Journal of the American Medical Informatics Association Advance Access published August 27, 2015

Data integration of structured and unstructured sources for assigning clinical codes to patient stays

RECEIVED 9 April 2015
REVISED 4 June 2015
ACCEPTED 29 June 2015

Elyne Scheurwegs[1], Kim Luyckx[2], Léon Luyten[2], Walter Daelemans[3], Tim Van den Bulcke[4]

ABSTRACT

Objective Enormous amounts of healthcare data are becoming increasingly accessible through the large-scale adoption of electronic health records. In this work, structured and unstructured (textual) data are combined to assign clinical diagnostic and procedural codes (specifically ICD-9-CM) to patient stays. We investigate whether integrating these heterogeneous data types improves prediction strength compared to using the data types in isolation.

Methods Two separate data integration approaches were evaluated. Early data integration combines features of several sources within a single model, and late data integration learns a separate model per data source and combines these predictions with a meta-learner. This is evaluated on data sources and clinical codes from a broad set of medical specialties.

Results When compared with the best individual prediction source, late data integration leads to improvements in predictive power (eg, overall F-measure increased from 30.6% to 38.3% for International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) diagnostic codes), while early data integration is less consistent. The predictive strength strongly differs between medical specialties, both for ICD-9-CM diagnostic and procedural codes.

Discussion Structured data provides complementary information to unstructured data (and vice versa) for predicting ICD-9-CM codes. This can be captured most effectively by the proposed late data integration approach.

Conclusions We demonstrated that models using multiple electronic health record data sources systematically outperform models using data sources in isolation in the task of predicting ICD-9-CM codes over a broad range of medical specialties.

Keywords: data integration, clinical coding, data mining, international classification of diseases, electronic health records

INTRODUCTION

In health care, electronic health records (EHRs) are becoming widely accepted as the de facto standard of storing medical information.[1,2] The information contained within these EHRs not only provides direct health information about patients, but is also used to monitor hospital activities for medical billing and population health management. Clinical coding can be defined as the assignment of procedural and diagnostic codes specified in a medical classification system (eg, ICD,[3] ICPC-2[4]). ICD-9-CM diagnostic and procedure codes[5] are mainly used for reporting and reimbursement purposes of health care providers, but are also a key factor for other research applications such as tracking patients with sepsis through ICD codes.[6,7] Currently, clinical codes are often attributed and registered manually by a specialized team of medical coders.

The primary objective of this paper is to assess whether the integration of structured and unstructured EHR data can improve automated predictions of clinical codes. Additionally, we analyze the informativeness of several data sources, both in isolation and combined, for multiple medical specialties. This is demonstrated by predicting procedural and diagnostic ICD-9-CM codes from various input sources (eg, letters, lab results, radiology reports).

BACKGROUND

Current automated clinical coding approaches for patient discharge files and radiology reports can be divided into handcrafted,[8,9] machine learning, and hybrid approaches,[10,11] with handcrafted and hybrid approaches being the most successful.[12] A literature review conducted by Stanfill *et al*[13] concluded that, while some systems show excellent results, most of them are used in controlled settings, often using normalized data and keeping a limited scope (eg, radiology reports). Secondly, many of the current approaches are not easily portable to other medical domains and different languages. Extending these approaches towards real-life EHR data and enabling these approaches to efficiently deal with high degrees of variability in terms of content, structure and language typical of clinical data, is an important challenge.

Perotte *et al*[14] exploited the hierarchy present in the ICD-9-CM classification system to predict diagnosis codes using (English) discharge files. Whereas most approaches focus on a smaller set of diagnosis codes, Perotte *et al* used the MIMIC-II database,[15] which contains a large set of 5030 distinct ICD-9-CM diagnosis codes. They achieved an F-measure of 39% in this specific setting.

Using both structured and unstructured data has already been shown valuable for multiple applications in the medical field. Abhyankar *et al*[16] used structured and unstructured data from the MIMIC-II database to identify a cohort of ICU patients who received dialysis. Their method, applying off-the-shelf information retrieval methods, not only allows for more effective cohort identification in comparison to using data sources in isolation, but is also attractive enough to be used by healthcare practitioners. Our research strengthens that conclusion and shows the value of combining structured and unstructured data for a different application, namely the assignment of clinical codes to an inpatient stay.

