International Journal of Medical Informatics 185 (2024) 105387

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf

Using machine learning to link electronic health records in cancer registries: On the tradeoff between linkage quality and manual effort

Philipp Röchner [a,b,*], Franz Rothlauf [b]

[a] Cancer Registry, Institute for Digital Health Data Rhineland-Palatinate, Große Bleiche 46, Mainz, 55116, Germany
[b] Information Systems and Business Administration, Johannes Gutenberg University, Jakob-Welder-Weg 9, Mainz, 55128, Germany

ARTICLE INFO

Keywords:
Record linkage
Data matching
Cancer registry
Electronic health records
Machine learning
Data quality

ABSTRACT

Background: Cancer registries link a large number of electronic health records reported by medical institutions to already registered records of the matching individual and tumor. Records are automatically linked using deterministic and probabilistic approaches; machine learning is rarely used. Records that cannot be matched automatically with sufficient accuracy are typically processed manually. For application, it is important to know how well record linkage approaches match real-world records and how much manual effort is required to achieve the desired linkage quality. We study the task of linking reported records to the matching registered tumor in cancer registries.

Methods: We compare the tradeoff between linkage quality and manual effort of five machine learning methods (logistic regression, random forest, gradient boosting, neural network, and a stacked method) to a deterministic baseline. The record linkage methods are compared in a two-class setting (*no-match/ match*) and a three-class setting (*no-match/ undecided/ match*). A cancer registry collected and linked the dataset consisting of categorical variables matching 145,755 reported records with 33,289 registered tumors.

Results: In the two-class setting, the gradient boosting, neural network, and stacked models have higher accuracy and F_1 score (accuracy: 0.968 – 0.978, F_1 score: 0.983 – 0.988) than the deterministic baseline (accuracy: 0.964, F_1 score: 0.980) when the same records are manually processed (0.89% of all records). In the three-class setting, these three machine learning methods can automatically process all reported records and still have higher accuracy and F_1 score than the deterministic baseline. The linkage quality of the machine learning methods studied, except for the neural network, increase as the number of manually processed records increases.

Conclusion: Machine learning methods can significantly improve linkage quality and reduce the manual effort required by medical coders to match tumor records in cancer registries compared to a deterministic baseline. Our results help cancer registries estimate how linkage quality increases as more records are manually processed.

1. Introduction

Identifying records that describe the same real-world entity is called record linkage, data deduplication, data matching, or entity resolution. Cancer registries collect and link information about cancer patients in a particular population. This information is used to: “1) define and monitor cancer incidence at the local, state, and national levels; 2) investigate patterns of cancer treatment; and 3) evaluate the effectiveness of public health efforts to prevent cancer cases and improve cancer survival” [1]. After matching reported records to individuals, cancer registries link incoming records to the corresponding registered tumor for that individual. Reported records match registered tumors either when all values describing the tumor are identical or when the data meet certain criteria. The latter case is of medical interest because tumors can change over time, different diagnostic methods describe the tumor with varying levels of accuracy, and medical experts may diagnose the same tumor differently [2]. Therefore, medical guidelines specify when two tumor records match and when they do not match (for the medical guidelines, see Section 2.2 and Appendix E). Our study focuses on linking reported records to registered tumors, as shown in Fig. 1.

We distinguish three different approaches to link records: deterministic, probabilistic, and machine learning approaches [3]. The choice of record linkage approach typically depends on the variable types of the data being linked. For string variables or categorical variables with

* Corresponding author at: Information Systems and Business Administration, Johannes Gutenberg University, Jakob-Welder-Weg 9, Mainz, 55128, Germany.
E-mail address: roechner@uni-mainz.de (P. Röchner).

https://doi.org/10.1016/j.ijmedinf.2024.105387
Received 24 March 2023; Received in revised form 5 October 2023; Accepted 20 February 2024
Available online 28 February 2024

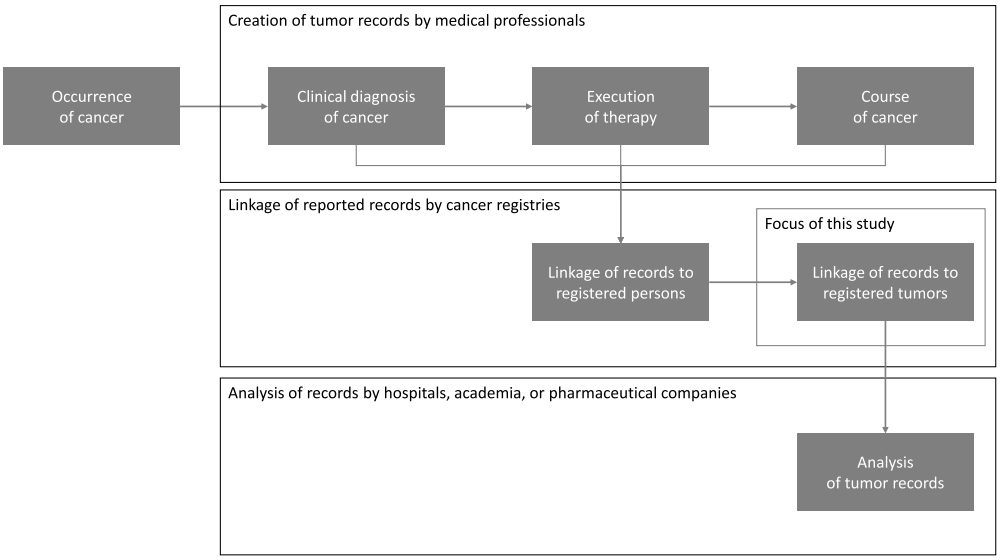

Fig. 1. A typical process used by cancer registries to collect, link, and analyze data on cancer patients. This study focuses on the second matching step that links reported records to matching registered tumors.

few values, deterministic and probabilistic approaches are often used (for data examined in related work, see Table 1 and Appendix A). For categorical variables with a large number of values, machine learning methods and deterministic rules are of interest. Since the reported records and registered tumors studied mainly contain categorical variables with a large number of values (for the data of this study, see Table 2 and Appendix C), we are investigating deterministic rules and machine learning methods.

Domain experts design deterministic rules to identify matching records. These rules can either compare attributes for exact matches [4] or use string distances and phonetic codes to measure similarities between strings [3,5]. If two records meet enough rules, they are matched. Machine learning methods used for record linkage learn the probabilities that records match in a supervised manner from records that have already been linked [6–9]. Although supervised machine learning methods can improve linkage quality [10,11], they have rarely been studied for record linkage, as discussed in Appendix A and shown in Table 1.

When possible, record linkage methods decide whether or not a reported record and a registered tumor match. Deterministic rules typically fail to link records if not enough rules are met. Machine learning methods fail if the probability of a match does not exceed a certain threshold. The remaining undecided records are processed manually. If a reported record does not match any tumor of the individual, a new tumor is registered. In general, the quality of matched records increases as more records that cannot be matched with sufficient certainty using record linkage methods are manually assigned. Thus, there is a tradeoff between the linkage quality of automated methods and the effort required to manually link records [12,13]. Linkage quality can be measured by classification metrics such as accuracy and F_1 score; manual effort can be measured by the number of records that cannot be decided by the record linkage method and are therefore processed manually.

