Computational and Structural Biotechnology Journal 24 (2024) 724–733

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

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

Research Article

Automated sample annotation for diabetes mellitus in healthcare integrated biobanking

Johannes Stolp [a,1], Christoph Weber [a,1], Danny Ammon [b], André Scherag [c], Claudia Fischer [c], Christof Kloos [d], Gunter Wolf [d], P. Christian Schulze [e], Utz Settmacher [f], Michael Bauer [g], Andreas Stallmach [h], Michael Kiehntopf [a,*,2], Boris Betz [a,*,2]

[a] Department of Clinical Chemistry and Laboratory Diagnostics and Integrated Biobank Jena (IBBJ), Jena University Hospital – Friedrich Schiller University Jena, Jena, Germany
[b] Data Integration Center, Jena University Hospital – Friedrich Schiller University Jena, Jena, Germany
[c] Institute of Medical Statistics, Computer and Data Sciences (IMSID), Jena University Hospital – Friedrich Schiller University Jena, Jena, Germany
[d] Department of Internal Medicine III, Jena University Hospital – Friedrich Schiller University Jena, Jena, Germany
[e] Department of Internal Medicine I, Jena University Hospital – Friedrich Schiller University Jena, Jena, Germany
[f] Department of General Visceral and Vascular Surgery, Jena University Hospital – Friedrich Schiller University Jena, Jena, Germany
[g] Department of Anesthesiology and Intensive Care Medicine, Jena University Hospital – Friedrich Schiller University Jena, Jena, Germany
[h] Department of Internal Medicine IV, Jena University Hospital – Friedrich Schiller University Jena, Jena, Germany

ARTICLE INFO

Keywords:
Diabetes mellitus (DM)
Machine learning (ML)
ICD-10
Electronic health record (EHR)
Biobanking
Logistic regression (LR)
Conditional inference forests (CIF)
Natural language processing (NLP)
Healthcare integrated biobanking (HIB)

ABSTRACT

Healthcare integrated biobanking describes the annotation and collection of residual samples from hospitalized patients for research purposes. The central idea of the current work is to establish an automated workflow for sample annotation, selection and storage for diabetes mellitus. This is challenging due to incomplete data at the time of sample selection. The study evaluates a machine learning (ML) and natural language processing (NLP) based two-step procedure for timely and precise sample annotation for diabetes mellitus. Electronic health record data of 785 persons were extracted from the hospital information system. In the first step, a conditional inference forest (CIF) model was trained and tested based on laboratory values from the first 72 h of the hospital stay using test- (n = 550) and training data sets (n = 235). Performance was compared with a simple laboratory cut-off classifier (LCC) and a logistic regression (LR) model. Algorithms based on laboratory values, ICD-10 codes or information from discharge summaries extracted by a natural language processing software (NLP-DS) were evaluated as a second (review) step designed to increase the precision of annotations. For the first step, recall/precision/F1-score/accuracy were 71 %/86 %/0.78/0.82 for CIF and 77 %/70 %/0.74/0.75 for LR compared to 73 %/68 %/0.70/0.72 for LCC. NLP-DS was the best-performing second (review) step (93 %/100 %/0.97/0.97). Combining first-step models with NLP-DS increased precision to 100 % for all procedures (66 %/100 %/0.80/0.85 for CIF&NLP-DS, 72 %/100 %/0.84/87.2 for LR&NLP-DS and 66 %/100 %/0.80/0.85 for LCC&NLP-DS). The number of samples removed by NLP-DS was higher for LR&NLP-DS and LCC&NLP-DS (removal rate 35 % and 38 % of initially selected samples) compared to CIF&NLP-DS (removal rate of 20 %). The developed two-step procedure is an efficient implementable method for timely and precise annotation of samples from diabetic hospitalized patients.

1. Introduction

Diabetes mellitus is a public health concern of growing importance due to increasing incidence and prevalence, a broad range of associated secondary medical conditions as well as high mortality as a result thereof [1]. Research in diabetes mellitus profits from large well-defined sample cohorts of diabetic patients. These cohorts can be built up by using residual blood samples of hospitalized patients in the context of

* Corresponding authors.
E-mail addresses: michael.kiehntopf@med.uni-jena.de (M. Kiehntopf), boris.betz@med.uni-jena.de (B. Betz).
[1] Christoph Weber and Johannes Stolp share first authorship
[2] Michael Kiehntopf and Boris Betz share senior authorship

https://doi.org/10.1016/j.csbj.2024.10.033
Received 8 August 2024; Received in revised form 20 October 2024; Accepted 20 October 2024
Available online 23 October 2024

healthcare integrated biobanking.

Healthcare integrated biobanking refers to the practice of incorporating biobanking systems—collections of biological samples and associated data—directly into healthcare settings and processes. This integration enables the collection, storage, and analysis of biological specimens alongside patient health records, facilitating personalized medicine, advanced research, and more precise diagnostics. By linking biobanking with healthcare delivery, clinicians can access and utilize these samples to enhance patient care, improve treatment outcomes, and drive innovations in medical research and personalized therapies.

In the context of healthcare integrated biobanking, prompt biopreservation of residual samples upon availability is crucial for minimizing pre-analytical effects [2,3]. Automated sample collection offers an attractive solution for reducing pre-analytical time. However, the identification and annotation of appropriate samples for biopreservation faces two challenges: I. Timing: Patient clinical phenotype data are frequently incomplete at the time of sample biopreservation. While laboratory values are available in most cases, ICD-10 billing codes and discharge summaries are in many cases not reviewed before patient discharge. II. Data completeness is often insufficient. Kufeldt et al. found that HbA1c measurements were ordered for no more than 15 % of inpatients during their stay in a maximum-care university hospital [4]. Therefore, identification and annotation approaches in healthcare integrated biobanking need to address these data gaps.

Our study introduces a novel two-step machine learning- and NLP based procedure to address the challenges of early identification and variations in data completeness. In the first step, potential samples for biopreservation are identified using laboratory data from the first 72 h after hospitalization. In the second step, these classifications undergo re-evaluation using data that would be available at patient discharge.

In a hospital workflow, samples automatically annotated by the first step as diabetic would be stored until patient discharge. Samples would then be re-evaluated and samples identified as false positive would be removed from storage. With low removal rates, this enables automation of sample annotation and collection without excess usage of storage resources by temporarily stored samples.

1.1. Related work

Many studies utilize diabetes mellitus identification algorithms in homogeneous study cohorts, such as the PIMA Indian Diabetes dataset or patients from an endocrinology ward, where missing values are minimal [5–9]. However, these cohorts and the features used for diabetes mellitus detection do not represent the data routinely collected from the general population during hospital stays, which is characterized by missing and inhomogeneous data. Consequently, the results of these studies are not directly applicable to the challenges faced in healthcare integrated biobanking.

Several studies have attempted to identify diabetes mellitus in heterogeneous communities or hospitalized cohorts similar to those investigated in our current study. However, most of these studies incorporate features related to anamnestic information (e.g. diabetic pedigree, hyperglycemic episodes), social health data (e.g. alcohol consumption, smoking, dietary habits), physical examination (e.g. hip-waist ratio, body mass index, blood pressure), or specialized laboratory tests (e.g. extended lipid profile, fasting plasma glucose) [10–15]. These features are often not readily available for automated decision-making regarding sample storage and may even be entirely missing during a patient's hospital stay. Thus, they are not ideal for usage in the context of healthcare integrated biobanking.