Correspondence to Elyne Scheurwegs, ADReM (Advanced Database Research and Modelling), Biomedical Informatics Research Center Antwerp (biomina), University of Antwerp, Middelheimlaan 1, (Office G2.33) 2020 Antwerpen; elyne.scheurwegs@uantwerpen.be; Tel: +32 (0)3 821.55.52

For numbered affiliations see end of article.

Pathak *et al* [17] employed a different approach by mapping structured and unstructured data from diverse EHR systems onto standardized vocabularies and ontologies to perform high-throughput phenotyping for patient cohort identification. This unified structured data view is a powerful resource to support secondary use and exchange of EHR data in general, and the goals set out by the SHARPn program in particular (ie, enhancing patient safety and improving patient outcome). From a technical point of view; however, there is a substantial dependency on standardized vocabularies or ontologies, which are not always available for every language. The availability of those resources for lesser-resourced languages such as Dutch remains an issue and mapping the (Dutch) hospital data we used onto standardized vocabularies or ontologies would require a substantial research effort. Additionally, some of the used structured data sources in our dataset (such as lab results) currently do not have a mapping onto a code system, which causes mapping to be an error-prone and time-intensive process. As a result, any additional mapping was out of the scope for this paper. The research presented in this paper distinguishes itself from the current state of the art on three different aspects, namely, portability, data integration, and use of real-life EHR data. First of all, our approach is portable towards different medical specialties. This work evaluates experiments on datasets from fourteen different specialties and with a large set of codes unique to each specialty. Portability is improved by using generalizable techniques instead of handcrafted rules and symbolic systems. Secondly, we not only use information found in discharge summaries, but we also maximize the available information in a specific context by means of *data integration* approaches, which yielded excellent results in other medical tasks.16–19 Our approach can use all available data sources, and can also work with any subset of these sources, depending on their local availability. Finally, our approach is able to use real-life, raw EHR data as its input which greatly increases its practical usability.

METHODS

The very nature of the data used in this study, being real-life data, presents a number of limitations and specific characteristics. First of all, the data cannot be considered to be complete. Medication prescriptions, for instance, are not always registered electronically and do not necessarily lead to actual medication intake. Secondly, the data used in this research originates from disparate data sources, which were created for varying purposes (eg, diagnostics, treatment follow-up, nurse handoff, billing, medical registration, and archiving). This required recreating some of the relations (eg, inpatient stay relations). Finally, the data contains errors (eg, spelling errors) and has missing values.

Dataset

Our dataset is derived from the clinical data warehouse at the Antwerp University Hospital (UZA, Belgium) and consists of a randomized subset of fully anonymized historical data with hospitalized patient stays, covering 2 years of data. Distinct sources in the dataset consist of both structured and unstructured (ie, textual) data. The extracted textual data sources are in Dutch. The dataset is divided into 14 medical specialties, as seen in Figure 1. Only sufficiently represented medical specialties are included in the dataset. Assigned procedure and diagnostic codes are observed to follow a Zipf distribution,[20] meaning that a few codes occur very frequently with a long tail of infrequently assigned codes.

The following types of data sources are used in our study (abbreviations are indicated with brackets):

Unstructured data sources:

- Surgery reports (Surgery Rpt): describing the details of a performed surgery.
- Letters (Letter): discharge letters and letters to direct a patient to a specialist.
- Notes (Note): day-to-day internal notes concerning a patient (eg, a progress report written by a nurse).
- Protocols (Protocol): textual representation of the results of certain procedures (eg, the textual interpretation of an Magnetic Resonance Imaging (MRI) scan).
- Attestations (Attestation): a letter validating a certain claim (eg, a medical leave of absence).
- Requests (Request): a formal letter asking for a medical appliance or service (eg, a request to receive an electric wheelchair for care at home).

