This is the published version of a paper published in .

Citation for the original published paper (version of record):

Case study: Feature engineering inspired by domain experts on real world medical data
*Intelligence-Based Medicine*, 8: 100110
https://doi.org/10.1016/j.ibmed.2023.100110

Access to the published version may require subscription.

N.B. When citing this work, cite the original published paper.

Permanent link to this version:
http://urn.kb.se/resolve?urn=urn:nbn:se:lnu:diva-125163
Case study - Feature engineering inspired by domain experts on real world medical data

Olof Björneld a,b,c,*, Martin Carlsson a,b, Welf Löwe a,b

a eHealth Institute, Department of Medicine and Optometry, Linnaeus University, S-391 82, Kalmar, Sweden
b Data Intensive Sciences and Applications (DISA-IDP), Department of Computer Science and Media Technology (CM), Faculty of Technology, Linnaeus University, S-391 82, Kalmar, Sweden
c Business Intelligence, IT Division, Region Kalmar County, S-392 32, Kalmar, Sweden

ARTICLE INFO

Keywords:
- Feature engineering
- Medical registry research
- Knowledge discovery in databases (KDD)
- Quantitative measures
- Electronic health record (EHR)
- Domain knowledge

ABSTRACT

To perform data mining projects for knowledge discovery based on health data produced in a daily health care workflow and are often built on vendor specific standards. EHRs usually have a substantial amount of non-structured information. That is, data not organised in a predefined manner. Although the data is very fine-grained it is hard to understand and mine without domain knowledge [1]. There is a wide range of medical research that use EHRs with the aim to discover, document, and mine without domain knowledge [2].

The process of finding new knowledge in real world registry data is called knowledge discovery. It is an iterative process that starts with a pattern or high-level scientific questions that include many diverse descriptive factors and relationships, which can result in unexpected information or new knowledge. The medical researcher is often a domain user or expert on the overall medical or physiological theory. The researcher has, for instance, knowledge about the relations between medical drugs, physiological effects, medical treatments etc., and seeks to test a medical research hypothesis.

To mine knowledge from real world data a data analyst is incorporated to perform data extraction, transformation, and presentation in close cooperation with domain experts and statisticians. This experience-driven iterative process to discover new knowledge from data, often described as knowledge discovery in databases (KDD), was

1. Introduction

In the included studies, EHR constitute the main data source. EHRs are used for documenting health care workflow and are often built on vendor specific standards. EHRs usually have a substantial amount of non-structured information. That is, data not organised in a predefined manner. Although the data is very fine-grained it is hard to understand and mine without domain knowledge [1]. There is a wide range of medical research that use EHRs with the aim to discover, document, and publish new knowledge [2].

The process of finding new knowledge in real world registry data is called knowledge discovery. It is an iterative process that starts with a pattern or high-level scientific questions that include many diverse descriptive factors and relationships, which can result in unexpected information or new knowledge. The medical researcher is often a domain user or expert on the overall medical or physiological theory. The researcher has, for instance, knowledge about the relations between medical drugs, physiological effects, medical treatments etc., and seeks to test a medical research hypothesis.

To mine knowledge from real world data a data analyst is incorporated to perform data extraction, transformation, and presentation in close cooperation with domain experts and statisticians. This experience-driven iterative process to discover new knowledge from data, often described as knowledge discovery in databases (KDD), was

* Corresponding author. eHealth Institute, Department of Medicine and Optometry, Linnaeus University, S-391 82, Kalmar, Sweden. E-mail addresses: olle.bjorneld@lnu.se, olle.bjorneld@regionkalmar.se (O. Björneld).

https://doi.org/10.1016/j.ibmed.2023.100110
Received 18 October 2022; Received in revised form 29 August 2023; Accepted 27 September 2023
Available online 28 September 2023
2666-5212/© 2023 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).
coined in 1989 at the first KDD workshop [3]. The related term of data mining, i.e., the use of specific methods or algorithms to find, extract useful and informative patterns in the data sources is, in our context, often used interchangeably with KDD [4,5].

Most often data mining generates new variables, also known as engineered features, to add data characteristics that are more directly able to predict an outcome of interest. In this study we claim that by understanding the implication between the original research questions and real-world health data in EHRs a faster and more valuable knowledge discovery process can be performed. Also, the result is more convincing since engineered features often have higher classification ability. We name this iterative cooperative process as “knowledge driven feature engineering” (KDFE). Referring to the knowledge pyramid; if KDFE was used, more knowledge could be gained from less data in a shorter time, and with fewer resources.

1.1. Research goal and questions

There is a vast body of literature that aims to support (medical) researchers in the knowledge discovery process, that is, to handle the input features and representation and assessment of knowledge. This might be overwhelming for a medical researcher whose protocol was just approved. In this paper we seek to support the medical researcher in the iterative data mining process with KDFE as it operationalises many approaches suggested in the literature. Ultimately, KDFE saves research time and increases outcome accuracy.

These claims boil down to the following research questions for the study:

Q1. Does the iterative data mining process that involves a medical researcher and a data analyst benefit from the engineering of new features (KDFE)?

Q2. Is it possible to automate the KDFE process?

2. Background

2.1. Knowledge discovery

Knowledge discovery is defined as “an interdisciplinary science whose goal is to extract useful and actionable knowledge from very large data repositories” [5] or as “describing the research efforts of turning raw data into useful knowledge for decision-making in scientific research” [6]. The process involves several steps that include data selection and pre-processing (data level), transformation and data mining (information level), and evaluation (knowledge level) [7]. Levels refer to the knowledge pyramid [8] which is used to describe the transition from data, to information, knowledge, and a final step of analysis. In this study, we focus on the data mining step of the knowledge discovery process.

