RESEARCH ARTICLE

A supervised machine learning approach to trace doctorate recipients' employment trajectories

Dominik P. Heinisch[1], Johannes Koenig[1,2], and Anne Otto[2]

[1]University of Kassel, Institute of Economics and INCHER-Kassel (Germany)
[2]Institute of Employment Research (IAB) Rhineland-Palatinate-Saarland (Germany)

Keywords: PhD, employment biographies, administrative data, record linkage, supervised machine learning

ABSTRACT

Only scarce information is available on doctorate recipients' career outcomes (BuWiN, 2013). With the current information base, graduate students cannot make an informed decision on whether to start a doctorate or not (Benderly, 2018; Blank et al., 2017). However, administrative labor market data, which could provide the necessary information, are incomplete in this respect. In this paper, we describe the record linkage of two data sets to close this information gap: data on doctorate recipients collected in the catalog of the German National Library (DNB), and the German labor market biographies (IEB) from the German Institute of Employment Research. We use a machine learning-based methodology, which (a) improves the record linkage of data sets without unique identifiers, and (b) evaluates the quality of the record linkage. The machine learning algorithms are trained on a synthetic training and evaluation data set. In an exemplary analysis, we compare the evolution of the employment status of female and male doctorate recipients in Germany.

1. RECORD LINKAGE OF INTEGRATED EMPLOYMENT BIOGRAPHY DATA

In recent years, the availability of comprehensive new administrative data sets on individual labor market biographies has enabled numerous studies in economics and other social sciences covering a wide range of labor market topics. However, administrative labor market records comprise a limited set of variables, thus narrowing the scope of potential research questions that can be addressed. Only scarce information is available about the career outcomes of doctorate recipients in Germany (BuWiN, 2013). This holds particularly for those doctorate recipients who pursue careers in the nonacademic sector. Knowing more about their labor market biographies is not only important for universities and policymakers. Without knowledge about potential career outcomes, students cannot make an informed decision on whether to start doctoral training or leave academia (Benderly, 2018; Blank et al., 2017).

The objective of the IAB-INCHER project of earned doctorates (IIPED) is to construct a comprehensive data set on labor market biographies of German doctorate recipients. The Integrated Employment Biographies (IEB) of the Institute for Employment Research (IAB) cover labor market records of about 80% of the German workforce. They comprise detailed individual-level information on sociodemographic characteristics, qualification levels, and job characteristics. However, there is no information about earned doctoral degrees. The catalog of the German National Library (DNB) provides this information. The DNB covers almost all German universities' doctorate recipients from 1970 to today. The DNB only provides

an open access journal

Check for updates

Citation: Heinisch, D. P., Koenig, J., & Otto, A. (2020). A supervised machine learning approach to trace doctorate recipients' employment trajectories. *Quantitative Science Studies*, *1*(1), 94–116. https://doi.org/10.1162/qss_a_00001

DOI: https://doi.org/10.1162/qss_a_00001

Received: 12 April 2019
Accepted: 3 August 2019

Corresponding Author: Johannes Koenig
Koenig@uni-kassel.de

Handling Editor: Ludo Waltman

The MIT Press

sufficient information for conventional record linkage (e.g., exact dates of birth) for a minority of individuals. To be able to link both data sets on a large scale, we apply a record linkage procedure that utilizes supervised machine learning algorithms, which are trained on a synthetic training and evaluation data set.

Numerous prior studies have used record linkage methods (Schnell, 2013) to supplement administrative labor market data. In many cases, the record linkage could be based on unique identifiers available in both data sets (e.g., name–surname combination, exact birth date, sex). If identifiers are incomplete or not fully reliable, more advanced "Merge Toolboxes" are available, which utilize string-comparison functions to calculate similarities between key words (e.g., employer's name) in both data sets (Schnell et al., 2004). Even if conventional approaches are able to successfully link two data sets, a proper evaluation of the linked data set's quality (in terms of recall and precision) would be advisable, rather than only reporting the number of final matched entities. Multiple matches between entries are another problem that our approach is able to take into account.

To overcome the limitations of existing record linkage methods, we develop and assess a set of supervised machine learning algorithms. This approach has several advantages: First, it is not restricted to data with high-quality identifiers. Second, the quality of the linked data set is assessable and comparable across different algorithms, as well as to conventional record linkage approaches. Third, our approach is applicable under strict data security requirements and ensures the rigorous anonymity of individual records, which are indispensable requirements in any use of social security data in Germany. Fourth, we utilize a synthetic training and evaluation data set, which allows us to evaluate the quality of the record linkage in the absence of external training and evaluation data.

Even though unique identifiers are absent in both data sets, the final linked data set meets high quality standards in terms of precision and recall. All tested supervised machine learning algorithms outperform heuristic (rule-based) approaches. Achieving a high recall rate not only allows researchers to address questions requiring larger and more complete samples, it also enables differentiation among subgroups. In addition, as the algorithm uses multiple features to predict true-positive matches, it is less likely to introduce bias into the sample. While the synthetic test and evaluation data set might by itself act as a source of bias, we do not find any distortions on observables. Depending on the parameter settings, the quality of the linked data sets can vary for each algorithm, which highlights the necessity of independent training and test data for selecting the best parameter specifications.

The obtained linked data set allows us to investigate the labor market trajectories of German doctorate recipients from 1975 to 2015 before, during, and after their graduation. As a practical application, we use the final data set to analyze the employment status of doctorate recipients at different points of time in their careers. In particular, we analyze gender-specific differences in the share of full-time and part-time employment during doctorate recipients' careers. We find that few doctorate recipients are unemployed after graduation. However, a substantial share of female doctorate recipients work part time. While female and male doctorate recipients show similar employment patterns during their graduation period, the share of part-time and full-time employed women diverges after that.

Our study is not solely limited to Germany. From a methodological point, the introduced method could be applied by further studies to improve the quality of record linkage approaches for the combination of micro data sets. From an empirical standpoint, Germany is one of the biggest "producers" of doctorate recipients among the OECD countries (OECD, 2018) and a huge labor market with a great variety of job positions for graduates.

Investigating the career trajectories of doctorate recipients in Germany contributes to increasing the required transparency for graduates' potential career outcomes in the academic and private sector. This evidence can thus also help students in other countries to make better informed decisions for the planning of their further careers.

The paper is structured as follows: In Section 2, the data sets of the record linkage approach are described. Section 3 presents the supervised machine learning algorithms in detail, as well as their implementation, and evaluates the different approaches we tested. In Section 4, the linked data set is used to investigate the employment status of doctorate recipients over time. In Section 5, we discuss some limitations of the proposed approach and draw implications for further research. Section 6 concludes.

2. DATA SOURCES

In this section, we introduce the two data sets that are integrated by record linkage: the Integrated Employment Biographies (German: Integrierte Erwerbsbiographien [IEB]) and the data set of doctorate recipients from the German National Library (Deutsche Nationalbibliothek [DNB]). Both data sets provide a nearly complete picture of the corresponding populations: The German workforce (subject to social security payments) is represented in the IEB and doctorate recipients who graduated from German universities are represented in the DNB. As a result, the DNB data provide a suitable supplement for the IEB, where information about tertiary education is incomplete. Both data sets are collected by public institutions following standardized procedures and regularities in the data preparation process, which makes them highly reliable and suitable for research purposes. While the DNB data have only been merged via record linkage with publication data (Heinisch & Buenstorf, 2018), the IEB data have been merged via record linkage with a number of external micro databases in the past (e.g., Antoni & Seth, 2012; Dorner et al., 2014; Wydra-Somaggio, 2015; Teichert et al., 2018).

2.1. Doctorate Recipients Data of the German National Library (DNB)

The DNB catalog covers almost the entire population of individuals who completed doctoral training at German universities—doctorate recipients—encompassing about one million authors of dissertations.[1] Two peculiarities lead to this. First, all German publications (published in Germany or by Germans) are held by the German National Library, which is "entrusted with the task of collecting, permanently archiving, bibliographically classifying and making available to the general public all German and German-language publications from 1913" (DNB, 2018). According to §§14 to 16 of the Act on the German National Library, media works are to be delivered to the library if a holder of the original distribution right has their registered office, a permanent establishment, or the main place of residence in Germany. Second, in Germany, doctoral students are obliged to publish their thesis in order to be awarded a doctorate from a German university, and the German National Library tracks thesis publications.[2]

Within the catalog of the German National Libraray, a separate note provides additional information on the type of publication, the year of submission, and the corresponding university name. Since data are selected by librarians for the purpose of archiving and classifying

[1] The German National Library makes its data accessible under the Creative Commons Zero license (CCO 1.0).

