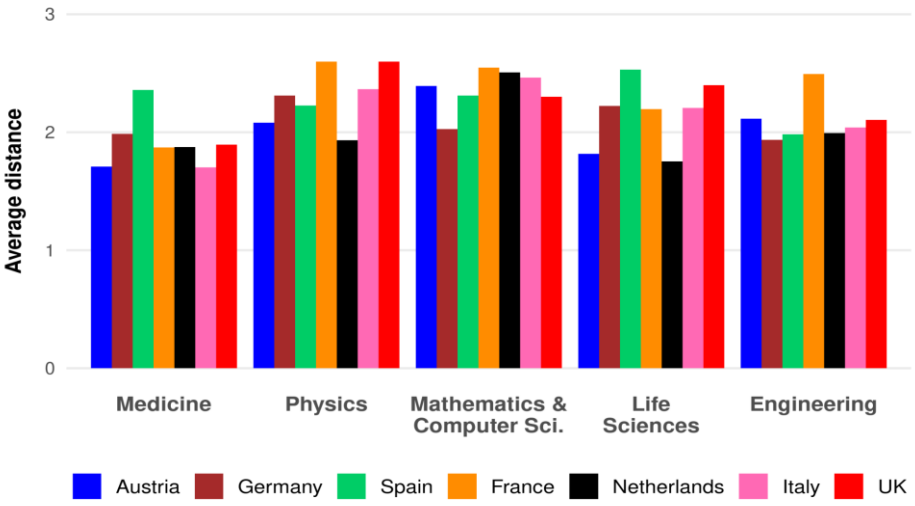
Figures 23 and 24 show differences across disciplines in the shares of graduates with at least one publication directly connected to a patent and the average distance from the technological frontier of the graduates' closest publication to a patent, respectively.

Figure 23. Percentage of graduates with at least one publication directly cited by a patent by discipline (all defence years)



NOTE: This graph is calculated for graduates with at least one publication connected to a patent at any distance during the doctoral training period, from t-3 to t+1, where t is the defence year. Data for the most recent years may be incomplete, as our patent and patent citation data are available only through 2024.

Figure 24. Average distance from the technological frontier of the graduates' closest publication to a patent by discipline (all defence years)



NOTE: This graph is calculated for graduates with at least one publication connected to a patent at any distance during the doctoral training period, from t-3 to t+1, where t is the defence year. Data for the most recent years may be incomplete, as our patent and patent citation data are available only through 2024.

Figures 25, 26, and 27 explore gender related trends in the share of graduates connected to the technological frontier at any distance, directly connected, and the average distance, respectively. Figure 28 breaks down the share of graduates directly connected to the technological frontier by country and gender.

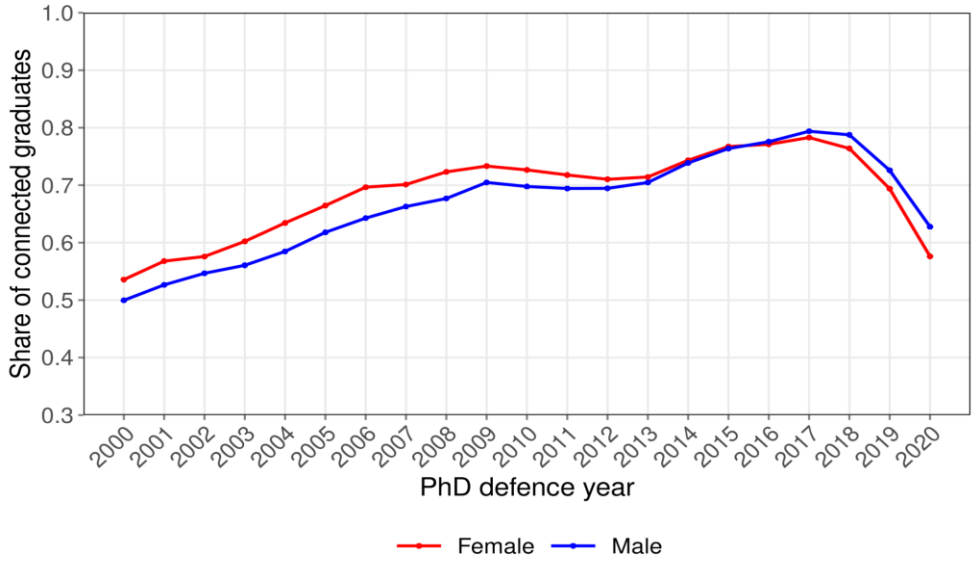
Figure 25 shows the shares of graduates with at least one publication connected to a patent, by gender and defence year. For female and male graduates, we observe over time an increasing share of graduates connected to the technological frontier. Female graduates are slightly more likely to be connected to the frontier than their male counterparts until 2013.

Figure 26 shows the shares of graduates with at least one publication directly cited by a patent, by gender and defence year. The share is calculated considering only graduates having at least one publication connected to a patent at any distance. Interestingly, male graduates are more likely to receive direct citations from patents to their publications than their female counterparts over the whole period considered.

Figure 27 shows the average distance from the technological frontier of the graduates' closest publication to a patent, by gender and defence year. The share is calculated considering only graduates having at least one publication connected to a patent at any distance. Interestingly, we observe, for both female and male graduates, that the average distance of the closest publication from the technological frontier is almost constant until 2013. From 2013, it seems to have increased for both females and males, denoting a graduate's research output that is farther from the technological frontier in more recent years. Female graduates are the farthest from the technological frontier.

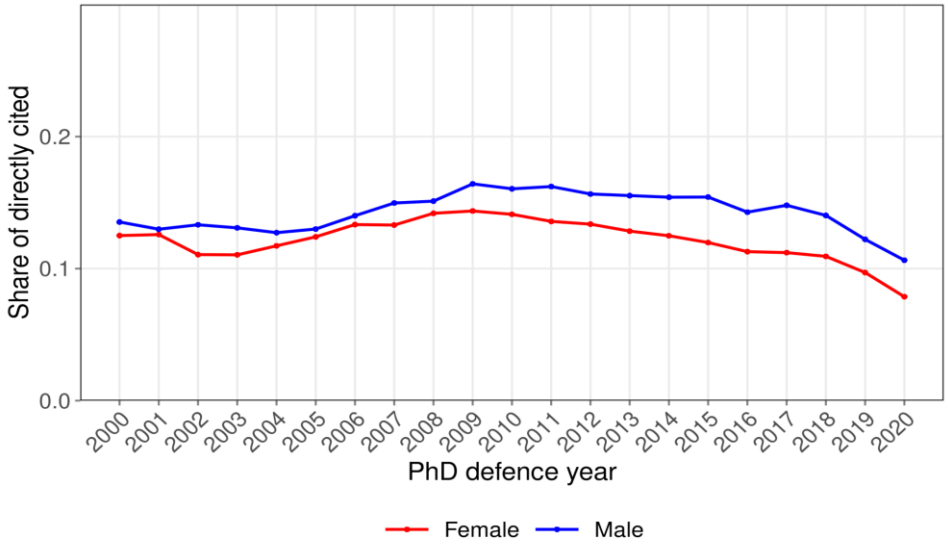
Figure 28 depicts the gender difference in the share of graduates with at least one publication directly cited by a patent, broken down by country. We observe that France shows no gender gap, while in other countries, male graduates are more likely to have a publication directly cited by a patent.

Figure 25. Share of graduates with at least one publication connected to the technological frontier by defence year and gender (all countries and disciplines)



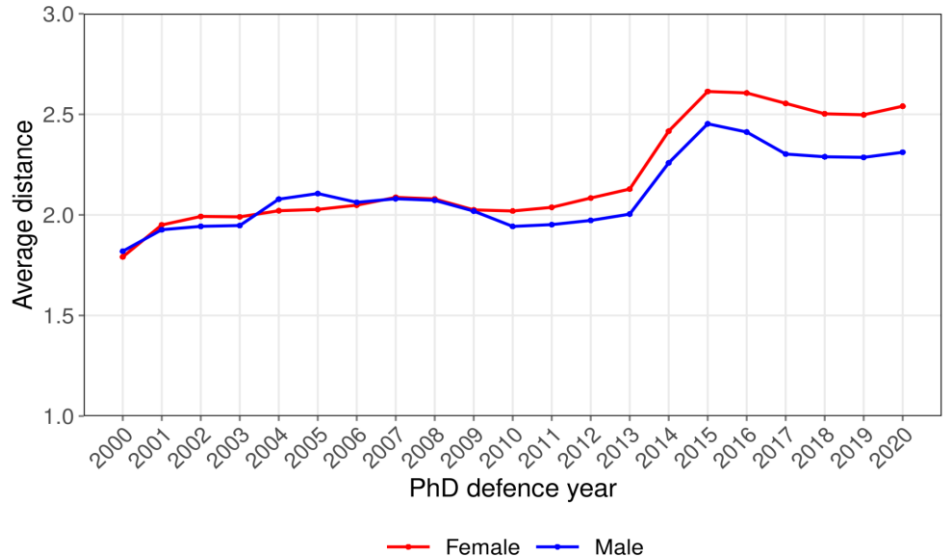
NOTE: This graph is calculated for graduates with at least one publication during the doctoral training period, from t-3 to t+1, where t is the defence year. Data for the most recent years may be incomplete, as our patent and patent citation data are available only through 2024.

