# Secondary Publication




**Finze, Nikola; Hoffmann, Philipp; Heinzl, Armin; u. a.**

## Supporting the Billing Process in Outpatient Medical Care : Automated Medical Coding Through Machine Learning

Date of secondary publication: 20.10.2026

Version of Record (Published Version), Conferenceobject

Persistent identifier: urn:nbn:de:bvb:473-irb-117658x

**Primary publication**

Finze, Nikola; Hoffmann, Philipp; Heinzl, Armin; u. a. (2022): Supporting the Billing Process in Outpatient Medical Care: Automated Medical Coding Through Machine Learning, in: Proceedings of the 30th European Conference on Information Systems (ECIS), Timișoara, Romania, Association for Information Systems, pp. 1–17

**Legal Notice**

This work is protected by copyright and/or the indication of a licence. You are free to use this work in any way permitted by the copyright and/or the licence that applies to your usage. For other uses, you must obtain permission from the rights-holders.

This document is made available under a Creative Commons license.




The license information is available online:
https://creativecommons.org/licenses/by/4.0/legalcode

# SUPPORTING THE BILLING PROCESS IN OUTPATIENT MEDICAL CARE: AUTOMATED MEDICAL CODING THROUGH MACHINE LEARNING

*Research Paper*

Luis Oberste, University of Mannheim, Germany, oberste@uni-mannheim.de

Nikola Finze, DaWaVision GmbH, Germany, nikola@finze.eu

Philipp Hoffmann, University of Mannheim, Germany, hoffmann@uni-mannheim.de

Armin Heinzl, University of Mannheim, Germany, heinzl@uni-mannheim.de

## Abstract

*Reimbursement in medical care implies significant administrative effort for medical staff. To bill the treatments or services provided, diagnosis and treatment codes must be assigned to patient records using standardized healthcare classification systems, which is a time-consuming and error-prone task. In contrast to ICD diagnosis codes used in most countries for inpatient care reimbursement, outpatient medical care often involves different reimbursement schemes. Following the Action Design Research methodology, we developed an NLP-based machine learning artifact in close collaboration with a general practitioner's office in Germany, leveraging a dataset of over 5,600 patients with more than 63,000 billing codes. For the code prediction of most problematic treatments as well as a complete code prediction task, we achieved F1-scores of 93.60 % and 78.22 %, respectively. Throughout three iterations, we derived five meta requirements leading to three design principles for an automated coding system to support the reimbursement of outpatient medical care.*

*Keywords: Action Design Research, Deep Learning, Medical Code Prediction, Medical Informatics, Natural Language Processing, Outpatient Care, Reimbursement*

## 1 Introduction

Reimbursement in medical care requires significant administrative effort for the medical personnel: On the one hand, it demands detailed documentation of the patients' medical history, their physical examination, and the treatments provided. On the other hand, the billing of these services involves assigning fine granular diagnosis and treatment codes to the patient records using standardized healthcare classification systems. Those coding systems are mostly based on a fee schedule that disaggregates the content and fees of various medical treatments and services (Homeyer et al., 2021).

For inpatient clinical care, reimbursement and financing in a fixed-fee-per-case model have been internationally widely adopted to explicitly encourage greater cost efficiency (Busse et al., 2011). In this context, payments are based on fixed prices per unit of activity according to patient characteristics. To specify the latter, country-specific modifications of ICD (International Statistical Classification of Diseases and Related Health Problems), a global standard for diagnostic health information maintained by the World Health Organization, are often used.

In contrast, reimbursement schemes in outpatient care often rely on cost-based payment models that incorporate every single medical treatment and service provided by the physician, i.e., the costs incurred by each patient. In Germany, statutory health insurance and private health insurance use

different reimbursement schemes for outpatient care that are based on the German Uniform Assessment Standard (EBM) and German Fee Schedule for Physicians, respectively. Comparably, similar payment models in outpatient care are used in other healthcare systems (e.g., TARMED in Switzerland or HCPCS in the US).

Various deep learning, natural language processing (NLP), and other machine learning (ML) techniques have been deployed to overcome inefficient manual coding by automatically generating codes (e.g., Jatunarapit et al., 2016; Li et al., 2019). Existing research on automated code assignment based on medical documentation has been mostly conducted on ICD codes, becoming a research topic on automated decision support in healthcare and, therefore, a flashpoint in medical AI research. Many promising results have been demonstrated by predicting ICD codes based on free-text diagnosis reports (e.g., Li et al., 2019) to assist human coders in the assignment of correct codes.

In contrast to ICD, the structure and steps of outpatient care billing differ. In Germany, certain EBM codes must be mutually exclusively assigned, include variants based on patient characteristics, or differentiate holiday and Sunday supplements. Moreover, the amount of EBM codes per patient is considerably higher for each visit to a general practitioner compared to ICD codes. Thus, the billing process for health services is challenging in practice, since it demands extra time for the medical personnel during and after each patient visit, as well as at the end of every day and quarter. Thereby, even basic fixed rates for a doctor's visit vary from patient to patient and have to be individually specified (Neumann-Grutzeck, 2020). Hence, setting up each billing list becomes a *time-consuming* and *error-prone* task. To the authors' best knowledge, ML-based EBM coding for the billing with German statutory health insurance has not been investigated in scientific work.

To consider the lack of ML research on EBM coding and to research how the context of cost-based outpatient billing shapes the development of an automatic coding system, we draw on an Action Design Research (ADR) project (Sein et al., 2011) with a general practitioner's office located in Germany, totaling a dataset associated with over 5,600 patients. Our research can be summarized by the following research question:

***How can NLP-based ML techniques be utilized for designing an automated coding system for outpatient care to support the general practitioner's office's billing process*?**

Thereby, we aim to (1) expose the benefits and limits of applying ML to automated EBM coding for outpatient billing, (2) deliver value for a research case of a statutory health insurance-accredited general practitioner by designing and developing a useful artifact for the billing process that supports automated code generation as well as (3) generalize our findings into design principles.

With this research paper, we aim to contribute to the literature on ML-based health information systems by considering the context of outpatient medical billing and the characteristics of the national healthcare system. Secondly, we shed light on the design of a system that ascertains state-of-the-art predictive performance for EBM coding but is also useful for end-users in the billing process.

# 2 Background and Related Work

Since reimbursement schemes in medicine vary all over the world, we briefly discuss the background of medical billing to highlight the national and international characteristics of the problem. Moreover, we present a synopsis of related scientific work on extracting medical codes to align the artifact design with state-of-the-art developments in ML-based medical code prediction.

## 2.1 Medical billing

Transforming medical records into standardized diagnosis codes serves as a foundation for insurance billing procedures and allows for better comparability, essential for worldwide health transparency and epidemiology (Janssen and Kunst, 2005; Névéol et al., 2017; Yan et al., 2010).

Reimbursement systems for *inpatient* care around the world often use fixed-fee-per-case classification systems, for which the coding of diagnoses and procedures is important since it delivers the basis for

the definition of patient groups and billing of medical treatments and services. ICD, although often used with country-specific modifications, is an important international standard that is widely used for *diagnosis* codes (Quan et al., 2005). As a result, the accuracy and efficiency of ICD coding have always been crucial for clinical practice (Yu et al., 2019). Until today, human coders in hospitals review each patient's medical history, including the diagnosis and further free-text medical notes, and add the respective ICD codes manually. To assist the manual assignment, automated ICD coding has become a fundamental task in clinical data mining.