Pragmatic data mining process models have been introduced in the industry. They include SEMMA (Sample, Explore, Modify, Model, and Access) and CRISP-DM (Cross Industry Standard Process for Data Mining) [9]. They have been compared and reviewed in relation to the overall KDD process concluding that CRISP-DM is more complete than SEMMA and that both methods can guide the implementation of data mining [10].

2.2. Data transformation

KDD starts with an ideation phase where a domain expert works in close collaboration with a data analyst to refine the initial research question. In particular, specific features that support the initial hypothesis are engineered [7]. A feature is equivalent to a variable or attribute; the term feature is used in this study. These engineered features, with alias FE, are designed to have a higher predictive ability than the original baseline features, named ORG. Features with high classification ability and informative features are used as interchangeable terms. The ORG features are based more on the exact wording of the research question. Iteratively, the medical researcher and the data analyst perform a manual process where new and refined features are constructed and added. The ORG features are combined with the engineered features resulting in the final evolved dataset ORGFE = ORG ∪ EF. Such workflow is described in general terms in Ref. [11].

2.2.1. Feature engineering

The process to transform and create new as well as more informative features is called feature engineering or sometimes feature generation when the focus is to create new features. Feature engineering often leads to better classification results, but it is time consuming and is better if performed in close cooperation with domain experts [12]. This cooperative iterative generation of new features based on domain knowledge involving a researcher and a data analyst is, in this study, referred to as knowledge driven feature engineering (KDFE). For each iteration the prediction ability of the ORGFE dataset improves until it reaches the theoretical maximum classification score. Therefore an automatic process benefits from a high degree of iteration [13].

So-called low level extractors can function as operators applied on features [12]. They can be built from atomic and “molecular” extractors that recursively combine results from other extractors. In a framework, information about train and test statistics of the extractors are collected and stored in a database for better feature engineering on new data.

Fig. 1 depicts the concept of low level extractors.

Moreover, information about extractor dependencies, intrinsic properties, and impact on prediction ability have a high value for the knowledge discovery pipeline. To utilise this information, a suitable framework is needed, also referred to as a feature store [14]. The concept of a feature store is now used by many artificial intelligence (AI) businesses to support their machine learning (ML) process [15].

Medical information usually contains longitudinal data, which are difficult to analyse but have great potential for informative features, especially, when sequences instead of single events are used [16]. Zhao et al. further states that learning temporal patterns leads to more informative features than domain knowledge did. Use of longitudinal data with high granularity is important for successful feature engineering [16].

2.2.2. Feature selection

Prior to feature engineering, feature selection (FS) should be performed on features. It is an important ML method and should be performed for several reasons [17] including faster model generation, prevention of overfitting, better classification and prediction results. Overfitting problems are often the case in high-dimensional settings [17].

In the medical domain there is no gold standard FS method. Studies use various types with similar results. Some examples are Gain ratio, Information Gain Coverage, Confidence coverage with negative feature mining [18] and Recursive Feature Elimination for Support Vector Machines, Lasso regulation, Information Gain, Correlation-Based, and even no FS, i.e., use of all features for classification or prediction [17]. No salient feature selection method exists, for optimal result several needs to be evaluated and the most often selected features should be used [19]. Often FS method is selected based on prediction performance such as AUROC, but the research group also need to take other aspects in consideration such as computation time and ability to reduce redundancy in the dataset [20].

2.2.3. Imbalanced target class

When ML is used to train prediction models the ratio of the target class is important. If imbalanced, actions are needed to prevent loss of prediction performance. This can be performed by either undersampling of majority or oversampling of a minority class. One frequently used method is synthetic minority oversampling technique (SMOTE) [21].
exact oversampling ratio is dependent on many factors, such as type of classification model, model size and dataset characteristics. SMOTE should only be performed on the training dataset, and a downside is that overfitting can occur due to duplication [22, 23]. For optimal performance fine tuning of over- and undersampling ratios is needed [23, 24].

2.2.4. EHR data characteristics

The research questions in the two studied cases were on a patient level, meaning from a data storage perspective that each patient is one index (row) with several dimensions, attributes, or measures. Descriptive dimensions and measures are named features. When machine learning should be performed this structure is very central, even for time series analysis. For example, when to perform SAX indexing the goal is a dataset on patient level, though the raw data contains time series [25]. Briefly, SAX indexing is a symbolic piecewise representation of a timeseries based on predetermined breakpoints. Datasets with the described characteristics are, according to us, named machine learning ready (MLR) datasets.

In our experience medical data is not stored in a MLR format, instead relational database structure is frequently used with $1:N$, $N:N$ cardinality. This observation has been made in other studies when examining clinical documentation systems, in this case regarding kidney donation [26]. See schematic example in Fig. 2.

Fig. 2 also describes the structure of a healthcare data warehouse (DWH), which is the natural environment for medical data analysts.

Implementation of a research question on information and data structure in DWH is about how to best apply the hypothesis, stated on a patient level, with the relational data in the DWH. This case study is about how to analyse knowledge discovery ability when feature engineering is driven by domain knowledge, defined as KDFE, on data stored in DWH.