Figure 26. Share of graduates with at least one publication directly cited by a patent by defence year and gender (all countries and disciplines)



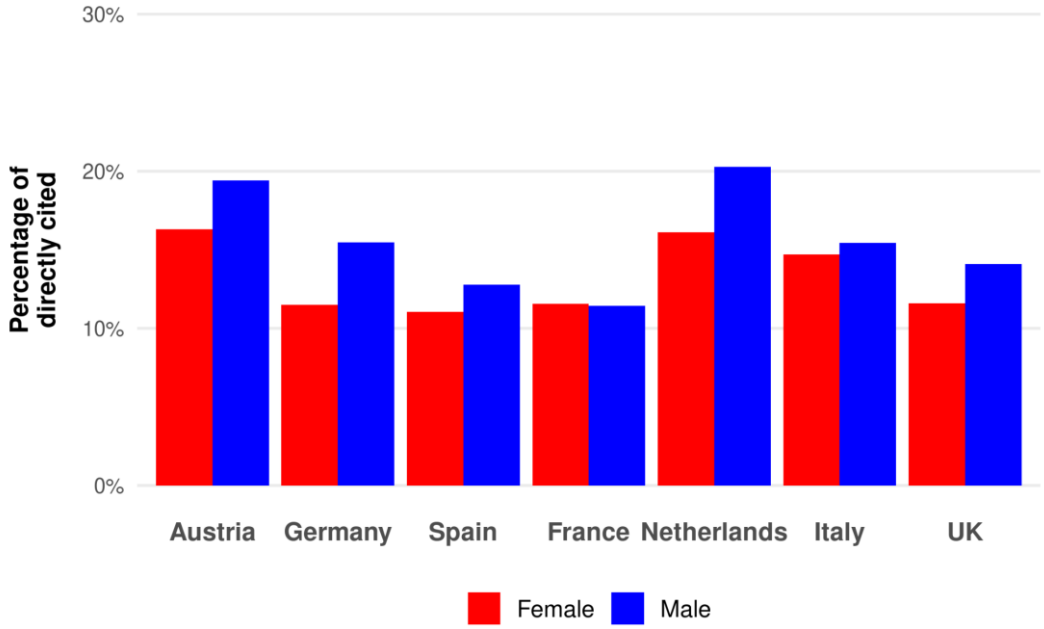
NOTE: This graph is calculated for graduates with at least one publication connected to a patent at any distance during the doctoral training period, from t-3 to t+1, where t is the defence year. Data for the most recent years may be incomplete, as our patent and patent citation data are available only through 2024.

Figure 27. Average distance from the technological frontier of the graduates' closest publication to a patent by defence year and gender (all countries and disciplines)



NOTE: This graph is calculated for graduates with at least one publication connected to a patent at any distance during the doctoral training period, from t-3 to t+1, where t is the defence year. Data for the most recent years may be incomplete, as our patent and patent citation data are available only through 2024.

Figure 28. Percentage of graduates with at least one publication directly cited by a patent by country and gender (all defence years and disciplines)



NOTE: This graph is calculated for graduates with at least one publication connected to a patent at any distance during the doctoral training period, from t-3 to t+1, where t is the defence year. Data for the most recent years may be incomplete, as our patent and patent citation data are available only through 2024.

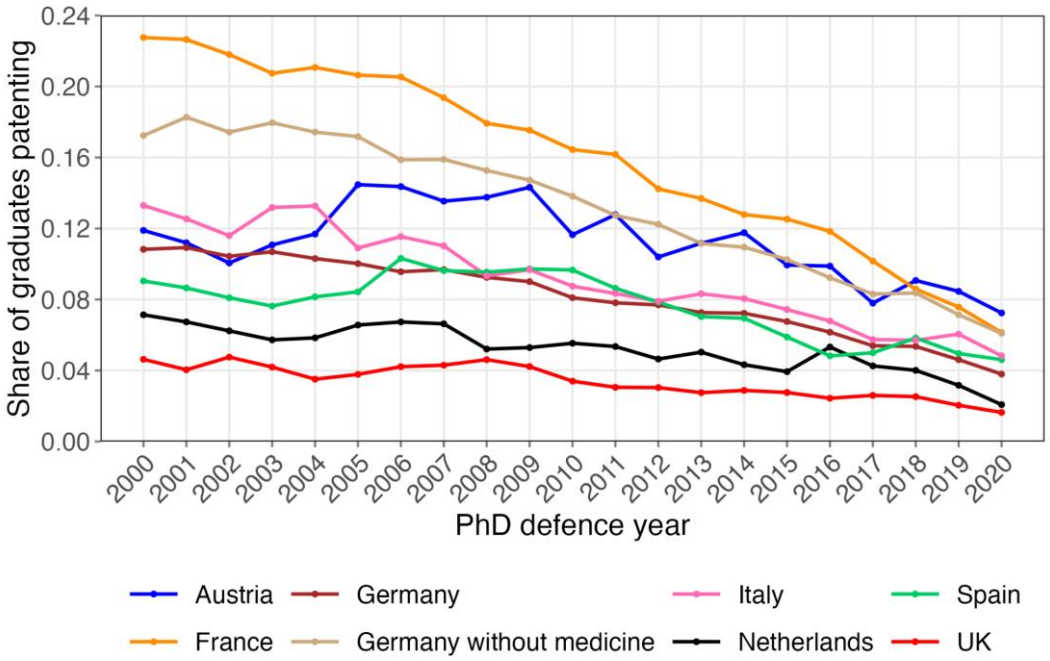
6.3 Doctoral graduates as inventors of EPO patent applications

The figures presented in this section provide insights into the involvement of doctoral graduates in inventive activity as measured by EPO patent applications. Figures 29-32 show general trends in the share of doctoral graduates appearing as inventors in EPO patents over time and across countries and disciplines. Figures 33-35 show gender disparities in the doctoral graduates' involvement in patenting. Together, they highlight that patenting among doctoral graduates is highly uneven across countries, disciplines, and genders. It should be noted that the apparent decline in the most recent years in some of the figures is partly due to the truncation of the data.

Figure 29 shows the share of doctoral graduates who appear as inventors on at least one EPO patent, broken down by defence year and country of doctoral degree-granting institutions. The trends highlight differences across countries. As with the publication data in Section 6.1, we show two different lines for Germany: with and without doctoral graduates in Medicine. As was the case for publications, France appears as the top country in the figure, with the highest rate of involvement of doctoral graduates in patenting for cohorts between 2000 and 2017, followed by Germany (without Medicine) and Austria, with the gap narrowing and even reversing in more recent years. About 22%

of French doctoral graduates[49] and 18% of German doctoral graduates (without medicine) of the early cohorts (2000-2005) have become inventors of EPO patents at some point (during or after their doctoral studies). The other countries have lower rates for those cohorts, between 5 and 14%, with the lowest share for the UK. This low patenting propensity may reflect a broader trend of low per-head patenting in the UK (e.g., CIIP, 2024; EPO, 2024). The share of doctoral graduates-inventors is lower for the most recent cohorts for all countries, very likely due to some truncation, as our patent data are available only through 2024.