[2] The DNB data set has been used for various analyses. For example, Buenstorf and Geissler (2014) studied advisor effects based on laser-related dissertations, and Heinisch and Buenstorf (2018) identified the doctoral advisors of doctorate recipients. Both studies confirm the high reliability and completeness of the DNB data.

Table 1. Illustration of the DNB data

dnb_id	name	surname	birth_year	gender	nationality	uni_name	publication_year	subject
87640472	Marta	Musterfrau	NA	female	German	Kiel	2010	Economics
12342124	Max	Maulwurf	1979	male	German	Jena	2008	Medicine
07986678	Martin	Mustermann	NA	male	Italian	Kassel	1993	Engineering

Note: The table provides fictitious examples of the DNB data set.

these publications, bibliographic information is documented with an overall high degree of accuracy.[3] The coverage is almost complete for all years and disciplines.

From 1995 to 1997 onwards, the DNB created the *Personennormdatei*, a data set comprising all authors as separated entities. This additional catalog improves the information available on authors. Beginning in 1997, the year of birth is recorded for the majority of authors in the data set, as well as additional information on authors' nationality. However, most of these variables cannot be used as identifiers (variables) for the linkage procedure, because the coverage rates vary strongly over time. A stylized example of the DNB data is provided in Table 1.

2.2. Integrated Employment Biographies (IEB)

The IEB unites data from five different historic data sources, each capturing a different segment of the German social security system.[4] It contains detailed information on all individuals who are liable to social insurance contributions in Germany (i.e., employees, unemployed individuals, job seekers, recipients of social benefits and participants in active labor market programs). Civil servants, self-employed, family workers, and doctoral candidates financed solely by scholarships etc. are not part of the social security system and therefore not reported in the IEB. Taken together, the data cover approximately 80% of the German workforce.

The IEB data comprise the starting and ending dates of all spells (i.e., periods of unemployment, benefit receipt, employment) for each individual (vom Berge et al., 2013). Additionally, for each individual a range of sociodemographic characteristics is documented (e.g., sex, date of birth, nationality, qualification level, job features [type of employment, occupation, industry affiliation, region of workplace]). While, although incomplete, information of vocational training certificates obtained, or bachelor's and master's degrees, is part of the IEB, no information on doctoral degrees exists. Information is available on a daily basis from 1975 to the most current

[3] Nevertheless, some effort was necessary to clean the data. The names of the individuals were standardized. For example, name information was coded in UTF-8 and separated by commas into first and last names. Further, all variants of misspelled university names were checked manually and assigned to the corresponding institution. Year information in the database was corrected for nonplausible cases. Electronic resources were also added to the database. In recent years, some dissertations have been included exclusively as electronic resources. However, many electronic doctorate theses are also listed as a physical book. Further, different versions of the same work are possible, such as university deposit copies and commercial publisher editions, with possible later new editions. The database was cleaned for these duplicates, which were identified by two different approaches. If a reference is made in the DNB's title holdings to an identical publication other than the original publication, these publication are considered identical. However, an explicit reference to identical publications is not given for all double-listed publications. Therefore, duplicates were detected based on title information. Titles and subtitles were standardized (i.e., punctuation marks, upper and lower case, spaces, etc. were removed) and cleaned (i.e., names and other nontitle information were removed) and a fuzzy string comparison was used to take care of small variations. Further, we excluded all authors with incomplete name information (e.g., entries with missing first name or surname). See also Heinisch and Buenstorf (2018) for further details.

[4] These five data sources are the Employee History, Benefit Recipient History, Unemployment Benefit II Recipient History, Participants-in-Measures History, and the Jobseeker History.

Table 2. Illustration of the IEB data

iab_id	employment	begin_date	end_date	place_work	school_degree	apprenticeship	class_econ_activity
92240472	Mini-job	01/01/1996	31/12/1996	Kiel	A level	No qualification	49.32 Taxi operation
92240472	Part-time	01/01/1997	31/12/1997	Kiel	A level	University degree	85.42 Tertiary education
92240472	Part-time	01/01/1998	31/12/1998	Kiel	A level	University degree	85.42 Tertiary education
92240472	Unemployed	01/01/1999	31/01/1999	Kiel	A level	University degree	
92240472	Full-time	01/02/1999	31/12/1999	Berlin	A level	University degree	72.11 Research and experimental development on biotechnology
92240472	Full-time	01/01/2000	31/12/2000	Berlin	A level	University degree	72.11 Research and experimental development on biotechnology
32134444	Mini-job	01/06/2003	31/08/2003	Buxtehude	No qualification	No qualification	55.20 Holiday and other short-stay accommodation
32134444	Mini-job	01/07/2004	31/09/2004	Jena	Primary School	No qualification	55.10 Hotels and similar accommodation
32134444	Part-time	01/01/2007	31/12/2007	Jena	A level	University degree	86.10 Hospital activities
32134444	Full-time	01/01/2008	31/12/2008	Halle	A level	University degree	86.10 Hospital activities
20347523	Part-time	01/08/1980	31/12/1980	Frankfurt	Primary School	Vocational training	4.11 Central banking
20347523	Full-time	01/01/1981	31/12/1981	Frankfurt	Primary School	Vocational training	66.11 Administration of financial markets

Note: The table provides fictitious examples of the IEB data set.

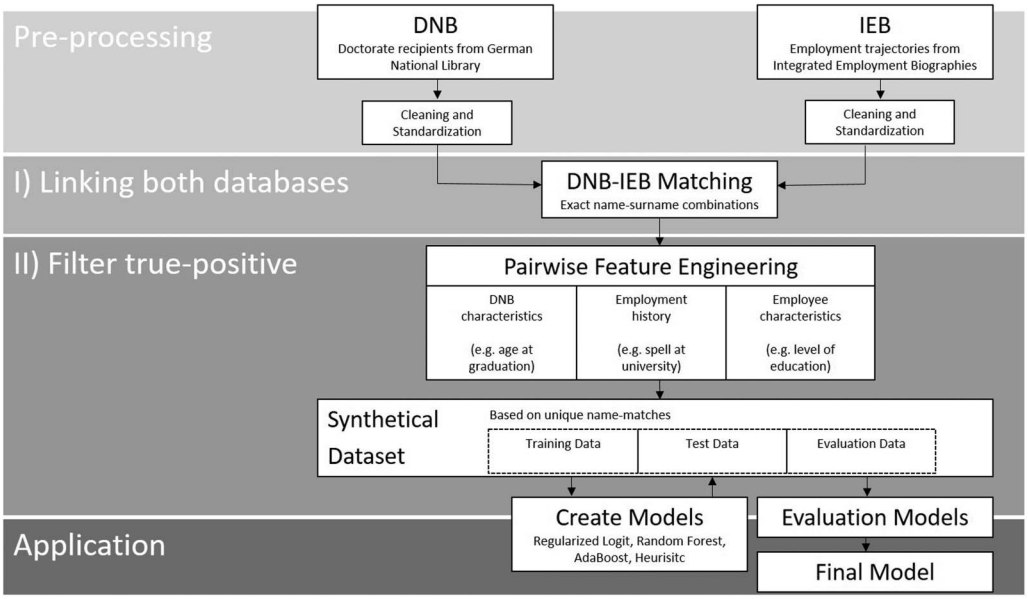

Figure 1. Overview of the data processing and record linkage procedure.

year for West Germany, and from 1993 for East Germany. Hence, the IEB enables labor market biographies of individuals in the public and the private sector to be tracked over time.

The IEB data are highly reliable for all variables that are directly relevant for social insurance contributions. However, some information in the data, such as information on secondary schooling, is less reliable, as it is transmitted by the employer solely for statistical purposes (Fitzenberger et al., 2005). Furthermore, some variables contain missing values, which vary over time (e.g., Antoni et al., 2016). Confidential information that would make individuals identifiable (e.g., name and address) is not accessible for researchers (Schnell, 2013). An anonymized system-independent individual identifier links social security registers and administrative data of the Federal Employment Agency (Dorner et al., 2014).[5] Table 2 shows a fictitious example of the preprocessed IEB data.