Two previous publications exhibit a high degree of comparability to the current research.

A study conducted by Cardozo et al. [16] exclusively uses real-world laboratory-derived parameters (along with age), which are the most reliably accessible information in the hospital information system shortly after hospitalization. This approach compares with the first step of our two-step procedure (timely identification).

In a study published by Lee et al. [17] the authors use EHR data including laboratory data, medication data, free-text notes and ICD-10 codes from a population of hospitalized individuals utilizing both machine learning and NLP algorithms. This approach can be well compared to the second step of our two-step procedure (precise identification). Both studies, however, did not take into consideration the timing of potential residual sample annotation and collection, which includes the aspect of restricted data availability. However, this is of primary importance in the context of healthcare integrated biobanking.

The novelty of the current work therefore lies in the combination of machine learning-based classification using laboratory values and NLP-based analysis of discharge summaries in a two-step procedure taking into account the timing of data availability in a routine hospital workflow.

2. Methods

2.1. Study design

We used retrospective EHR data collected from a single tertiary healthcare institution for the construction and evaluation of a two-step procedure enabling the fast buildup of well-defined sample cohorts for research by automating the decision to store samples based on automated sample annotation for disease entities (diabetes mellitus in the current work).

For the first step of the procedure (early annotation), the performance of ML-based approaches was compared with a cut-off classifier to categorize patients as diabetic or not. The approach is based on laboratory values from the first 72 h of the hospital stay.

For the second step of the procedure (precise annotation), an NLP-based analysis of discharge summaries was compared to algorithms using ICD-10 codes and laboratory values. The approach is based on EHR data from the complete hospital stay – generally available only after patient's discharge (Fig. 1).

2.2. Dataset

The dataset of this retrospective study was derived from the 3000PA text corpus of the Smart Medical Information Technology for Healthcare consortium (SMITH) [18–20]. The dataset consists of electronic health record (EHR) data from 785 persons of European descent, who had an index hospital stay at a ward of Jena University Hospital for at least five days between 2010 and 2015. Only documents from deceased patients were included in the 3000PA text corpus for data privacy reasons [18].

The data extracted for each patient from the laboratory information system and the hospital information system of Jena University Hospital, included age, sex, length of hospital stay, dates and times of hospital admission, measured laboratory parameters, laboratory measurement results, dates and times of laboratory measurements, the Charlson Comorbidity Index (dichotomized ≥ 3 or < 3), the presence or absence of comorbidities in the categories of malignancy, liver disease, and chronic kidney disease, ICD-10 codes assigned to the patient's case, and discharge summaries.

Discharge summaries contained free-text notes on patient history, anamnesis, clinical examination, diagnostics, and treatment including medication.

In addition to data from the index hospital stay, data from previous hospital stays was collected, where it was available, and used for imputation but not for model training or testing.

To grant patient de-identification each case and all data pertaining to it were assigned a unique seven-digit alphanumeric string obtained via a hash function.

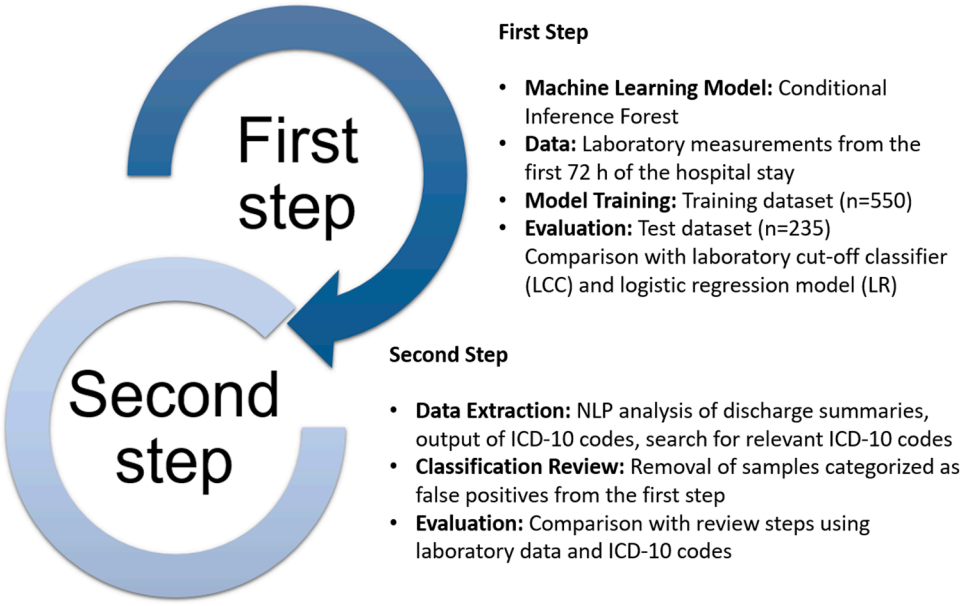

Fig. 1. Illustration of the study design: For the first step, ML-based approaches were evaluated and compared with a simple cut-off classifier. For the second step an NLP-based approach was compared with classifiers based on laboratory data and ICD-10 codes for diabetes mellitus.

2.3. Compliance with ethical standards and case classification

The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013) and was approved by the local ethics committee of the Friedrich Schiller University of Jena at the Faculty of Medicine. The ethics review committee waived informed consent of participants at an individual level. The study was also approved by the data protection officer of Jena University Hospital.

Manual classification of cases as diabetic and not diabetic was carried out independently by two physicians, who labeled cases as positive or negative for diabetes mellitus based on discharge summaries, ICD-10 billing codes, and laboratory values. The physicians used data from several (previous and future) hospital stays and hence had access to more data per patient than the ML algorithms which made classifications based on data from only the index hospital stay.

In cases where the classification differed between the two experts, the cases and reasons for the respective decisions were discussed and reevaluated to come to a final classification. This was necessary in less than 1 % of cases.