In contrast, reimbursement systems differ for *outpatient* care (Busse et al., 2011). For example, medical services are provided by around 148,000 private practitioners in Germany, which are members of a regional health insurance association by law. For statutorily assured patients, the physicians receive reimbursement from their respective association, being solely allowed to invoice treatments and services that appear in the EBM coding catalog. This coding catalog is based on the mix of treatments and services delivered, comparable to other European healthcare systems (e.g., TARMED model in Switzerland). Similar to ICD coding, EBM coding depends on the patient's medical documentation and is a highly time-consuming, laborious, and error-prone task that requires medical knowledge and know-how of coding rules.

## 2.2 Medical code prediction

To improve the quality and accuracy of the ICD coding process, on the one hand, and to lower healthcare expenditures, on the other hand, researchers all over the world have developed methods for computational ICD coding for more than three decades (Kaur and Ginige, 2018). Typically, the medical code prediction task is formulated as a classification problem. In multi-label classification, multiple labels from a set of potential, distinct medical codes are assigned to each hospital visit. Multi-class problems in this context can be found for specific medical questions, e.g., predicting the primary cause of death according to ICD-10 (Mujtaba et al., 2017).

In many cases, the resulting annotation task represents a large-scale multi-label classification problem, resulting from a high number of labels per instance and an extremely large label set, e.g., over 68,000 ICD-10 codes (Atutxa et al., 2018; Baumel et al., 2018). Consequently, previous work on the multi-label classification of data from electronic health records (EHR) focused on a relevant subset of codes. Xu et al. (2019), for instance, selected the 32 most representative ICD-10 codes. Other authors abstracted single medical codes using the coding scheme to significantly collapse them into their groups (Eslami et al., 2020; Zweigenbaum and Lavergne, 2016). Large-scale multi-label classification, however, has primarily been of scientific interest outside the healthcare domain (Baumel et al., 2018).

As input for the multi-label classification, extant work approaches structured, semi-structured, or unstructured data. Structured data can include patient demographics, laboratory results, and vital signs, but also past ICD codes which are used to predict future ICD codes. Semi-structured data occurs, for instance, in description phrases written by physicians to a particular diagnosis (Xu et al., 2019). Extant research on ICD coding utilized unstructured text data contained in diagnostic reports (e.g., Goldstein et al., 2007; Crammer et al., 2007; Farkas and Szarvas, 2008), discharge summaries (e.g., Baumel et al., 2018; Xu et al., 2019), or death certificates (e.g., Zweigenbaum and Lavergne, 2017; Seva et al., 2017). To transform the text into a structured representation to be fed into ML-based classifiers, bag-of-words feature extraction was commonly applied. Bigrams and trigrams were also included for contiguous medical terms. In summary, medical text classification results in complex NLP problems due to multilingualism and long documents (e.g., discharge summaries with over 1,900 words), while being written under time pressure (Atutxa et al., 2018; Baumel et al., 2018).

Historically, rule-based or dictionary-based classifiers are applied within automated or semiautomated assignments of ICD codes. As these require manual, time-consuming crafting which also limits them in capturing synonyms (Kaur and Ginige, 2018; Kavuluru et al., 2013), supervised ML excelled at the large-scale processing of medical text corpora. Influential initiatives such as the Cross-Language Evaluation Forum hosted several shared clinical information extraction tasks for medical document analysis (Kelly et al., 2014). As a result, support vector machine (SVM)-based classifiers were

established as the most performant and extensively used ones throughout ICD coding studies. In this regard, Perotte et al. (2014) proposed to leverage the hierarchy of ICD codes to train per-label SVM classifiers. In a top-down fashion, the classifiers were applied in case its parent code has been predicted. Marafino et al. (2014) trained binary SVM classifiers for four procedures and diagnoses individually, that are associated with a higher risk of rehospitalization. Koopman et al. (2015) trained binary SVM classifiers in a cascaded architecture, i.e., the first one determined the presence of cancer and multiple ones on the second level identified the type of cancer according to ICD codes. Zweigenbaum and Lavergne (2016) investigated hybrid classification, combining multi-class ML and dictionary projection so that others carried on the additional use of lexical resources (Ebersbach et al., 2017; Zweigenbaum and Lavergne, 2017).

During the last years, deep learning has been established as an important part of text classification methods and has begun to surpass traditional ML-based algorithms in several tasks (Minaee et al., 2020). Recurrent Neural Network-based models which learn representations of sequential data for word-to-code or code-to-code learning were successfully applied (Catling et al., 2018; Duarte et al., 2018; Miftakhutdinov and Tutubalina, 2017). Yet, a dominant deep learning architecture known from computer vision and speech recognition are Convolutional Neural Networks (CNN), which have been applied to character-level text classification by Zhang et al. (2015). Karimi et al. (2017) assessed different CNN models for ICD code generation based on radiology reports and achieved superior performance to SVMs. Mullenbach et al. (2018) employed a CNN for ICD code prediction based on discharge summaries and applied a per-label attention mechanism so that the model could be explained based on the most relevant text segments for each code.

### 2.3 Synthesizing design knowledge of code prediction for medical billing

Based on the findings from the literature, it becomes evident that major improvements have been made in supporting ICD-related coding in inpatient medical care through machine and deep learning. For the design of ML-based coding systems, a data mining pipeline involves effortful data cleansing and preprocessing before applying classification algorithms (Burkul et al., 2020). These steps come with a range of crucial healthcare-related challenges, including the availability of patient data, proprietary and heterogeneous data forms, as well as data anonymization (Pumplun et al., 2021). Unstructured or semi-structured forms of underlying data, which is the case in medical text fields, require NLP-based techniques to extract features, such as bag-of-words and n-grams, or novel deep learning architectures, such as CNN. By applying a character-level CNN for multi-class classification, abnormal character combinations like misspellings can be learned (Hausmann, 2018). CNNs can outperform classification models for automated coding of medical text, whereas SVMs still serve as state-of-the-art baselines throughout many contributions in the ICD coding literature (Li et al., 2019; Xie et al., 2019), whereas logistic regression and stochastic gradient descent are commonly studied for other ML-based text classification tasks (Muneer and Fati, 2020). Regarding the definition of the target space in medical code prediction, the problematic number of labels was stressed which consequently needs to be considered during automated coding systems design. However, while machine and deep learning show promising results in automated medical code prediction based on ICD codes, we could not find any research on non-ICD-related coding settings such as German outpatient medical care. Hence, our work on automated EBM coding will be informed by the challenges, the ML pipeline, extraction and preprocessing techniques, as well as applied algorithms available in ICD-related research.

## 3 Research Approach

### 3.1 Conceptual approach

Designing and evaluating artifacts have been parts of the Information Systems research discipline since its beginnings (Peffers et al., 2018) to address and solve heretofore unsolved organizational problems (Hevner et al., 2004). We followed ADR by Sein et al. (2011). In contrast to other Design

Science Research (DSR) genres like Design Science Research Methodology (Peffers et al., 2007) or Design-oriented IS research (Winter, 2008; Österle et al., 2011), where the evaluation of each design cycle is relegated to a downstream phase after artifact construction, the evaluation in ADR takes place as a continuous interaction between researchers and practitioners throughout each cycle (Sein et al., 2011).