Figure 29. Share of graduates appearing as inventors on at least one EPO patent by defence year and country (all disciplines)



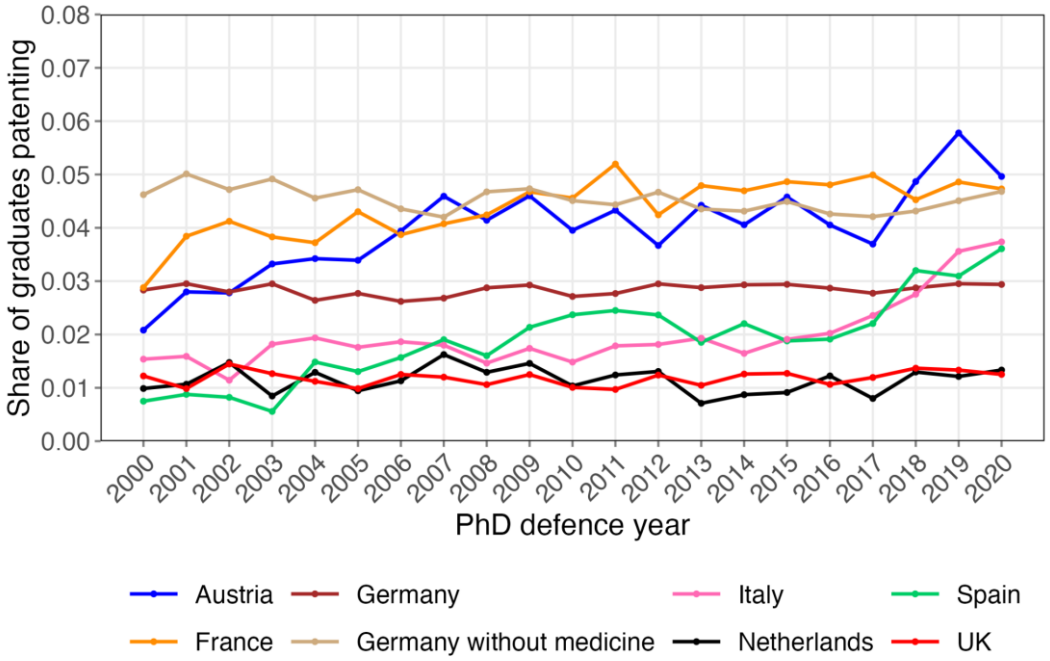
NOTE: Data for the most recent years may be incomplete, as our patent data are available only through 2024.

The picture is more homogeneous across countries when we narrow our focus to patents filed during the doctoral training period in Figure 30. Here, truncation issues are alleviated because we focus on a stricter time window around the defence date per cohort. In general, less than 5% of doctoral graduates in all countries become inventors during their doctoral studies (with some exceptions in 2011 in France and 2019 in Austria), and we observe an increasing trend, with graduates from more recent cohorts

[49] This higher rate for France could be again (as it was the case for the publications matching) due to the higher quality of the French ETD repository, which enables a higher recall rate, or to the specificity of French names, which are easier to match with PATSTAT due to the smaller number of homonymous names. Yet, we cannot exclude substantive reasons, such as the prevalence of doctoral graduates in France in early cohorts active in areas more prone to patenting, as we notice that the gap separating France from the other countries reduces over time.

being more likely than those from older ones to appear as inventors in patents filed during the doctoral training period, particularly for Austria, Italy, and Spain. France, Germany (without Medicine), and Austria have the highest rates, standing out with a share between 4 and 6% over the last fifteen years of our time window.

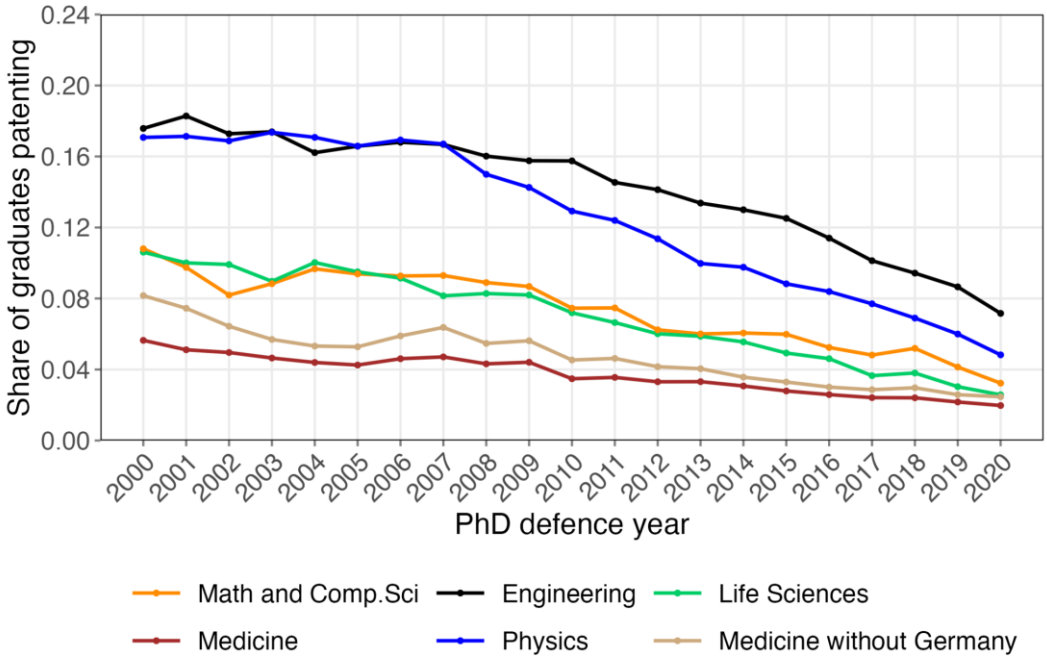
Figure 30. Share of graduates appearing as inventors on at least one EPO patent filed *during their doctoral studies* by defence year and country (all disciplines)



NOTE: The period during doctoral studies spans from three years before to one year after the PhD defence. Data for the most recent years may be incomplete, as our patent data are available only through 2024.

Figure 31 disaggregates by discipline, illustrating that patenting activity is far more common among graduates in Engineering and Physical Sciences rather than in Medicine and other disciplines. This disciplinary divide is consistent with the proximity of certain research areas to industrial applications. Figure 32 combines the country and discipline perspectives to compare two doctoral graduates' cohort periods (defence years 2000–2010 and defence years 2011–2020). This allows the observation of both structural differences and recent changes in how doctoral graduates from different disciplines and national systems engage in patenting. The high rate of inventive activity among French, Austrian, and German graduates in Physical Sciences and Engineering in the first period is remarkable (between 16 and 25%) and could be one of the reasons driving the results observed in Figure 29.

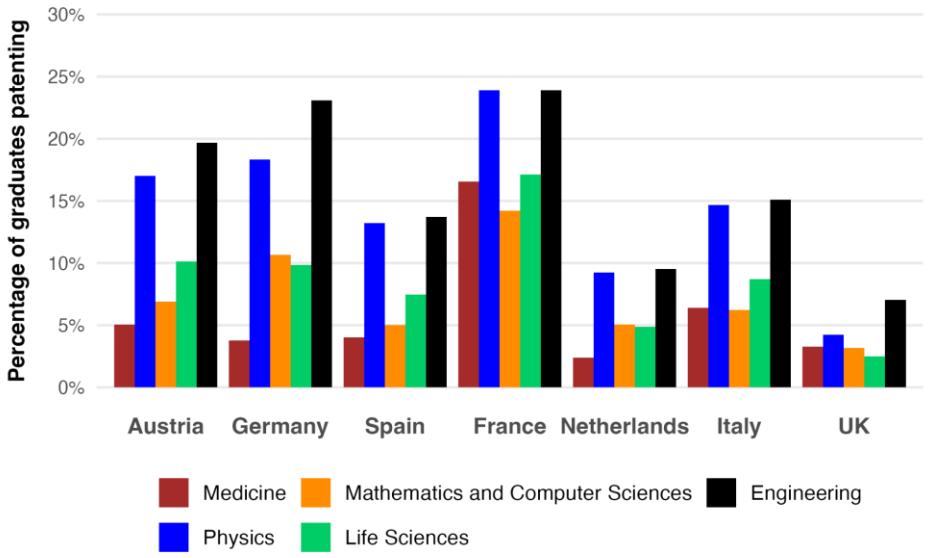
Figure 31. Share of graduates appearing as inventors on at least one EPO patent by defence year and discipline (all countries)