3. CLASSIFYING DOCTORATE RECIPIENTS IN THE GERMAN LABOR MARKET DATA

3.1. Problem Description

In this section, we first describe the general record linkage problem, and then expand on it in terms of its applicability to social security data, where researchers have to deal with large volumes of highly sensitive data. The record linkage procedure aims at identifying as many entries in both data sets that belong to the same entity. This target function is optimized under the constraint of keeping the number of incorrect matched entries as low as possible. To achieve this target, a two-step procedure is applied: First, entries in both data sets are matched by using an imperfect identifier (i.e., the names of individuals). Second, falsely matched combinations are eliminated. Figure 1 presents an overview of the record linkage approach described in this section.

[5] The IEB and its scientific use file have been extensively discussed in the past. See, for example, Dorner et al. (2010) for a brief discussion of the IEB, Oberschachtsiek et al. (2009) for a more detailed description of the IEB sample, and Zimmermann et al. (2007) for the scientific use file.

The first step aims to match as many entries as possible of both data sets that might belong to one entity. In other words, in the first step, the data sets are actually linked. This can be achieved, for example, by exact string matching between entries' names, or by calculating distances between the entries' names using a fuzzy string matching algorithm. The second step aims to identify as reliably as possible true linked entities among the matched entry pairs. In other words, in the second step, correctly linked entries that belong to one entity are filtered from incorrectly matched entries. As social security data comprise large volumes of data with many homonyms (in our case the entire German workforce), the filtering of true-positive matched entries is a more serious problem, in particular, as incorrectly spelled names are less frequent in administrative data. Therefore, this paper is primarily focused on improving the second step of the record linkage procedure.

The linked entries of both data sets by a specific identifier will result in 0-to-*n* possible combinations of matched entries, of which 0-to-1 combinations truly belong to one entity. In those cases, where multiple entries match into one entity, many-to-many (*n*-to-*m*) matched entries occur. Identifying the true matched entities in a set of *n*-to-*m* matched entries can be described as a classification problem. The following description of the classification problem is based on Gareth et al. (2013) and Bishop (2006). Formally, the classification task is to find a function $f(X)$ that correctly classifies two matched entries of both data sets as one entity. With a quantitative response variable $Y \in c(Same, Different)$ and using a set of p different predictors,

$$X = (X_1, X_2, \ldots, X_p)$$

$$Y = f(X) + \epsilon,$$

where ϵ is the error term.

In practice, there are numerous restrictions that complicate the estimation of the classification function f: Unique entity identifiers (or keys) and reliable predictors such as combinations of name, birthday, and birthplace may be lacking. Even if the available data are generally of high quality, information may be imprecise, misreported, or incomplete for individual entries. Even in cases where reliable predictors exist, privacy requirements may restrict the number of predictor variables X that are accessible to researchers.

If the reliability of a single or multiple predictors cannot be ascertained, or if only a set of weak predictors is available, machine learning algorithms can improve the record linkage quality. Machine learning algorithms have been applied to a number of record linkage problems and several solutions are available (e.g., Christen, 2012b). In this paper, we use machine learning algorithms to solve the classification problem described above in accurately filtering true matched entries. In this case the classification problem can be described as the best combination of available input variables X that predict $\hat{Y}$:

$$\hat{Y} = \hat{f}(X),$$

with $\hat{Y}$ as classification output and $\hat{f}$ as our estimation equation for the classification function f. The accuracy of $\hat{Y}$ depends on two aspects, as the following equation shows: the reducible and irreducible errors:

$$E\left(Y - \hat{Y}\right)^2 = E\left[f(X) + \epsilon - \hat{f}(X)\right]^2 = \left[f(X) - \hat{f}(X)\right]^2 + Var(\epsilon).$$

The reducible error $\left[f(X) - \hat{f}(X)\right]^2$ results from $\hat{f}$ not being a perfect estimation for f. As the name implies, the reducible error can be reduced by more sophisticated statistical learning

methods or by increasing the input variables' X predictive power. In contrast, the irreducible error $Var(\epsilon)$ would persist even if $\hat{f}$ were a perfect approximation of f. The set of input variables X entering into function f cannot predict ϵ by definition, as they result from errors in measuring X. A suitable classification procedure identifies the best functional relation of X in $\hat{f}$ that approximates f, by minimizing the reducible error $\left[f(X) - \hat{f}(X)\right]^2$.

Solving classification problems is a traditional field of application for machine learning techniques. Machine learning algorithms can help to find suitable approximations of the classification function f (Christen, 2012a). However, these approaches have not found much use in research using administrative labor market data. Record linkage procedures used in this context have mostly been based on heuristic approaches. Data are linked by calculating similarities between names (Schnell, 2013) and "rules-based" heuristics (e.g., information on whether two entries originate from the same or different regions). Applying heuristic approaches requires high-quality data. Even then, heuristic approaches do not exploit the full potential of the data, because they do not use the optimal functional form of $\hat{f}$ or the best representation of X.

A wide selection of sophisticated classification algorithms is available to estimate $\hat{f}(X)$. These can broadly be categorized into deterministic, probabilistic, and (machine) learning-based approaches (Christen, 2012b). Higher predictive power can be expected for supervised machine learning techniques. Supervised machine learning algorithms require training data to approximate the best representation of f by a specific representation of the input variables X. A wide variety of machine learning algorithms have been developed, and the choice of specific algorithms involves a trade-off between classification quality and computational demands. In addition, not all algorithms are implemented in the statistical software packages available in the settings where administrative data may be accessed.[6] Reflecting these considerations, our approach utilizes three well-known machine learning algorithms: regularized logistic regressions, AdaBoost, and Random Forests.[7]

For these machine learning algorithms, we do not know the best model specification for our classification problem a priori. Therefore, our machine learning algorithms need to be tuned to discover the parameter setting that results in the most powerful prediction to correctly classify entities. Different ways exist to identify the best tuning parameters. Here, our approach is based on trial and error. A regularized logistic regression estimates a logistic regression model with an additional penalty term to avoid overfitting. This requires ex ante specification of both the penalty parameter and a threshold probability value above which estimated matches are classified as belonging to the same entity. The Random Forest algorithm uses decision trees for classification. By randomly selecting a set of m variables, a specific number of n decision trees is constructed. Each decision tree uses these m variables to split the data set-specific thresholds to classify the data into matches and nonmatches. A sequence of multiple splits divides the data into distinct decision regions. A majority vote over the n decision trees decides on the class of each entry in the matched data set. The number of randomly drawn variables (m) and

[6] The respective administrative data can only be used on secured machines available at IAB. More advanced methods, such as multilayer neuronal networks, are computationally intensive and their application is not technically feasible in our case.

[7] All algorithms used are available as R Packages. We used the programming language R Version: 3.3.2 (R Core Team, 2017) and the following R packages: for AdaBoost the package ada (Culp et al., 2006), for regularized logistic regressions the package glmnet (Friedman et al., 2010), and for Random Forest the package randomForest (Liaw and Wiener, 2002).

the number of trees (n) must be specified ex ante. AdaBoost is a boosting method developed for binary outcome variables. Similar to Random Forests, it is based on decision trees, but the classifiers are trained sequentially. After each iteration, the classification output is weighted by its classification success, giving a higher weight to misclassified matched entries in the next iteration. After converging, all decision trees give a majority vote on the matched entries class. The number of iterations and weights must be set as parameters ex ante.[8]

In our approach, these machine learning algorithms are tested against a heuristic (rule-based) classification. For the heuristic classification approach the number of variables considered in classification needs to be specified ex ante. For the heuristic, we generated all possible combinations of the matching variables used. As a result, we get a number of possible decisions where only one of the possible matching variables, up to all of these matching variables, needs to take the value 1. The heuristic then classifies pairs of entries by comparing one or several variables. For example, one relative restrictive heuristic approach could classify entries as belonging together if they have the same name and surname in both databases, they were employed in the university region 5 years before or after graduation, they had a proper education, and they are working at a university or research institue. A less restrictive approach could link all entries with same name-surname combination and a proper age. The common objective of all these approaches is to develop a function $\hat{f}$ that accurately separates the spaces of same versus different entities in both data sets. Applying different model specifications enables us to select from a range of models with different properties. The aim of this task is to find an optimum between precision and recall; that is, to link as many entries of both data sets as possible (high recall) while minimizing the number of false classification decisions (high precision).