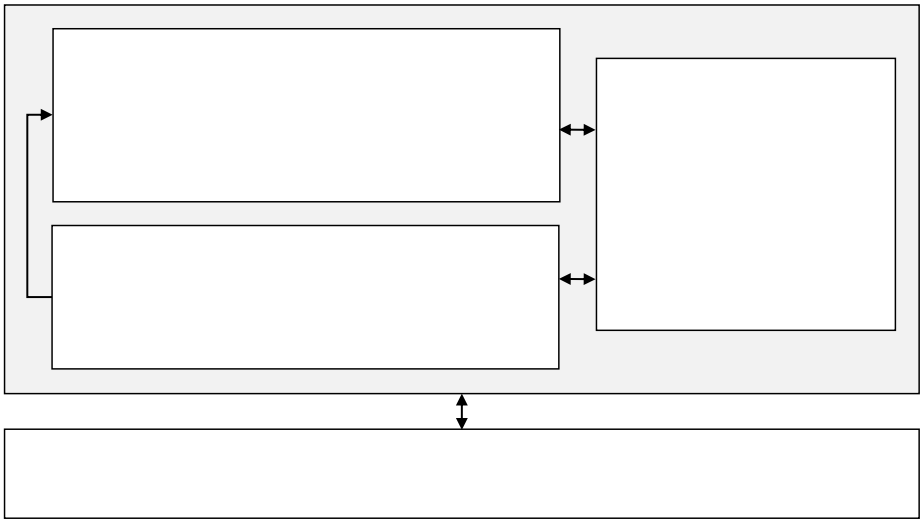


*Figure 1. Action design research process (adapted from Sein et al., 2011).*

ADR includes the same important elements of the DSR methodology (Figure 1), which "emphasizes the design and construction of applicable artifacts" (Peffers et al., 2007, p. 131) and builds in its core on a continuous 'Building, Intervention, and Evaluation' (stage 2). It is accompanied by a preceding 'Problem Formulation' (stage 1) as well as paralleled by the 'Reflection and Learning' stage (stage 3). Lastly, 'Formalization of Learning' (stage 4) abstracts artifact-related results to generalized outcomes.

Considering the lack of ML research on EBM coding, the ADR methodology allowed us to study the organizational context of a general practitioner's office and how this affects the development and use of an automated EBM coding system. At the same time, we could incorporate stakeholders and end-users in the research process to design and evaluate the artifact concurrently in interaction. In combination with the support of ML-based medical coding literature, it let us produce theoretical outputs and derive prescriptive design knowledge in the form of design principles, generalized to cost-based billing of outpatient care.

## 3.2 Problem formulation

Following the structure of the ADR approach, we first formulated the problem. Derived from a real-world problem observed within a general practitioner's office, our case serves as an instantiation of the generalization, which relates to a broader class of problems.

Following Sein et al. (2011), the problem formulation includes the identification and conceptualization of the research object through ingraining two main principles: practice-inspired research and a theory-ingrained artifact. To conduct practice-inspired research, we leverage a knowledge-creation opportunity at the intersection of the technological and organizational domain of the billing process for outpatient medical care at a German general practitioner's office. Therefore, the project was conducted in close collaboration between the team of the general practitioner's office including the physician (practitioner within the ADR team) and the medical personnel (end-users) as well as the authors of the paper (researchers). To permeate the problem, job shadowing in combination with semi-structured interviews was conducted by the research team, such that guiding questions concerning time-consuming administrative tasks, difficulties related to those tasks, as well as key resources and activities to manage the tasks and their difficulties were asked after shadowing several patient

treatments. Thereby, the interviews allowed for open questions to let the participants come up with new ideas (Yin, 2018).

To ensure the foundation in scientific knowledge for a theory-ingrained artifact, literature about different approaches to ML-based automated ICD coding was reviewed. We conducted a systematic search following a structured approach of Webster and Watson (2002). The search was performed on PubMed, ScienceDirect, ResearchGate, AIS eLibrary, and Google Scholar, using the terms 'ICD code', 'ICD classification', 'diagnosis codes', 'medical coding', 'computer-assisted coding', 'machine learning', 'ICD automation', along with syntactical variants and combinations of them. Articles were included in case of computationally approaching medical coding, such as based on ICD. Through this search, 25 relevant papers were identified.

The gathered knowledge helped us to structure the problem, to identify solution possibilities as well as guide the design of our artifact (Sein et al., 2011). This included an understanding of the challenges related to harmonization and anonymization of medical data; key aspects in the ML pipeline for medical code prediction, such as determining the feature and target space; extraction and preprocessing techniques including bag-of-words and n-grams; as well as the appropriate algorithms SVM and CNN. Based on the results of the literature review, we were able to confirm that NLP approaches stemming from ICD code assignment provide a promising potential to support the general practitioner's billing process. Moreover, this helped us to ensure that upcoming findings derived from our artifact instantiation can be applied to the class of medical outpatient billing problems rather than the specific general practitioner's office studied in this work.

## 3.3 Building, intervention, and evaluation

By leveraging the acquired knowledge of the first stage, the second one is an iterative phase including building the artifact By the research team, intervention by the physician, as well as evaluating not only the artifact but also the problem itself (Sein et al., 2011). Since our approach focused on innovative technological design rather than innovation through organizational intervention, we found ourselves on the IT-dominant side of the research design continuum of ADR. However, we also leveraged one key element of the organization-dominant BIE (building, intervention, and evaluation) research design in that we additionally gathered feedback from the end-users (medical assistants) for further ongoing refinement. This was due to time constraints and the availability of the physician but in contrast to a distinct organization-dominant BIE research design, we did not deploy the artifact in the organization during early design iterations (Sein et al., 2011). Based on initial job shadowing and interviews in the previous stage, we developed a conceptual prototype in January 2021, which we refined during the first iteration. We concluded this BIE cycle with an on-side presentation at the general practitioner's office. The second BIE cycle took place between February and March 2021. We closed this BIE cycle with a small field test (alpha version) at the general practitioner's office with the medical personnel involved. During the third BIE cycle in April and May 2021, we addressed the field test findings and adjusted the artifact accordingly. Next, we presented a working interface in a second field test to the general practitioner (beta version).

## 3.4 Reflection and learning

The reflection and learning stage parallels the first two stages. According to Sein et al. (2011), it ensures to move the learnings from building a solution for a particular instance towards a broader class of problems through conscious reflection on the problem and the artifact. Thus, through 'guided emergence', the relevant contributions to knowledge have been identified (Sein et al., 2011).

From our literature review, we derived design knowledge to incorporate the most recent body of scientific ML approaches as well as the findings regarding different medical coding systems for billing. The constant comparison of our artifact towards those two literature streams helped us to embody our solution within the generalized class of the problem. Moreover, we iteratively evaluated the artifact within a field test at the general practitioner's office with the corresponding medical personnel to guide us for the next BIE cycle. We leveraged the constant reflections to abstract case-

specific learnings into meta requirements (cf. Walls et al., 2004; Walls et al., 1992), acting as generic requirements from theorized attributes of the artifact, to derive design principles in turn (in stage 4).

### 3.5 Formalization of learning

After three BIE cycles, we formalized our findings from the previously derived meta requirements. As suggested by Sein et al. (2011), we conceptually moved from the 'specific-and-unique' to the 'generic-and-abstract' by three levels of generalization: After generalizing the problem space already in the first ADR stage, we secondly generalized the solution space based on the body of existing literature. Thirdly, we derived 'materiality oriented' design principles (Chandra et al., 2015) from the meta requirements inferred from the reflection on findings of our IT-artifact instantiation. Our design principles are presented based on the 'aim of a user in a context' schema by Gregor et al. (2020).

## 4 Results of the ADR Process of Designing an Automated Medical Coding System

### 4.1 EBM coding for the billing process of a general p 's office

Based on the literature review and the practical insights through the job shadowing and the interviews, we gathered important knowledge about the billing process in the general practitioner's office. This knowledge was the foundation for the formulation of the research problem. The findings emphasized the differences in reimbursement coding in outpatient care, in contrast to other medical code predictions (e.g., ICD coding). Those differences helped us to define the problem as an instance of a new class of problems. Based on this, we argue that while having different approaches in different countries, the overall coding procedure remains relatively similar so that we can ensure with our instantiation the generalizability of our findings to a broader class of billing-related problems.