2.3. Knowledge representation and measurement

Qualitative and quantitative domain knowledge are distinguished [27]. The former is based on observation, inference, and induction. It requires significant engagement from persons with deep knowledge. The knowledge relies to a high degree on subjective interpretation. By use of ML methods quantitative domain knowledge can be validated by large amounts of raw data. The latter is defined and expressed in mathematical language, which enables a quantitative analysis by use of ML. There are models that can structure and quantify qualitative knowledge by Fuzzy Cognitive Maps, but incorporating them with quantitative evidence requires additional algorithms [28].

The discovered knowledge in health care can be represented in various ways. It can be expressed as a set of hypotheses whereby knowledge discovery is the process that produces new hypotheses evaluated by domain experts [29]. It can also be represented in (ML) models, such as decision trees, Bayesian networks, Markov models, but also rules or others [30]. Many systems and methods exist, although there is, as yet, no consensus on a standard [30].

In general, there is a tradeoff between interpretability/explainability and prediction performances, such that one comes at the expense of the other. However, eXplainable AI (XAI) is gradually changing this paradigm, as the reasoning underlying accurate black box models is becoming more accessible. The tradeoff is highly dependent [31]. This was confirmed in a stratified review showing that XAI performance is dependent on the context, the characteristics of the users, and their different evaluation goals [32]. The issue of interpretability of a ML model is central to the medical domain. To describe the most informative features SHAP technique can be used [33].

Regardless of the knowledge discovery process, a measure to estimate the knowledge inherent in information or data is central in data mining. The goal of knowledge metrics is to compare two information sets to identify the one that can generate the most knowledge. For classification tasks, especially ML-based classification, the AUROC measure is prominent [34].

2.4. Medical registry studies

Research on personal information must be carried out in compliance with applicable legislations, such as the GDPR [35] and other national legislation on personal data. Based on our experience, not just the two included cases, a detailed schematic overview of the medical research...
process, with explanatory text and examples, was produced. The overview can be found in the Appendix. In short, it starts with a research question, data that supports the hypothesis is collected based on ethical approval, and the process ends with publication of discovered knowledge.

2.5. Domain knowledge driven feature engineering (KDFE)

Both studied projects had a complex multifactorial research question with a non-trivial explanatory model. Therefore, many potential engineering features were hand crafted and a direct result of the collaborative process (KDFE), which is a central part of the data mining process. For more details about the datasets and the engineered potential features see section 4.1 and the Appendix.

The main objective of this study was not to look at individual performances of the FS method or classification algorithms. Instead, it was to (i) prove that the KDFE process is meaningful and creates value to the researcher, and (ii) lay the foundation for an automated KDFE.

An interesting and emerging field related to automated KDFE is autoML, meaning high level solutions for automated machine learning. There are big possibilities in healthcare for autoML, reasons are a lot of available data and often lack of competence in machine learning. However, work is still needed for better efficiency on larger retrospective datasets [36]. Other challenging aspects in healthcare data are large sample size, high imbalance, and feature limitations. Future potential is significant [37].

3. Methodology

3.1. Study method

This case study involved two cases of medical research projects (P1 and P2). The focus was not on specific subprocesses or to look at the cases from a holistic view. According to Ref. [38] this study is defined as a multiple embedded case study. An elaborated definition of case study is “an empirical inquiry that investigates a contemporary phenomenon (the ‘case’) in depth and within its real-world context”, and embedding in this context means to analyse an interesting or prominent aspects of the cases [38]. This study focused on the KDFE process and not the entire KDD process.

To estimate the discovered knowledge, we used a machine learning pipeline that included imputation, normalisation, feature engineering, feature selection and, finally, supervised learning classification with optimisation of model parameters. The target class in P2 was imbalanced, therefore SMOTE was used. The knowledge discovery in the form of gained knowledge was represented by classification metrics from the selected features and their ability to predict the outcome variable. The classification results for the baseline dataset (ORG) were compared to the evolved dataset (ORGFE).

Classification ability (AUROC) of the feature datasets (ORG/ORGFE) was the main measure for gained knowledge. For a deeper analysis, accuracy, precision, recall and F1 scores were also collected.

3.2. Experiment pipeline

To quantify the knowledge discovery associated with the KDFE process we executed an ML pipeline on datasets (ORG, ORGFE) for P1 and P2. An overview of activities performed is presented below in Fig. 3.

The data mining process started with (i) a research question, (ii) data was then collected from DWH, (iii) KDFE produced features with high classification ability, and (iv) classification ability was then evaluated. Activities in the pre study sections are important to a medical research project but not part of this study, for more details see the Appendix concerning the medical research process. At the bottom of Fig. 3 are references to the different levels in the knowledge pyramid.

To fully describe the details of the ML pipeline that was used for the evaluation of KDFE see Fig. 4. In short, (i) the data was split into a train and a test set, (ii) FS was performed on the train set w/wo normalisation, (iii) the top five features were selected in each generated dataset, (iv) another train test split was performed with the intention to perform an extra validation of the cross validation and grid search for best model hyperparameters. However, the extra validation was never executed. (v) SMOTE was performed at ratio 1:2 if the target value was imbalanced, (vi) best model parameters for each data set were used for evaluation on test dataset, (vii) classification metrics were collected.

3.3. Data preparation

3.3.1. Imputation by use of mean

As Jager et al. have pointed out, real world data often has much missing data. To compensate, imputations can improve the prediction performance both for regression and classification tasks [39]. In Ref. [39] different imputation methods were used and evaluated, imputation by mean/mode worked well as a baseline though dependent on missing ratio and missingness pattern. The aim of this study was not to fine tune for optimal performance, instead to show that KDFE works. Therefore, imputation by mean/mode as imputation method was used in the study. Mean value was used for float values and mode for categorical.