We study the tradeoff between linkage quality and manual effort of record linkage methods on a real-world categorical dataset. The dataset was collected and labeled by a cancer registry, matching 145,755 reported records with 33,289 registered tumors. We compare machine learning methods for record linkage (logistic regression, random forest, gradient boosting, neural network, and a stacked approach) to a deterministic baseline designed by domain experts.

To our knowledge, the task of assigning records to tumors in cancer registries has not been discussed in the literature. Moreover, the quality of record linkage methods is often studied only on synthetic datasets with limited ability to generalize to real-world datasets [3]. Finally, the tradeoff between linkage quality and manual effort is rarely discussed.

2. Materials and methods

We performed the experiments in the programming language R and trained the machine learning models using the package H2O [20].

2.1. Dataset

The dataset of our registry-based study consists of 145,755 reported records and 33,289 registered tumors from 31,902 individuals residing in the federal state of Rhineland-Palatinate, Germany. The tumors studied were diagnosed between January 2019 and August 2020, reported before February 2021, and include all reportable tumor localizations according to national and regional laws in Rhineland-Palatinate of individuals older than 18 years. Table 2 summarizes the characteristics of the variables describing the reported records and registered tumors, and gives examples of their values; the variables and their preprocessing are explained in more detail in Appendix C and Appendix D. Fig. 1 and Appendix B describe the data collection and ground truth generation.

Fig. 2 shows how the binary labeled dataset used to train and test the machine learning methods is created. To train the machine learning methods, we make pairs of all reported records and all registered tumors for each individual. Linked pairs of reported records and registered tumors are labeled *match*; unlinked pairs are labeled *no-match*. Thus, for a reported record of an individual with n tumors, we create one pair labeled *match* and $n-1$ pairs labeled *no-match*. As a result, 92.4% of the reported and registered tumor pairs were labeled *match* and 7.6% were labeled *no-match*. The final dataset consists of 157,756 pairs of reported records and registered tumors.

Newly reported records are processed by pairing them with all of a patient's registered tumors. Record linkage methods are then used to determine whether the pair of reported record and registered tumor matches, does not match, or if the pair cannot be decided.

2.2. Record linkage methods

Deterministic rules The deterministic rules compare pairs of reported records and registered tumors, as shown in the right part of Fig. 2. The output of the deterministic rules for a record-tumor pair is one

Table 1
Related work comparing record linkage methods. Publications are sorted by year of publication. If an article studied multiple record linkage tasks, we report the number of records from the datasets with the most records in the last two columns.

Reference	Year	Methods	Domain	Data origin	Variable type	# Reported records	# Registered entities
[14]	2010	Deterministic Probabilistic Machine learning	Bibliographic E-commerce	Real-world	String Date	2,616	64,263
[15]	2011	Deterministic Probabilistic	Medical	Simulation	Categorical Date	10,000	10,000
[16]	2015	Deterministic Probabilistic	Medical	Simulation	Categorical Date	10,000	200,000
[17]	2016	Deterministic Probabilistic	Medical	Real-world	String Categorical Date	1,587,120	1,587,120
[18]	2019	Deterministic Probabilistic	Medical	Real-world	String Categorical Date	69,523	176,154
[19]	2020	Deterministic Probabilistic	Medical	Real-world Simulation	String Categorical Date	2,000	17,415
This study	2024	Deterministic Machine learning	Medical	Real-world	Categorical Date	33,289	145,755

Table 2
Variables describing reported records and registered tumors with the number of unique values, missing rate, example values, and explanation of the variable.

		# Unique values		Missing rate			
Variable	Type	Reported records	Registered tumors	Reported records	Registered tumors	Example	Explanation
ICD-10 code	Categorical	500	449	0.02	0.00	C18.9	Classification of disease
ICD-O topography	Categorical	365	301	0.04	0.00	C50.9	Location of tumor
Tumor laterality	Categorical	6	6	0.04	0.00	L	Site location for paired organs
ICD-O morphology	Categorical	492	427	0.39	0.06	8140/3	Cell type and behavior of tumor
Diagnosis date	Date	562	1,509	0.00	0.00	21.01.2020	Date of tumor diagnosis

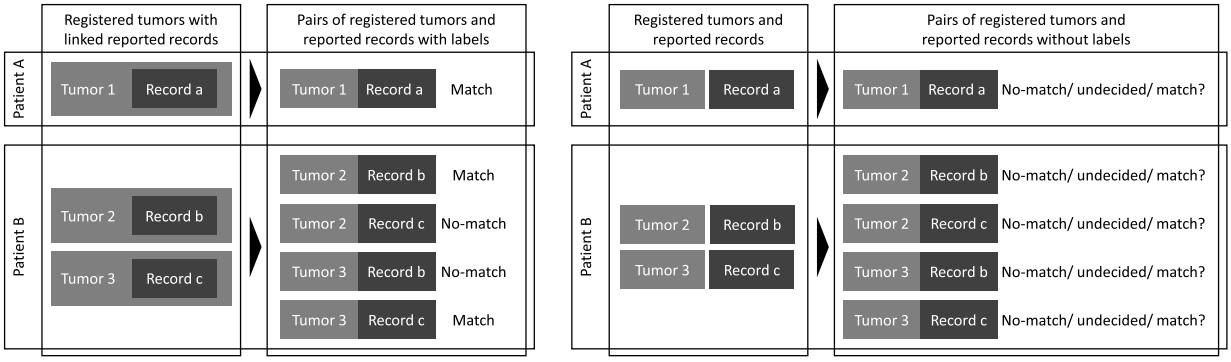

Fig. 2. Label generation to train and test the machine learning methods (left) and pairwise comparison of unmatched reported records with registered tumors (right). Reported records are paired with all registered tumors of a patient. To train the machine learning methods (left), pairs of linked reported records and registered tumors are labeled *match*. Record-tumor pairs that are not linked are labeled *no-match*. To process unmatched reported records (right), the record linkage methods decide whether the reported record and the registered tumor match, do not match, or cannot be decided.

of three values: *match* if the reported record and the registered tumor match, *no-match* if they do not match, or *undecided* if the reported record cannot be decided and is processed manually. Appendix E describes the corresponding medical rules.

Machine learning methods To link the reported records to the registered tumors, we use logistic regression, random forest, gradient boosting, neural networks, and a stacked method because these machine learning methods represent a variety of approaches commonly used for tabular data [21,22]. The machine learning methods are trained on previously linked data, as shown in Fig. 2. The trained models then return a matching score for each pair of reported records and registered tumors. Appendix F and Table 3 present the machine learning methods and their hyperparameters. We selected the hyperparameters by 2-fold cross-validation on the training set (80% of the entire dataset), evaluated the machine learning method on a separate test set (20% of the entire dataset), and repeated the training and evaluation approach 10 times with different random splits for the training and test sets. Appendix G discusses the detailed training and evaluation approach.

2.3. Evaluation approach

We compare the deterministic rules and the machine learning models in a two-class (*no-match/ match*) and a three-class setting (*no-match/ undecided/ match*). The literature often discusses record linkage with two classes, as shown in Table 1 and discussed in Appendix A; the three-class setting typically appears in applications.

Two-class setting evaluation approach To convert the three output classes of the deterministic rules (*no-match/ undecided/ match*) into two classes, we restrict the evaluation to record-tumor pairs in the test set that the deterministic rules evaluate as *match* or *no-match*. A classifi-

Table 3
Hyperparameters of the machine learning methods. We have highlighted in bold the hyperparameter configurations with the highest mean F_1 score during cross-validation. The entries in brackets are the values for the corresponding hidden layer of the neural network.