While ICD coding is challenging in hospitals due to a high amount of diagnoses per patient (Rios and Kavuluru, 2013), a patient in a general practitioner's office usually has only one new diagnosis or an already existing long-term diagnosis. EBM coding, on the contrary, was mentioned to be heavily complex. The general practitioner office's billing strongly relies on the correct assignment of EBM codes since it affects the payments by its regional health insurance association. If an EBM code of a treatment was left out, the physician does not get the expense reimbursed. On the other hand, if the physician invoices EBM codes that do not match a patient's treatment, they get fined. Thus, to ensure the correct assignment of the EBM codes, they are assigned by the medical assistants and the physicians themselves before, during, and after the patient treatment. In addition to the daily check, all patients' EBM codes are being checked manually at the end of every quarter before the file is being exported and sent to the respective health insurance association. This usually takes a week to check for the respective employees. For these reasons, the EBM coding process is not only a highly time-consuming and error-prone task but also directly affects profit. This way, the support of ML seems highly beneficial and insights into automated ICD coding research helped gain information about how automated medical coding can be applied as potential solution possibilities (Sein et al., 2011).

### 4.2 Building, intervention, and evaluation

#### 4.2.1 First iteration – conceptual prototype

Based on the results of the problem formulation stage, we decided to explore the EBM coding problem with NLP-based multi-label classification. The first BIE stage was initiated by extant automated ICD coding approaches as a basis for a conceptual prototype. With this prototype, we interviewed medical personnel to gather feedback and find out which EHR data is relevant for assigning EBM codes in practice. Figure 2 depicts the schema of the conceptual EBM coding system prototype.

Relevant information consisted of patient data like age and gender, textual documentation written by the physician during anamnesis, consultation, sonography, and other therapeutical interventions, as well as ICD codes. Given the patient data and textual input, the system was intended to predict all accruing EBM codes to support the whole patient journey in a general practitioner's office. This starts with a patient entering the doctor's office and the import of data from the insurance card, for which already basic EBM codes are assigned. During a consultation, text fields like anamnesis, general notes, and procedure information are filled in by the physician. Next, the patient is either treated by the physician for a sonographic examination or further treated by medical assistants for vaccinations, bandages, or others. Every step within this patient process can but must not include an EBM code being set and thus, we trained the system on all potential EBM codes.

| | Gender | Age | Treatment Day | Text Field | Text Field Content | EBM Codes |
|---|---|---|---|---|---|---|
| Example Data | w | 52 | 20200728 | Impf | 2.Influenza li.OA | 03003, 89111 |
| | m | 63 | 20201018 | ICD | R53 G Ersch"pfung szustand | 03004, 35110 |
| Statistics | w: 57 %<br>m: 43 % | Range: 0–101<br>Mean: 54 | Min: 20200102<br>Max: 20201230 | Most common: Documentation (32%), Note (18%)<br>Avg. # of fields per visit: 2.8 | Avg. # of words: 25 | # of codes: 63,789 (166 unique)<br>Avg. # of codes per visit: 2.5 |
| | # of patients: 5604 | | | | | |

*Figure 2. Details of the dataset with exemplary input data (grey) for automated EBM coding.*

Further evaluating the concept with the medical personnel, we found that groups of EBM codes exist that must be considered at certain points in time during the billing process:

> *"[…] such that we would only use a system that proposes vaccination codes while documenting a vaccination indication, not when documenting a sonographic procedure."*

This was an important insight for us and affected the following design and development iteration strongly. Together with the medical personnel, we collated all relevant EBM code groups together with respective text fields, i.e., the parts of the application at which the medical personnel would use the system to find a specific set of EBM codes. We reflected this requirement and divided codes into contextual groups, which was essential to ensure the usefulness of the prediction. That way, only vaccination billing codes are suggested when entering information into the vaccination text field.

In addition to the process-based support in the daily work routine, the physician mentioned the challenge of ensuring the completeness of the billing numbers before sending them to the health care insurance at the end of every quarter:

> *"Every quarter, we have to close the office for a week in order to check all the patient's EBM codes and ensure the completeness of the billing list."*

Hence, even though it is important to group the codes according to the text fields used in certain situations of the daily work routine, this approach would not help in ensuring the completeness of the overall quarterly billing. For this reason, a second approach spanning the whole range of EBM codes was necessary. Based on this feedback, the remaining text fields that were found important by the medical personnel were also added to the input data.

### 4.2.2 Second iteration – artifact customization

The findings of the evaluation with the medical personnel were the foundation for the second iteration. Thereby, initial requirements were reviewed and adjusted. On the one hand, an EBM code prediction approach taking into account all outcomes at once was too broad to support medical personnel during the daily treatment administration. On the other hand, quarterly billing required validation of the whole range of EBM codes across all textual descriptions. We developed these approaches in parallel, referred to as *daily code prediction on treatment description* (artifact A) and *quarterly code verification on full text* (artifact B), as shown in Figure 3.

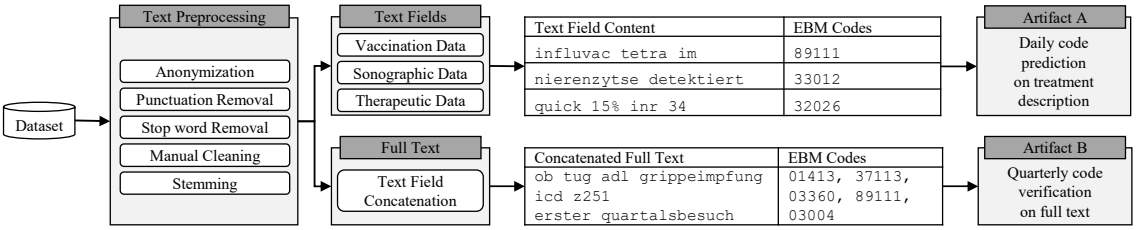



*Figure 3. Customized artifact variants for EBM code prediction.*

Artifact A aims at EBM code prediction during a patient's visit. Hereby, it was necessary to create smaller groups of coherent EBM codes based on the patient journey within the general practitioner's process. The medical personnel mentioned vaccinations as the most frequently forgotten code group since it is in direct interaction and not using a computer at this time. Therefore, we focused on the EBM code group of vaccinations for this artifact. Artifact B is to establish ML models based on the overall available EHR data of the system. Thereby, no differentiation on what text fields might account for which steps in the patient journey is being made since this approach shall help the medical personnel find missing or wrong EBM codes at the end of every quarter. For both artifacts, a preceding NLP-based ML pipeline had to be implemented.

The first step was to acquire and preprocess the data from the doctor's office system. In total, the descriptions stem from 28 different text fields, most frequently anamnesis and general notes. Data were anonymized by deleting names, locations, and dates identified using the SpaCy open-source named entity recognition library. For both artifacts, the text data were preprocessed by tokenization, stop-word removal, and stemming. We leveraged bag-of-words feature extraction and implemented the system using the scikit-learn ML library in python. We split the data into 70 %, 20 %, and 10 % for training, testing performance of different ML adjustments, and validation of final performance, respectively. We discussed both artifact customizations with the physician and received the following feedback:

> *"After indication and administration of the vaccination, we bill it with a single EBM code."*

Thus, we set up artifact A (daily code prediction) as a multi-class classification problem. For vaccinations (8 unique codes), we achieved a micro F1-score of 90.62 % with a linear SVM classifier.

For quarterly code verification, ML was applied as multi-label classification. Artifact B, however, suffered from the extreme classification problem. To counteract this, we excluded EBM codes that occurred less than 20 times in total and, thus, were negligible according to the physician, removing only 1.28 % of the data. Since the general practitioner does not assign codes for billing that were externally provided by a laboratory, we could exclude further codes. With 75 EBM codes remaining, we achieved a micro F1-score of 56.57 %.