3.3.2. Normalisation of dataset

The ML pipeline included an evaluation of normalisation (yes/no) of the datasets (ORG, ORGFE). For more information see the Appendix.

3.4. Feature selection (FS)

Four feature selection methods were used to select the top five features with the best classification ability for each FS method dataset. FS

---

Fig. 3. Study activity overview.
method datasets are the datasets used as input to the classification models. They are more specific than one for each FS method. Instead, the granularity is FS method, normalisation (yes/no) and feature group (ORG/ORGFE). The FS implementation was done in a python notebook. We used the following FS methods, cf. Table 1. For more information see the Appendix and [40].

3.5. Evaluation of selected features

To evaluate the performance of the FS methods, classification metrics were collected by implementation in a python notebook [40]. This study is similar to Ref. [34], using medical data from EHR for classification, therefore we also use the AUROC as primary performance evaluation metric.

The classification was performed by splitting the datasets into a train and a test subset with ratio 0.7. In [41] a ratio of 1:1, 1:2 and “no SMOTE” was evaluated, the results showed small differences depending on studied measure. Based on best practice from an experience report we used SMOTE with ratio 1:2 in this study [42].

Four different FS methods were used according to section 2.7, all with and without normalisation. Classification models can give different results with the same input dataset. Therefore, five different models were used, and the relevant metrics were collected. The models were: Logistic regression, Random forest, Gaussian Naives Bayes (Gaussian NB), Gradient Boosting (XGB) and Artificial neural net (Neuralnet). Grid search with ten outer loops and three inner loops on each parameter search space was used for optimal hyperparameter settings. Search space can be found in the Appendix with more details of the used models. As seen in Fig. 4 a second unnecessary train/test split was performed, the original intension was to perform an extra validation of the succeeding subprocesses SMOTE, and nested cross validation.

3.6. Performance metrics

The feature selection process used different FS methods. The
selection resulted in a list of top five features with the highest classification ability. The FS methods have their own methods to select the best features. The used metric for classification was AUROC for the FS datasets, the implemented metrics for each model in python/Sklearn were used. For more details about AUROC see Ref. [43] and for Sklearn [44]. Since the ML pipeline included ten outer grid search loops, four FS methods and two normalisation modes (yes/no) a total of 80 sets of metrics for each model were collected. The corresponding numbers for each FS method was 100 due to fewer FS methods compared to classification models.

3.8. Ethical considerations

The research was approved by the Regional Ethical Review Board in Linköping (no. 2018/154-31 and 2013/404-31). Informed consent of participants was not required as the retrospective studies did not affect the healthcare of included patients. Social security numbers (Swedish personnummer) were not revealed during the work, instead a pseudonymised code represented each unique patient.

4. Case study and results

The aim of the study was to analyse the classification ability of features engineered by KDFE. The focus was not on individual methods, models, or metrics, but on the KDFE-process. Since datasets from P1 and P2 came from real world data, the number of included individuals in each dataset differed, as well as the number of features. In this study, we performed no analysis of this discrepancy.

Comparison of feature selection and classification ability of ORG and ORGFE datasets were performed. The study focused on the ability of the KDFE process to improve classification accuracy.

To measure and evaluate KDFE four feature selection methods were used, as described in the Methodology section. Five classification models were applied to the FS datasets, and metrics were collected. Selected features were not ranked by their importance in either the python notebook experiments or in included graphics. Analyses were performed and classification metrics for baseline (ORG) versus KDFE (ORGFE) for each project (P1, P2) were compared.

4.1. Datasets in the case study

For identification of the datasets in the study the following nomenclature was used. Datasets started with the medical research project ID (P1, P2) followed by a suffix, “-ORG” for baseline features, “-FE” for engineered features and “-ORGFE” for the evolved features (ORGFE = ORG ∪ FE).

The structure and relations of the collected data from the DWH used in the two medical research projects were stored in the structure according to Fig. 5. The patient table with one patient per row is connected to a care event table with all included care events, including the columns Patient ID, Registration date, Event type ID and Value (date, decimal, char) depending on the type of event. For filtering and handling of metadata a table with all event types was used.

4.1.1. Medical research project 1 (P1)

The medical research question was to identify factors that made people fall and lead to fractures. The approach was related to blood HbA1c concentrations, osteoporosis, gender, age, and bone preserving medication. HbA1c stands for blood glycohemoglobin test. There were 82,742 individuals included from the Region of Kalmar län in Sweden. The ratio of the target class was 0.5, since it was a case control study.

The number of features in P1 were P1-ORG = 11, P1-FE = 26 and P1-ORGFE = 37. The described individual features can be found in the Appendix.

4.1.2. Medical research project 2 (P2)

The cohort was patients who were prescribed antiepileptic drugs (AED). Known side effects were bone structure defects that lead to fractures and osteoporosis. In this study, healthcare events related to the medication of AED and their negative effects were collected. The focus was on the type of AED; all AEDs were divided into two groups (inducing/non-inducing) by their known negative effect on bone status.

The study generated clinical values regarding risk assessment for osteoporosis for patients with antiepileptic medications. Drugs belonging to “inducing-AED” had a statistically negative side effect on bone mineralization, while the “non-inducing AED” did not [45]. 23,396 individuals were included from the Region of Kalmar län in Sweden. The ratio of the target class was 0.145.