2.4. Data preparation

To obtain a machine learning-ready data structure the collected data was processed as follows:

1. For each patient, laboratory data were evaluated for three time intervals: measured within the first day (0–24 h), the second day (24–48 h) and the third day (48–72 h) of the index hospital stay. Laboratory values measured for less than 50 % of patients during the first 72 h of the hospital stay were removed from the variables used for model training and testing. (44 parameters were measured in 50 % or more than 50 % of patients and are listed in the supplemental) An exception from this rule was made for HbA1c, which was measured in less than 50 % of patients but was not removed due to its high importance as a marker for diabetes mellitus.
2. If multiple values for the same laboratory variable were measured within one time interval for a patient, the value selected for each time interval was the one with the greatest absolute distance from the center of the reference range of the respective laboratory variable.
3. If certain parameters were not measured for a patient during a specific time interval, missing values were filled by
   31. Last Observation Carried Forward followed by
   32. First Observation Carried Backward Strategy
   33. Patients with no measurements during the index hospital stay were imputed using the median of this laboratory variable during the last 3 years. In 588 patients (75 %), this rule was applied at least once, with an average of 9 variables being fitted this way per patient.
   34. Still missing values were imputed via predictive mean matching [21] as implemented in the mice R package (with standard parameter settings except for method = "pmm" and seed = 1 (for reproducibility purposes)). In 521 patients (66 %), this rule was applied at least once, with an average of 5 variables being fitted this way per patient.
4. Of pairs of correlating laboratory parameters (correlation $\geq$ 0.8, MCH and MCV, hematocrit and hemoglobin, ASAT and ALAT) one parameter was removed.
5. Recursive feature elimination (RFE), a simple backwards selection algorithm, was used to select the smallest number of laboratory parameters necessary for the creation of an accurate statistical model. For this task, the rfe function from the caret R package was adjusted to specify a random forest, with 5 times repeated 10-fold cross-validation as the resampling method, and accuracy as a summary metric to evaluate and select the optimal variable subset for the model. See supplemental for the relative contribution of each variable and resampling results for the variable subset. Eight different numerical variables: glucose, HbA1c, age, creatinine, urea, mean corpuscular hemoglobin concentration (mchc), hematocrit and mean corpuscular volume (mcv) were selected by RFE (Fig. 2).

The data preprocessed via the above steps contains case IDs, reference standard classification (diabetes mellitus/ no diabetes mellitus) for each ID, patient age, and the measured or imputed values of the seven laboratory parameters for each of the three-time intervals (0–24 h, 24–48 h and the third day 48–72 h).

2.5. Classification models for the first step

After obtaining a machine learning-ready data structure via the data

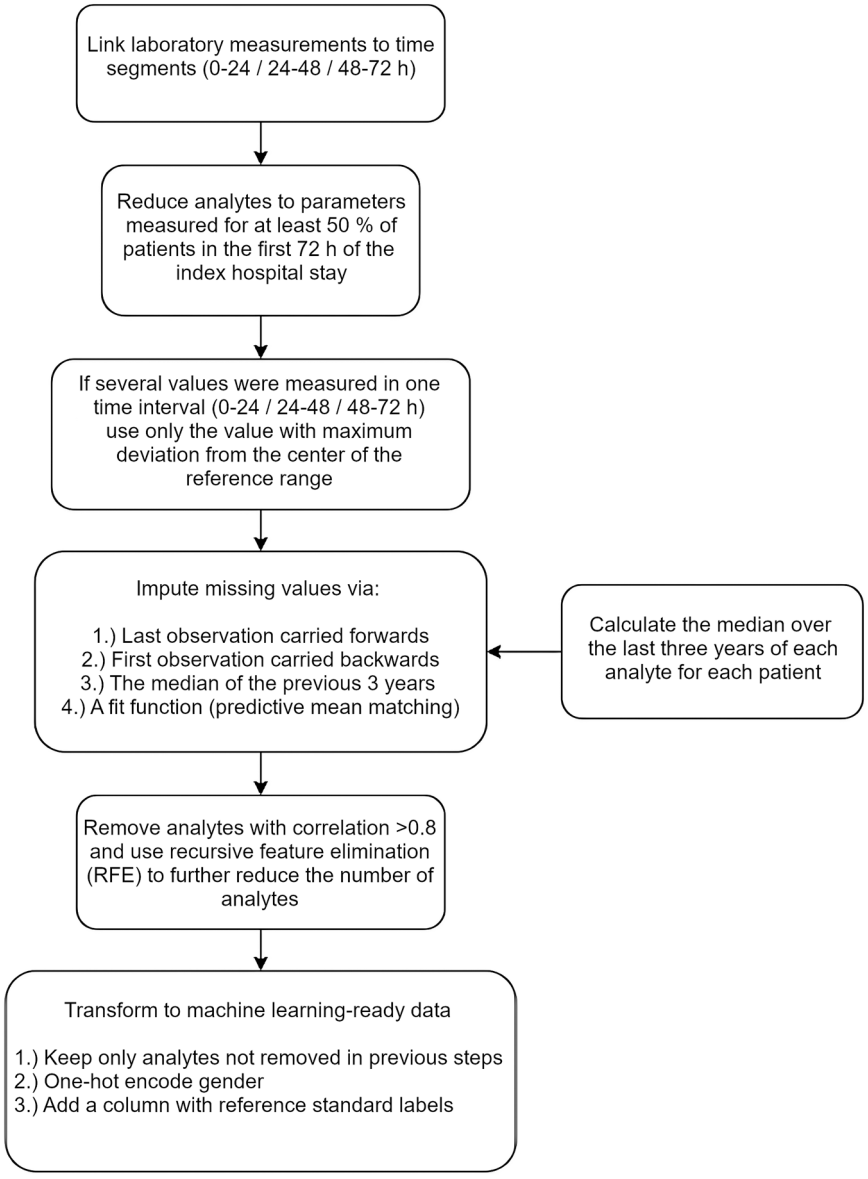

Fig. 2. Flow chart detailing the data preparation steps of laboratory data.

preparation steps described above the study dataset was divided into 70 % training data (n = 550) and 30 % test data (n = 235) stratified by the diabetes outcome to grant similar ratios of diabetes mellitus and non-diabetes mellitus in both. Spot-checking for 19 machine-learning models was performed to identify the best-performing machine learning models for the first step of the two-step procedure. A list of spot-checked models and their performance is available in the supplemental.

The main performance metric evaluated for spot-checking was accuracy. In addition to this, Cohen's Kappa was assessed as a second parameter.

Based on spot-checking results the conditional inference forest model (CIF) was chosen for the implementation of the first classification step due to the highest accuracy and Cohen's Kappa of all machine learning classification models. In addition to this logistic regression (LR) was used as a baseline model to contextualize performance result (LR being a common choice for baseline models in ML research) and a *l*aboratory *c*ut-off *c*lassifier (LCC) was used to compare the performance of CIF with a non-ML, rule-based classifier.

The LCC was a simple algorithm written in R with the following cut-off rules using the laboratory values for HbA1c and blood glucose over the first 72 h of the hospital stay. Diabetes mellitus was assigned if any of the HbA1c values during this time period were $\geq$ 6.5 % or if any of the random glucose blood test values were $\geq$ 11.1 mmol/l.

Models were trained and evaluated using five separate 10-fold cross-validations as a resampling scheme to optimize accuracy. The final accuracy of each model was averaged over all resamples.

Hyperparameter tuning was performed for the selected models using the grid search algorithm of the caret package in R.

2.6. Review algorithms for the second step

EHR information from discharge summaries analyzed by NLP software, registered ICD-10 billing codes, and laboratory values (HbA1c and blood glucose) from the complete index hospital stay were used to develop three respective classifiers for the identification of diabetes mellitus.

The three classifiers developed for the review of classifications (second step) were:

*L*aboratory *p*arameters (LP*):* To detect diabetes mellitus based on laboratory values, the same cut-off rules as for LCC were applied over

the entire index hospital stay. Records missing HbA1c and glucose values were treated as negative for diabetes mellitus as diagnosis could not be made.

ICD classifier *(*ICD*):* Registered ICD-10 billing codes from patients were screened for diabetes mellitus according to the code list provided by Quan et al. based on the Charlson Comorbidity Index [22].