We again reflected on results with the general practitioner and gained new insights as follows. For artifact A, not only each type of vaccination must be billed separately but also the stage and recipient of the vaccination need to be differentiated (whether it is initial, follow-up, or booster), and whether the patient is younger or older than 60 years). Next to vaccination, we reassured other billing mechanisms based on the patient journey for artifact A and the physician pointed to further therapeutical steps for which EBM code recommendations would be useful. Following this, the patient process within the general practitioner's office was examined in more detail and two further types of treatment could be categorized. It was found that *sonographic* examinations as well as *therapeutic procedures*, in case they were performed, were noted by the medical personnel separately.

The feedback regarding artifact B constituted important additional knowledge that could be leveraged to comply with billing rules. An insurance rate 0300X, for instance, is set whenever a patient visits the doctor's office for the first time in a quarter and varies depending on the age of the patient. Attending health programs such as cardiology, on the other hand, requires an electrocardiogram being monitored and thus, a corresponding EBM code being set. Such interdependencies of codes were collected during collaboration with the physician. As a result, both artifact variants allowed to include further text fields or engineer features to let the system learn legal dependencies, for instance, of disease programs.

### 4.2.3 Third iteration – end-user refinement

Based on the evaluation of the artifact customization in the second iteration, more contextual information was included in the artifacts and additional feature engineering was performed. Next to bag-of-words, we tested bag-of-n-grams (with $n = \{2, 3, 4\}$) and TF-IDF-based feature extraction. We selected models based on the literature, including linear SVM classifier, LightGBM, stochastic gradient descent, random forest, logistic regression, adaBoost, *k*-nearest neighbors, and multinominal naive Bayes. In addition, we tested a character-level CNN implemented in TensorFlow.

When implementing the changes for the daily code prediction on *vaccination* treatment descriptions (artifact A), an additional feature indicating whether the patient is older than 60 increased the F1-score of the linear SVM classifier to 96.65 %. For CNN, we tried out different hyperparameter settings. The final setup included an input length of 96 and convolution width of 7 characters, with a resulting F1-score of 97.98 % for vaccination codes, outperforming the linear SVM classifier. Since the medical personnel perceived it as practically very useful if the vaccination codes were extended with supplemental information to differentiate initial, follow-up, or booster vaccinations, we extended the feature space accordingly. This increased the number of EBM codes to 19 and reduced the performance to 93.60 %. However, the ML model learned to recognize even minor differences in the vaccination notes, predicting the extended billing codes for vaccination types in a most useful way.

Since the personnel also mentioned sonography and therapeutical procedure codes belonging to certain problematic EBM code groups, we also trained our system on respective text fields. Age and gender information of patients were additionally included since physical characteristics differ between the sexes and different examinations are performed depending on the age. The final CNN models for predicting *vaccination*, *sonography*, and *therapeutical procedure* codes ultimately achieved F1-scores of 93.60 %, 94.16 %, and 81.18 %, respectively.

For the quarterly code verification on full text (artifact B), we examined codes with high occurrences but low prediction scores with the medical personnel, to find potentially missing system information. This way, code prediction for disease programs was improved by adding further data such as electrocardiogram documentation. We binned age data into five categories according to EBM code regulations. A second feature indicated whether it is the first time a patient visits the practice in a quarter since this affected an often-coded fixed insurance rate (0300X). After that, a micro F1-score of 78.22 % could be achieved with bag-of-words feature extraction and multi-class LightGBM.

## 4.3 Reflection and learning

Initial job shadowing, different perspectives on the use of the system during the billing process, as well as different user and coding goals all provided a practice-inspired, realistic assessment that helped us to inform and understand the iterative design of the artifact. We continuously reflected on the specific artifact features during the previous stages, from which we collected meta requirements (MRs), as can be seen in Table 1, which in turn helps us infer generalized knowledge about designing an automated coding system in outpatient care.

During the first iteration, we observed that outpatient billing refers to different situations during a patient journey where an EBM code can but must not be set, such that the EBM coding system must allow assigning EBM codes spanning the whole practice patient journey (MR 1). Furthermore, we can reflect that – much different from extant ICD coding approaches that usually aim at solving large-scale multi-label classification problems –, sometimes only a specific set of billing codes must be suggested depending on how the system is used for specific treatments, e.g., directly after a vaccination. Hence, the utility of the coding system had to be evaluated for each specific use in the billing context (MR 2).

In iteration 2, we established artifact variants: Artifact A specialized in code prediction for groups of codes that are problematic and often forgotten in medical routine, as in the case of vaccination. Artifact B generated predictions of a large range of most frequently used codes at the end of a quarter, to reduce the workload of medical personnel. This showed that retrospective validation is a special characteristic of our outpatient field tests requiring careful design considerations (MR 3).

| ID | Meta Requirement (MR) | Description |
|---|---|---|
| MR 1 | Billing process | The coding system must allow coding of the whole patient journey. |
| MR 2 | Workflow | The coding system must ensure utility along with billing process steps. |
| MR 3 | Retrospective | The coding system must support quarterly completeness checks. |
| MR 4 | Billing regulations | The coding system must follow EBM coding rules. |
| MR 5 | Engineering | The coding system must apply practice-driven ML engineering. |

*Table 1. Meta requirements for the design of a coding system to support medical billing.*

From the third iteration, we could reflect that additional features had to be engineered to meet rules that exist in the EBM-based reimbursement model. These occurrences were identified after examining low prediction scores on a code-level together with billing-experienced personnel. For this reason, we learned that granular refinement of the system is necessary to follow coding rules which do not exist for pure diagnosis code prediction but in the outpatient setting (MR 4).

Finally, much of the feedback gathered through ongoing intervention with the physician has been driving the development of ML components. According to the feedback, for instance, that single codes are desired to be predicted, we applied multi-task classification in contrast to common ICD coding approaches. Thus, the feature engineering was inspired heavily by information gathered through collaboration with the physician, ensuring that the ML can learn from meaningful data. From that, we argue that such a coding system must closely perform practice-driven ML engineering (MR 5).

## 4.4 Formalization of learning

Next, we outline how we have captured knowledge from our reflection and learning stage as operational principles that are transferable to cases belonging to the same problem class of automated coding for medical billing. In Table 2, we highlight three design principles that were important for the design of our EBM coding system, based on the meta requirements reflected during the ADR process.

| ID | Design Principle | Addressed MRs |
|---|---|---|
| DP 1 | In order to support different perspectives in the billing process, the system should differentiate individual use variants in practice. | MR 2, MR 3 |
| DP 2 | In order to support diverse coding goals, the system should align ML components with the correctness and completeness trade-off. | MR 3, MR 4, MR 5 |
| DP 3 | In order to support legal aspects, the system should carefully consider the specific characteristics of different medical reimbursement models. | MR 1, MR 4 |

*Table 2. Design principles for automated coding to support outpatient billing.*

**DP 1:** We propose to examine the way the medical personnel uses the system at different points in time during the billing process. In our case, we identified variants in the creation of billing lists which led to a partitioning of the system for different groups of EBM codes (MR 2), according to the use of certain text fields in individual situations. Another variant of the artifact was explicitly designed to retrospectively check the completeness which differs from the way previous artifacts have been used (MR 3). Ultimately, a differentiation of the coding perspective increases utility for the general practitioner compared to a system that is independent of the variants performed in the billing process.