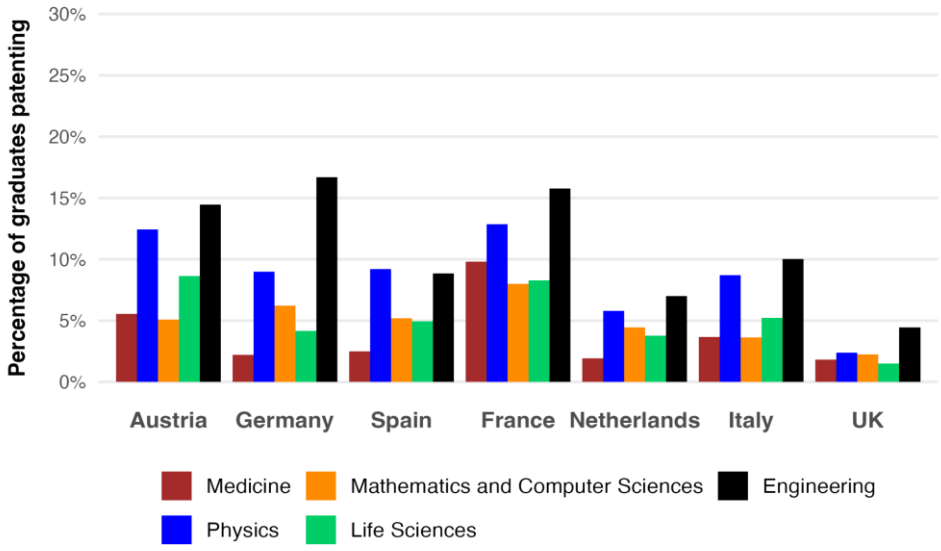
NOTE: Data for the most recent years may be incomplete, as our patent data are available only through 2024.

Figure 32. Percentage of graduates appearing as inventors on at least one EPO patent by country and discipline

(a) Defence years 2000-2010



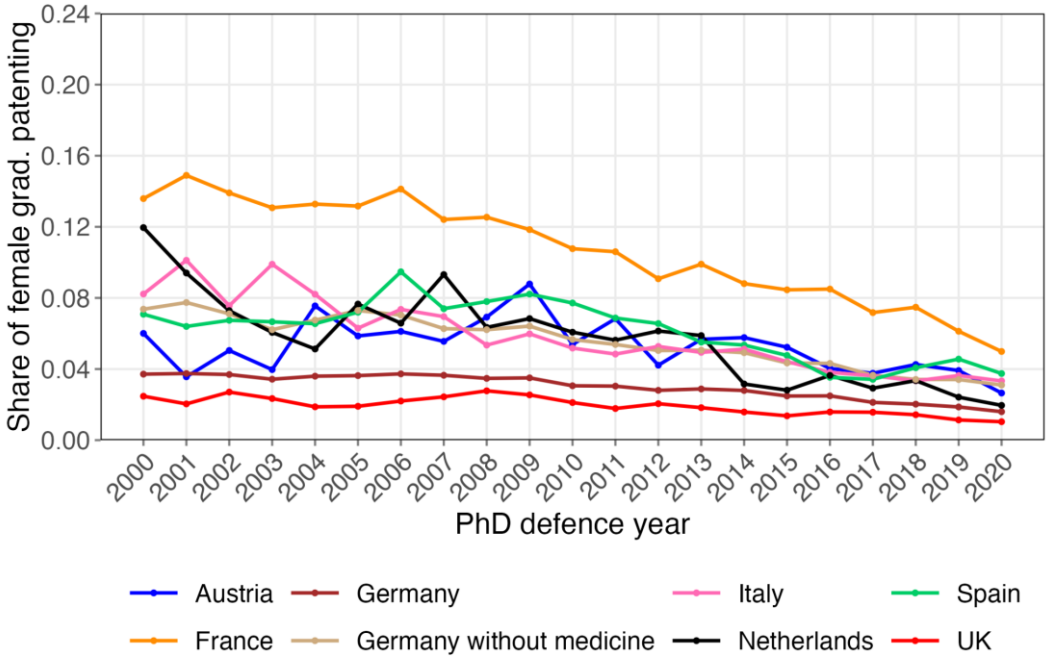
(b) Defence years 2011-2020



NOTE: Data for the most recent years may be incomplete, as our patent data are available only through 2024.

Figures 33, 34, and 35 examine gender disparities. Figure 33 shows the share of female doctoral graduates appearing as inventors, relative to the population of female graduates, by defence year and country. France is the country with the highest rate of involvement of women, between 12 and 15% for the early cohorts. Spain, Italy, Austria, the Netherlands, and Germany (without medicine) exhibit similar levels over the time period considered, altough the Netherlands and Austria have more irregular trends (because of their smaller size). The UK shows the lowest share of patenting at less than 3% for all cohorts.

Figure 33. Share of female graduates appearing as inventors on at least one EPO patent by defence year and country (all disciplines)



NOTE: Data for the most recent years may be incomplete, as our patent data are available only through 2024.

Figure 34 compares the overall patenting rates between men and women across different countries, and Figure 35 compares patenting rates between men and women across different disciplines; in both cases, rates are computed relative to the population of graduates of the same gender. All in all, these results suggest that gender gaps in the doctoral graduates' involvement in patenting are partly linked to the uneven representation of women across disciplines with higher inventive output, such as Engineering and Physical Sciences.

Figure 34. Percentage of graduates appearing as inventors on at least one EPO patent by gender and country (all disciplines)

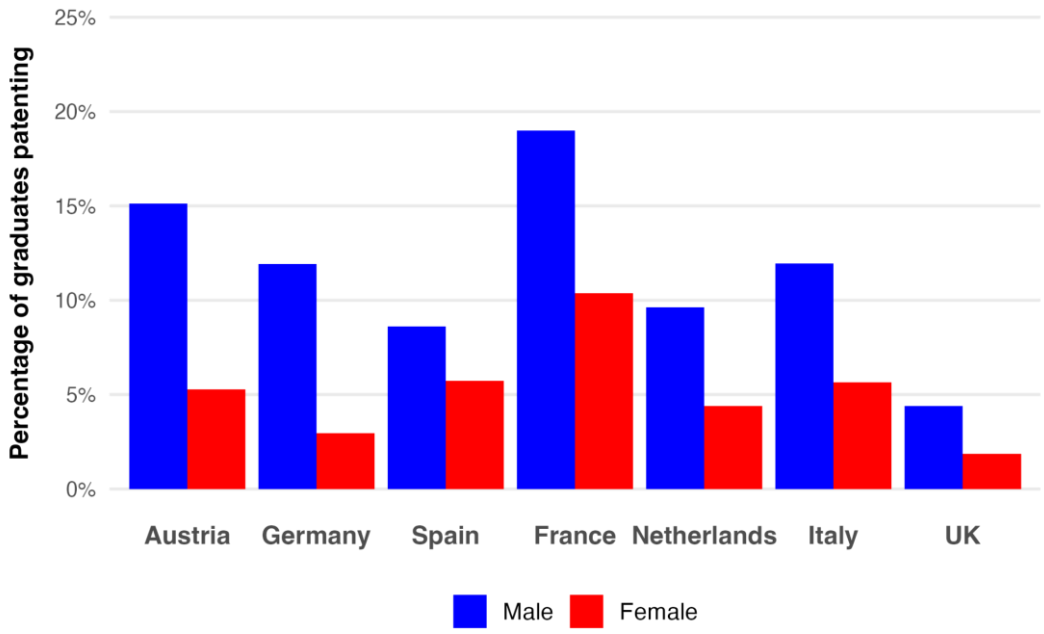
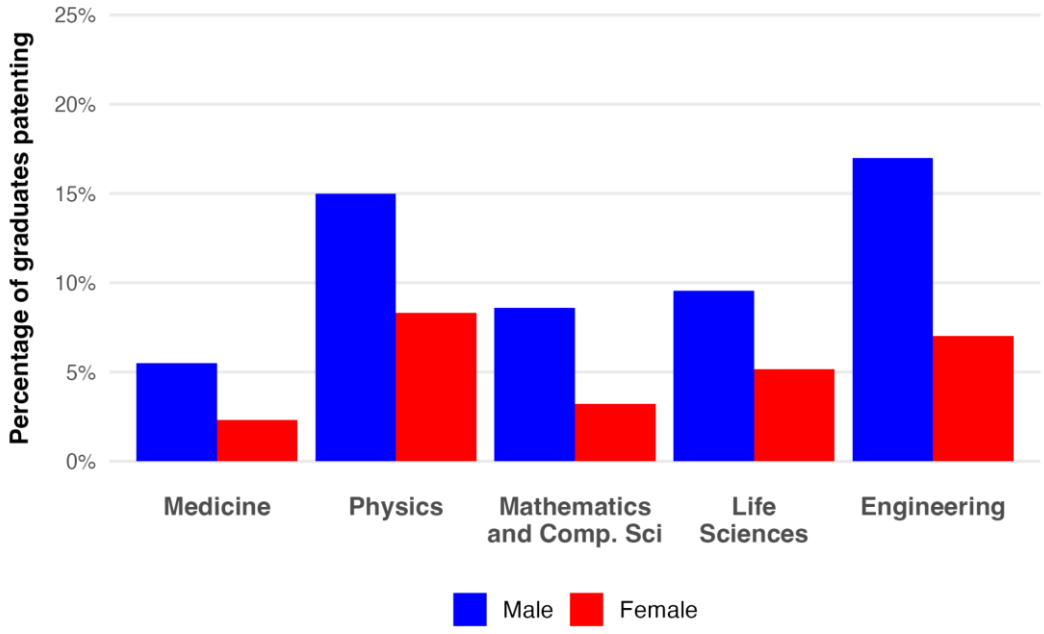