The number of features in P2 were P2-ORG = 12, P2-FE = 42 and P2-ORGFE = 54. The individual features and a description can be found in the Appendix.

4.2. Results of feature selection

A summary of all selected features for each project, regardless of feature selection method and normalisation (yes/no) is shown in Fig. 6. For more details about the feature selection methods and normalisation, see the Methodology section. One obvious result is that engineered features (FE) are more often selected than baseline features (ORG) when feature selection is performed on the two evolved datasets (P1-ORGFE, P2-ORGFE)

4.3. Classification metrics

Classification metrics for (i) all models, (ii) all FS methods, (iii) w/wo normalisation, (iv) feature group (ORG/ORGFE) were collected and analysed. Collected metrics are presented in tables. In the box plots below the following are depicted, (i) average value represented as a ‘x’, (ii) median values represented as a small line, (iii) start of first quartiles and end of third represented as the filled box, (iv) max/min values represented as whiskers and, (v) outliers represented as dots outside the filled box.

Several classification metrics from the test dataset on (i) feature

![Fig. 5. Database relational scheme of clinical research database.](image-url)
The collected scores for several metrics versus KDFE are depicted in Table 2 for P1 and P2, and in Figs. 7 and 8.

4.3.2. Classification scores (AUROC) for FS methods

Collected AUROC scores for FS methods for baseline versus KDFE are depicted in Table 3 for P1 and P2, and in Figs. 9 and 10.

4.3.3. Classification score (AUROC) for classification models

Collected AUROC scores for classification models for baseline versus KDFE are depicted in Table 4 for P1 and P2, and in Figs. 11 and 12.

4.4. Independent evaluation of FS method “no feature selection”

As can be seen in Table 3 a very good FS method was “No feature selection”, that is, to use all features, which seems a bit odd according to Ref. [17]. To make sure no data leakage was present we performed a small alternative classification experiment on P1 where we used one model with high scores. The same split-ratio (0.7) was used, and since P1 was balanced SMOTE was not needed. The used model was RF (max_depth = 10) which resulted in test AUROC of 1.0.

To visualize the classification ability of the features a SHAP value analysis was performed. As seen in Fig. 13, where FE19 meaning engineered feature #19, the feature had a very high classification ability.

4.5. Grid search for optimal hyperparameters

A grid search was performed for each classification model. Each model was executed 10 × 3 times for each feature selection dataset as described in section 2.9. Optimal hyperparameters for the three inner grid loop iterations were not recorded. This information was, on the other hand, collected for the ten outer loops. Since each outer loop can generate optimal hyperparameters, in theory 1600 optimal hyperparameter settings could exist, distributed over all feature selection datasets. However, many feature selection datasets received the same optimal hyperparameter setting. The number of unique combinations was 561. A sample for project = P1, feature group = ORG, normalisation = Yes, FS-model = KBEST MUTUAL, and model = Logistic regression can be found in Table 5.

5. Related work

As stated in Ref. [46] there are pragmatic mining methods not based on a theoretical model, but many are on a higher level without a description of the details in medical registry research. Our experiment used the CRISP-DM model as a base but described and evaluated the process in more detail from two real world examples.

In [47] an interactive visual system for pattern mining based on temporal data was developed. The system also offered the possibility to bin lab results or other clinical events by known thresholds into more relevant events. Examples are HbA1c normal or HbA1c high [47], also performed feature engineering where healthcare events were binned into new features and discovered knowledge was visualised. Our experiment explored and evaluated binned features and temporal patterns, but on medical data from different contexts and in more complex

---

Table 2

<table>
<thead>
<tr>
<th>METRIC</th>
<th>P1</th>
<th>P1</th>
<th>P1</th>
<th>P2</th>
<th>P2</th>
<th>P2</th>
<th>P1 ORGFE</th>
<th>P1 ORGFE</th>
<th>P2 ORGFE</th>
<th>P2 ORGFE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ORG</td>
<td>ORGFE</td>
<td>ORG</td>
<td>ORG</td>
<td>ORG</td>
<td>ORG</td>
<td>ORG</td>
<td>ORG</td>
<td>ORG</td>
<td>ORG</td>
</tr>
<tr>
<td>NUMBER OF SCORES</td>
<td>400</td>
<td>400</td>
<td>–</td>
<td>400</td>
<td>400</td>
<td>–</td>
<td>400</td>
<td>400</td>
<td>–</td>
<td>400</td>
</tr>
<tr>
<td>AUROC</td>
<td>0.62</td>
<td>0.82</td>
<td>3.9E-79</td>
<td>0.61</td>
<td>0.89</td>
<td>1.4E-91</td>
<td>0.62</td>
<td>0.82</td>
<td>3.9E-79</td>
<td>0.61</td>
</tr>
<tr>
<td>ACCURACY</td>
<td>0.62</td>
<td>0.82</td>
<td>4.4E-73</td>
<td>0.80</td>
<td>0.91</td>
<td>3.1E-84</td>
<td>0.62</td>
<td>0.82</td>
<td>4.4E-73</td>
<td>0.80</td>
</tr>
<tr>
<td>PRECISION</td>
<td>0.64</td>
<td>0.84</td>
<td>5.9E-80</td>
<td>0.31</td>
<td>0.67</td>
<td>1.6E-103</td>
<td>0.64</td>
<td>0.84</td>
<td>5.9E-80</td>
<td>0.31</td>
</tr>
<tr>
<td>RECALL</td>
<td>0.59</td>
<td>0.83</td>
<td>4.7E-50</td>
<td>0.25</td>
<td>0.68</td>
<td>1.4E-85</td>
<td>0.59</td>
<td>0.83</td>
<td>4.7E-50</td>
<td>0.25</td>
</tr>
<tr>
<td>F-1</td>
<td>0.58</td>
<td>0.81</td>
<td>7.7E-61</td>
<td>0.26</td>
<td>0.65</td>
<td>3.1E-89</td>
<td>0.58</td>
<td>0.81</td>
<td>7.7E-61</td>
<td>0.26</td>
</tr>
</tbody>
</table>