Structured data sources:

- Lab results (Lab results): consisting of a test id, a numeric or categorical value, a unit, and sometimes a conclusion. Lab test naming uses an in-house naming convention.
- Inpatient medication prescriptions (ATC): set of prescribed medications in the form of ATC codes (Anatomical Therapeutic Chemical classification system).[21]
- Oncological pathology codes (CODAP): in the form of CODAP codes,[22] a code system describing abnormal tissue growth, analyzed after biopsy.
- Medical Specialty (DEPT): describing the associated medical specialties for a particular stay. (eg, a patient being treated in the cardiology department is being followed-up by a doctor from the gastroenterology department and is therefore associated with both medical departments).
- Demographic data (DEM): year of birth and gender.
- Procedure codes (RIZIV): describing a medical procedure or intervention (eg, MRI scan). These are registered automatically or manually. The specific nomenclature used is the Belgian RIZIV standard,[23] which is strongly linked with ICD-9-CM procedure codes.

The prediction space consists of: (1) ICD-9-CM procedure codes (ICD-9-CM Proc), which describe procedures performed on a patient during his/her hospital stay and (2) ICD-9-CM diagnosis codes (ICD-9-CM Diag), which contain the primary and secondary diagnoses of a patient.[5] These codes have been assigned manually by a specialized team of coders, which is a potentially error-prone process.[12,24] Additionally, the Belgian government only requires a subset of procedural ICD-9-CM codes to be coded, namely all codes between indices 0 and 87 and a selection of codes in the 87–99 "Miscellaneous diagnostic and therapeutic procedures" range (ICD-9-CM Volume 3). This leads to 1636 unique required codes in our dataset. The not required codes are sometimes but not consistently assigned and were therefore left out of the training data. For diagnostic ICD-9-CM codes, the entire set of codes was included.

Data representation

Linking specific data elements – both structured and unstructured – to one or more procedure or diagnosis codes, requires a process of one-to-many and many-to-one data mapping. This mapping is performed by projecting the data elements to a single level.

An inpatient stay is reported in multiple data elements, originating from a disparate set of databases (Figure 2A). These databases were developed for different purposes, causing them to have different ways of representing inpatient stays. When a patient is transferred from one medical specialty to another during an inpatient stay, we refer to both stays as *partial stays*. We use this *partial stay* (linked to a single medical specialty) as the single level to which we link all data elements.

In Figure 2B, an example structure of an inpatient stay can be seen. Data elements are linked directly on different levels within the inpatient stay. An indirect link between a more detailed level in the inpatient stay and a data element can be made by looking at the given date linked to

Figure 1: The top graph shows the number of patient records in the available datasets per medical specialty; the bottom graph shows the number of unique ICD-9-CM codes (procedural: left, diagnostic: right) per specialty.