Figure 35. Percentage of graduates appearing as inventors on at least one EPO patent by gender and discipline (all countries)



7. Conclusions

The DOC-TRACK project has produced a novel and accurate methodological approach for:

- Linking the entire population of doctoral graduates to the authors of scientific publications and patent inventors, which relies on using a Random Forest classification algorithm –trained and tested with manually curated datasets– to provide reliable matches.
- Evaluating the relevance of the knowledge produced by doctoral graduates –measured by publications during their doctoral studies– for the inventive activity of all inventors, as measured by their citation distance to any EPO patent applications filed within five years of the publication date.

The project has also produced a number of descriptive results obtained by applying its methodology to all dissertations produced from 2000 to 2020 in seven European countries: Austria, France, Germany, Italy, the Netherlands, Spain, and the United Kingdom. The key takeaways from these results are:

- A large and increasing share of doctoral graduates produce at least one publication during their doctoral studies, ranging from about 40% to over 70% for the 2000 cohort to 60% to almost 90% in 2020, depending on the country and discipline. While levels vary by country and discipline, the increasing trend is common to all of them. This trend is consistent with the expansion of the number of papers and journals over our time window, which contributed to changing the structure of opportunities.
- An increasing number of publications were produced during the doctoral studies, from about 2 to 4 for the 2000 cohort to 4 to 6 (9 for Italy) in 2020. Here again, cross-country level differences are remarkable, but trends are common. Also, it appears that countries in which doctoral students are more productive are not the same in which a majority of them publish at least once, which suggests that publishing activity of doctoral graduates has also changed due to a mix of requirements for graduating and research opportunities.
- When looking at doctoral graduates' contributions to patenting, indirect contributions emerge as major ones. While the share of graduates who appear on a patent application during their doctoral studies ranges from about 1% to 5%, depending on the country and regardless of the graduation cohort, the share of those with at least one PhD publication directly or indirectly cited by a patent filed within 5 years of the publication is much higher and increasing over time. It ranges from 45-60% for the 2000 cohort to 75-90% for the 2018 cohort (depending on the country and discipline). Once again, cross-country and cross-disciplinary differences concern levels, not trends. Also interesting, the share of graduates with at least one PhD publication directly cited by a patent presents a

flat trend over the time window considered. This suggests that the value of doctoral work for inventions is increasing, but cannot be measured by the most traditional indicators.

- Gender differences are remarkable. Men are more likely to publish or patent than women both during and after the doctoral studies, and their publications during the doctoral studies are more likely to be cited directly by a patent (albeit less strongly); but when it comes to indirect contribution (having a publication ultimately cited by a patent, via a chain of citations by other publications) women perform better than men from 2000 to 2014. Some of the gender differences are due to composition effects, with large variations across disciplines and countries, which deserve further exploration.

Our methodology and data can become important tools for scholars and stakeholders alike, including policymakers. Indeed, the methodology's replicability across countries allows for cross-country comparisons and investigations into how diverse national contexts, institutional environments, and educational policies may influence doctoral graduates' performance.

We believe that a wise use of our data may reshape the public perception of science in general and doctoral studies in particular. Among others, it may cast light on the importance of doctoral education not just for academic careers, but also for inventive activity and technology transfer, which often goes under- or unappreciated.

Our data also look promising for social and economic research on science and technology. First, they add to the evidence base on gender bias in science and innovation. Second, they lend themselves to be used for studying the scientific and technological impact of other specific categories of doctoral graduates, such as foreign-educated ones (immigrants to Europe and/or mobile within Europe). Third, they enable in-depth analyses of the tension between measuring individual careers and the growing role of teams, organisations, and collaboration, as well as the consequences for measurement and attribution. All this will shed light on the degree of efficiency in the use of human capital and, consequently, encourage the design of policies aimed at increasing it.

REFERENCES

Ahmadpoor, M. and B. Jones. 2017. The dual frontier: Patented inventions and prior scientific advance. Science, 357, 583-587.

Baccini, A., De Nicolao, G. and E. Petrovich. 2019. Citation gaming induced by bibliometric evaluation: A country-level comparative analysis. PLoS ONE 14(9): e0221212.

Black, G. C. and P. E. Stephan. 2010. The Economics of University Science and the Role of Foreign Graduate Students and Postdoctoral Scholars. In American Universities in a Global Market, edited by Charles T. Clotfelter, 129–161. Chicago: University of Chicago Press.

Brischoux, F. and F. Angelier. 2015. Academia's Never-Ending Selection for Productivity. Scientometrics. 103(1), 333–36.

Buenstorf, G. and D.P. Heinisch. 2020. When do firms get ideas from hiring PhDs?, Research Policy, Volume 49, Issue 3, 103913, ISSN 0048-7333.

Buenstorf, G., Koenig, J. and A. Otto. 2023. Expansion of doctoral training and doctorate recipients' labour market outcomes: Evidence from German register data. Studies in Higher Education, 48, 1216-1242.

Callaert, J., Du Plessis, M., Grouwels, J., Lecocq, C., Magerman, T., Peeters, B., Song, X., Van Looy, B. and C. Vereyen. 2011. Patent statistics at Eurostat: Methods for regionalisation, sector allocation and name harmonisation. Eurostat Methodologies and Working Papers, Luxembourg: European Union.

Chawla, N. V., Bowyer, K. W., Hall, L. O. and W. P. Kegelmeyer. 2002. Smote: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16:321-357.

CIIP, Cambridge Industrial Innovation Policy. 2024. UK Innovation Report 2024. IfM Engage. Institute for Manufacturing, the University of Cambridge.

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

Corsini, A., Martínez, C. and E. Miguélez. 2025. Structuring PhD Programs: More control, better outcomes? IPP Working Paper. Institute of Public Goods and Policies, CSIC, Madrid, Spain.

Di Iasio, V., Lissoni, F., Miguelez, E., Tarasconi, G., Ménière, Y., Grilli, M. and I. Rudyk. 2022. Women's participation in inventive activity: Evidence from EPO data. European Patent Office. Munich.

Diez, C., Arkenau, C. and F. Meyer-Wentrup. 2000. The German medical dissertation—Time to change?. Academic Medicine, 75(8), 861-863

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

Enders, J. 2002. Serving many masters: The PhD on the labour market, the everlasting need of inequality, and the premature death of Humboldt. Higher education, 44(3), 493-517.

EPO, European Patent Office. 2024. The role of European universities in patenting and innovation, European Patent Office.

Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B. and F. Herrera. 2018. Learning from Imbalanced Data Sets. Springer.

Gareth, J., Witten, D., Hastie, T. and R. Tibshirani. 2013. An Introduction to Statistical Learning, Springer.

He, H. and E.A. Garcia. 2009. Learning from Imbalanced Data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263–1284.

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

Horta, H. and J. M. Santos. 2016. The Impact of Publishing during PhD Studies on Career Research Publication, Visibility, and Collaborations, Research in Higher Education 57, 28–50.

Koenig, J. 2025. Costs and benefits of a formal academic qualification beyond the PhD. Higher Education, 90, 613–647.

Konsortium BuWiK. 2025. Bundesbericht Wissenschaftlerinnen und Wissenschaftler in einer frühen Karrierephase 2025. Statistische Daten und Forschungsbefunde zu Promovierenden und Promovierten in Deutschland.wbv Publikation.

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

Lax-Martínez, G., Juano-i-Ribes, H.S. de, Yin, D., Feuvre, B.L., Hamdan-Livramento, I., Saito, K. and J. Raffo. 2021. Expanding the World Gender-Name Dictionary: WGND 2.0.