Method	Fixed hyperparameters	Variable hyperparameters
Logistic regression		Penalty mix: 0.0, 0.5, **1.0**
		Penalty strength: optimized automatically
Random forest	Number of trees: 150	Variables per split: 2, 4, **6**
(base learner)	Sample size ratio: 0.632	Minimal records per node: **20**, 40, 60
Gradient boosting	Number of trees: 150	Tree depth: 2, 4, **6**
		Learning rate: 0.05, 0.1, **0.2**
Neural network	Number of hidden layers: 3	Number of units: [100, 100, 100], [150, 150, 150], **[200, 200, 200]**
	Activation function: rectified linear unit	Dropout ratio: [0, 0, 0], [0.25, 0.25, 0.25], **[0.5, 0.5, 0.5]**
	Optimizer: Adadelta	
	Loss: binary cross-entropy	
	Batch size: 1024	
	Epochs: 20 with early stopping	
Random forest	Number of trees: 50	
(meta learner)	Variables per split: 4	
	Sample size ratio: 0.632	
	Minimal records per node: 50	

cation threshold converts the matching scores of the machine learning methods into two classes. We chose the classification threshold for each machine learning method to maximize the mean F_1 score on the validation folds during cross-validation.

We measure the linkage quality by accuracy, balanced accuracy, F_1 score, precision, recall, specificity, area under the precision-recall curve (AUPRC), and area under the receiver operating characteristic curve (AUROC). The mean and standard deviation of the linkage quality measures from 10 runs are reported.

In the two-class setting, we test the statistical significance of linkage quality using a pairwise Wilcoxon rank-sum test [23]. Appendix H discusses the details of the statistical test.

Three-class setting evaluation approach To compare the record linkage methods in a three-class setting (*no-match/ undecided/ match*), we transform the matching scores of the machine learning models into three classes by classifying all records close to the classification threshold of the two-class setting as *undecided*. This means that for a machine learning model

$$f : \mathcal{X} \to [0,1]$$

that predicts matching scores, we define a model that predicts three classes by

$$\hat{f} : \mathcal{X} \to \{\text{no-match},\ \text{undecided},\ \text{match}\} \text{ with}$$
$$\hat{f}(x) := \begin{cases} \text{no-match} & \text{if } f(x) \in \left[0, t - \frac{l}{2}\right), \\ \text{undecided} & \text{if } f(x) \in \left[t - \frac{l}{2}, t + \frac{l}{2}\right], \\ \text{match} & \text{if } f(x) \in \left(t + \frac{l}{2}, 1\right], \end{cases} \tag{1}$$

for the set of record-tumor pairs $\mathcal{X}$ as described in Section 2.1, a record-tumor pair $x \in \mathcal{X}$, a classification threshold $t \in [0,1]$, and an interval length $l \in [0,1]$.

As in the two-class setting, we choose for each method the classification threshold t that maximizes the mean F_1 score on the validation folds during cross-validation. For an interval length l equal to zero, the machine learning methods decide all record-tumor pairs as in the two-class setting. The percentage of undecided records increases with the interval length l. When the interval length l is large enough, all record-tumor pairs are evaluated *undecided*.

In the three-class setting, we evaluate decided and undecided records separately. A record is decided if it is labeled *no-match* or *match*. Thus, each record is either decided or undecided depending on the interval length l in Equation (1). To evaluate the linkage quality of the undecided records, each undecided record is labeled *no-match* or *match* as in the two-class setting or equivalently for an interval length l equal to zero in Equation (1). The number of decided and undecided records are mutually dependent, since if $x\%$ of the records are decided, then $1 - x\%$ of the records are undecided, and vice versa. The number of undecided (or decided) records determines the manual effort since all undecided records are reviewed by domain experts.

We measure linkage quality in the three-class setting using accuracy and F_1 score averaged over 10 runs.

3. Results

3.1. Two-class setting results

First, we discuss the statistical significance of linkage quality in the two-class setting. Table 4 shows the adjusted p-values for the Wilcoxon rank-sum test under the null hypothesis, that the linkage quality measure of two record linkage methods is *not* significantly different. The alternative hypothesis is that the linkage quality measure of the two linkage methods is significantly different. For clarity, we only show adjusted p-values above 0.01.

Overall, 109 of the 120 linkage quality measures compared pairwise were significantly different at the 0.01 significance level. In contrast, none of the linkage quality measures for the neural network and gradient boosting models were statistically different at the 0.01 significance level. In addition, the balanced accuracy of the neural network, gradient boosting, stacked models, and the recall of the random forest and logistic regression were not statistically different at this significance level.

Next, we compare the linkage quality measures of the discussed methods in the two-class setting. Table 5 shows the mean accuracy, balanced accuracy, F_1 score, precision, recall, specificity, AUPRC, and AUROC and their standard deviations over 10 runs. The gradient boosting and neural network models have the highest mean accuracy, balanced accuracy, F_1 score, and recall. The stacked method performs best in terms of balanced accuracy, precision, specificity, AUPRC, and AUROC. All linkage quality measures of the deterministic rules exceed those of the logistic regression and the mean accuracy, F_1 score, and recall of the random forest. The mean recall of the deterministic rules also exceeds that of the stacked method. For all reported measures, logistic regression has the lowest value.

3.2. Three-class setting results

In the three-class setting (*no-match/ undecided/ match*), we investigate how the accuracy and F_1 score of the studied record linkage approaches depend on the percentage of decided (*no-match/ match*) and

Table 4
Adjusted p-values from pairwise Wilcoxon rank-sum test. P-values are adjusted using the Holm-Bonferroni correction. For clarity, we only show rows with adjusted p-values above 0.01. In total, 109 of the 120 pairwise compared linkage quality measures were significantly different at a significance level of 0.01. Rows are sorted by decreasing adjusted p-values.

Method 1	Method 2	Linkage quality measure	Adjusted p-value
Neural network	Gradient boosting	Balanced accuracy	0.481
Neural network	Gradient boosting	AUPRC	0.436
Neural network	Gradient boosting	AUROC	0.315
Neural network	Stacked method	Balanced accuracy	0.210
Gradient boosting	Stacked method	Balanced accuracy	0.130
Random forest	Logistic regression	Recall	0.105
Neural network	Gradient boosting	Specificity	0.105
Neural network	Gradient boosting	Precision	0.089
Neural network	Gradient boosting	Recall	0.023
Neural network	Gradient boosting	Accuracy	0.015
Neural network	Gradient boosting	F_1 score	0.011

Table 5
Mean and standard deviation of linkage quality in the two-class setting over 10 runs. We highlighted the highest mean linkage quality values and results that are not statistically different from the highest linkage quality at a 0.01 significance level. Record linkage methods are sorted by decreasing accuracy.