The Charlson Comorbidity Index is a tool used to predict the ten-year mortality risk for patients based on the presence and severity of various comorbid conditions. It assigns weighted scores to a range of diseases—such as heart disease, diabetes, and cancer—where the cumulative score helps estimate the likelihood of mortality.

To enable the application of the index in administrative hospital discharge data, coding algorithms using the International Classification of Diseases were developed. We used the ICD assignments for diabetes mellitus from the Charlson Comorbidity Index to single out diabetic cases based on ICD-10 codes [22].

The relevant codes were all codes from E10.x-E14.x. Patients were labeled as not diabetic if none of these codes were present and as diabetic if one or more of these codes were present.

Classifier based on NLP-analysis of discharge summaries (NLP-DS): To extract evidence of diabetes mellitus from discharge summaries, the NLP text mining tool HEALTH DISCOVERY v5.7.0 by AVERBIS (https://health-discovery.io/) was applied using the discharge pipeline with default settings to extract anamnestic and clinical information (for detailed information, see the AVERBIS HEALTH DISCOVERY User Manual Version 5.7, 04/12/2018). The input data consisted of the discharge summary free-text notes in ASCII format. The output data consisted of a JSON file of ICD-10 codes, assigned to the case IDs by the NLP software based on the NLP analysis of the free-text notes. After analysis and annotation of ICD-10 codes, a Python script was used on the output JSON file to search it for relevant ICD-10 codes. For the detection of diabetes mellitus, ICD-10 codes according to Quan et al. were used as described above [22] (Fig. 3).

2.7. Statistical analysis

Models were evaluated using sensitivity/recall, specificity, positive predictive value (PPV)/precision, negative predictive value (NPV), accuracy (ACC), area under the receiver-operating curve (AUROC), F1-score, and removal rate.

In the context of the present work, we defined removal rate as an important metric for the performance of the two-step procedure. We defined it as the total rate of samples (mainly false positive, but also true positive) that would be stored after step one and would be removed from the biorepository by step two.

A low removal rate is desirable to increase the overall efficiency of the process of building well-defined sample cohorts, as less storage and material resources would be needed for the temporary storage of samples that are later identified as false positive and removed.

Data analysis was performed using RStudio (version 1.4.1103) [23], R software (version 4.0.3) [24] and the caret package for all machine learning models [25]. Descriptive statistics were presented as the mean (SD) or median (first quartile to third quartile) for continuous variables, and absolute numbers (percentages) for categorical variables.

3. Results

The study cohort included 785 individuals, of which 45.35 % had diabetes mellitus according to the reference standard. Individuals with diabetes mellitus were, on average, older (75.5 vs. 74 years), had worse kidney function (eGFR of 46.2 vs. 54.5 ml/min per 1.73m^2), had more concomitant diseases and had a longer hospital stay compared to participants without diabetes mellitus. Testing for HbA1c and blood glucose was performed in more individuals in the diabetic group (47.8 % and 91 %, respectively) compared to the non-diabetic group (24.7 % and 87.9 %, respectively). Diabetes mellitus was associated with fewer liver diseases but with a higher Charlson Comorbidity Index (Table 1).

During the initial phase of a hospital stay, only laboratory values are commonly available for sample identification, while discharge summaries and ICD-10 billing codes are not obtained until later. Therefore, we developed machine-learning models (CIF and LR) using a training dataset of 550 individuals, utilizing seven laboratory parameters frequently measured within the first 72 h. The performance of these models was then compared with a simple laboratory values cut-off classifier (LCC).

In the test dataset, the classification via CIF and LR reached areas

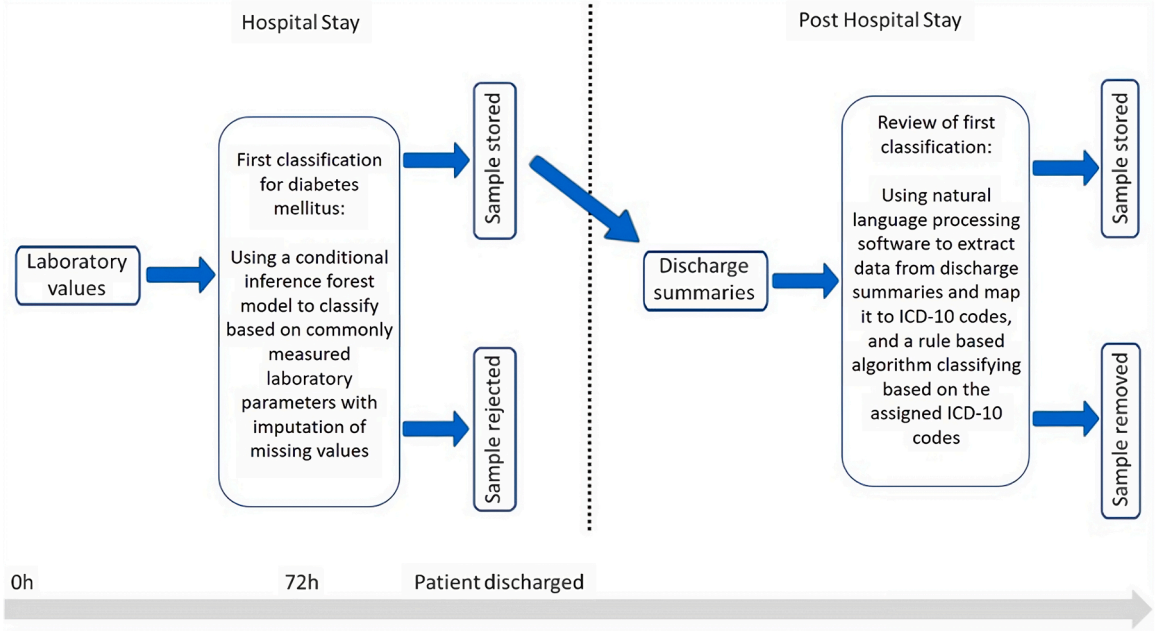

Fig. 3. Illustration of the suggested two-step procedure for implementation into the smart hospital workflow with healthcare integrated biobanking. A conditional inference forest (CIF) algorithm classifies the sample directly upon arrival as the first classification step. Discharge summary analysis (NLP-DS) is performed after patient discharge as the review step.

Table 1
Characteristics of all individuals, individuals with diabetes mellitus, and individuals without diabetes mellitus identified using the reference standard. * indicates a difference between groups with $p_{unadjusted} \leq 0.05$ calculated with two-sided t-test or chi-square test.