**DP 2:** Assessing benefits and limits during the development of the ML system while keeping the studied practical context under consideration, was another key principle in our artifact design. In our case, we were faced with a trade-off between correctness and completeness due to using ML for large-scale text classification. However, for each situation in the billing process identified through DP 2, completeness or accuracy are favored differently and thus, treated as separate tasks to which ML components were developed accordingly (MR 3). At the same time, we have learned that we had to refine the system to follow coding regulations (MR 4) for improving the system. The physician's

feedback influenced the whole ML pipeline including feature engineering, preprocessing, model setup, as well as optimization (MR 5). In total, this had a higher efficiency for the physician compared to a holistic ML application that would compromise between completeness and correctness, on the performance for predicting interdependent codes, or on the fit to diverse coding goals.

**DP3:** As the ADR process has shown, the initial problem formulation relied on understanding the outpatient billing process of the doctor's office as well as the points of contact of EBM codes in the patient journey (MR 1). However, the reimbursement model in this context had also led to system refinements that are at the core of code regulations, which even the medical personnel had to acquire (MR 4). From that, we can conclude that carefully considering not only practical routine but also characteristics of reimbursement models, help to design the system that applies to other medical billing processes. From a German general practitioner, this system would easily be transferable to other medical specialties, such as orthopedics, which also get reimbursed based on EBM codes. It would also be reasonable to apply the artifacts to other billing-related classification systems that are based on a mix of treatments and services delivered, such as for privately assured patients. By considering reimbursement models during the design, the artifact can be characterized and generalized systematically to international contexts. For instance, outpatient care in Switzerland follows a similar reimbursement model for statutorily assured patients based on EBM-like service and treatment codes.

# 5 Discussion

This study profoundly contributes to the design of medical coding systems by expanding the research area with a new topic, automated EBM coding for outpatient billing. Compared to prior research on medical text classification for diagnosis code assignment (e.g., Baumel et al., 2018; Xu et al., 2019), we have been able to demonstrate for outpatient billing purposes, that an automated coding is differently applied (DP 1), such as daily code prediction and quarterly code verification.

Foremost, ML-based artifacts in healthcare face several challenges (Pumplun et al., 2021). Although the availability of patient data was given within the ADR collaboration, proprietary and heterogeneous data forms led to an effortful process also for EBM coding. This was present throughout the ADR iterations since refinements after practical feedback required acquiring additional system data, such as electrocardiogram descriptions. Data anonymization, in contrast, could be resolved by recognizing and omitting names, locations, and dates via open-source libraries. Nevertheless, we conclude that there are further persisting challenges in the limelight when employing these systems for medical billing. Stemming from ICD-related ML research, the high number of labels typical in medical coding is challenging for ML models (Baumel et al., 2018). This emerges also in EBM coding for outpatient billing so that coding trade-offs need to be spotted (DP 2), even though an ADR-based approach could realize important refinements thanks to practice-inspired research. Instead of using top-k code labels (Xu et al., 2019) or more abstract group-level codes as done in ICD coding (e.g., Eslami et al., 2020), we could sufficiently counteract the extreme classification problem through expert knowledge, by removing problematic instances that were seldom and at the same time negligible from the viewpoint of the physician or that were provided by a third party in practice. The medical outpatient context also provided the hitherto unexplored opportunity to collapse treatment groups according to steps in the billing process. Lastly, the domain-specific feature engineering was as crucial as commonly known for traditional healthcare ML, whereas the CNN directly achieved promising results on full-text data (Miotto et al., 2018). However, the studied context provided knowledge beyond medical expertise with the help of reimbursement characteristics (DP 3), for instance, for the binning of age similar to EBM rules. Therefore, a close collaboration with the physician was crucial to identify weak spots in the EBM code prediction and achieve performance improvements.

We anticipate that the study results can serve as a reference for medical service providers. From a practical perspective, the designed daily code prediction artifact based on treatment description (artifact A) can pave the way for replacing the repetitive work of daily reimbursement coding from EHR data for the medical practice. It could therefore enable a shift of working-hour capacities from

administrative to patient-oriented work as well as an increase in employee satisfaction by displacing possibly unsatisfactory work for the medical assistants.

Due to the nature of ML applications, the outcome of a model highly depends on the amount and quality of available data. Therefore, one limitation of this research is that we investigated the EBM coding system in the context of one general practitioner's office, with data limited to the fiscal year 2020. A prior study identified the problem of inter-rater agreements in the case of diagnosis coding (Gobeill and Ruch, 2018). Since the general practitioner's office in this work has employed three physicians and six assistants, coding experiences might be heterogeneous within the personnel. Therefore, additional effort should be spent to include further fiscal years as well as review and rule out errors in the data to mitigate personal bias and ensure data correctness.

EBM coding for outpatient billing was found to be associated with rules, interdependencies, and high complexity. Infusing such information, driven by domain experts, is required to establish a practically useful and reliable system (Mujtaba et al., 2017; Pestian et al., 2007). This may not be limited to feature engineering or cascading classifiers (e.g., Koopman et al., 2015). Instead, combining data-driven and knowledge-based systems in deep learning has demonstrated to improve robustness and performance (Chari et al., 2020; Garcez et al., 2019). Recent research has proposed methods such as neural-symbolic computing to integrate rules into neural networks (Rueden et al., 2021). This allows the models to conform to constraints or comply with guidelines given by the medical billing context.

Future research should also further investigate the deployment in outpatient billing processes. One important step is to analyze the economic benefits of automated coding systems in this context, particularly those that result from process improvements.

As another extension of this study, we consider utilizing pre-trained word embeddings, in particular of the quarterly code verification that lacks in accuracy most. With the recent advent of transformer architectures, a promising path is to utilize deep contextualized word representations, such as 'BERT' (Devlin et al., 2018) or medical text-specialized variants, to increase performance and make results more stable across medical providers and heterogeneous writing styles. Next to that, attention mechanism is likely to be worth exploring to highlight relevant words of an EBM code prediction and make the artifact explainable (Mullenbach et al., 2018), which may further increase practical utility.

# 6 Conclusion

The billing procedure in general practitioner's offices is a time-consuming and error-prone task that lacks system support. Leveraging the benefits of ML-based coding for outpatient billing, this paper has followed ADR to develop an artifact that contributes to a practical concern of physicians (Susman & Evered, 1978). The ADR includes important elements of the DSR methodology, which "emphasises the design and construction of applicable artefacts" (Peffers et al., 2007, p. 131). This approach has been fundamental in responding to the necessity of combining the theoretical contributions while assisting in solving the current problems of outpatient billing. The practical feedback served as an essential step in creating the problem-solving automated EBM coding system.

We have been able to indicate that EBM codes can be predicted with considerable success based on EHR data. During the development of the artifact, we established coherent EBM code groups, achieved a highly performant classification, and provided utility through parallelizing complete code prediction as well as most problematic code prediction tasks. Along this vein, we are able to conclude that it is crucial to examine coding perspectives in the billing process, align coding goals with the benefits and limits of ML components, as well as carefully consider reimbursement characteristics.

This research represents an important first step toward better support for medical personnel in the billing process for outpatient reimbursement and demonstrates that automated code prediction can leverage the benefits of ML. The findings of this study reveal that the system can learn to predict EBM codes from available medical documentation and this way, let human encoders identify and focus on problematic patient entries. For the general practitioner, these text field-based approaches not only draw one's attention to billing-relevant information but also help set the correct EBM code.

## References


Atutxa, A., Casillas, A., Ezeiza, N., Fresno-Fernández, V., Goenaga, I., Gojenola, K., Martínez, R., Oronoz-Anchordoqui, M., and Perez-de-Viñaspre, O. (2018). "IxaMed at CLEF eHealth 2018 task 1: ICD10 coding with a sequence-to-sequence approach," in: CLEF eHealth 2018 (ed.) *Lab Overview CLEF eHealth 2018*.