Method	Accuracy	Balanced accuracy	F_1 score	Precision	Recall	Specificity	AUPRC	AUROC
Gradient boosting	**0.978 ± 0.002**	**0.961 ± 0.003**	**0.988 ± 0.001**	0.995 ± 0.000	**0.981 ± 0.002**	0.94 ± 0.006	0.999 ± 0.000	0.989 ± 0.001
Neural network	**0.976 ± 0.002**	**0.961 ± 0.003**	**0.987 ± 0.001**	0.996 ± 0.000	**0.978 ± 0.003**	0.944 ± 0.006	0.999 ± 0.000	0.988 ± 0.001
Stacked method	0.968 ± 0.004	**0.963 ± 0.002**	0.983 ± 0.002	**0.997 ± 0.000**	0.969 ± 0.004	**0.957 ± 0.005**	**0.999 ± 0.000**	**0.992 ± 0.001**
Random forest	0.953 ± 0.004	0.936 ± 0.005	0.974 ± 0.002	0.993 ± 0.001	0.956 ± 0.005	0.917 ± 0.012	0.998 ± 0.000	0.983 ± 0.002
Logistic regression	0.915 ± 0.007	0.713 ± 0.008	0.954 ± 0.004	0.958 ± 0.001	0.949 ± 0.009	0.477 ± 0.024	0.982 ± 0.001	0.847 ± 0.004
Deterministic rules	0.964 ± 0.001	0.904 ± 0.003	0.98 ± 0.000	0.987 ± 0.000	0.974 ± 0.001	0.834 ± 0.006	0.992 ± 0.000	0.904 ± 0.003

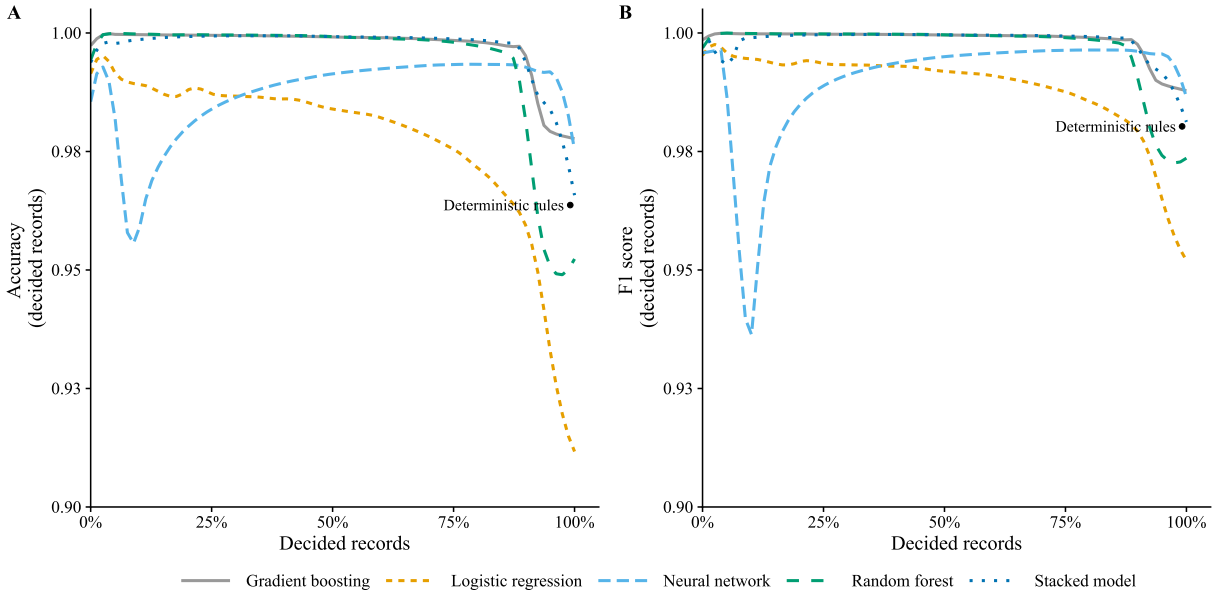

Fig. 3. Linkage quality of decided records (only *match* or *no-match*) over the percentage of decided records (depending on l in Equation (1)) in the three-class setting; the percentage of decided records measures manual effort (since the more records are decided, the fewer records are undecided and the other way around). Accuracy (A) and F_1 score (B) measure linkage quality. We evaluate the models on the test set and determine the records to be decided separately for each model. Accuracy and F_1 score are averaged over 10 runs. Note that the vertical axis does not start at the origin.

undecided records since the manual effort increases with the number of undecided records and decreases with the number of decided records.

Linkage quality of decided records First, we examine the linkage quality for the records decided by the approaches. Fig. 3 plots the mean linkage quality of decided records against the percentage of decided records in the test set.

The deterministic rules decide on average 99.11% of all records, leaving 0.89% of all records undecided; they have a mean accuracy of 0.964 and a mean F_1 score of 0.980. Thus, gradient boosting and the stacked approach have higher accuracy and F_1 scores than the deterministic rules for all percentages of undecided records (even if they decide 100% of all records).

Comparing the machine learning methods, the neural network performs best if more than approximately 90% of the records are decided. If less than approximately 90% of the records are automatically decided, gradient boosting, random forest, and the stacked method have the highest mean accuracy and F_1 score. As expected, the mean accu-

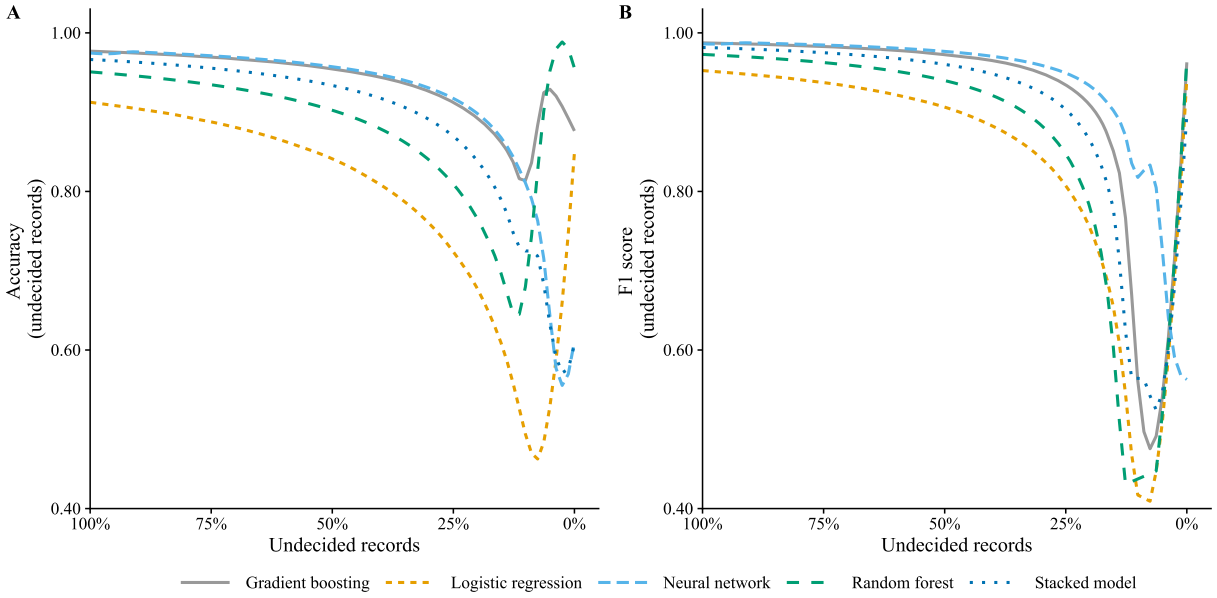

Fig. 4. Linkage quality of undecided records over the percentage of undecided records (depending on l in Equation (1)). Linkage quality is measured by accuracy (A) and F_1 score (B); the percentage of undecided records measures manual effort. We evaluate the models on the test set and determine the undecided records separately for each model. Accuracy and F_1 score are averaged over 10 runs. Note that the vertical axis does not start at the origin. The x-axis is inverted to match the x-axis of Fig. 3.

racy and F_1 score of all machine learning methods, except the neural network, decrease as the number of decided records increases. The neural network behaves differently, with a minimum accuracy and F_1 score at approximately 10% of decided records.