Characteristic	All individuals	Diabetes Mellitus	No Diabetes Mellitus
n	785	356	429
Age [years], mean (SD)	74.6 (12.2)	75.5 (10.2)	74 (13.6)
Sex, male, n (%)	476 (60.6 %)	214 (60.1 %)	262 (61.1 %)
HbA1c measured, n (%)	276 (35.2 %)	170 (47.8 %)*	106 (24.7 %)*
HbA1c [%], median (Q1-Q3[#])	6.3 (5.7 - 6.9)	6.9 (6.3 - 7.5)	5.8 (5.5 - 6.1)
Glucose measured, n (%)	701 (89.3 %)	324 (91 %)*	377 (87.9 %)*
Glucose [mmol/l], median (Q1-Q3[#])	7.1 (5.8 - 9.2)	8.6 (6.5 - 11.9)	6.4 (5.6 - 7.9)
eGFR measured at admission, n (%)	780	355 *	425 *
eGFR [ml/min per 1.73 m^2], median (Q1-Q3[#])	49.6 (28.6 - 77.3)	46.2 (25.4 - 68)	54.5 (32.4 - 82.4)
Charlson Comorbidity Index category ($\geq$ 3), n (%)	387 (49.3 %)	246 (69.1 %)*	141 (32.9 %)*
Any malignancy, n (%)	125 (15.9 %)	48 (13.5 %)	77 (17.9 %)
Liver disease, n (%)	137 (17.5 %)	54 (15.2 %)*	83 (19.3 %)*
Chronic renal disease, n (%)	303 (38.6 %)	166 (46.6 %)	137 (31.9 %)
Length of hospital stay [days], median (Q1-Q3[#])	10.7 (5.4 - 12.8)	11.1 (5.8 - 13.5)*	10.4 (4.9 - 12.8)*

[#] First and third quartiles

under the receiver operating characteristic of 0.84 (0.787–0.894) and 0.817 (0.763–0.872), respectively (p = 0.11; Fig. 4).

The optimal threshold for both models was determined by Youden's index. The LR model had the highest recall when compared to CIF and LCC (Table 2). The CIF model had the highest precision, F1-score and accuracy, followed by LR and LCC.

To enhance precision, we introduced the second-step review process for samples identified by using the aforementioned algorithms. This review step, which reexamines samples using EHR data, is not time-critical in terms of sample quality. It can be performed after patient discharge, when a comprehensive patient electronic health record (EHR) dataset is accessible.

Laboratory values from the hospital stay (LP), ICD-10 billing codes (ICD-10), and discharge summary information (NLP-DS) were utilized to identify individuals with diabetes mellitus. No single source identified all diabetic individuals in the study cohort (Fig. 5). The highest precision, recall, and accuracy were observed with NLP-DS, followed by ICD and LP. Combining LP and/or ICD with NLP-DS, considering a positive case as identified by at least one source, improved recall but reduced precision compared to NLP-DS alone.

Consequently, combining data sources only marginally improved accuracy in comparison to NLP-DS alone (Table 3).

Incorporating NLP-DS as a secondary review step in the two-step procedure effectively identified and eliminated all false-positive samples selected by the first-step algorithms. The two-step procedure with NLP-DS reached a precision of 100 % regardless of the algorithm used in the first step (Table 4). However, some samples that were correctly classified by the first-step models were also removed during the NLP-DS analysis, resulting in a decrease in overall recall compared to the different first-step algorithms (Table 2 and Table 4).

Notably, the number of samples removed by the NLP-DS analysis was substantially higher for LR&NLP-DS and LCC&NLP-DS (removal rate $\geq$ 35 % of initially collected samples) compared to CIF&NLP-DS (removal rate of around 20 %). Fig. 6 illustrates the relationship between sensitivity of the two-step procedures and their respective removal rates in the test dataset. (Fig. 6)

4. Discussion

This study demonstrates the effectiveness of a two-step procedure utilizing machine learning and NLP algorithms in optimizing sample collection for healthcare integrated biobanking. After the second (review) step, the procedure achieved 100 % precision in timely collecting residual samples.

Among the machine learning algorithms used in the first step, the CIF model achieved the highest F1-score, slightly surpassing the

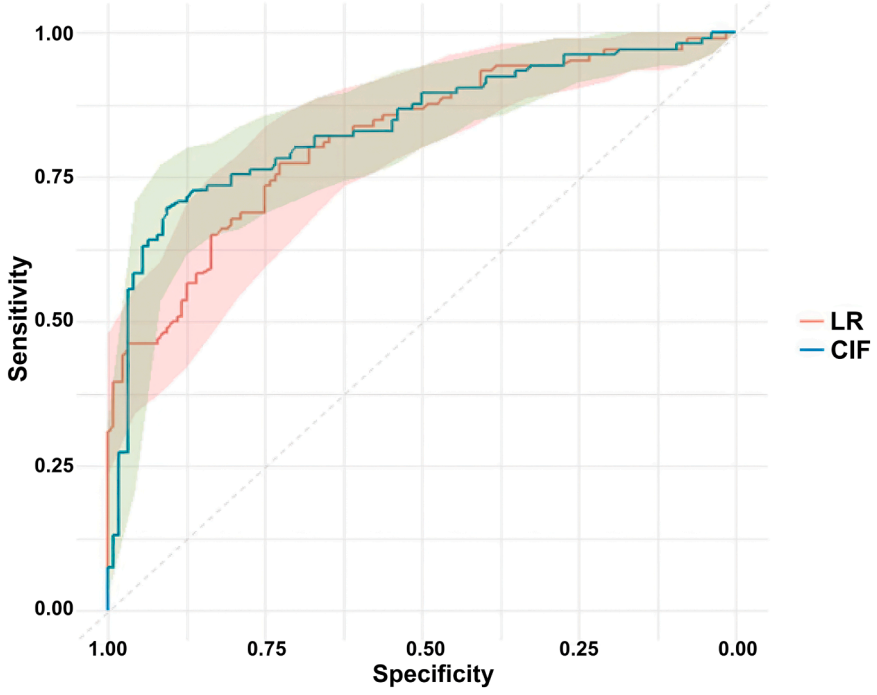

Fig. 4. Receiver operating characteristic (ROC) for logistic regression (LR) and conditional inference forest (CIF) algorithms based on multiple laboratory variables measured during the first 72 h of hospital stay for the identification of diabetic individuals.

Table 2
Performance of logistic regression (LR) and conditional inference forest (CIF) models based on multiple laboratory variables and laboratory cut-off classifier (LCC) for the identification of individuals with diabetes mellitus in the test dataset (n = 235 with n = 106 diabetic cases). TP: number of true positive cases; Spec: specificity; NPV: negative predictive value; ACC: accuracy; CI: confidence interval.

Model	Cases (n) selected	TP (n)	Recall [95 % CI]	Spec [95 % CI]	Precision [95 % CI]	NPV [95 % CI]	ACC [95 % CI]	F1-score [95 % CI]
CIF	87	75	0.708 [0.611 - 0.792]	0.906 [0.842 - 0.951]	0.862 [0.771 - 0.927]	0.789 [0.714 - 0.852]	0.816 [0.761 - 0.864]	0.777 [0.682 – 0.854]
LR	117	82	0.774 [0.682 - 0.849]	0.727 [0.641 - 0.802]	0.701 [0.609 - 0.782]	0.795 [0.710 - 0.864]	0.748 [0.687 - 0.802]	0.736 [0.643 – 0.814]
LCC	113	77	0.726 [0.631 - 0.809]	0.719 [0.632 - 0.795]	0.681 [0.587 - 0.766]	0.760 [0.674 - 0.833]	0.722 [0.660 - 0.779]	0.703 [0.608 – 0.787]

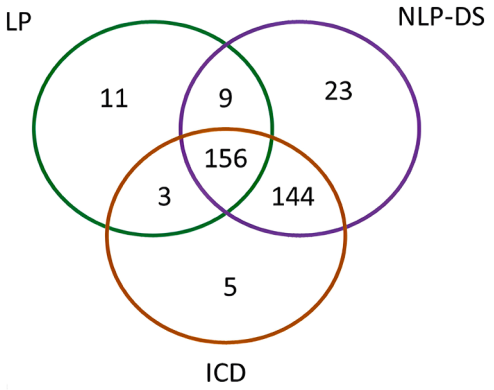

Fig. 5. Venn diagram comparing identification of diabetic individuals (n = 356) in the total study cohort by the different EHR data sources: laboratory parameters HbA1c and glucose (LP), information extracted from discharge summaries by NLP based software (NLP-DS) or ICD-10 billing codes (ICD). A total of five individuals were recognized by none of the three data sources, but by manual review only.