Baumel, T., Nassour-Kassis, J., Cohen, R., Elhadad, M., and Elhadad, N. (2018). "Multi-label classification of patient notes. A case study on ICD code assignment," in: *Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence*. URL: http://arxiv.org/pdf/1709.09587v3.

Burkul, P., Umapathy, K., Asaithambi, A., and Huang, H. (2020). "Data mining pipeline for performing decision tree analysis on mortality dataset with ICD-10 codes,". In *SAIS 2020 Proceedings*. URL: https://aisel.aisnet.org/sais2020/28.

Busse, R., Geissler, A., Quentin, W., and Wiley, M. M. (2011). *Diagnosis-Related Groups in Europe. Moving Towards Transparency, Efficiency and Quality in Hospitals.* Maidenhead: Open University Press.

Catling, F., Spithourakis, G. P., and Riedel, S. (2018). "Towards automated clinical coding," *International journal of medical informatics* 120, 50–61.

Chandra, L., Seidel, S., and Gregor, S. (2015). "Prescriptive knowledge in IS research: Conceptualizing design principles in terms of materiality, action, and boundary conditions," in: *2015 48th Hawaii International Conference on System Sciences*: IEEE, pp. 4039–4048.

Chari, S., Gruen, D. M., Seneviratne, O., and McGuinness, D. L. (2020). *Directions for Explainable Knowledge-Enabled Systems*. URL: http://arxiv.org/pdf/2003.07523v1.

Crammer, K., Dredze, M., Ganchev, K., Talukdar, P. P., and Carroll, S. (2007). "Automatic code assignment to medical text," in: Cohen, K. B., Demner-Fushman, D., Friedman, C., Hirschman, L., & Pestian, J. P. (eds.) *Proceedings of the Workshop on BioNLP 2007 Biological, Translational, and Clinical Language Processing - BioNLP '07*, Morristown, NJ, USA: Association for Computational Linguistics, pp. 129–136. URL: https://www.researchgate.net/publication/234795167_Automatic_code_assignment_to_medical_text (visited on 01/10/2021).

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*.

Duarte, F., Martins, B., Pinto, C. S., and Silva, M. J. (2018). "Deep neural models for ICD-10 coding of death certificates and autopsy reports in free-text," *Journal of biomedical informatics* 80, 64–77.

Ebersbach, M., Herms, R., and Eibl, M. (2017). "Fusion methods for ICD10 code classification of death certificates in multilingual corpora," in: CLEF Conference and Labs of the Evaluation Forum (ed.) *Information Access Evaluation meets Multilinguality, Multimodality, and Visualization*.

Eslami, S., Adorjan, P., and Meinel, C. (2020). "SehMIC: Semi-hierarchical multi-label ICD code classification," in: *CLEF (Working Notes)*.

Farkas, R. and Szarvas, G. (2008). "Automatic construction of rule-based ICD-9-CM coding systems," *BMC bioinformatics* 9 Suppl 3, S10.

Garcez, A. d., Gori, M., Lamb, L. C., Serafini, L., Spranger, M., and Tran, S. N. (2019). *Neural-Symbolic Computing: An Effective Methodology for Principled Integration of Machine Learning and Reasoning*. URL: http://arxiv.org/pdf/1905.06088v1.

Gobeill, J. and Ruch, P. (2018). "Instance-based learning for ICD10 categorization," in: CLEF eHealth 2018 (ed.) *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum*: 2125. URL: http://ceur-ws.org/Vol-2125/paper_149.pdf (visited on 05/20/2021).

Goldstein, I., Arzrumtsyan, A., and Uzuner, O. (2007). "Three approaches to automatic assignment of ICD9-CM codes to radiology reports," *AMIA Symposium* 2007, 279–283.

Gregor, S., Kruse, L., and Seidel, S. (2020). "Research perspectives: The anatomy of a design principle," *Journal of the Association for Information Systems* 21, 1622–1652.

Hausmann, G. (2018). "Von der Rechnung zur Abrechnung," *iX Developer 2018 - Machine Learning: Verstehen, verwenden, verifizieren*, 100–105.

Hevner, A. R., R, A., March, S., T, S., Park, J., Ram, and Sudha (2004). "Design science in information systems research," *Management Information Systems Quarterly* 28, 75.

Homeyer, A., Lotz, J., Schwen, L. O., Weiss, N., Romberg, D., Höfener, H., Zerbe, N., and Hufnagl, P. (2021). "Artificial intelligence in pathology: From prototype to product," *Journal of pathology informatics* 12, 13.

Janssen, F. and Kunst, A. E. (2005). "ICD coding changes and discontinuities in trends in cause-specific mortality in six European countries, 1950-99," *Bulletin of the World Health Organization* 82 (12), 904–913.

Jatunarapit, P., Piromsopa, K., and Charoenlap, C. (2016). "Development of thai text-mining model for classifying ICD-10 TM," in: *2016 8th International Conference on Electronics, Computers and Artificial Intelligence (ECAI)*, pp. 1–6.

Karimi, S., Dai, X., Hassanzadeh, H., and Nguyen, A. N. (2017). "Automatic diagnosis coding of radiology reports: A comparison of deep learning and conventional classification methods," in: *BioNLP 2017*, Vancouver, Canada: Association for Computational Linguistics, pp. 328–332. URL: https://www.aclweb.org/anthology/W17-2342.

Kaur, R. and Ginige, J. A. (2018). "Comparative analysis of algorithmic approaches for auto-coding with ICD-10-AM and ACHI," *Studies in health technology and informatics* 252, 73–79.

Kavuluru, R., Han, S., and Harris, D. (2013). "Unsupervised extraction of diagnosis codes from EMRs using knowledge-based and extractive text summarization techniques," *Advances in artificial intelligence. Canadian Society for Computational Studies of Intelligence. Conference* 7884, 77–88.

Kelly, L., Goeuriot, L., Suominen, H., Schreck, T., Leroy, G., Mowery, D. L., Velupillai, S., Chapman, W. W., Martinez, D., Zuccon, G., and others (2014). "Overview of the share/clef ehealth evaluation lab 2014," in: *International Conference of the Cross-Language Evaluation Forum for European Languages*, pp. 172–191.

Koopman, B., Zuccon, G., Nguyen, A. N., Bergheim, A., and Grayson, N. (2015). "Automatic ICD-10 classification of cancers from free-text death certificates," *International journal of medical informatics* 84 (11), 956–965.

Li, M., Fei, Z., Zeng, M., Wu, F.-X., Li, Y., Pan, Y., and Wang, J. (2019). "Automated ICD-9 coding via a deep learning approach," *IEEE/ACM transactions on computational biology and bioinformatics* 16 (4), 1193–1202.

Marafino, B. J., Davies, J. M., Bardach, N. S., Dean, M. L., and Dudley, R. A. (2014). "N-gram support vector machines for scalable procedure and diagnosis classification, with applications to clinical free text data from the intensive care unit," *Journal of the American Medical Informatics Association : JAMIA* 21 (5), 871–875.

Miftakhutdinov, Z. and Tutubalina, E. (2017). "KFU at CLEF eHealth 2017 task 1: ICD-10 coding of English death certificates with recurrent neural networks," in: CLEF Conference and Labs of the Evaluation Forum (ed.) *Information Access Evaluation meets Multilinguality, Multimodality, and Visualization*.

Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., and Gao, J. (2020). *Deep Learning Based Text Classification: A Comprehensive Review*. URL: http://arxiv.org/pdf/2004.03705v3.

Miotto, R., Wang, F., Wang, S., Jiang, X., and Dudley, J. T. (2018). "Deep learning for healthcare: review, opportunities and challenges," *Briefings in bioinformatics* 19 (6), 1236–1246.