Linkage quality of undecided records Next, we study the linkage quality of the machine learning methods for undecided records. Fig. 4 plots the mean linkage quality for undecided records over the percentage of undecided records in the test set. The x-axes of Figs. 3 and 4 correspond to each other as if x% of the records are decided than $1 - x$% of the records are undecided. We do not report results for the deterministic rules because it is not possible to generate the labels *match* or *no-match* for the undecided records of the deterministic rules.

The accuracy and F_1 scores of the undecided records are lower than the results for decided records. This is as expected, since the undecided records are closer to the classification threshold t in Equation (1) than the decided records. Furthermore, the linkage quality steadily decreases with a smaller number of undecided records until a global minimum is reached between approximately 10% and 0% of undecided records. For a smaller number of undecided records, both measures of linkage quality increase again.

Comparing the machine learning methods, the neural network and gradient boosting have the highest mean accuracy and F_1 score on the undecided records if more than approximately 10% of the records remain undecided. If there are less than approximately 10% of undecided records, or even if all records are decided, gradient boosting and random forest perform best.

4. Discussion

We find the following: First, in the two-class setting, the gradient boosting, neural network, and stacked model outperformed the deterministic rules. Second, in the three-class setting, these machine learning methods can automatically process all reported records and still have higher accuracy and F_1 scores than the deterministic baseline. Finally, the linkage quality of the machine learning methods, except for the neural network, increases with a higher percentage of manually processed records. Based on these findings, machine learning can improve linkage quality and reduce the manual effort required by medical coders to match tumor records in cancer registries.

The deterministic rules have a fixed linkage quality and manual effort (manual processing of on average 0.89% of all records). In contrast, one can adjust the linkage quality and manual effort of machine learning methods. Thus, cancer registries can implement machine learning methods depending on the resources available for manual processing or the required linkage quality. For example, increasing the manual effort from 0.89% to 8.90%, which is a factor of 10 increase compared to the manual effort of the deterministic rules, increases the accuracy of the gradient boosting model from 0.978 to 0.994 and the F_1 score from 0.988 to 0.997. To achieve an accuracy of 0.999 with gradient boosting, domain experts must manually process 40.00% of the reported records; to achieve an F_1 score of 0.999, it is necessary to manually process 18.60% of all records.

Fewer models were trained for the logistic regression, random forest, gradient boosting, and neural network approaches than for the stacked method. For logistic regression, we trained three different hyperparameter configurations using 2-fold cross-validation, and the final model was fitted to the entire training dataset, for a total of seven models. For random forest, gradient boosting, and neural network, we fitted two models to subsamples of the training data during cross-validation for each of the nine grid search hyperparameter configurations, and the final model to the entire training data, for a total of 19 models trained for each approach. For the stacked method, we fit four different methods and the meta learner, for a total of 65 models. Thus, if computational resources are not limited and high precision, specificity, AUPRC, and AUROC are required, a stacked method (mean F_1 score: 0.983, mean AUROC: 0.992) can be implemented. Otherwise, gradient boosting (mean F_1 score: 0.988, mean AUROC: 0.989) or neural networks (mean F_1 score: 0.987, mean AUROC: 0.988) are good choices.

If machine learning methods automatically process almost all records in the three-class setting, we observe lower linkage quality. This suggests that machine learning models may have problems processing records close to the classification threshold t in Equation (1). To improve the linkage quality of such records, we intend to investigate active learning approaches for record linkage in future research [24,25]. That is, records that are difficult for machine learning models to de-

cide will be sent to domain experts for evaluation. We will then use these evaluated records to improve the machine learning models. Active learning approaches can also reduce the number of records used to train the machine learning models and thus reduce the manual effort of domain experts to label records [26].

5. Conclusions

Cancer registries link large numbers of electronic health records using deterministic rules and probabilistic approaches; machine learning methods are rarely used. Typically, not all records can be processed automatically and are therefore passed on to medical coders. The linkage quality and how it depends on the manual effort to link those records that cannot be automatically decided by record linkage methods is of great interest. Therefore, this study investigates how linkage quality depends on manual effort for five machine learning methods and a deterministic baseline. We compare the record linkage methods in a two-class (*no-match/ match*) and a three-class (*no-match/ undecided/ match*) setting, where the record linkage methods do not automatically decide a predefined percentage of records.

We find that gradient boosting, neural network, and stacked models outperform the deterministic baseline in both settings. Even if the machine learning methods automatically process all reported records, the accuracy and F_1 scores of these three approaches are higher than the deterministic baseline. Moreover, the linkage quality of the machine learning methods, except for the neural network, increases as the number of manually processed records increases. Our results help cancer registries estimate the increase in linkage quality for different amounts of manual effort spent on processing undecided records.

Overall, machine learning can improve linkage quality and reduce the manual effort of medical coders to match tumor records in cancer registries compared to a deterministic baseline.

List of abbreviations

AUPRC: area under the precision-recall curve; AUROC: area under the receiver operating characteristic curve; ICD-10: international statistical classification of diseases and related health problems, 10th revision; ICD-O: international classification of diseases for oncology, third edition.

Summary table

What was already known:

- Deterministic and probabilistic approaches, rarely machine learning, are used for record linkage.
- Machine learning methods for record linkage can outperform non-learning approaches.
- Records that cannot be linked automatically are usually processed manually.

What this study adds to our knowledge:

- Machine learning methods can improve the linkage quality of tumor records in cancer registries.
- Machine learning methods can reduce the manual effort required to link tumor records in cancer registries.
- Cancer registries can estimate how linkage quality increases as more manual effort is spent on manually processing records.

Ethical approval and consent to participate

The Joint Ethics Committee of the Faculty of Economics and Business Administration of the Goethe University Frankfurt and the Johannes Gutenberg University Mainz has classified the research project as ethically unobjectionable.

The consent to participate is based on the German Cancer Early Detection and Registry Act (Krebsfrüherkennungs- und -registergesetz, § 65c Sozialgesetzbuch V) and its regional implementation (Landeskrebsregistergesetz).

The Cancer Registry of the Institute for Digital Health Data Rhineland-Palatinate (IDG Institut für digitale Gesundheitsdaten RLP gGmbH Geschäftsbereich Krebsregister) has approved administrative access to anonymized patient data.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used Grammarly and DeepL Write to check spelling and grammar. After using these services, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Funding

This research has not received any specific grant from any public, commercial, or nonprofit funding agency.

CRediT authorship contribution statement

PR processed, analyzed, and interpreted the data. All authors drafted, revised, read, and approved the article for publication.

Declaration of competing interest

The authors declare that they have no conflicts of interest.

Availability of data and materials

The data supporting the results of this study are available from the authors upon reasonable request and with the permission of the Cancer Registry of the Institute for Digital Health Data Rhineland-Palatinate.

Acknowledgements

We thank the Cancer Registry of the Institute for Digital Health Data Rhineland-Palatinate for providing data and medical expertise, and Martin Briesch for his insightful study on the linkage of tumor records in cancer registries.