Lax-Martínez, G., Raffo, J. and K. Saito. 2016. Identifying the Gender of PCT inventors (No. 33), WIPO Economic Research Working Papers. World Intellectual Property Organization - Economics and Statistics Division.

Magerman, T., Van Looy, B. and X. Song. 2006. Data Production Methods for Harmonized Patent Statistics: Patentee Name Harmonization. SSRN Electronic Journal. RePEc.

Maraut, S. and C. Martínez. 2014. Identifying author-inventors from Spain: methods and a first insight into results. Scientometrics, 101(1), 445-476.

Marx, M. and A. Fuegi. 2020a. Reliance on science by inventors: Hybrid extraction of in-text patent-to-article citations. National Bureau of Economic Research.

Marx, M. and A. Fuegi. 2020b. Reliance on Science: Worldwide Front-Page Patent Citations to Scientific Articles. Strategic Management Journal, 41, 1572-1594.

Miguelez, E., Raffo, J. D. Chacua, C., Coda-Zabetta, M., Yin, D., Lissoni, F. and G. Tarasconi. 2019. Tied in: The Global Network of Local Innovation. World Intellectual Property Organization. Economic Research Working Paper Series No. 58.

Mikolov, T., Chen, K., Corrado G. and J. Dean. 2013. Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781 [Cs], January.

Murmann, J.P. 2013. The coevolution of industries and important features of their environments. Organization Science, 24(1), 58-78.

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

Rose, M.E. and J.R. Kitchin. 2019. pybliometrics: Scriptable bibliometrics using a Python interface to Scopus SoftwareX, 10 (2019), Article 100263.

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

Shibayama, S. 2019. Sustainable Development of Science and Scientists: Academic Training in Life Science Labs. Research Policy 48: 676–92.

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

Storti, C. 2019. Il deposito, la valorizzazione e la conservazione delle tesi di dottorato nell'esperienza di Magazzini digitali: un contributo per la ricerca e l'accesso. JLIS.it: Italian Journal of Library and Information Science 10(4): 114–124.

Appendix A: ETD data features present in each repository

		Theses. FR	DissOnline. DE	TESEO. ES	Narcis. NL	ONB Catalogue.AT	OPAC BNCF.IT	EThOS.BL .UK
Doctoral thesis	Title	yes	yes	yes	yes	yes	yes	yes
	Abstract	yes	Yes	yes	yes	-	-	yes
	Keywords	yes	yes	yes	yes	yes	-	-
	Discipline	yes	yes	yes	-	yes	yes	yes
	Doctoral degree-granting institution	yes	yes	yes	yes	yes	yes	yes
	Date of defence	yes	yes	yes	yes	yes	yes	yes*
Graduate	Name	yes	yes	yes	yes (initials)	yes	yes	yes
	Surname	yes	yes	yes	yes	yes	yes	yes
Supervisor /Jury members	Recorded	yes	yes	yes	yes	yes	yes	yes
	Role	yes	-	yes	-	yes	yes	-
	Name	yes	yes	yes	yes (initials)	yes	yes	yes
	Surname	yes	yes	yes	yes	yes	yes	yes

* For the UK, “date” refers to dissertation publication year, which may be later than the year of viva (defence).

Appendix B: Doctoral graduates - publication authors matching, Random Forest classification variables

			Country						
Group	**Variable**	**Description**	**DE**	**FR**	**ES**	**AT**	**NL**	**IT**	**UK**
Social proximity	Supervisor is a co-author	Continuous. Maximum similarity among the Levenshtein similarities between the full names of the doctoral graduate's supervisors and co-supervisors and the full names of the Scopus author's co-authors, considering all the full names of the Scopus author's co-authors associated with the same Scopus author ID. The Levenshtein similarity is calculated as: 1 – (Levenshtein_distance / max_length_between_the_two_full_names)		X	X	X	X		X
	Supervisor/Committee is a co-author	Continuous. Maximum similarity among the Levenshtein similarities between the full names of the doctoral graduate's supervisor and co-supervisor and jury committee members and the full names of the Scopus author's co-authors, considering all the full names of the Scopus author's co-authors associated with the same Scopus author ID. The Levenshtein similarity is calculated as: 1 – (Levenshtein_distance / max_length_between_the_two_full_names)	X					X	
	Only initial for the supervisor's name	Dummy. Value 1 if the only information available for the supervisor's name is the initial letter; 0 otherwise.					X		
	No info supervisor	Dummy. Value 1 if no information on the doctoral graduate's supervisor is available; 0 otherwise.		X	X	X	X		X
	No info supervisor and committee	Dummy. Value 1 if no information on the doctoral graduate's supervisor and committee members is available; 0 otherwise.	X						
Geographic alignment	Affiliation: city/institute	Dummy. Value 1 if the institute or city where the doctoral graduate defended the thesis appears in the affiliation of at least one Scopus author's article that is published during the graduate's doctoral studies and the subsequent year; 0 otherwise.[50]	X	X	X	X	X	X	X
	Affiliation: country	Dummy. Value 1 if the country where the doctoral graduate defended the thesis appears in the affiliation of at least one Scopus author's article that is published during the graduate's doctoral studies and the subsequent year; 0 otherwise.	X	X	X	X	X		X
Topical similarity	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	X	X	X	X	X	X*	X

[50] We included an additional year beyond the doctoral studies to be less restrictive and improve recall.

		Scopus authors' articles published during the graduate's doctoral studies and the subsequent year.[51]							
	Same discipline (share)	Continuous. Share of the Scopus author's articles published in journals having a Scopus ASJC (All Science Journal Classifications) discipline that coincides with the graduate's thesis discipline.	X	X	X	X	X		X
	No info discipline	Dummy. Value 1 if no information on the thesis discipline is available; 0 otherwise.	X	X	X	X	X		X
	Same keywords (share)	Continuous. Share of the Scopus author's articles that contain a keyword coinciding with a keyword in the graduate's thesis.				X			
	No info keywords	Dummy. Value 1 if no information on the thesis keywords is available; 0 otherwise.				X			
Name similarity	Name similarity	Continuous. Maximum similarity among the Levenshtein similarities between the doctoral graduate's full name and the Scopus author's full names, considering all the Scopus author's full names associated with the same Scopus author ID. The Levenshtein similarity is calculated as: 1 – (Levenshtein_distance / max_length_between_the_two_full_names)	X	X	X	X	X	X°	X
	Only initial for the graduate's name	Dummy. Value 1 if the only information available for the graduate's name is the initial letter; 0 otherwise.					X		
Publication productivity	Normalised productivity	Continuous. Number of normalised articles published by the Scopus author during the graduate's doctoral studies and the subsequent year. The productivity is normalised by dividing it by the average researchers' productivity in the same country, discipline, and time period corresponding to the graduate's doctoral training period.	X	X	X	X	X		X
	Zero productivity	Dummy. Value 1 if no articles were published by the Scopus author during the graduate's doctoral studies and the subsequent year; 0 otherwise.	X	X	X	X	X	X♦	X

* For Italy, topical similarity was calculated by embedding the titles of doctoral theses and Scopus-indexed articles using the SPECTER2 model from HuggingFace, which generates contextualized document embeddings optimized for measuring semantic similarity between scientific texts.

° In addition to measuring name similarity between each doctoral graduate and Scopus author candidate, the Random Forest model for Italy incorporates a score representing the relative commonness or rarity of an individual's name among all authors (both thesis authors and Scopus authors) in the dataset.

♦ The Random Forest model for Italy includes a variable representing the time difference, in years, between the Scopus candidate author's first publication and the doctoral thesis defence year.

[51] We rely on the EPO's translation engine to translate the titles of theses and articles into English. For Italy and the Netherlands, translation of titles into English was made using the DeepL API.