Fig. 6. Selected features for P1, P2 and TOTAL (P1+P2) for all FS methods on P1-ORGFE and P2-ORGFE.
patterns based on domain expertise. As described in Ref. [16] medical longitudinal data has a high potential for informative features.

If data mining is performed with feature engineering in an iterative manner it generates better results [12]. The study presents interesting results but lacks detailed information and statistical evidence that feature engineering benefits from the KDFE process.

[48] describes a process similar to KDFE but without the use of ML methods, probably due to low use of ML at the time of the study design. In a more recent study domain experts are used in close cooperation for the diagnosis and treatment of diabetes [49]. The discovered knowledge was represented in rules. The domain users were involved for basic concepts and to obtain useable and large datasets related to diabetes. Compared to our study no evaluation of used feature selection method was employed, and domain experts were not fully utilised since they did not participate in a KDFE-process.

The involvement of domain experts is also studied with emphasis on its importance [50], but lacks knowledge about how to involve domain experts in the data mining process.

As [16] described, temporal information can generate features with high classification ability, but it is difficult and complex to engineer informative features from such sources. Use of drug medication and diagnostic events to mine temporal-based features for patient phenotyping by use of vector representation can be rewarding [51]. Compared to our study [51] did not evaluate feature selection methods and descriptions of the impact of involving domain users in the feature engineering process.
In [52] temporal patterns based on triplets (lab result - clinical event - lab result) were used for classification models. By comparing datasets of feature engineered data on different classification models the study had many similarities to our study, but with some important differences [52]. does not describe the iterative data mining process leading to more informative features and only a limited context and operators were used. The study used lab results before and after clinical events.

To summarise related work there are many studies that have methodically similar parts or functions as KDFE, but we have found none that uses parts from them all. One of the main aims in this study was to confirm that KDFE is worth using and could be automated. An important aspect of future automation is that other studies have successfully shown that KDFE-like functions are important and produce new knowledge.

6. Discussion

As explained in the introduction, the aim of the study was to justify that a process involving a domain expert, data scientist and statistician (KDFE) is worth utilising. The knowledge discovery process starts with general interest in a medical area based on a research hypothesis. When the process iterates the dataset evolves, grows, and is refined, which leads to more informative features and higher classification performance.

6.1. Explanatory model for the KDFE process

The long-term goal is to provide the medical researcher with augmented computer science knowledge; the data analyst will not be
unemployed but less needed. This augmented intelligence is far from artificial intelligence where the aim is often to construct a high-performing algorithm or model that uses little human assistance.

The aetiology for the KDFE process originates in the gap between the medical researcher’s healthcare perspective and the granularity of the data in EHRs and DWHs. Medical staffs are highly skilled to identify patterns from a variety of sources, and to read subclinical signs, often non-quantitative. These vague assumptions lead to new research questions.

First, these research questions need to be transformed when data is requested from DWH, and secondly the data in the database has a high complexity, resolution, and granularity [1]. The data analyst can help the medical researcher sharpen the initial research question and create more knowledge from the same data.

6.2. Quantitative comparison with inherent baseline

To compare the results from our case study in quantitative terms to other studies is difficult since many parts are unique, and no comparable studies without great bias considerations were found. Gained experience from historically performed knowledge discovery projects led to the experimental setup using research projects (P1 and P2) with an inherent baseline (P1-ORG, P2-ORG). Baselines (or the baseline if singular) were compared to the evolved dataset (P1-ORGFE, P2-ORGFE) to see how KDFE performs in comparison to an ordinary medical registry project. Since the experiment was carried out within the same project bias was minimised.

6.3. Does the KDFE process create value (Q1)?

As seen in Fig. 6 the feature selection methods select the engineered features (FE) to a much greater extent than the baseline features (ORG). But is it worth investing time and effort in feature selection methods? From a knowledge discovery perspective our results show that feature engineering is worth the additional effort since the engineered features were selected more often.

The evolved datasets ORGFE showed a much higher classification accuracy (AUROC) compared to the baseline datasets ORG. As shown in Table 4, the AUROC scores for logistic regression, random forest, Gaussian Naive Bayes, XGBoost, and neural network models are significantly higher for the evolved datasets compared to the baseline datasets.

In Fig. 11, the box plots for P1 show that the AUROC scores for logistic regression, random forest, Gaussian Naive Bayes, XGBoost, and neural network models are significantly higher for the evolved datasets compared to the baseline datasets.

In Fig. 12, the box plots for P2 show that the AUROC scores for logistic regression, random forest, Gaussian Naive Bayes, XGBoost, and neural network models are significantly higher for the evolved datasets compared to the baseline datasets.