Appendix A. Related work

Record linkage methods and their quality on different datasets are actively studied. Table 1 summarizes related work by record linkage method, problem domain, and characteristics of the selected datasets.

Six of the studies shown in Table 1 discuss record linkage tasks in a medical context; only study [14] discusses two other domains. In the first task, they matched scientific publications provided by different bibliographic sources based on their title, authors, location, and year. In the second task, products available on different e-commerce platforms are linked based on their product name, description, manufacturer, and price. In the medical context, people are often matched based on their name, address, gender, and date of birth. This is of interest in hospital admissions to identify people who have been treated at the hospital before [17,18]. Another application is the collection of epidemiological data. In [19], data from people with HIV and syphilis are linked. In the simulated scenarios of [15] and [19], people are matched. In [16], the authors simulated the registration of claims for hospital events. Because most research studies primarily patients in the medical domain, record linkage methods are compared on datasets with strings, dates, and categorical variables with few values.

The datasets studied are either collected in the real world or simulated. In [15] and [16], the datasets are created synthetically by drawing values for records from predefined distributions. Noise and overlap are simulated to create different test scenarios. How well results on simulated datasets generalize to real-world datasets can be unclear. For simulated datasets, the matching records are known. Studies based on real-world datasets use different approaches to evaluate the quality of record linkage methods. If available, the true matches are linked by a unique identifier [18,14]. Otherwise, the ground truth is generated by manual verification [17,14], an ensemble of the compared methods, or a combination of both [19]. The number of records of the investigated datasets varies between 2,000 and 1,587,120.

Existing studies extensively discuss deterministic and probabilistic approaches [15–19]. Deterministic approaches either link records that match in all compared variables [19], match in a subset of compared variables [15], or when the sum of weights for matching variables exceeds a certain threshold [17]. In the case of numeric or string variables, these matches can be either exact [16] or inexact based on similarity measures between the variables [18].

Several studies also include probabilistic approaches based on the Fellegi-Sunter method [14,16–19,27]. The authors of [15] and [18] discuss an approach that uses expectation maximization to determine the threshold for the Fellegi-Sunter method. The study of [19] includes a Bayesian approach that allows to drop the independence assumption of the individual records pairs in the case of the Fellegi-Sunter method. In the experiments conducted, this leads to a better detection of mismatching records. The results of [15] show that probabilistic approaches outperform deterministic rules in the studied settings. In [16] and [19], the authors also discuss the impact of data quality on record linkage quality. They show that if data quality is low, probabilistic approaches outperform deterministic rules. If data quality is high, they perform similarly. In addition to the previously mentioned methods, the study of [14] discusses machine learning approaches such as support vector machines and decision trees. The authors find that machine learning approaches outperform non-learning approaches in their experiments.

Appendix B. Data collection

Fig. 1 shows the data collection process in cancer registries. After an individual develops cancer, hospitals, pathologists, and physicians collect information during clinical diagnosis, treatment, and disease progression, such as recovery, cancer recurrence, or death, and report this information electronically to cancer registries.

Cancer registries first use probabilistic approaches to match incoming records to the correct individual. Deterministic rules then link reported records to the matching tumor. If existing probabilistic and deterministic approaches fail to match reported records, domain experts manually process them. In addition, domain experts can adjust the decisions of the probabilistic and deterministic approaches. In the Cancer Registry of the Institute for Digital Health Data Rhineland-Palatinate, approximately 15 medical coders with at least three years of training perform these tasks. Physicians specializing in cancer care review ambiguous records. Hospitals, researchers, or pharmaceutical companies then analyze the collected records to improve existing therapies or develop new ones.

This study focuses on the second matching step in Fig. 1 that links reported records to matching registered tumors. Thus, the records studied are already linked to the matching individual.

Appendix C. Variables

Five variables describe each reported record and each registered tumor: ICD-10 code, ICD-O topography, ICD-O morphology, tumor laterality, and diagnosis date. The ICD-10 code is the International Statistical Classification of Diseases and Related Health Problems published by the World Health Organization [28]. It is a general system for classifying diseases. The ICD-10 codes for tumor diseases consist of four characters separated by a period: the first three characters categorize the disease in general, and the fourth character describes the disease in more detail. The ICD-O topography and morphology are part of the International Classification of Diseases for Oncology and are also published by the World Health Organization [29]. Both ICD-O codes are tumor-specific classification systems. The ICD-O topography describes the tumor's location in the person's body. The ICD-O morphology characterizes the tumor's cell type and behavior. Tumor laterality indicates the side of the paired organs involved: either right or left, both sides, centered, not applicable, or unknown. Table C.6 shows the five most frequent values for each variable, and Table C.7 shows patient demographics. The Cancer Registry of the Institute for Digital Health Data Rhineland-Palatinate does not collect information on patient comorbidities, ethnicity, or socioeconomic status.

Appendix D. Variable preprocessing

For the diagnosis date of a reported record and the diagnosis date of a registered tumor, we calculate their time difference in days. We then group each time difference into four categories: either both tumors were diagnosed on the same day, the time difference in days between the reported and registered tumor is greater than zero and less than or equal to 92 days, the time difference in days is greater than 92 days, or at least one of the diagnosis dates is unknown. The threshold of 92 days is motivated by guidelines for tumor registration [30].

We fill missing values with a dummy value, rather than completely excluding incomplete records or variables. After preprocessing, each variable is categorical, and its number of unique values varies between 6 and $1,509$. We one-hot encode the variables for the machine learning methods.

Appendix E. Deterministic rules

Domain experts designed the deterministic rules to match reported records to registered tumors based on the International Rules for Multiple Primary Cancers of the International Association of Cancer Registries and the International Agency for Research on Cancer [31]. Moreover, we considered adaptations and comments from the Association of Epidemiological Cancer Registries in Germany [30].

The rules for linking reported records to registered tumors depend on the specific tumor characteristics and the diagnosis date. In general, one primary tumor per organ and tissue type is registered regardless of time. In addition, dependent tumors are matched. For example, recurrences, metastases, and tumors that have spread from one organ to another are linked to one primary tumor. In contrast, tumors of paired organs, skin or intestine, and early stages of later invasive tumors with different therapies are counted separately [30]. Further detailed guidelines are defined in [32] and [33].

A decision tree with a total of 33 nodes and a depth of 28 nodes represents the deterministic rules. Each node can have one logical rule for one variable up to 12 logical rules for multiple variables. If possible, missing ICD-O morphology or ICD-O topography codes are completed based on the ICD-10 code of the corresponding record. Furthermore, special rules process records with ICD-O topography codes of paired organs, ICD-10 codes of death certificates, and ICD-O morphology codes of systemic diseases that are difficult to distinguish in clinical practice.

Appendix F. Machine learning methods

The investigated machine learning models calculate a matching score for each pair of reported records and registered tumors of an individual. If the matching score is above (or below) a certain threshold, then the record-tumor pair is labeled *match* (*no-match*), as described in

Table C.6
The five most frequent values of reported records and registered tumors per variable with the number (#) and percentage (%) of occurrences.