Overfitting is a serious risk when the best algorithm is selected. Overfitting means that the prediction function $\hat{f}$ follows the error term $Var(\epsilon)$, generating estimates for f that are as close as possible to the observed training data, but not allowing accurate estimates for new observations outside the training data. In this case, the trained algorithm is useless, as the trained model is an exact representation of the training data but cannot be generalized to other data. This would fail the task of finding a function $\hat{f}$ that predicts our outcome variable Y as well as possible: $Y \approx \hat{f}(X)$ for any observation.

To overcome overfitting, out-of-bag predictions are used to evaluate the algorithms' classification success. These require an independent data set that has not been used in training the algorithms. The training data are split into several data sets that are specifically used first for training, second for identification of the correct parameters, and third for evaluation. For training and evaluation, data are required for which true outcomes of the quantitative response variable $Y \in c(Same, Different)$ are known to the researcher.

3.2. Preprocessing and Record Linkage

In this section, we discuss the application of the record linkage procedure described in section 3.1 to classify correctly dissertation authors from the DNB data set in the IEB data set.

[8] The regularized logistic regression was estimated with values for the penalty parameter of 0, 0.3, 0.5, 0.7, and 1. For the threshold probability, we selected values ranging from 0.1, 0.2, 0.5, and 0.6 to 0.8. For the Random Forest algorithm, the number of randomly drawn variables (m) was set to 2, 3, and 5. The number of trees (n) was specified as 20, 100, 200, and 500. For AdaBoost, the number of iterations used for estimation were set to 50, 100, 250, and 500 and the weights were set as parameters to 0.01, 0.2, 0.5, 0.9, 1.

3.2.1. Data preprocessing

Even though both data sets are of high quality, several preprocessing steps were required before the actual record linkage (see footnote 3). The cleaned dissertation data set includes 984,359 doctorate recipients. In a second step, the DNB data set is merged with the IEB data. For this step, confidential name-surname information is required, which is both not contained in the anonymized IEB and not accessible for researchers. In the IAB, it is only possible to use this information for data linkage with a reasoned data request and if the data linkage is conducted in a secured technical environment, assuring data protection of the confidential information (Dorner et al., 2014).[9] For this reason, the Data Information Management (DIM) Department of the IAB, which fulfills these technical prerequisites, is working as a data trustee for the data linkage.[10] First, the data linkage was conducted for exact name-surname combinations. Unlike other data sets (e.g., patent data), both data sets are of comparable high quality regarding the spelling of names, including spellings using German umlauts. We therefore used a naïve string-matching algorithm to minimize the number of false-positive matched pairs. With naïve string matching for 876,927 entries, at least one corresponding individual with the same name-surname combination was identified in the DNB data, with 18,787,699 corresponding entries in the IEB. The IEB includes only individuals covered by the German social security system, but not others such as civil servants or students receiving scholarships, which could explain why some names of doctorate recipients do not match with any entry in the IEB (see above).

To ensure data security, each researcher working with the IEB is only allowed to use a restricted sample of the IEB. For this reason, the maximum number of multiple matched entries to individuals was limited to not more than 300 namesakes in the IEB. This excludes doctorate recipients with very common name-surname combinations (e.g., "Werner Müller"). If we had included all matches that exceed the threshold in the matching process, it would have been necessary to use an extraordinarily large sample of the IEB, since some doctorate recipients had up to 73,212 name twins. The final data set is further limited to doctorate recipients who graduated between 1975 and 2015. East German doctorate recipients graduating before 1990 also had to be excluded because reliable IEB employment periods are only available for East Germany beginning in 1993. To save computational power and reduce the number of false-positive matched pairs, we deleted all matched pairs aged below 20 in the year of submission. In Germany individuals usually receive their doctoral degree at the age of 32.5 years. If an entry in the DNB database is connected to a number of entries in the IEB database while some of them are aged below 20 in the year of submission, these entries most certainly do not belong to the same entity. Summing up, the final database contains information about 687,979 doctorate recipients from the DNB and the corresponding 15,468,638 IEB entries.

[9] The IAB as a whole fulfills the legal requirements for data security, as it is a department of the Federal Employment Agency in Germany, which in turn is obliged to ensure data security as a social service provider in accordance with the standards of §78 Social Security Code X.

[10] The DIM Department carried out the record linkage using individual identifiers (e.g., first name, surname) in both data sets, and it alone stores this information. Then, the DIM Department pseudonymized the personal data according to the legal definition of §3 para. 6a Federal Data Protection Act and replaced them with identification numbers. The correspondence tables of this data linkage were only provided to the researchers as anonymized data sets. The subsequent steps of data processing and matching were carried out only based upon this anonymized data. The risk of restoration of the personal reference is countered by administering the confidential personal data, which are required for the identification of the cases, only from the data trustee. In the end, the researcher only has access at IAB to the final anonymized data set for further scientific work. When publishing results, care is taken to ensure that only sufficiently large case numbers that do not allow conclusions to be drawn about individuals are presented.

3.2.2. Generation of synthetic test and training data

Supervised (machine) learning algorithms require training data to approximate the best predictive model. As a result, for training and evaluation of the algorithm, a set of reliable observations is necessary where matched entries belonging to one entity (true-positive matches) can be distinguished from false-positive matched entries (true-negative matches). Several strategies can be applied to identify a "gold standard" sample that can be used to train and evaluate the algorithm (Christen, 2012a). An ideal solution would require surveying a selection of doctorate recipients asking about their realized career paths, or asking them to identify which career trajectory belongs to them among all the matched entries. The responses would provide the "gold standard" data set, which can be generalized to predict other matched entries. However, data security and practical reasons make this infeasible. First, social security data are subject to stringent data privacy requirements. The data are strictly anonymized, and contacting individuals based on their private addresses is restricted as well. Second, even if individuals could be directly asked, mistakes as well as low response rates might reduce the representativeness of the sample obtained. Therefore, we created a synthetic training and evaluation data set from the available data. One important aspect in creating a synthetic training and evaluation data set is its representativeness of the overall (matched) population. It should contain the same variables, which should moreover follow a similar frequency distribution and similar error characteristics. In our approach, we use name-surname combinations, as we believe the frequencies of name-surname combinations are independent of the variables used as classifiers.

For training the algorithm, we need both true-positive matches and true-negative matches. For our synthetic training and evaluation data set, our true-positive matches ($Y \in c(Same)$) are based on unique name-surname combinations. These are doctorate recipients whose name-surname combination appears only once in both databases: the Integrated Employment Biography Data and the data set received from the catalog of the German National Library. Since both data sets cover the underlying populations almost completely, these matched entries are expected to belong to the same entity.[11] For this approach, it is of only limited importance that the IEB data only contain information for individuals that are liable to social insurance contributions in Germany. Since it is expected that during their employment trajectories the overwhelming majority of people are captured at least once in one of the different segments of the German social security system, potential pairs are collected from the almost complete underlying population. As a result, entries that are linked based on name-surname combination in both databases and where exactly one-to-one possible name-surname combination occurs, can be expected to be very likely to belong to the same entity.

For our true-negative matches, these uniqe DNB entries were merged with a random set of entries from the IEB data set. As the name of an individual is highly gender dependent, we limit the randomly matched sample to entries with the same name but different surname. This leads to a sample where individuals were linked on same surname but different name. This procedure leads to a large number of wrongly matched entries. To specify a representative number of true-negative matched entries, we follow the overall distribution of matched entries and randomly draw a similar number of matched entries for each wrongly matched DNB entry.

[11] We performed a number of plausibility checks, which provided support to our conjecture. For example on an aggregated level, we investigated the career paths of this unique name-surname combinations for different subjects, gender, and years and compared their career paths to known career paths of doctorate recipients from previous studies (e.g., BuWiN, 2017). The identified career trajectories indicate plausibility of these matches on an aggregated level.