### Table 4

<table>
<thead>
<tr>
<th>Classification</th>
<th>P1</th>
<th>P1</th>
<th>P1</th>
<th>P2</th>
<th>P2</th>
<th>P2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>ORG ORGFE</td>
<td>ORGFE – ORG</td>
<td>ORGFE – ORG</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average AUROC</td>
<td>0.57 0.84</td>
<td>6.9E-26 0.59</td>
<td>0.73 6.7E-9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>0.68 0.88</td>
<td>5.6E-23 0.55</td>
<td>0.88 3.1E-28</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Random Forest</td>
<td>0.58 0.83</td>
<td>3.5E-24 0.59</td>
<td>0.84 4.0E-26</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gaussian NB</td>
<td>0.68 0.88</td>
<td>9.4E-23 0.57</td>
<td>0.88 2.0E-26</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>XGB</td>
<td>0.59 0.65</td>
<td>8.6E-03 0.57</td>
<td>0.74 1.6E-15</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NeuralNet</td>
<td>0.57 0.65</td>
<td>8.6E-03 0.57</td>
<td>0.74 1.6E-15</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Fig. 11. Box plot of test scores for P1 by classification model.

Fig. 12. Box plot of test scores for P2 by classification model.
6.3.2. Is the KDFE process of value to the researcher?

The feature engineered datasets (FE) comprise more knowledge than the datasets with the baseline features (ORG). For this question as well, the statistical analysis shows strong support to reject the no difference in the analysed classification metric (see Tables 2–4). However, the gained knowledge must be placed in relation to the amount of time spent on the process by the researcher.

The true value of the discovered knowledge to the researcher using the KDFE process cannot be described in quantitative terms. On the one hand, the KDFE process can help the researcher if a low enough p-value for support of the initial research question is provided, which could result in a publishable paper. It is not to run the statistical analysis with exactly the same research question and hope for better sampling and a lower p-value as described in Ref. [58]. Instead, a more specified dataset is generated with better utilisation of the potential in the collected medical raw data. This transition can be analysed if classification scores for baseline features (ORG) compared to engineered features (FE) for each project are compared.

On the other end we have the computer scientist that can be very pleased with small enhancements regarding performance or classification scores based on models with $>100$ features without direct reference to daily healthcare, even if the results are difficult to interpret.

This conflict can manifest itself as follows: Apparent new knowledge with a great quantitative outcome will be useless to domain experts if the results have no direct use in regular work, or is very hard to explain or trust [32]. To a medical researcher discovered knowledge must be understandable and trustworthy, otherwise it is not discoverable = useable knowledge. Rather, it is noise that drowns out other more useful results.

6.3.3. Quantitative improvements

A central issue for knowledge discovery is to define a measure for discovered knowledge that takes both sides into consideration. A measure must include both quantitative and qualitative components. In this study we prove that we obtain quantitative improvements, at least from a computer scientific perspective. If a KDFE process is used it will result in more effective knowledge discovery.

Qualitative aspects, like usefulness, transparency and interpretability of the discovered knowledge needs to be analysed. That is beyond the scope of this study, but a data structural foundation has been established that will deal with qualitative issues (see sections Conclusion and Can the KDFE-process be automated (Q2) for future work).

6.4. Can the KDFE process be automated (Q2)?

The described KDFE process is time consuming, though we proved that it adds quantitative value to the researcher. If results and gained experience were collected from this study a knowledge foundation based on feature store [14] could be created that would support the development of an automated KDFE process. Qualitative aspects will be central in the design of the knowledge database that automated KDFE uses.

### Table 5

<table>
<thead>
<tr>
<th>PROJECT</th>
<th>FEATURE GROUP</th>
<th>FS METHOD</th>
<th>NORMALISATION</th>
<th>MODEL</th>
<th>OPTIMAL HYPER PARAMETERS</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>ORG</td>
<td>KBestMutual Yes</td>
<td>Logistic regression</td>
<td>$C = 0.31622776601683794$, max_iter $= 1000$, penalty $= ‘l1’$, solver $= ‘liblinear’$</td>
<td></td>
</tr>
<tr>
<td>P1</td>
<td>ORG</td>
<td>KBestMutual Yes</td>
<td>Logistic regression</td>
<td>$C = 0.31622776601683794$, max_iter $= 1000$, penalty $= ‘l2’$, solver $= ‘liblinear’$</td>
<td></td>
</tr>
<tr>
<td>P1</td>
<td>ORG</td>
<td>KBestMutual Yes</td>
<td>Logistic regression</td>
<td>$C = 10000.0$, max_iter $= 1000$, penalty $= ‘l1’$, solver $= ‘liblinear’$</td>
<td></td>
</tr>
<tr>
<td>P1</td>
<td>ORG</td>
<td>KBestMutual Yes</td>
<td>Logistic regression</td>
<td>$C = 56.234432431903491$, max_iter $= 1000$, penalty $= ‘l1’$, solver $= ‘liblinear’$</td>
<td></td>
</tr>
<tr>
<td>P1</td>
<td>ORG</td>
<td>KBestMutual Yes</td>
<td>Logistic regression</td>
<td>$C = 56.234432431903491$, max_iter $= 1000$, penalty $= ‘l2’$, solver $= ‘liblinear’$</td>
<td></td>
</tr>
<tr>
<td>P1</td>
<td>ORG</td>
<td>KBestMutual Yes</td>
<td>Logistic regression</td>
<td>$C = 56.234432431903491$, max_iter $= 1000$, penalty $= ‘l1’$, solver $= ‘liblinear’$</td>
<td></td>
</tr>
<tr>
<td>P1</td>
<td>ORG</td>
<td>KBestMutual Yes</td>
<td>Logistic regression</td>
<td>$C = 56.234432431903491$, max_iter $= 1000$, penalty $= ‘l1’$, solver $= ‘liblinear’$</td>
<td></td>
</tr>
</tbody>
</table>
6.5. Limitations

This is a multiple embedded case study that investigates specific subareas of two medical research projects, outside a holistic perspective. Issues that are only briefly accounted for are imputation methods, dataset normalisation, running times and performances of individual FS methods or classification scores.