	Reported records			Registered tumors		
Variable	Value	#	%	Value	#	%
ICD-10 code	C61	19,032	13.06%	C61	5,051	15.17%
	C50.4	13,014	8.93%	C50.4	2,191	6.58%
	C50.9	6,376	4.37%	C34.1	1,298	3.90%
	C34.1	6,041	4.14%	C20	1,205	3.62%
	C20	5,868	4.03%	C50.8	938	2.82%
ICD-O topography	C61.9	19,172	13.15%	C61.9	5,057	15.19%
	C50.4	13,424	9.21%	C50.4	2,283	6.86%
	C50.9	6,990	4.80%	C34.1	1,302	3.91%
	C34.1	6,095	4.18%	C20.9	1,244	3.74%
	C20.9	5,870	4.03%	C50.8	995	2.99%
Tumor laterality	T	50,919	34.93%	T	15,684	47.11%
	R	37,922	26.02%	R	7,675	23.06%
	L	35,844	24.59%	L	7,304	21.94%
	U	7,593	5.21%	U	1,243	3.73%
	B	5,019	3.44%	B	871	2.62%
ICD-O morphology	8140/3	28,430	19.51%	8140/3	9,814	29.48%
	8500/3	15,458	10.61%	8500/3	4,083	12.27%
	8070/3	3,766	2.58%	8070/3	1,603	4.82%
	8520/3	2,736	1.88%	8520/3	685	2.06%
	8380/3	2,511	1.72%	8380/3	624	1.87%
Diagnosis date (year-month)	2019-1	15,027	10.31%	2019-1	2,350	7.06%
	2019-2	13,689	9.39%	2019-5	2,202	6.61%
	2019-3	12,572	8.63%	2019-2	2,136	6.42%
	2019-4	11,912	8.17%	2019-3	2,092	6.28%
	2019-5	11,839	8.12%	2019-7	1,998	6.00%

Table C.7
Gender breakdown in absolute numbers (#) and in percentage (%) and patient demographics with mean and standard deviation of age at diagnosis.

	Patients		Age at diagnosis (in years)	
Gender	#	%	Mean	Standard deviation
Male	16,128	50.6%	68.4	12.4
Not reported	35	0.1%	61.9	14.7
Female	15,739	49.3%	64.3	15.0
All	31,902	100.0%	66.4	13.9

Section 2.3. If the machine learning models cannot process a record-tumor pair with sufficient certainty, the record-tumor is marked as *undecided* and forwarded for manual processing (for the undecided records of the machine learning methods, see Equation (1)).

Table 3 shows the fixed and variable hyperparameters of the investigated machine learning methods. Appendix G describes the optimization of the variable hyperparameters during training.

Logistic regression fits a sigmoid function to the data. During training, the parameters of the sigmoid function are adjusted to minimize a cost function that includes a penalty term to avoid overfitting. We varied the strength and mixture of the L^1 and L^2 penalties.

Random forests and gradient boosting models are both ensembles of decision trees. Random forests train multiple decision trees on bootstrapped data samples and consider only a random subset of input variables for each split. Both reduce the dependence of individual decision trees on each other and avoid overfitting. A majority vote of all trees gives the final classification [34]. We fixed the number of decision trees and the sample size ratio for hyperparameter optimization. The number of randomly sampled variables for each split and the minimum number of records per node to keep the decision tree growing were varied.

Unlike the parallel construction of random forests, gradient boosting grows decision trees sequentially by multiplying the output of each decision by a learning rate and adding them up. Each decision tree reduces the residuals of the previous model between the predicted matching scores and the actual labels. The sequentially grown decision trees have a low depth to avoid overfitting [35]. We used the same number of decision trees for gradient boosting as for random forest and varied the depth of each decision tree and the learning rate.

Feed-forward neural networks predict by iteratively passing data through a sequence of affine linear transformations and nonlinear activation functions. The neural network adjusts the parameters of the affine linear transformations during training to minimize the observed classification loss. Each neural network is trained for 20 epochs or until the average loss over two epochs on the validation set decreases less than 5% compared to the average loss over the previous two epochs. We use dropout to reduce overfitting by replacing a randomly selected fraction of the affine linear transformations with the zero function after each parameter adaptation [36]. During hyperparameter optimization, we vary the number of nodes in the hidden layers and the dropout ratio.

Stacked models predict in two steps: First, different models, the base learners, predict labels for the records. Second, an additional model, the meta learner, predicts the final label based on the predictions of the base learners [37]. We use logistic regression, random forest, gradient boosting, and neural network models as base learners. Our meta learner is a random forest model. The random forest meta learner has the same hyperparameters as the random forest base learner, but we did not optimize the hyperparameter values of the meta learner during training.

Appendix G. Training approach

The machine learning methods are trained and evaluated on different datasets to measure the linkage quality independently of the data used for training. Before data preprocessing, we randomly divided the dataset into a training set with 80% and a test set with 20% of the data.

Since the studied dataset contains more record-tumor pairs labeled *match* than record-tumor pairs labeled *no-match* (for label generation,

see Section 2.1), we balance the labels in the training set. Therefore, we randomly up-sample the few record-tumor pairs labeled *no-match* and down-sample the many record-tumor pairs labeled *match*. Overall, the sampling does not change the size of the training set. We do not sample the test set to ensure a realistic evaluation of the record linkage methods.

We fit each machine learning method with different hyperparameters on the training set (for the hyperparameters, see Appendix F). For each approach, the hyperparameter configurations are generated by the values of the fixed and all combinations of the variable hyperparameters of Table 3. We evaluate three different hyperparameter configurations for logistic regression and nine for the random forest, gradient boosting, and neural network approaches. 2-fold cross-validation evaluates each hyperparameter configuration by dividing the training set into two subsets of equal size. We train each model on one subset and evaluate it on the other. This process is repeated by switching the subsets for training and evaluation. Finally, we select the model with the highest mean F_1 score over these two evaluations. The stacked approach fits the meta learner to the predictions of the four base learners on the training set generated during cross-validation. The hyperparameters of the meta learners are not optimized.

In the literature, 5-fold or 10-fold cross-validation is often used [38, 39]. On the one hand, increasing the number of folds used for cross-validation also increases the number of trained models in our experiments. On the other hand, increasing the number of folds used for cross-validation also increases the size of the subsamples used for training. Consequently, the performance of a model trained on a larger cross-validation subsample is a better estimate of the model's performance fitted to the entire training set. However, the number of folds used for cross-validation may have little effect on our results because the dataset studied is relatively large. In our opinion, a 2-fold cross-validation is suitable to evaluate the performance of the models while limiting the computational effort of our experiments.

To evaluate the influence of randomness due to the test and training split and the initialization of the machine learning methods, the training and evaluation approach is repeated 10 times.

Appendix H. Statistical significance

In the two-class setting, we test the statistical significance of linkage quality using a pairwise Wilcoxon rank-sum test [23]. This test ranks the runs of two record linkage methods by a measure of linkage quality, such as accuracy. Then, the ranks of the runs of each record linkage method are summed and compared under the null hypothesis that the linkage quality of the two methods is not significantly different. If the null hypothesis holds, then the test randomly ranks the runs of both methods. We test the significance of all examined linkage quality measures and each pair of record linkage methods.

To address the multiple testing problem, which can increase the number of statistically significant hypotheses, we apply the Holm-Bonferroni correction separately to all hypotheses concerning a linkage quality measure [40]. The Holm-Bonferroni correction increases the p-value of each null hypothesis, depending on the number of hypotheses tested and the initial p-values.

References

[1] M.C. White, F. Babcock, N.S. Hayes, A.B. Mariotto, F.L. Wong, B.A. Kohler, H.K. Weir, The history and use of cancer registry data by public health cancer control programs in the United States, Cancer 123 (2017) 4969–4976.