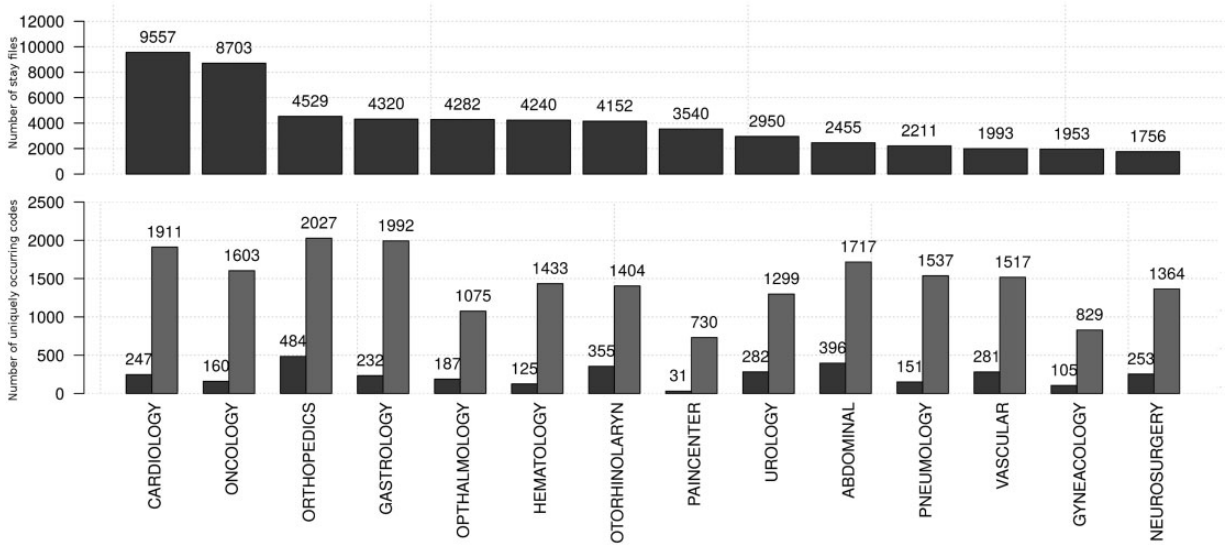
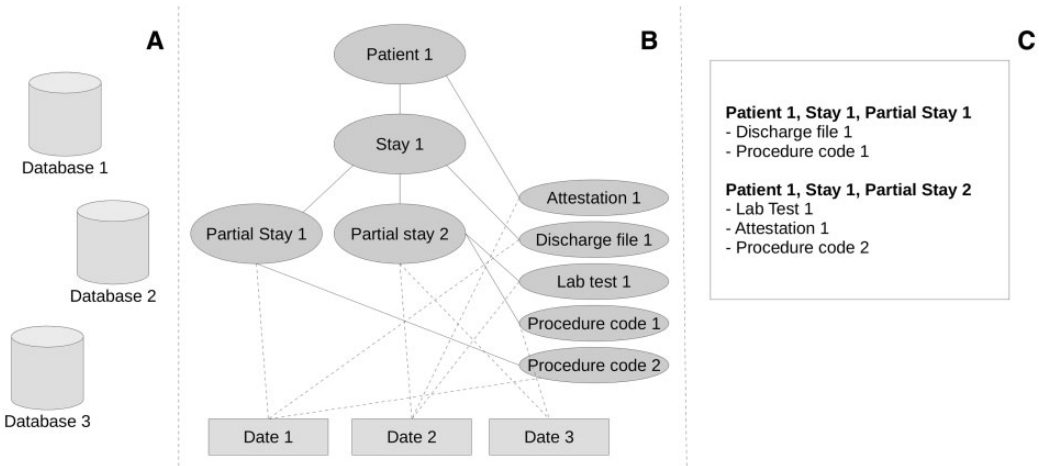


Figure 2: Visualization of an example set of structured and unstructured data types. (**A**) The disparate databases. (**B**) An example structure of a patient medical record, with full lines representing connections of data types and their level and dashed lines to their given date. (**C**) The mapping on the partial stay level.



these elements. This allows flattening the data elements to a representation on the *partial stay* level, as seen in Figure 2C.

This approach results in a list of partial stays with a mean of 8 distinct source types and a standard deviation of 2.5 sources. This means that while most patient files have multiple data sources linked to them, they seldom have all sources. An overview of the presence of these sources can be seen in Figure 3.

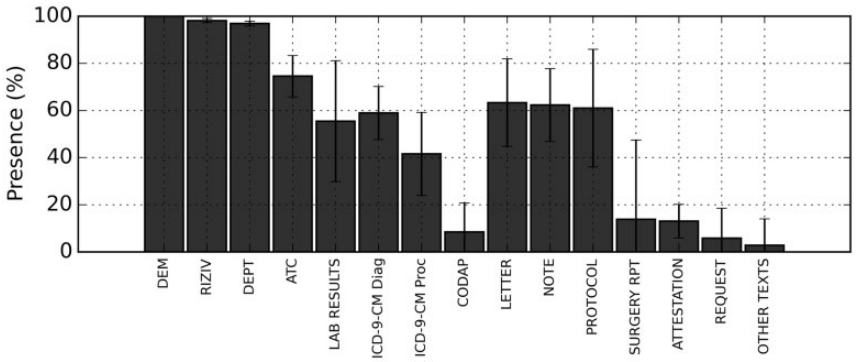
The classification task this paper focuses on is the assignment of clinical codes to *partial stays*, which in turn represent EHR data collected during routine clinical processes within a medical specialty, or within the time interval of that partial stay. Fifty-six thousand six hundred and forty-one distinct partial stays can be found in the entire dataset. Figure 1 presents the number of partial stays per specialty. Partial stays that do not have any codes assigned are not excluded from the dataset, since this is the result of human error that we do not expect to cause bias in the results.