Table 3
Performance of the EHR data sources, ICD-10 codes (ICD), the laboratory parameters (LP) HbA1c and glucose, and information extracted from discharge summaries by natural language processing software (NLP-DS) individually and in combination for the identification of individuals with diabetes mellitus in the complete dataset (n = 785). Spec: specificity; NPV: negative predictive value.

	Recall	Spec	Precision	NPV	Accuracy	F1-score
LP	0.503	0.942	0.877	0.695	0.743	0.639
NLP-DS	0.933	1.000	1.000	0.947	0.969	0.965
ICD	0.865	0.993	0.990	0.899	0.935	0.924
LP or NLP-DS[#]	0.972	0.942	0.933	0.976	0.955	0.952
LP or ICD[#]	0.921	0.935	0.921	0.935	0.929	0.921
NLP-DS or ICD[#]	0.955	0.993	0.991	0.964	0.976	0.973
LP or NLP-DS or ICD[#]	0.986	0.935	0.926	0.988	0.958	0.955

[#]If any of the rules indicated diabetes mellitus, then the individual was classified as having diabetes mellitus.

performance of the best algorithm reported by Cardozo et al. [16] — another study that exclusively relied on laboratory parameters for diabetes mellitus identification (without the 72-hour limitation aspect though).

The limited performance of simple cut-off classifiers using HbA1c and blood glucose has also been reported in previous studies [17,26,27]. Two potential factors contribute to this:

1) Only half of the individuals with diabetes mellitus received HbA1c measurements during their hospital stay, and approximately 75 % of diabetic individuals not detected by laboratory values lacked documented HbA1c measurements. Imputation strategies were necessary for HbA1c in a significant number of patients and probably affected the performance. Routine HbA1c measurement upon admission for all hospitalized patients could lead to a higher number of diabetes diagnoses using laboratory parameters alone [4].
2) In individuals with well-controlled or even over-treated diabetes mellitus [28], blood glucose and HbA1c levels may not exceed the threshold criteria for diabetes mellitus. On the other hand, transient elevation of blood glucose levels due to medication and inflammation in non-diabetic individuals might reduce specificity of cut-off classifiers based on laboratory values.

For the second step of our procedure, we used information extracted from discharge summaries, which outperformed laboratory parameters (HbA1c and glucose) and ICD-10 codes regarding identifying hospitalized individuals with diabetes mellitus. These findings align with results from other studies [17,26,31]. ICD-10 billing codes are commonly utilized in epidemiological and clinical studies for identification of diabetic patients [29]. However, relying only on ICD-10 codes can result in inaccurate estimation of diabetes mellitus prevalence due to factors such as undiagnosed individuals, omission of codes, or incorrect coding [30]. Evaluation metrics achieved through NLP-based software analysis of discharge summaries matched or surpassed the performance of recent studies in diabetes mellitus identification [10–12,14,17,32]. In comparison to other studies, this approach requires only one data element (discharge summary), which should be available for every patient, at least after their hospital discharge.

Considering the frequent unavailability of discharge summaries during the decision-making process of preserving residual biosamples from hospitalized individuals, the implementation of an automated two-step procedure presents a viable solution for smart hospitals. The combined two-step procedure exhibited slightly lower accuracy/F1-score compared to recent studies regarding diabetes identification in the general population [10–14] or hospitalized patients [32]. This may result from variations in study cohort characteristics and in machine learning approaches applied. In addition, the data utilized in those studies (such as specialized laboratory values and detailed epidemiological data) are seldom accessible during routine hospital care within the initial 72 h after admission.

Few studies focusing on identification of diabetes mellitus have presented imputation strategies, while many tend to exclude incomplete datasets from analysis [6,12,13]. However, this approach is not feasible in the framework of healthcare integrated biobanking, where there is considerable variation in the extent of laboratory parameters measured per patient. Exclusion of incomplete datasets would result in the exclusion of many samples. In our study, the use of seven routine laboratory parameters, along with age, allowed for the implementation of an imputation strategy, eliminating the need for exclusions of individuals due to a lack of data availability.

Although LR and LCC combined with NLP-DS slightly outperformed the CIF&NLP-DS algorithm in terms of overall F1-score in the two-step procedure, CIF&NLP-DS may be the preferred choice for routine system implementation due to a significantly lower sample removal rates.

Precision and removal rate are the most important metrics to judge

Table 4
Performance of the two-step procedure combining conditional inference forest (CIF), logistic regression (LR), or laboratory cut-off classifier (LCC) models with discharge summaries analyzed with NLP software (NLP-DS) as a second (review) step in the test dataset (n = 235 with n = 106 diabetic cases). TP: number of true positive cases; ACC: accuracy.

Procedure	1st step[#]	2nd step*	TP (n)	Removal rate (%)	Recall [95 % CI]	Precision [95 % CI]	ACC [95 % CI]	F1-Score [95 % CI]
CIF & NLP-DS	87	70 (17)	70	19.5 %	0.660 [0.562 - 0.750]	1.000 [0.972 - 1.000]	0.846 [0.793 - 0.890]	0.795 [0.712 – 0.857]
LR & NLP-DS	117	76 (41)	76	35.0 %	0.717 [0.621 - 0.800]	1.000 [0.953 - 1.000]	0.872 [0.822 - 0.912]	0.835 [0.752 – 0.889]
LCC & NLP-DS	113	70 (43)	70	38.1 %	0.660 [0.562 - 0.750]	1.000 [0.972 - 1.000]	0.846 [0.793 - 0.890]	0.795 [0.712 – 0.857]

[#]Samples (n) selected in the first step by CIF, LR, or LCC
*Samples (n) selected in the second (review) step by NLP-DS, with number of samples removed in brackets.