Table 3. Variables for machine learning

Name	Description	Source
spell_research	Dummy, value 1 if individual has/had a spell at a university or research institute. European statistical classification for economic activities was used. Values were extended by record linkage for research institutions and universities.	IEB
spell_hospital	Dummy, value 1 if individual has a spell in a hospital/medical practice. European statistical classification for economic activities was used.	IEB
prop_educ	Dummy, value 1 if education of individual belongs to university entrance qualification.	IEB
age_sub	Continuous, age in submission year.	IEB/DNB
right_age	Dummy, value 1 if individual is between 25 and 40 years old in submission year. Used for heuristic approach instead of age_sub.	IEB/DNB
same_ror_y5	Dummy, value 1 if individual was employed in university region 5 years before/after graduation.	IEB/DNB
first_spell_before	Continuous, first year in IEB subtracted from year of submission.	IEB/DNB
right_first_spell_before	Dummy, value 1 if first_spell_before is between −10 and 5. Used for heuristic approach instead of first_spell_before.	IEB/DNB
year_diss	Continuous, year of submission.	DNB
eastern	Dummy, value 1 if individual graduated in new federal states.	DNB
social science	Dummy, value 1 if individual graduated in social science.	DNB
natural science	Dummy, value 1 if individual graduated in natural science.	DNB
engineering	Dummy, value 1 if individual graduated in engineering.	DNB
medicine	Dummy, value 1 if individual graduated in medicine.	DNB
law/economics	Dummy, value 1 if individual graduated in economics/business studies/law.	DNB
nbr	Continuous, number of common namesakes in IEB Data.	IEB

Using this strategy, we obtain a synthetic training and evaluation data set, for which the true matching status is known and which is representative of the overall matched population.[12]

3.2.3. Classification variables

Three types of variables are created that are used as classifications. The first set of variables contains information on entries in the IEB data set (e.g., an employment spell at a university); the second one contains information on entries in the DNB data set (e.g., the year of submission), and the third one contains information calculated from both data sets (e.g., the lag between dissertation submission and the first employment spell). Table 3 gives an overview of the classification variables X, which are used to predict $\hat{Y}$. In Table 4, a stylized sample illustrates the final data set. Tables A1, A2, A3 (in Appendix A) provide descriptive statistics for an

[12] The creation of the artificial training and evaluation database was technically executed by the DIM Department of the IAB, which was working as a data trustee. See also footnote 10.

Table 4. Illustration of DNB-IAB record linkage

dnb_id	iab_id	spell_research	spell_hospital	prop_educ	age_sub	same_ror_y5	first_spell_before	year_diss	eastern	social s.	natural s.	engineering	law/economics	medicine	nbr
12342124	92240472	1	0	1	40	0	−11	2007	1	0	0	0	0	1	3
12342124	32134444	0	1	1	29	1	−5	2007	1	0	0	0	0	1	3
12342124	20347523	0	0	0	45	0	−27	2007	1	0	0	0	0	1	3
87640472	08898092	0	0	0	66	0	5	2010	0	0	0	0	1	0	2
87640472	90980983	1	0	1	31	1	−10	2010	0	0	0	0	1	0	2

Note: The table shows the stylized IAB-DNB linkage in fictitious examples.

Table 5. Descriptive statistics for the classification variables in the synthetic training and evaluation data separated for true-negative and true-positive

Variable	Same	Median	Mean	Min	Max
spell_research	1	1	0.6368	0	1
spell_research	0	0	0.0657	0	1
spell_hospital	1	0	0.3745	0	1
spell_hospital	0	0	0.1008	0	1
prop_educ	1	1	0.9507	0	1
prop_educ	0	0	0.3238	0	1
age_sub	1	31	32.5199	20	91
age_sub	0	36	37.8844	20	102
right_age	1	1	0.8996	0	1
right_age	0	0	0.4546	0	1
same_ror_y5	1	1	0.7297	0	1
same_ror_y5	0	0	0.0156	0	1
first_spell_before	1	−6	−6.9672	−40	37
first_spell_before	0	−11	−11.4541	−45	39
right_first_spell_before	1	1	0.7112	0	1
right_first_spell_before	0	0	0.4242	0	1

Note: Descriptive statistics on the distribution of features used to classify true-positive matched entries in the IEB and DNB data in the synthetic training and evaluation data set. The data are split into two samples: true-positive matches based on unique name-surname combinations and true-negative matches based on entries with the same name, but different surname. The true-positive matches are indicated by "Same" = 1.

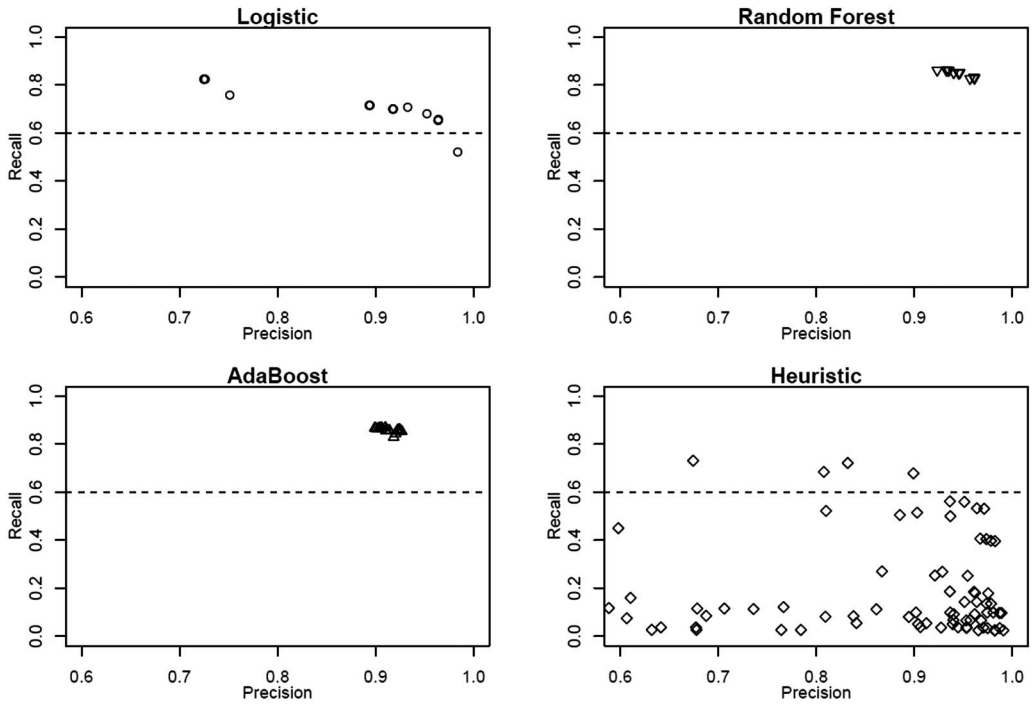

Figure 2. Recall-precision plots for estimated algorithms under different tuning parameters.

assessment of the representativeness of the synthetic training data set and the full (matched) population.

Table 5 reports the general descriptive statistics for the classification variables separately for the true-positive and true-negative matched entries in the synthetic training and evaluation data set. For example, about 63.68% of the individuals in the true-positive sample had one employment spell at a university or other research institution (spell_research), as compared to 6.57% of the individuals in the true-negative sample, indicating a high predictive power for the spell_research variable. This synthetic training and evaluation data set contains some 50,000 matched doctorate recipients with up to 300 potential matched IEB entries. We divided this data set into two equal parts: a training data set and an evaluation data set. A block randomization was applied to divide the data set into the two subsets. Block randomization is a technique that reduces bias and balances the allocation of individuals into different subsets. This increases the probability that each subset contains an equal number of multiple matched entries.