Appendix C: Doctoral graduates - patent inventors matching, Random Forest classification variables

Group	**Variable**	**Description**
Social proximity	Supervisor is a co-inventor	Continuous. Maximum similarity among the Jaccard similarities between the full names of the doctoral graduate's supervisors and co-supervisors and the full names of the inventor's co-inventors. The Jaccard string-matching algorithm is based on 3-grams.
Geographic alignment	Affiliation: city	Dummy. Value 1 if the city where the doctoral graduate defended the thesis appears in the address of at least one inventor's patent application; 0 otherwise.
	Affiliation: country	Dummy. Value 1 if the country where the doctoral graduate defended the thesis appears in the address of at least one inventor's patent application; 0 otherwise.
Topical similarity	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 inventor's patent applications.[52]
	Same discipline (share)	Continuous. Share of inventor's patents classified in the technological sector (from patent classifications) associated with the scientific discipline that matches the doctoral graduate's thesis discipline. To link technological sectors to scientific disciplines, we use the publicly available Reliance on Science dataset (Marx and Fuegi, 2020a and 2020b), which connects patents to publications through citations. The scientific disciplines are identified using Scopus ASJC (All Science Journal Classifications) codes for publications and the corresponding discipline classifications of the theses.
Name similarity	Name similarity	Continuous. Maximum similarity among the Jaccard similarities between the full name of the doctoral graduate and the different versions of the inventor's full name. The Jaccard string-matching algorithm is based on 3-grams.
Patenting activity	Normalised productivity	Continuous. Number of normalized patents filed by the inventor, adjusted for truncation due to recent cohorts by dividing the count of patents by the number of years available for the graduate to file patents (from thesis defence until the last year observed in the Patstat data, 2024).
	Productivity before defence	Dummy. Value 1 if the inventor filed a patent before the graduate's doctoral defence year; 0 otherwise.
	Years to first patent	Continuous. Number of years elapsed between the graduate's doctoral defence and the inventor's first patent application.
Patent academic features	Inventor's academic title	Dummy. Value 1 if the inventor's name on any of their patent applications contains an academic title (e.g., PhD, Dr., Prof., Professor, etc.); 0 otherwise.
	Co-inventor's academic title	Dummy. Value 1 if the name of at least one co-inventor on the inventor's patent applications contains an academic title (e.g., PhD, Dr., Prof., Professor, etc.); 0 otherwise.
	University applicant	Dummy. Value 1 if at least one applicant on any of the inventor's patent applications is classified as a university; 0 otherwise.
	Broad academic applicant	Dummy. Value 1 if at least one applicant on any of the inventor's patent applications is classified in the broad academic sector, which includes universities, hospitals, non-profits, and public research organizations; 0 otherwise. The sector allocation of patent applicants as university or within a

[52] We rely on the EPO's translation engine to translate the titles of theses and articles into English.

		broadly defined academic sector is based on the KUL EEEPAT dataset, version of February 2025 (Callaert et al., 2011).
Thesis country	Thesis France	Dummy. Value 1 if the doctoral graduate defended the thesis in France; 0 otherwise.
	Thesis Germany	Dummy. Value 1 if the doctoral graduate defended the thesis in Germany; 0 otherwise.
	Thesis Spain	Dummy. Value 1 if the doctoral graduate defended the thesis in Spain; 0 otherwise.
	Thesis Austria	Dummy. Value 1 if the doctoral graduate defended the thesis in Austria; 0 otherwise.
	Thesis Netherlands	Dummy. Value 1 if the doctoral graduate defended the thesis in the Netherlands; 0 otherwise.
	Thesis Italy	Dummy. Value 1 if the doctoral graduate defended the thesis in Italy; 0 otherwise.
	Thesis UK	Dummy. Value 1 if the doctoral graduate defended the thesis in the UK; 0 otherwise.

Appendix D: A more in-depth examination of scientific productivity

This appendix offers a more granular examination of the scientific productivity of doctoral graduates across countries and fields. Whereas Section 6 of the main report presented average publication levels, the analysis here explores the *entire distribution* of publication outcomes. This broader perspective provides a more nuanced understanding of cross-country differences and of the extent to which productivity varies within each national doctoral system. Figures D1 to D5 present boxplots summarizing the distribution of publications per doctoral graduate (ETD) by country and scientific field. For each discipline, the upper panel considers publications produced during the doctoral period (from three years before the defence to one year after), while the lower panel tracks output in the subsequent years, starting in the second year after completion. Because publication counts are highly dispersed, the y-axis is expressed on a logarithmic scale to facilitate comparison.

Each boxplot conveys several distributional features:

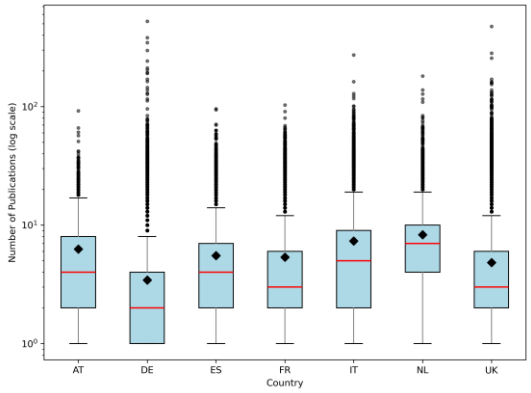
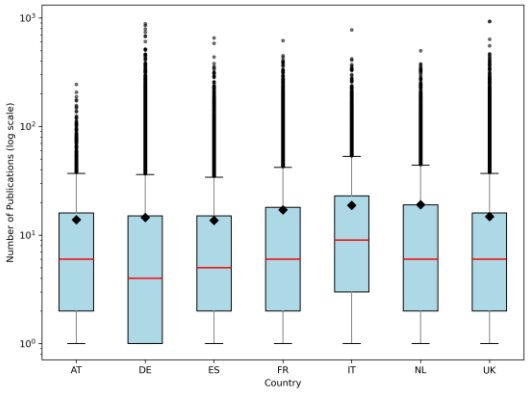
- Interquartile Range, IQR (the box): The box spans the 25th (Q1) to the 75th (Q3) percentile. This measure is robust to extreme values and provides an indication of heterogeneity within each system. Wider boxes reflect greater dispersion in publication activity.
- Median (the horizontal red line): The red line inside the box identifies the median publication count. It represents the value that splits the distribution into two parts, with 50 per cent of graduates publishing more and 50 per cent publishing fewer publications. Because it is not influenced by outliers, the median offers a reliable indication of the "typical" publication output for a given group.
- Mean (the black diamond): The diamond represents the average number of publications per graduate. Unlike the median, the mean is sensitive to very prolific individuals. When the diamond lies far above the median, it signals a highly skewed distribution in which a small number of graduates publish substantially more than the rest.
- 1.5 × IQR range (the whiskers): The whiskers extend from the edges of the box to the smallest and largest values that fall within 1.5 times the IQR below Q1 or above Q3. They summarize the range within which most graduates' publication counts lie. Short whiskers indicate a compact distribution with limited variability, while longer whiskers point to broader differences in productivity.
- Outliers (the dots outside the whiskers): Observations that fall beyond the 1.5 × IQR thresholds appear as individual points. These correspond to graduates whose publication counts are unusually high, or unusually low, relative to their peers in the

same country and field. Outliers help identify exceptionally prolific researchers as well as those with minimal publication activity.

Overall, the analysis of scientific productivity during and after the PhD reveals clear and persistent differences across countries and fields. Although each discipline displays its own characteristic level of publication intensity, the ranking of countries is remarkably stable. In nearly every field, graduates trained in Italy, Spain, and the Netherlands tend to publish more during their PhD and maintain higher levels of productivity afterwards. By contrast, Germany and France consistently record lower medians, with Austria and the UK generally occupying intermediate positions.

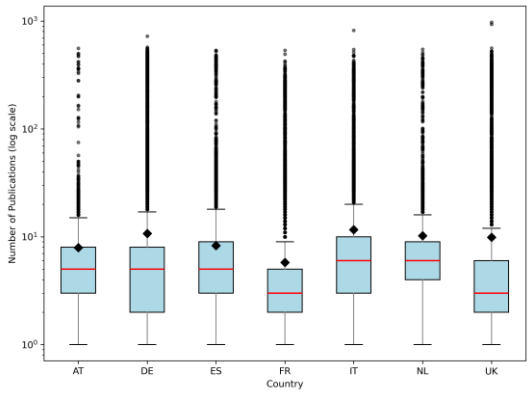
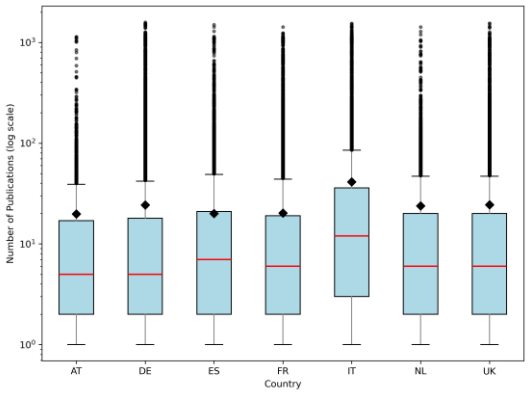
These regularities are already visible when examining publication behavior during the doctoral period. Most fields show a moderate level of activity already during the PhD, but the extent of that activity differs widely. In Medicine (Figure D1), for instance, publication during the PhD is particularly intense in the Netherlands and Italy, where the median doctoral graduate produces around seven and five publications, respectively. Austria and Spain follow with medians of four, while France and the UK record typical values closer to three. Germany stands out with substantially lower medians, around two publications, reflecting a markedly less publication-oriented structure of medical training. These cross-country differences widen after the PhD.