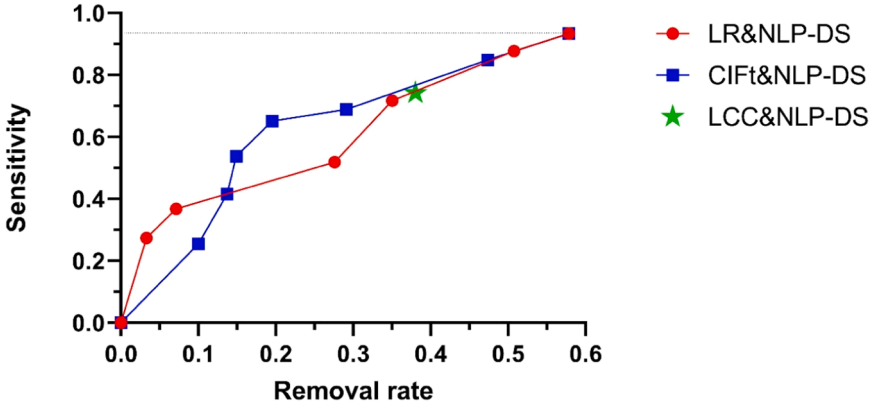

Fig. 6. Relationship between sensitivity for the identification of diabetic individuals and the sample removal rate after the second (review) step using different threshold cut-offs for the machine learning models. LR: logistic regression; CIF: conditional inference forest; NLP-DS: discharge summaries analyzed by natural language processing software; LCC: laboratory cut-off classifier.

the performance of an algorithm suited for the task of automatically annotating and storing samples from patients with a specific pathology (diabetes mellitus). A high performance in these two metrics avoids the excessive use of storage resources for temporarily stored samples and grants high quality of annotation for sample cohorts, which are positive for a certain pathology.

Based upon these metrics we suggest the combination of CIF classification of laboratory values from the first 72 h of hospital stay and the analysis of discharge summary free-text notes via NLP analysis (NLP-DS) as the best candidate for practical implementation into the routine workflow of a hospital.

If a clinical data warehouse (cDWH) is implemented in a hospital, the data needed for automated storage decisions may be accessible upon hospital admission due to data from a previous hospital stay with a diagnosis of diabetes mellitus.

In this scenario, the developed procedure could be used to analyze data retrieved from the cDWH from previous hospital stays.

In addition to this, it can — in cases where available cDWH data does not indicate the presence of diabetes mellitus — prove beneficial to apply the procedure to data collected during the hospital stay as it is possible that a new diagnosis is made during any hospital stay.

The procedure would also prove useful if a patient has no data stored in the cDWH yet, does not consent to the use of data collected prior to the current hospital stay or the stored data is incomplete e.g. due to missing information in discharge letters or ICD coding.

In this study ICD-10 codes from the Charlson Comorbidity Index were used for the definition of diabetes mellitus for both the ICD-classifier and the NLP-DS-classifier. Adding further ICD-10 codes to define cases of diabetes mellitus, other than the ones included in the Charlson Comorbidity Index might further improve the sensitivity of both the ICD-classifier and the NLP-DS-classifier. Such additional codes could, for instance, refer to secondary pathologies. However, in the vast majority of cases, patients with codings for secondary diseases stored in the EHR or determined by the NLP software were also assigned at least one code from the list of codes indicative of diabetes mellitus according to the Charlson Comorbidity Index. In this respect, the addition of further codes may be useful in cases where ICD-coding or information in discharge letters is incomplete in respect to a patient's known pathologies but is unlikely to improve the algorithm where coding and discharge letters are complete.

The strength of this study lies in the thorough characterization of the study cohort, using a comprehensive dataset consisting of laboratory values, discharge letters, and ICD-10 codes, and the reliable and systematic manual classification of all individuals by independent physicians as the reference standard, the comparative application of machine learning and NLP algorithms as well as the consideration of the order of data availability in a routine hospital workflow and therefore the direct applicability to healthcare integrated biobanking.

5. Limitations, future work and conclusion

Certain limitations should be acknowledged. Only two machine-learning algorithms (CIF and LR) were optimized for step 1 due to their favorable spot-checking performance and only one natural language processing tool was used for discharge summary analysis in step 2. While this approach simplified implementation for proof of concept, we cannot exclude that exploring alternative NLP tools and algorithms might further improve performance and future studies could compare additional algorithms to assess their performance.

The applied single imputation methods like median imputation are straightforward for handling missing laboratory values in ML data. However, it is possible that other more intricate imputation strategies

could further improve imputation results.

The sample size for training machine-learning classifiers was relatively small and the patient population primarily consisted of older patients in advanced disease stages. Further training and independent validation with larger datasets and datasets from demographically diverse cohorts and different hospitals are warranted to proof generalizability and further improve model performance.

Furthermore, access to additional data might have improved performance. However, we chose to use only sources that are reliably accessible during a single hospital stay.

Finally, we did not differentiate between type 1 and type 2 diabetes due to the limited number of participants.

In conclusion, our study's results provide a crucial foundation for the future implementation of annotation and selection procedures in hospital information systems. These procedures aim to detect and classify samples from individuals with diabetes mellitus and other specific pathologies with high precision. This will facilitate the formation of high quality, well-defined comparative study cohorts in the context of healthcare-integrated biobanking within a smart hospital. However, additional validation and confirmation of the current results in larger, independent cohorts are warranted to ensure robustness and generalizability.

Compliance with ethical standards

This study was conducted in accordance with the Declaration of Helsinki (as revised in 2013) and was approved by the local ethics committee of the Friedrich Schiller University of Jena at the Faculty of Medicine (Ethik-Kommission der Friedrich-Schiller-Universität Jena, Bachstraße 18, 07740 Jena, Germany, ethikkommission@med.uni-jena.de, 4639–12/15). The ethics review committee waived informed consent of participants at an individual level. The study was also approved by the data protection officer of Jena University Hospital. At the time of retrospective data collection, every individual was deceased.

Funding

This work was supported by the Deutsche Forschungsgemeinschaft (DFG) under grant KI 564/2-1 and HA 2079/8-1 within the STAKI^{2}B^2 project ("Semantic Text Analysis for Quality-controlled Extraction of Clinical Phenotype Information within the Framework of Healthcare-Integrated Biobanking"); the SMITH consortium and the Jena University Hospital – Friedrich Schiller University Jena is supported by the German Federal Ministry of Education and Research (01ZZ1803C). We further acknowledge support by the German Research Foundation Projekt-Nr. 512648189 and the Open Access Publication Fund of the Thueringer Universitaets- und Landesbibliothek Jena.

CRediT authorship contribution statement

Johannes Stolp: Writing – review & editing, Writing – original draft, Visualization, Methodology, Investigation, Formal analysis. **Christoph Weber:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation. **Danny Ammon:** Writing – review & editing, Resources. **André Scherag:** Writing – review & editing, Resources, Funding acquisition. **Claudia Fischer:** Writing – review & editing, Validation, Software, Methodology, Investigation, Formal analysis, Data curation. **Christof Kloos:** Resources. **Gunter Wolf:** Resources. **P. Christian Schulze:** Resources. **Utz Settmacher:** Resources. **Michael Bauer:** Resources. **Andreas Stallmach:** Resources. **Michael Kiehntopf:** Writing – review & editing, Writing – original draft, Visualization, Supervision, Project administration, Investigation, Funding acquisition, Conceptualization. **Boris Betz:** Writing – review & editing, Writing – original draft, Validation, Supervision, Methodology, Investigation, Formal analysis, Conceptualization.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2024.10.033.