Table 6. Classification results – best parameter settings (on training data set)

	+1 (best parameter)				+1 (min recall 0.6)			
Model	Precision	Recall	F1	Accuracy	Precision	Recall	F1	Accuracy
Logistic	0.9328	0.7099	0.8062	0.9860	0.9644	0.6558	0.7807	0.9848
Random Forest	0.9457	0.8520	0.8964	0.9919	0.9616	0.8287	0.8902	0.9916
AdaBoost	0.9246	0.8602	0.8912	0.9914	0.9268	0.8534	0.8886	0.9912
Heuristic	0.8991	0.6786	0.7734	0.9826	0.8991	0.6786	0.7734	0.9826

Table 7. Evaluation of the classification results – best parameter settings

	+1 (best parameter)			
Model	Precision	Recall	F1	Accuracy
Logistic	0.9410	0.7018	0.8040	0.9847
Random Forest	0.9584	0.8337	0.8917	0.9910
AdaBoost	0.9196	0.8605	0.8891	0.9904
Heuristic	0.9110	0.6742	0.7749	0.9825

3.2.4. Model selection and evaluation

For model selection, each classification algorithm was trained and tested for various parameter specifications. Algorithms were trained on three quarters of the training data set and evaluated (by recall and precision) on the remaining quarter. The results are shown in Figure 2, which shows the recall-precision curve separately for alternative classification algorithms and model specifications. Table 6 shows the best training results for our evaluation measures.

All algorithms achieve satisfactory classification results and would generally be applicable. The heuristic approach also achieves sufficiently high values in terms of precision. In some specifications it outperforms most of the more advanced and computationally demanding algorithms.[13] However, the more demanding algorithms outperform the heuristic approach in that they reach comparable rates of precision but achieve substantially higher recall. Depending on the parameter settings, the classification success of the specific algorithms varies substantially (e.g., results for the logit model vary from a recall/precision of 0.5683/0.8805 to 0.9840/0.5219). This illustrates the advantage of using a supervised learning approach, as it allows the evaluation of the record linkage quality not only by how many individuals are linked, but also by the achieved quality of linked entities.

We next selected those specifications of the algorithms that achieved the highest average values in recall and precision and those with the highest precision and a recall of at least 0.6. For the evaluation, we took the best parametrized models and trained them again on the full training data set. Then we evaluated the trained models on the evaluation data set. Table 7 shows the further evaluation results. All models show qualitatively similar results. The Random Forest algorithm outperforms the other algorithms. The best performing algorithm was then used to classify true-positive matched entries in the full (matched) data set.

Based on the approach outlined above, the Random Forest algorithm identifies 552,459 individuals as $\hat{Y} = c(Same)$. If the Random Forest algorithm identifies more than one entry in the IEB that matches one entry in the DNB (or vice versa), then we decided to exclude respective cases from the final data set. Hence, the final data set for the IAB-INCHER project of earned doctorates (IIPED) consists of a total of 447,606 doctorate recipients, and the overall matching quota amounts to 45.47%.

[13] For example, one heuristic classified matched entries as belonging to the same entity if a matched IEB entry had a spell in a hospital/doctor's office, a spell at a university/research institute, one spell in the university region at least −5/5 years before/after submission, is aged between 25 and 40 at submission, and has a labor market entry at least 10 years before or at least 5 years after submission. This heuristic reached a precision of 0.9889. However, while being very precise, the heuristic is only able to link a very selective sample of doctorate recipients with the IEB data set, with a recall of 0.0962.

Table 8. Additional quality assessment

	Same value in IEB and DNB data	Different value in IEB and DNB data
year of birth	95.33%	4.67%
gender	99.08%	0.92%

4. APPLICATION

In this section, we evaluate data from the IAB-INCHER project of earned doctorates (IIPED) in two ways. First, we assess how representative the linked data set is of the total population of doctorate recipients in Germany. Second, we present an exemplary analysis of the employment status of female and male doctorate recipients over time. This example is used to check whether the empirical results obtained with the linked data set are consistent with existing empirical evidence. In doing so, we explore whether the data can be used to analyze research questions related to the labor market biographies of doctorate recipients in Germany.

4.1. The Labor Market Sample of Doctorate Recipients

Figure B1 depicts the share of linked doctorate recipients in the total population of doctorate recipients over time. This share increases strongly from 34.51% in the starting year 1975 to 61.70% in 2015. For doctorate recipients in the period before and after German reunification, the matching quota lies at 39.61% and 57.43% respectively. At 33.08%, the share of female doctorate recipients in the merged database is comparable to the 33.51% share in the population of doctorate recipients received from the DNB. Reliable information on domestic and foreign doctorate recipients is available for selected years in the DNB catalog. In 2013, the share of domestic doctorate recipients in the DNB was 85.37%, while the respective share in the merged database is 87.62%, indicating that domestic-born doctorate recipients are slightly overrepresented. Figure B2 illustrates the average shares of merged doctorate recipients by discipline over the entire observation period. Overall, average matching rates vary across fields, with values ranging from 42.81% for sports to 60.88% for sciences and mathematics. As additional evidence of matching quality, we compared variables in both data sets (IEB and DNB) that were not employed in the matching procedure. Table 8 depicts the consistency of linked entries for year of birth and gender, which were both not used as classification variables because of limited coverage in the DNB data set. Both variables indicate high accuracy for our record linkage procedure on an aggregated level. Nevertheless, in some cases the identified linked entries were not correctly matched.

4.2. The Employment Status of Doctorate Recipients

We now investigate how the employment status of doctorate recipients changes before, during, and after their doctoral studies. We differentiate among five types of employment status: full-time job, part-time job, mini-job,[14] vocational training, and unemployment. Figure 3 shows the employment status of all linked doctorate recipients in the final data set at different points in time throughout their careers. As the exact date of graduation is unknown, our point of reference (year 0) is the final day of the year in which the dissertation was published. Most doctorate recipients hold full- or part-time positions, with only small shares of graduates being unemployed, in vocational training, or holding mini-jobs at any point in time. Doctoral

[14] The monthly income in a mini-job does not exceed € 450, and the number of working hours is limited to 15 per week.

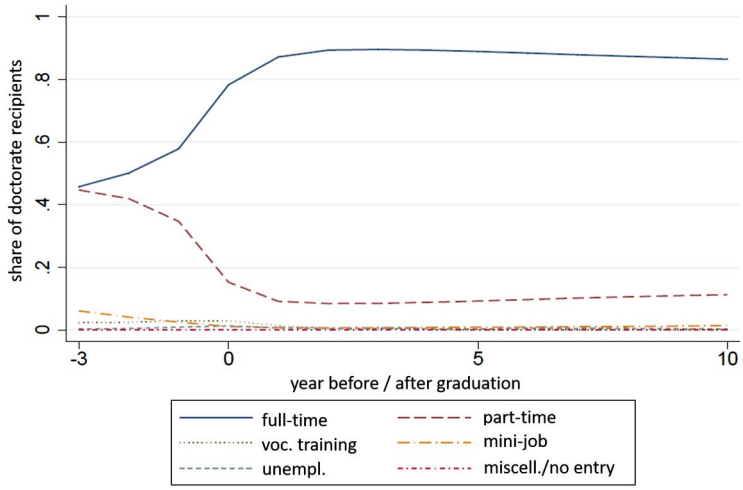

Figure 3. Employment status over time before/after graduation.

students are often employed in part-time positions at universities or public research organizations. The shares of part-time employment range between 44.71% and 34.70% for 3 years to 1 year before graduation, whereas postsubmission employment changes from part-time to full-time positions in academia, other parts of the public sector, or the private sector.

The share of full-time jobs increases from 78.29% in year 0 to a maximum of 89.59% 3 years later, and then diminishes to 86.46% in year 10 after graduation. In turn, the share of part-time employment increases from 8.50% 3 years after to 11.28% 10 years after graduation. This change can be explained by male and female doctorate recipients following different career patterns over time (see Figure 4).[15] While the majority of male graduates constantly work full-time after their doctorate education, a larger share of women also have a part-time position after graduation. This gender-specific full-time gap increases over time. While 94.34% of men are full-time employed 10 years after graduation, the corresponding share among female doctorate recipients declines to 62.51% after 10 years. These results are in line with existing evidence on gender-specific employment patterns, where female part-time employment is often attributed to an uneven distribution of family-related responsibilities, such as childcare and care of elderly family members among men and women (Wanger, 2015). These results clearly demonstrate that the data from IIPED is representative of the overall doctorate recipient population entering the German labor market, particularly in more recent cohorts. The exemplary analysis of doctorate recipients' employment status over time is in line with previous findings. This data set can therefore be employed to study a wide range of research questions related to the postdoctoral careers of doctorate recipients.