[2] W.A. Wells, P.A. Carney, M.S. Eliassen, A.N. Tosteson, E.R. Greenberg, Statewide study of diagnostic agreement in breast pathology, J. Natl. Cancer Inst. 90 (1998) 142–145.

[3] O. Binette, R.C. Steorts, (Almost) all of entity resolution, Sci. Adv. 8 (2022), eabi8021.

[4] A.L. Potosky, G.F. Riley, J.D. Lubitz, R.M. Mentnech, L.G. Kessler, Potential for cancer related health services research using a linked Medicare-tumor registry database, Med. Care 31 (1993) 732–748.

[5] W.W. Cohen, P. Ravikumar, S.E. Fienberg, A comparison of string distance metrics for name-matching tasks, in: Proceedings of the 2003 International Conference on Information Integration on the Web, 2003, pp. 73–78.

[6] N. Kooli, R. Allesiardo, E. Pigneul, Deep learning based approach for entity resolution in databases, in: Asian Conference on Intelligent Information and Database Systems, Springer, 2018, pp. 3–12.

[7] D. Nasseh, J. Stausberg, Evaluation of a binary semi-supervised classification technique for probabilistic record linkage, Methods Inf. Med. 55 (2016) 136–143.

[8] S. Rong, X. Niu, E.W. Xiang, H. Wang, Q. Yang, Y. Yu, A machine learning approach for instance matching based on similarity metrics, in: International Semantic Web Conference, Springer, 2012, pp. 460–475.

[9] M. Ebraheem, S. Thirumuruganathan, S. Joty, M. Ouzzani, N. Tang, Distributed representations of tuples for entity resolution, Proc. VLDB Endow. 11 (2018) 1454–1467.

[10] M.J. Bailey, C. Cole, M. Henderson, C. Massey, How well do automated linking methods perform? Lessons from us historical data, J. Econ. Lit. 58 (2020) 997–1044.

[11] S. Mudgal, H. Li, T. Rekatsinas, A. Doan, Y. Park, G. Krishnan, R. Deep, E. Arcaute, V. Raghavendra, Deep learning for entity matching: a design space exploration, in: Proceedings of the 2018 International Conference on Management of Data, 2018, pp. 19–34.

[12] K. Scheel, T. Franke, A. Weikert, M. Dick, A. Walter, J. Zeidler, T. Hartz, Record linkage in clinical cancer registration: experiences and findings from lower Saxony, in: German Medical Data Sciences: Bringing Data to Life, IOS Press, Amsterdam, 2021, pp. 101–109.

[13] W. Oberaigner, W. Stühlinger, Record linkage in the cancer registry of Tyrol, Austria, Methods Inf. Med. 44 (2005) 626–630.

[14] H. Köpcke, A. Thor, E. Rahm, Evaluation of entity resolution approaches on real-world match problems, Proc. VLDB Endow. 3 (2010) 484–493.

[15] M. Tromp, A.C. Ravelli, G.J. Bonsel, A. Hasman, J.B. Reitsma, Results from simulated data sets: probabilistic record linkage outperforms deterministic record linkage, J. Clin. Epidemiol. 64 (2011) 565–572.

[16] Y. Zhu, Y. Matsuyama, Y. Ohashi, S. Setoguchi, When to conduct probabilistic linkage vs. deterministic linkage? A simulation study, J. Biomed. Inform. 56 (2015) 80–86.

[17] A. Waldenburger, D. Nasseh, J. Stausberg, Detecting duplicates at hospital admission: comparison of deterministic and probabilistic record linkage, Stud. Health Technol. Inform. 226 (2016) 135–138.

[18] A.F. Karr, M.T. Taylor, S.L. West, S. Setoguchi, T.D. Kou, T. Gerhard, D.B. Horton, Comparing record linkage software programs and algorithms using real-world data, PLoS ONE 14 (2019) e0221459.

[19] T. Avoundjian, J.C. Dombrowski, M.R. Golden, J.P. Hughes, B.L. Guthrie, J. Baseman, M. Sadinle, Comparing methods for record linkage for public health action: matching algorithm validation study, JMIR Public Health Surveill. 6 (2020) e15917.

[20] H2O.ai, h2O: R Interface for H2O, H2O.ai, http://www.h2o.ai, 2020.

[21] R. Shwartz-Ziv, A. Armon, Tabular data: deep learning is not all you need, Inf. Fusion 81 (2022) 84–90.

[22] L. Grinsztajn, E. Oyallon, G. Varoquaux, Why do tree-based models still outperform deep learning on typical tabular data?, Adv. Neural Inf. Process. Syst. 35 (2022) 507–520.

[23] F. Wilcoxon, Individual comparisons by ranking methods, Biometrics 1 (1945) 80–83.

[24] T. Enamorado, Active learning for probabilistic record linkage, 2018, Available at SSRN 3257638.

[25] P. Ren, Y. Xiao, X. Chang, P.-Y. Huang, Z. Li, B.B. Gupta, X. Chen, X. Wang, A survey of deep active learning, ACM Comput. Surv. 54 (2021) 1–40.

[26] J. Kasai, K. Qian, S. Gurajada, Y. Li, L. Popa, Low-resource deep entity resolution with transfer and active learning, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 5851–5861.

[27] I.P. Fellegi, A.B. Sunter, A theory for record linkage, J. Am. Stat. Assoc. 64 (1969) 1183–1210.

[28] W.H. Organization, ICD-10: international statistical classification of diseases and related health problems: tenth revision, 2nd ed., World Health Organization, 2005.

[29] W.H. Organization, International Classification of Diseases for Oncology (ICD-O), 3rd ed., 1st revision ed., World Health Organization, 2013.

[30] C. Stegmaier, S. Hentschel, F. Hofstädter, A. Katalinic, A. Tillack, M. Klinkhammer-Schalke, Das Manual der Krebsregistrierung, Zuckschwerdt, Munich, 2018.

[31] W.G. Report, International rules for multiple primary cancers (ICD-O third edition), Eur. J. Cancer Prev. 14 (2005) 307–308.

[32] C. Martos, E. Crocetti, O. Visser, B. Rous, F. Giusti, et al., A proposal on cancer data quality checks: one common procedure for European cancer registries, Publications Office of the European Union, Luxembourg, 2014.

[33] A. Gavin, B. Rous, R. Marcos-Gragera, R. Middleton, E. Steliarova-Foucher, M. Maynadie, R. Zanetti, O. Visser, Towards optimal clinical and epidemiological registration of haematological malignancies: guidelines for recording progressions, transformations and multiple diagnoses, Eur. J. Cancer 51 (2015) 1109–1122.

[34] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32.

[35] J.H. Friedman, Greedy function approximation: a gradient boosting machine, Ann. Stat. (2001) 1189–1232.

[36] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (2015) 436–444.
[37] M.J. Van der Laan, E.C. Polley, A.E. Hubbard, Super learner, Stat. Appl. Genet. Mol. Biol. 6 (2007).
[38] Y.S. Abu-Mostafa, M. Magdon-Ismail, H.-T. Lin, Learning from Data, vol. 4, AML-Book New York, 2012.
[39] T. Hastie, R. Tibshirani, J.H. Friedman, J.H. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, vol. 2, Springer, 2009.
[40] S. Holm, A simple sequentially rejective multiple test procedure, Scand. J. Stat. (1979) 65–70.