Table D1: Publications per doctoral graduate – Medicine

During PhD	*After PhD*
	

Italian medical graduates reach a median of nine publications in the postdoctoral period, while in most other countries, the typical graduate produces between five and six, and Germany remains at around four. The boxplots make it clear that these differences are not only a matter of central tendency: Italy and the Netherlands display longer upper tails and greater dispersion, reflecting a sizeable group of highly productive researchers who continue publishing at a sustained pace well after completing their degree.

Table D2: Publications per doctoral graduate – Physical Sciences

During PhD	*After PhD*
	

A similar picture emerges in Physical Sciences (Figure D2), though at a higher overall level of scientific output. During the PhD, Italian and Dutch graduates again show the highest typical publication rates, with medians of six publications, followed closely by Austria, Germany and Spain, where the median is around five. France and the UK display more modest medians, around three publications. The differences become more pronounced after graduation. Italian physicists, in particular, experience a sharp increase in productivity, reaching a median of twelve publications after the PhD, while Spain follows at seven, and France, the Netherlands, and the UK cluster around six. Austria and Germany remain around five. Physical Sciences also displays the most extreme outliers: the boxplots reveal individuals exceeding several hundred publications, and in some countries, more than a thousand. These high-performing researchers strongly influence the mean values, which in Italy and Spain are several times larger than the medians.

Patterns in Life Sciences (Figure D3) mirror those in Medicine, albeit at slightly lower levels of productivity. During the PhD, the Netherlands clearly stands out with a median of five publications, while Italy and Spain follow at around four. Austria, Germany, and France show medians of three, and the UK is slightly lower. After the PhD, publication intensity increases across the board, but the ranking remains similar: Italy reaches a median of seven publications, followed by Spain and France at around six, and the Netherlands, Austria, and the UK at around five to six. Germany again records the lowest typical values, at around four. The distributions reveal substantial heterogeneity in Italy, Spain, and the Netherlands, where interquartile ranges are wide and upper-tail outliers are frequent.

The Engineering field (Figure D4) presents another clear case of marked differences across countries. Spain and Italy exhibit the highest levels of publication activity during the PhD, with medians around five publications, closely matched by the Netherlands. Austria follows with a median of four, while Germany and the UK record medians of three, and France around two. The postdoctoral phase amplifies these contrasts: Spanish and Italian graduates reach medians of eight publications, while the Netherlands and the UK produce around five, Austria four, and Germany and France around three. Engineering shows some of the widest distributions among the fields studied, especially in Italy and Spain, where the upper quartile and extreme values rise sharply.

In Mathematics & Computer Science, differences are somewhat less pronounced during the PhD but become clearer afterwards. Austria, Spain, and the Netherlands show similar median publication levels during the PhD, around four publications, while

Germany and Italy follow with medians of three. France and the UK remain lower, with medians of about two publications during the doctoral phase.

Table D3: Publications per doctoral graduate – Life Sciences

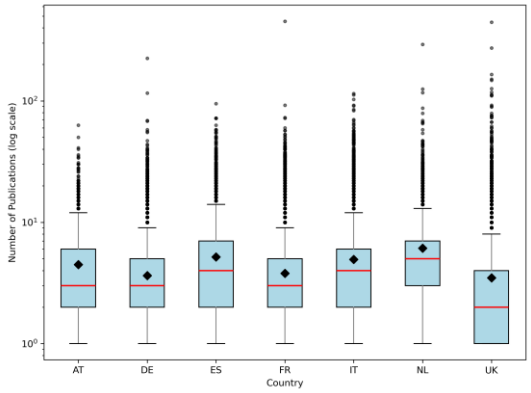
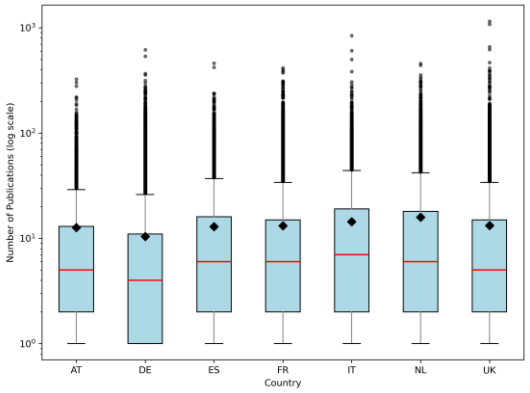
During PhD	*After PhD*
	

Table D4: Publications per doctoral graduate – Engineering

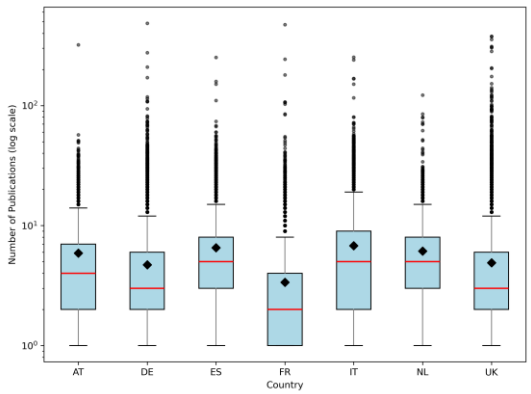
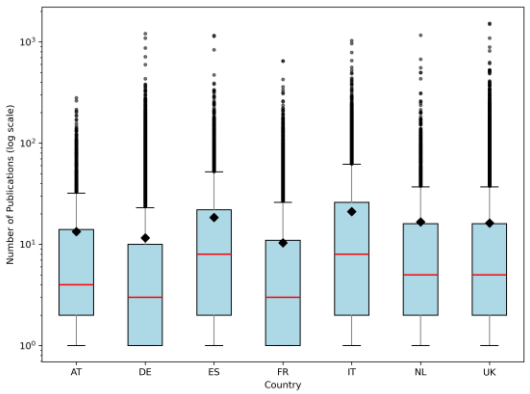
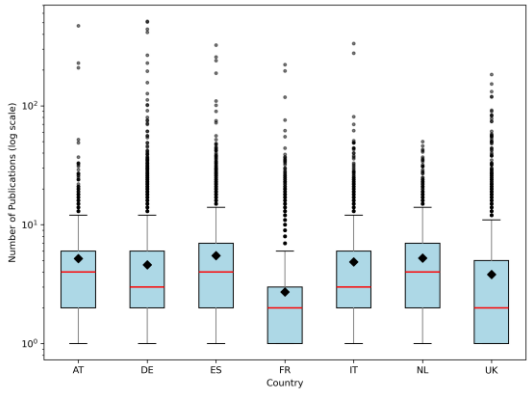
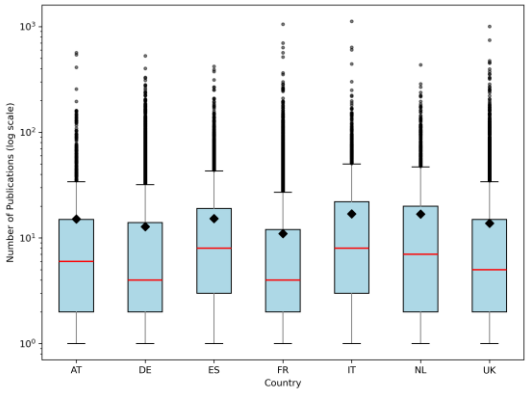
During PhD	*After PhD*
	

Table D5: Publications per doctoral graduate – Mathematics and Computer Science

During PhD	*After PhD*
	

After the PhD, Spain and Italy reach the highest typical levels, with medians of eight publications, followed by the Netherlands at seven and Austria at six. The UK records a median of five, while Germany and France remain at about four. As in the other fields, Spain and Italy exhibit particularly long right tails and wide interquartile ranges.

Across all disciplines, two overarching findings stand out. First, the country ranking is highly consistent: Italy, Spain, and the Netherlands almost always lead; Austria and the UK generally occupy middle positions; and Germany and France tend to show lower publication output during and after the PhD. Second, these differences begin to emerge already during the doctoral period, and the postdoctoral stage tends to widen them rather than narrow the gap.