5. LIMITATIONS

As shown above, machine learning provides a suitable approach to overcome the limitations of traditional record linkage methods. However, machine learning comes with limitations of its

[15] For the analysis, we used a sample of the fully linked data set, but we imposed some restrictions on the data. Since data were collected for administrative purposes, we had to correct some spell information in the data (Kaul et al., 2016) to construct the sample for the subsequent analysis. Further, we dropped unreliable very short (un-) employment episodes (below seven days). For the analysis, we use information on all graduates at the end of a given year (December 31) for 3 years prior to and 10 years after the publication year of the dissertation.

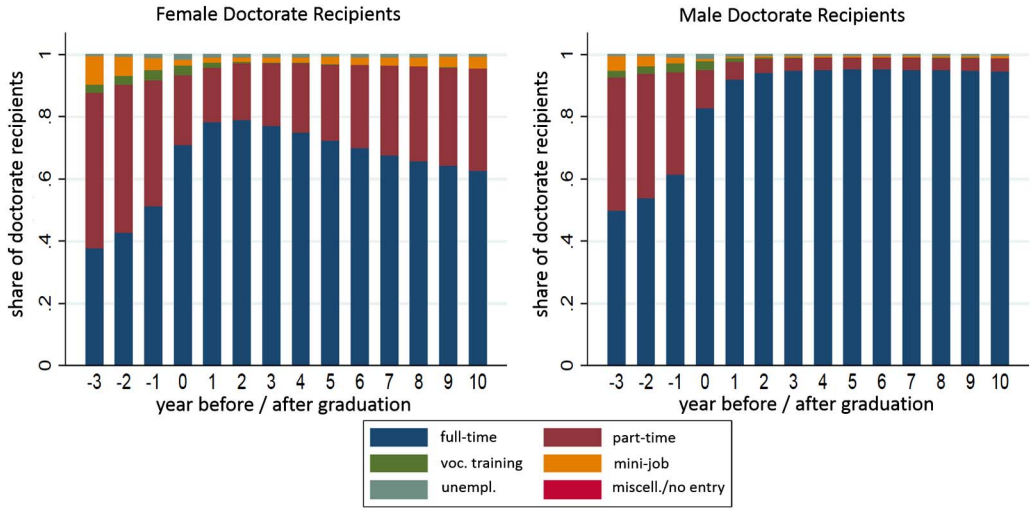

Figure 4. Employment status over time before and after graduation, separately for male and female doctorate recipients.

own, which are the focus of this section. Most importantly, as noted above, the linkage is based on a synthetic training and evaluation data set. Here, unique name-surname combinations were merged with individuals sharing the same name but a different surname to receive the true-negative sample of matched entries. While this method allows us to create a database for training the algorithm that is as close as possible to the original database, this method is biased if the characteristics of surnames are dependent on (some of) the classification variables.

Moreover, we carefully controlled the plausibility of the linked data for the unique name-surname combinations. Nevertheless, this check was only possible at an aggregated level of different disciplines and years before and after graduation. While the results were comparable to other findings about the labor market trajectories of doctorate recipients at these aggregate levels of analysis (for example to information from BuWiN, 2013, 2017), the chosen approach could nevertheless lead to misclassifications in individual cases. In addition, the algorithm was only used for doctorate recipients with equal to or fewer than 300 namesakes. Even if it is expected that the algorithm would work sufficiently well for more than 300 potential matches for each entity, more linkage variables X would be advisable for training the function $\hat{f}$ for a precise classification.

Moreover, the IEB does not capture individuals who are not liable to social security contributions (e.g., civil servants, self-employed individuals, and family workers). Therefore, the final database may be biased towards those doctorate recipients who are part of the German social security system. For instance, certain occupations (e.g., physicians and lawyers) are traditionally self-employed or employed as civil servants (e.g., pastors, teachers). These graduate groups are underrepresented in the database. Furthermore, the DNB only contains records of published doctoral theses for German universities, while foreign doctorate recipients from non-German universities are not covered.

6. CONCLUSIONS

In this paper, we describe our approach using machine learning techniques to improve the record linkage of two sets of administrative data: a list of almost all German doctorate

recipients collected in the catalog of the German National Library (DNB), and the Integrated Employment Biographies (IEB) of the Institute for Employment Research (IAB). Linking these data sets was motivated by an interest in studying the labor market trajectories of German doctorate recipients at different stages of their careers. We show that supervised machine learning algorithms can be fruitfully applied to the linkage of social security data with other data. The proposed method has several advantages over traditional methods. On the one hand, its application is not restricted to micro data with overall high quality (where, for example, name-surname combinations and exact birth dates or social security numbers are available as unique identifiers). In addition, the quality of the matching algorithm can be assessed and compared to simple heuristics. On the other hand, the approach is applicable to contexts with strong privacy requirements, as is the case for anonymous social security data.

Bearing in mind a number of limitations, an evaluation of the method provides the following insights, which may help inform further work. First, machine learning algorithms can be trained on a synthetic training and evaluation data set if a "gold standard" sample is not feasible, and a supervised machine learning algorithm can be used for classifying individuals in administrative data. Second, in our specific application, simple heuristics (as have been used in prior record linkage approaches for German social security data) reach sufficiently high rates of precision. However, machine learning algorithms combine comparably high precision with drastically improved recall. Third, depending on the tuning parameters used, each algorithm can have a number of potential classification outcomes. This indicates a need to evaluate results from different algorithms.

The final database allows us to investigate the labor market trajectories of German doctorate recipients before, during, and after their graduation from 1975 up to 2015. A first evaluation of the database provides the following insights: While only a few doctorate recipients are unemployed, we find a substantial share of female doctorate recipients working part time. While female and male doctorate recipients show similar employment states during their graduation period, the shares of part-time and full-time employment diverge over the career paths of men and women.

ACKNOWLEDGMENTS

We thank Guido Bünstorf and the entire WISKIDZ-Team, Rasmus Bode, Tom Hanika, Andreas Rehs, and Igor Asanov for their valuable and constructive suggestions during the planning and development of this research work, as well as Judith Heinisch for her helpful comments.

AUTHOR CONTRIBUTIONS

Dominik P. Heinisch: Conceptualization; Methodology; Software; Visualization; Writing—original draft; Writing—review & editing. Johannes König: Formal analysis; Investigation; Methodology; Project administration; Software; Validation; Visualization; Writing—original draft; Writing—review & editing. Anne Otto: Data curation; Resources; Software; Visualization; Writing—original draft; Writing—review & editing.

COMPETING INTERESTS

The authors have no competing interests.

FUNDING INFORMATION

We gratefully acknowledge support from the German Federal Ministry of Education and Research (BMBF) under grant number 16FWN016.

DATA AVAILABILITY

Data used in this manuscript are subject to strict requirements on social data protection in Germany and cannot be made available in a data repository. For further details, see Section 3.2.

REFERENCES

Antoni, M., & Seth, S. (2012). ALWA-ADIAB—Linked individual survey and administrative data for substantive and methodological research. *Schmollers Jahrbuch*, *132*(1), 141–146.

Antoni, M., Ganzer, A., & vom Berge, P. (2016). Sample of integrated labour market biographies (SIAB) 1975–2014. FDZ-Datenreport, 4/2016.

Benderly, B. L. (2018). A trend toward transparency for Ph.D. career outcomes? *Science*. https://doi.org/10.1126/science.caredit.aat5250

Bishop, C. M. (2006). *Pattern recognition and machine learning*. New York, NY: Springer.

Blank, R., Daniels, R. J., Gilliland, G., Gutmann, A., Hawgood, S., Hrabowski, F. A., Pollack, M. P., Price, V. & Schlissel, M. S. (2017). A new data effort to inform career choices in biomedicine. *Science*, *358*(6369), 1388–1389.

Buenstorf, G., & Geissler, M. (2014). Like doktorvater, like son? Tracing role model learning in the evolution of German laser research. *Jahrbücher für Nationalökonomie und Statistik*, *234*(2–3), 158–184.

Christen, P. (2012a). *Data matching: Concepts and techniques for record linkage, entity resolution, and duplicate detection*. New York, NY: Springer.

Christen, P. (2012b). A survey of indexing techniques for scalable record linkage and deduplication. *IEEE Transactions on Knowledge and Data Engineering*, *24*(9), 1537–1555.