The focus of this study was on the answers to the questions stated in the introduction: Q1 (Does the iterative data mining process involving a medical researcher and a data analyst benefit from the engineering of new features based (KDFE)?) and Q2 (Is it possible to automate the KDFE-process?).

The cost of resources in the KDFE process cannot easily be measured and are highly individual. Therefore, that aspect was excluded in this case study. Our goal was to prove that the KDFE process creates new or deeper knowledge. We use classification ability (AUROC) to represent discovered knowledge.

6.5.1. High classification scores

Some classification methods for P1 resulted in very high AUROC (0.99–1.00). This was especially the case when the FS method “No feature selection” was used. The results indicate that there was an overfitting problem or information leakage in the used ML pipeline. To investigate the finding an alternative independent ML pipeline was implemented displaying the same high result. See section 4.4. Still, not all classification models for P1 and FS method = ”No feature selection” datasets resulted in a very high AUROC, which indicated there was no systematic data leakage. Instead, it showed how much ML models can vary with the same input. If more data was used for the model training the models should probably have found this strong relation. One can argue that FE19 should have been removed to get a better evaluation of FS methods and classification models, but that was not the main goal of the study. The intention was to run an experiment as similar to a real medical registry research project as possible.

Since this study did not focus on classification performance, instead the KDFE process, features with high strong dependency and weak classification models, are not excluded. By presenting average and median values in figures regarding test AUROC scores the effect of outliers can be analysed in box plots in section 3. The statistical analysis, including both high and low scores, showed that engineered features resulted in higher classification scores (p-values << 0.001).

7. Conclusion

As reviewed and dissected in previous sections there is considerable published material regarding feature engineering in the field of medical research, but to our knowledge there is no study that by quantitative measures proves that KDFE creates knowledge by comparing it with an inherent baseline. This study, based on real world medical data, provides a clear answer to the question. “Does the iterative data mining process involving a medical researcher and a data analyst benefit from the engineering of new features based (KDFE)?” – The answer is in the affirmative.

The described KDFE process produces quantitative knowledge. KDFE also generates medical assessments and clinical value, which was proved in an earlier medical study within the domain of osteoporosis treatment.

Researchers uncertain about whether to invest effort and resources in advanced data mining and feature engineering or not can now see that there is great potential. Individuals either in healthcare or computer science can use the results from this study to better understand the medical research process from a knowledge-discovery perspective.

Humans-in-the-loop in ML projects are important, produce more knowledge, but are time consuming. This study has explored the interaction between domain experts and data scientists, which results in greater understanding and insights.

The different findings from the study can now be combined to construct the foundation for an automatic KDFE that minimises the involvement of domain experts and at the same time maximises discovered knowledge.

7.1. Improvements

By not using the test dataset from the second split prior to SMOTE, data was thrown away. However, the result still supported the aims of the study. If all potential data was used even better results could have been achieved.

Even if we have shown that the KDFE process should be used there are aspects that need to be investigated and further developed to optimise its value to the researcher.

Explainability, interpretability and qualitative aspects of knowledge discovery are important when new knowledge and technology should be implemented in daily healthcare. Which kind of metadata and features can provide both a high level of knowledge discovery and be accepted by daily healthcare operations?

Apart from the higher classification ability, KDFE must be cost effective for the researcher in a wider context. Trust and support from an XAI perspective must be considered as well as invested time and resources. Boosting trust and supporting more metrics can be added to discovered knowledge. To best learn from performed experiments, domain-oriented features and the machine learning process metadata can be collected in a feature store framework and feed back into the KDFE process. As an examples, here is metadata for feature FE3 - “Latest HbA1c value before fracture” (project P1).

- Temporal-type: Latest
- Information domain: Lab analysis
- Measure type: Decimal
- Vital parameter: HbA1c
- Main diagnosis: Diabetes
- Physiology: Endocrine
- Organ: Liver
- Positive influencers: Food, exercise, weight, and age
- Value-stability: Medium (does not change too much over time)

If detailed metadata are collected and stored in a knowledge database, a solution for knowledge prospecting can be implemented. The knowledge prospect solution should serve as guidance or augmentation to researchers regarding features or information areas with a high potential for new knowledge based on collected metrics. For example, a heat map can visualize undiscovered knowledge represented by metadata that can be used for ideation of new informative features.

A natural expansion is to use more feature engineering techniques such as temporal components or intrinsic feature information such as statistical information or information entropy [16,47]. The most valuable improvement is an automation of the KDFE process, meaning that an algorithm will do the feature engineering, knowledge prospecting and discovery based on learned knowledge from this study as well as collected experience. Results from automatic KDFE could be benchmarked against established autoML solutions.

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.

Acknowledgements

This research received no specific grants from funding agencies in the public, commercial, or not-for-profit sectors.
Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ibmed.2023.100110.

References