Data Availability

The data supporting the results of this study are subject to restrictions due to the Thuringian Hospital Act and are therefore not publicly available. The data are available from the authors (B.B.) on reasoned request and with the approval of the local ethics committee of the Friedrich Schiller University of Jena at the Faculty of Medicine (Ethik-Kommission der Friedrich-Schiller-Universität Jena, Bachstraße 18, 07740 Jena, Germany, ethikkommission@med.uni-jena.de) and the data protection officer of Jena University Hospital (Zentrum für Gesundheits- und Sicherheitsmanagement, Beauftragte für Datenschutz, Am Klinikum 1, 07747 Jena, Germany, Datenschutzbeauftragter@med.uni-jena.de).

References

[1] World Health Organization. Global report on diabetes: WHO Press, 2016.
[2] Haslacher H, Bayer M, Fiegl H, Gerner M, Hofer P, Korb M, et al. Quality management at the national biobanking level - establishing a culture of mutual trust and support: the BBMRI.at example. Clin Chem Lab Med 2019(12);57:e301–5.
[3] Knutti N, Neugebauer S, Scherr F, Mathay C, Marchese M, Henry E, et al. Introduction of BD Vacutainer® Barricor™ tubes in clinical biobanking and application of amino acid and cytokine quality indicators to Barricor plasma. Clin Chem Lab Med 2022;60(5):689–700.
[4] Kufeldt J, Kovarova M, Adolph M, Staiger H, Bamberg M, Häring HU, et al. Prevalence and distribution of diabetes mellitus in a maximum care hospital: urgent need for HbA1c-screening. Exp Clin Endocrinol Diabetes 2018;126(2):123–9.
[5] Howlader KC, Satu MS, Awal MA, Islam MR, Islam SMS, Quinn JMW, et al. Machine learning models for classification and identification of significant attributes to detect type 2 diabetes. Health Inf Sci Syst 2022;10(1):2.
[6] Nadeem MW, Goh HG, Ponnusamy V, Andonovic I, Khan MA, Hussain M. A fusion-based machine learning approach for the prediction of the onset of diabetes. Healthcare 2021;9(10):1393.
[7] Rahman M, Islam D, Mukti RJ, Saha I. A deep learning approach based on convolutional LSTM for detecting diabetes. Comput Biol Chem 2020;88:107329.
[8] Kanimozhi N, Singaravel G. Hybrid artificial fish particle swarm optimizer and kernel extreme learning machine for type-II diabetes predictive model. Med Biol Eng Comput 2021;59(4):841–67.
[9] Olisah CC, Smith L, Smith M. Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective. Comput Methods Prog Biomed 2022;220:106773.
[10] Dinh A, Miertschin S, Young A, Mohanty SD. A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. BMC Med Inform Decis Mak 2019;19(1):211.
[11] Ijaz MF, Alfian G, Syafrudin M, Rhee J. Hybrid prediction model for type 2 diabetes and hypertension using DBSCAN-based outlier detection, synthetic minority over sampling technique (SMOTE), and random forest. Appl Sci 2018;8(8):1325.
[12] Kopitar L, Kocbek P, Cilar L, Sheikh A, Stiglic G. Early detection of type 2 diabetes mellitus using machine learning-based prediction models. Sci Rep 2020;10:11981.
[13] Xue M, Su Y, Li C, Wang S, Yao H. Identification of potential type II Diabetes in a large-scale chinese population using a systematic machine learning framework. J Diabetes Res 2020;2020:6873891.
[14] Zhang L, Wang Y, Niu M, Wang C, Wang Z. Machine learning for characterizing risk of type 2 diabetes mellitus in a rural Chinese population: the Henan Rural Cohort Study. Sci Rep 2020;10:4406.
[15] Zou Q, Qu K, Luo Y, Yin D, Ju Y, Tang H. Predicting diabetes mellitus with machine learning techniques. Front Genet 2018;9:515.
[16] Cardozo G, Pintarelli GB, Andreis GR, Lopes ACW, Marques JLB. Use of machine learning and routine laboratory tests for diabetes mellitus screening. Biomed Res Int 2022;2022:8114049.
[17] Lee S., Martin E.A., Pan J., et al. Exploring the reliability of inpatient EMR algorithms for diabetes identification. BMJ Health Care Inform 2023;30:e100894.
[18] Hahn U, Matthies F, Lohr C, Löffler M. 3000PA-towards a national reference corpus of German clinical language. Stud Health Technol Inform 2018;247:26–30.
[19] Lohr C, Luther S, Matthies F, Modersohn L, Ammon D, Saleh K, et al. CDA-compliant section annotation of German-language discharge summaries: guideline

development, annotation campaign, section classification. AMIA Annu Symp Proc 2018:770–9.
[20] Winter A, Stäubert S, Ammon D, Aiche S, Beyan O, Bischoff V, et al. Smart medical information technology for healthcare (SMITH). Methods Inf Med 2018;57(S01): e92–105.
[21] van Buuren S, Groothuis-Oudshoorn K. mice: multivariate imputation by chained equations in R. J Stat Softw 2011;45(3):1–67.
[22] Quan H, Sundararajan V, Halfon P, Fong A, Burnand B, Luthi JC, et al. Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. Med Care 2005;43(11):1130–9.
[23] R Core Team. R: A Language and Environment for Statistical Computing, 2021, R Foundation for Statistical Computing: Vienna, Austria.
[24] RStudioTeam. RStudio. Integrated for R. Boston, MA: PBC; 2020. RStudio.
[25] Kuhn M., Wing J., Weston S., Williams A., Keefer C., Engelhardt A., et al. The caret package - classification and regression training, 2022 version. Misc functions for training and plotting classification and regression models.
[26] Upadhyaya SG, Murphree Jr DH, Ngufor CG, Knight AM, Cronk DJ, Cima RR, et al. Automated diabetes case identification using electronic health record data at a tertiary care facility. Mayo Clin Proc Innov Qual Outcomes 2017;1(1):100–10.
[27] Chamany S, Silver LD, Bassett MT, Driver CR, Berger DK, Neuhaus CE, et al. Tracking diabetes: New York City's A1C registry. Milbank Q 2009;87(3):547–70.
[28] Lipska KJ, Ross JS, Miao Y, Shah ND, Lee SJ, Steinman MA. Potential overtreatment of diabetes mellitus in older adults with tight glycemic control. JAMA Intern Med 2015;175(3):356–62.
[29] Khokhar B, Jette N, Metcalfe A, Cunningham CT, Quan H, Kaplan GG, et al. Systematic review of validated case definitions for diabetes in ICD-9-coded and ICD-10-coded data in adult populations. BMJ Open 2016;6(8):e009952.
[30] Horsky J, Drucker EA, Ramelson HZ. Accuracy and completeness of clinical coding using ICD-10 for ambulatory visits. AMIA Annu Symp Proc 2017;2017:912–20.
[31] Wei WQ, Teixeira PL, Mo H, Cronin RM, Warner JL, Denny JC. Combining billing codes, clinical notes, and medications from electronic health records provides superior phenotyping performance. J Am Med Inform Assoc 2016;23(e1):e20–7.
[32] Muhammad LJ, Algehyne EA, Usman SS. Predictive supervised machine learning models for diabetes mellitus. SN Comput Sci 2020;1(5):240.