Culp, M., Johnson, K., & Michailidis, G. (2006). Ada: An R package for stochastic boosting. *Journal of Statistical Software*, *17*(2), 1–27.

Deutsche Nationalbibliothek (DNB). (2018, November 13). The German National Library in brief. Retrieved from http://www.dnb.de/EN/Wir/ueberblick

Dorner, M., Bender, S., Harhoff, D., Hoisl, K., & Scioch, P. (2014). The MPI-IC-IAB-Inventor data 2002 (MIID 2002): Record-linkage of patent register data with labor market biography data of the IAB. FDZ-Methodenreport, 06/2014.

Dorner, M., Heining, J., Jacobebbinghaus, P., & Seth, S. (2010). The sample of integrated labour market biographies. *Schmollers Jahrbuch*, *130*(4), 599–608.

Fitzenberger, B., Osikominu, A., & Völter, R. (2005). Imputation rules to improve the education variable in the IAB employment subsample. ZEW-Centre for European Economic Research Discussion Paper (05-010).

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*(1), 1–22.

Gareth, J., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning*. New York, NY: Springer.

Heinisch, D. P., & Buenstorf, G. (2018). The next generation (plus one): An analysis of doctoral students' academic fecundity based on a novel approach to advisor identification. *Scientometrics*, *117*(1), 351–380.

Kaul, A. Neu, N., Otto, A., & Schieler, M. (2016). Karrierestart, Mobilität und Löhne von Absolventen der Informatik. IAB-Regional—Berichte und Analysen aus dem Regionalen Forschungsnetz. 03/2016.

Konsortium Bundesbericht Wissenschaftlicher Nachwuchs (BuWiN). (2013). *Bundesbericht Wissenschaftlicher Nachwuchs 2013*. Bielefeld: W. Bertelsmann Verlag.

Konsortium Bundesbericht Wissenschaftlicher Nachwuchs (BuWiN). (2017). *Bundesbericht Wissenschaftlicher Nachwuchs 2017*. Bielefeld: W. Bertelsmann Verlag.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News*, *2*(3), 18–22.

Oberschachtsiek, D., Scioch, P., Seysen, C., & Heining, J. (2009). Stichprobe der Integrierten Erwerbsbiografien. Handbuch für die IEBS in der Fassung 2008. FDZ-Datenreport, 03/2009.

Organisation for Economic Co-operation and Development (OECD). (2018). *Education at a Glance 2018: OECD Indicators*. Paris: OECD.

R Core Team. (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing.

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

Schnell, R. (2013). Getting big data but avoiding big brother. German Record Linkage Center Working Paper Series, 2/2013.

Teichert, C., Niebuhr, A., Otto, A., & Rossen, A. (2018). Graduate migration in Germany: New evidence from an event history analysis. IAB-Discussion Paper, 3/2018.

Vom Berge, P., König, M., & Seth, S. (2013). Sample of integrated labour market biographies (SIAB) 1975–2010. FDZ-Datenreport, 1/2013.

Wanger, S. (2015). Frauen und Männer am Arbeitsmarkt: Traditionelle Erwerbs- und Arbeitszeitmuster sind nach wie vor verbreitet. IAB-Kurzbericht, 4/2015.

Wydra-Somaggio, G. (2015). Das Ausbildungspanel Saarland: Dokumentation der Datenaufbereitung. IAB-Regional. IAB Rheinland-Pfalz-Saarland, 3/2015.

Zimmermann, R., Kaimer, S., & Oberschachtsiek, D. (2007). Dokumentation des "Scientific Use Files der Integrierten Erwerbsbiographien" (IEBS-SUF V1), Version 1.0. FDZ Datenreport, 1/2007.

APPENDIX A: ASSESSMENT OF THE TRAINING DATA SET

To assess the representativeness of the synthetic training and evaluation data set, we present descriptive statistics for both data sets. Results for the number of multiples matched entries per entity can be seen in Table A1. Table A2 shows descriptive statistics of the variable distributions for the synthetic training and evaluation data set. Table A3 shows descriptive statistics of the variable distribution for the full (matched) data set.

Table A1. Distributions for multiple matches of the synthetic training and evaluation data set and of the full (matched) data set

Feature	Min	1stQ	Median	Mean	3rdQ	Max
Artificial training/evaluation data set	1	1	4	22.0889	20	296
Full (matched) data set	1	1	4	22.4841	20	299

Table A2. Descriptive statistics for the synthetic training and evaluation data set

Feature	Median	Mean	Min	Max
spell_research	0	0.0911	0	1
spell_hospital	0	0.1130	0	1
prop_educ	0	0.3517	0	1
age_sub	35	37.6489	20	102
same_ror_y5	0	0.0475	0	1
first_spell_before	−11	−11.2539	−45	39
year_diss	2001	2000	1975	2015
eastern	0	0.1658	0	1
nbr	90	103.1859	1	296
social science	0	0.1048	0	1
natural science	0	0.2564	0	1
engineering	0	0.0833	0	1
medicine	0	0.4001	0	1
law/economics	0	0.1187	0	1

Table A3 Descriptive statistics for full (matched) data set

Feature	Median	Mean	Min	Max
spell_research	0	0.0846	0	1
spell_hospital	0	0.0964	0	1
prop_educ	0	0.3319	0	1
age_sub	35	37.3718	20	115
same_ror_y5	0	0.0573	0	1
first_spell_before	−10	−10.2697	−62	40
year_diss	1999	1998	1975	2015
eastern	0	0.1677	0	1
nbr	94	106.8058	1	299
social science	0	0.0855	0	1
natural science	0	0.2550	0	1
engineering	0	0.0882	0	1
medicine	0	0.4171	0	1
law/economics	0	0.1118	0	1

APPENDIX B: ASSESSMENT OF MERGED DATA SET

The following figures were created to check the quality of the matched IIPED data.

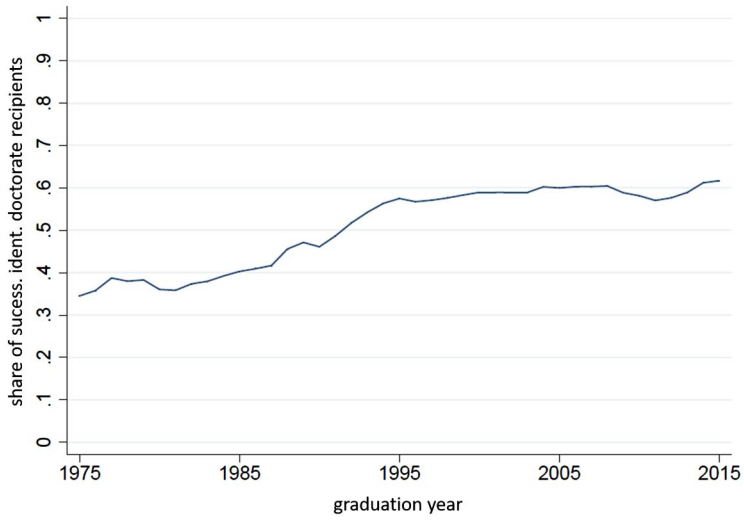

Figure B1. Successfully identified doctorate recipients by graduation year.

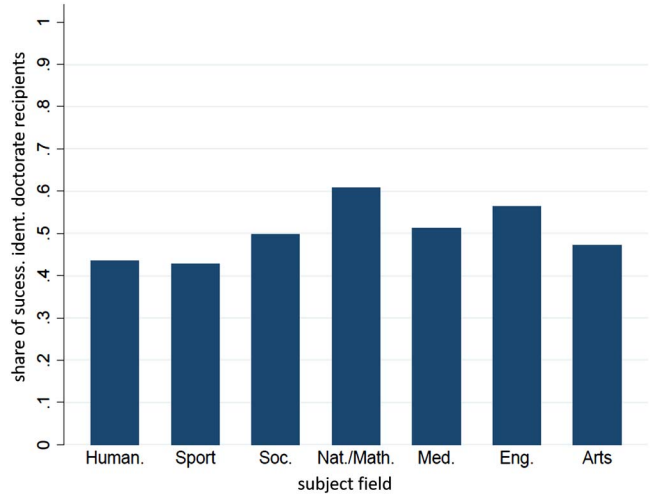

Figure B2. Successfully identified doctorate recipients by subject field.