Mujtaba, G., Shuib, L., Raj, R. G., Rajandram, R., Shaikh, K., and Al-Garadi, M. A. (2017). "Automatic ICD-10 multi-class classification of cause of death from plaintext autopsy reports through expert-driven feature selection," *PloS one* 12 (2), e0170242.

Mullenbach, J., Wiegreffe, S., Duke, J., Sun, J., and Eisenstein, J. (2018). "Explainable prediction of medical codes from clinical text," in: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, New Orleans, Louisiana: Association for Computational Linguistics, pp. 1101–1111. URL: https://www.aclweb.org/anthology/N18-1100.

Muneer, A. and Fati, S. M. (2020). "A comparative analysis of machine learning techniques for cyberbullying detection on Twitter," *Future Internet* 12 (11), 187.

Neumann-Grutzeck, C. (2020). "GOÄ auf dem besten Weg zu einem EBM : Neufassung der privaten Gebührenordnung," *MMW Fortschritte der Medizin* 162 (18), 35.

Névéol, A., Robert, A., Anderson, R., Cohen, K. B., Grouin, C., Lavergne, T., Rey, G., Rondet, C., and Zweigenbaum, P. (2017). "CLEF eHealth 2017 multilingual information extraction task overview: ICD10 coding of death certificates in English and French," in: CLEF Conference and Labs of the Evaluation Forum (ed.) *Information Access Evaluation meets Multilinguality, Multimodality, and Visualization*.

Österle, H., Becker, J., Frank, U., Hess, T., Karagiannis, D., Krcmar, H., Loos, P., Mertens, P., Oberweis, A., and Sinz, E. J. (2011). "Memorandum on design-oriented information systems research," *European Journal of Information Systems* 20 (1), 7–10.

Peffers, K., Tuunanen, T., and Niehaves, B. (2018). "Design science research genres: introduction to the special issue on exemplars and criteria for applicable design science research," *European Journal of Information Systems* 27 (2), 129–139.

Peffers, K., Tuunanen, T., Rothenberger, M. A., and Chatterjee, S. (2007). "A design science research methodology for information systems research," *Journal of Management Information Systems* 24 (3), 45–77.

Perotte, A., Pivovarov, R., Natarajan, K., Weiskopf, N., Wood, F., and Elhadad, N. (2014). "Diagnosis code assignment: models and evaluation metrics," *Journal of the American Medical Informatics Association : JAMIA* 21 (2), 231–237.

Pestian, J. P., Brew, C., Matykiewicz, P., Hovermale, D. J., Johnson, N., Cohen, K. B., and Duch, W. (2007). "A shared task involving multi-label classification of clinical free text," in: *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, USA: Association for Computational Linguistics, pp. 97–104.

Pumplun, L., Fecho, M., Islam, N., and Buxmann, P. (2021). "Machine learning systems in clinics – how mature is the adoption process in medical diagnostics?," in: Bui, T. (ed.) *Proceedings of the 54th Hawaii International Conference on System Sciences*: Hawaii International Conference on System Sciences.

Quan, H., Sundararajan, V., Halfon, P., Fong, A., Burnand, B., Luthi, J.-C., Saunders, L. D., Beck, C. A., Feasby, T. E., and Ghali, W. A. (2005). "Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data," *Medical care* 43 (11), 1130–1139.

Rios, A. and Kavuluru, R. (2013). "Supervised extraction of diagnosis codes from EMRs: Role of feature selection, data selection, and probabilistic thresholding," *IEEE International Conference on Healthcare Informatics. IEEE International Conference on Healthcare Informatics* 2013, 66–73.

Rueden, L. von, Mayer, S., Beckh, K., Georgiev, B., Giesselbach, S., Heese, R., Kirsch, B., Walczak, M., Pfrommer, J., Pick, A., Ramamurthy, R., Garcke, J., Bauckhage, C., and Schuecker, J. (2021). "Informed machine learning - a taxonomy and survey of integrating prior knowledge into learning systems," *IEEE Transactions on Knowledge and Data Engineering*, 1.

Sein, M., Henfridsson, O., Purao, S., Rossi, M., and Lindgren, R. (2011). "Action design research," *MIS Quarterly* 35 (1), 37–56.

Seva, J., Kittner, M., Roller, R., and Leser, U. (2017). "Multi-lingual ICD-10 coding using a hybrid rule-based and supervised classification approach at CLEF eHealth 2017," in: CLEF Conference and Labs of the Evaluation Forum (ed.) *Information Access Evaluation meets Multilinguality, Multimodality, and Visualization*.

Walls, J. G., Widermeyer, G. R., and El Sawy, O. A. (2004). "Assessing information system design theory in perspective: how useful was our 1992 initial rendition?," *Journal of Information Technology Theory and Application* 6 (2), 43–58.

Walls, J. G., Widmeyer, G. R., and El Sawy, O. A. (1992). "Building an information system design theory for vigilant EIS," *Information Systems Research* 3 (1), 36–59.

Webster, J. and Watson, R. (2002). "Analyzing the past to prepare for the future: Writing a literature review," *MIS Quarterly* 26 (2), xiii–xxiii.

Winter, R. (2008). "Design science research in Europe," *European Journal of Information Systems* 17 (5), 470–475.

Xie, X., Xiong, Y., Yu, P. S., and Zhu, Y. (2019). "EHR coding with multi-scale feature attention and structured knowledge graph propagation," in: Zhu, W., Tao, D., Cheng, X., Cui, P., Rundensteiner, E., Carmel, D., He, Q., & Xu Yu, J. (eds.) *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, New York, NY, USA: ACM, pp. 649–658.

Xu, K., Lam, M., Pang, J., Gao, X., Band, C., Mathur, P., Papay, F., Khanna, A. K., Cywinski, J. B., Maheshwari, K., Xie, P., and Xing, E. P. (2019). "Multimodal machine learning for automated ICD coding," in: Finale Doshi-Velez, Jim Fackler, Ken Jung, David Kale, Rajesh Ranganath, Byron Wallace, & Jenna Wiens (eds.) *Proceedings of the 4th Machine Learning for Healthcare Conference*, Ann Arbor, Michigan: PMLR, pp. 197–215. URL: http://proceedings.mlr.press/v106/xu19a.html (visited on 02/01/2021).

Yan, Y., Fung, G., Dy, J. G., and Rosales, R. (2010). "Medical coding classification by leveraging inter-code relationships," in: Rao, B., Krishnapuram, B., Tomkins, A., & Yang, Q. (eds.) *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '10*, New York, USA: ACM Press, p. 193. URL: https://dl.acm.org/doi/pdf/10.1145/1835804.1835831 (visited on 01/19/2020).

Yin, R. K. (2018). *Case study research and applications*. *Design and methods,* Sixth edition: Sage.

Yu, Y., Li, M., Liu, L., Fei, Z., Wu, F.-X., and Wang, J. (2019). "Automatic ICD code assignment of Chinese clinical notes based on multilayer attention BiRNN," *Journal of biomedical informatics* 91, 103114.

Zhang, X., Zhao, J. J., and LeCun, Y. (2015). "Character-level convolutional networks for text classification," *CoRR* abs/1509.01626.

Zweigenbaum, P. and Lavergne, T. (2016). "Hybrid methods for ICD-10 coding of death certificates," in: Grouin, C., Hamon, T., Névéol, A., & Zweigenbaum, P. (eds.) *Proceedings of the Seventh International Workshop on Health Text Mining and Information Analysis*, Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 96–105.

Zweigenbaum, P. and Lavergne, T. (2017). "Multiple methods for multi-class, multi-label ICD-10 coding of multi-granularity, multilingual death certificates," in: CLEF Conference and Labs of the Evaluation Forum (ed.) *Information Access Evaluation meets Multilinguality, Multimodality, and Visualization*.