Features

Several types of features are derived from data types, separately for structured and unstructured input. For structured input, we have a number of codes assigned to a patient (eg, ATC-codes, codes representing lab results). For lab results, we have a value and unit (eg, 25 mg) or a state (eg, positive, negative, strong reaction) in addition to the code representing a certain test.

A first set of features is derived by counting the number of occurrences of distinct assigned codes. A second set of features is specific

Figure 3: Presence (in %) of the different data types over all datasets. The main bar shows the average presence, the error bars show the standard deviation between the different medical specialties. Abbreviations are explained in the "dataset" section.



to the source itself. For lab results, this consists of the observed absolute values and deviation from the reference range (as derived from reference guides within the hospital for specific tests). For RIZIV codes, we use additional information found in its descriptive thesaurus[22] and make counts of specific types of codes occurring (eg, the number of codes that involve a general action, such as "attachment of an infusion"). A third set of features consists of multiple meta-features including the average amount of assigned codes per day and the total amount of (unique) codes per stay.

For unstructured input, we use a combined bag of words (BoW) of all documents of a certain type (eg, notes) associated to a specific partial stay. These individual BoWs are preprocessed with Frog,[25] a morpho-syntactic analyzer and dependency parser for Dutch. This natural language processing (NLP) tool uses a pretrained model to perform sentence detection, tokenization, and part-of-speech tagging. From this last set of features, words tagged as nouns and verbs were used as features.

This is a relatively superficial level of NLP, but has the advantage that easy adaptation to other languages is possible, as opposed to using deeper text understanding techniques. Experiments with different text representations based on a BoW (consisting of the words themselves, the lemmas, or the nouns and verbs for each document) yielded comparable results to the chosen approach (results not shown).

Experimental Setup

For each clinical code, a separate classifier is trained. Training is done in parallel, so that a single partial stay can have multiple codes assigned. Ten-fold cross-validation was applied to generate robust models. For each fold, the feature selection (see below) was performed to reduce the feature space.

In the machine learning phase, multiple general classifiers were evaluated to account for differences in performance among certain data types. Naive Bayes[26] and Random Forests[27] were used for prediction and compared afterward using the WEKA software package.[28] The Naive Bayes implementation is multinomial with Laplace smoothing. Random Forest was capped at 100 trees generated.

Feature Selection

The exhaustive list of derived features also generates a large amount of uninformative features. This list is reduced by applying feature selection algorithms.

Feature selection for features originating from unstructured data was performed by Term Frequency - Inverse Document Frequency (TF-IDF).[29] TF-IDF is a widely used feature selection method, chosen for its ability to do feature selection in linear time with respect to the number of features. Features from a structured input source were first filtered through a weak gain ratio filter and further filtered via minimum Redundancy - Maximum Relevance (mRMR).[30] The mRMR algorithm minimizes redundancy and maximizes relevance of the selected features, but is only computationally feasible for smaller numbers of features. We have compared mRMR to gain ratio and information gain feature selectors, with mRMR selecting less features while yielding a similar F-measure (results not shown). Within a cross-validation fold, the feature selection thresholds were optimized by performing multiple evaluations for different threshold choices.

Data Integration

Two methods for data integration were evaluated for this paper: early and late data integration.[31] Early data integration (cf, Figure 4A) consists of integrating the features of different sources before training the model. Model training is then performed on the entire feature space (after feature selection).

Late data integration (Figure 4B) is an ensemble method in which the prediction results from separate models, trained on each distinct source, are used as input for a second (meta-) classifier or by means of composite methods such as voting, weighing, stacking, or averaging. We opted for training a meta-classifier that takes the predictions and class probabilities of the individual models as input for classification within the same fold. This second classifier is a Bayesian network, structured learning is performed with hill climbing.[32] This approach proved to perform consistently better than Random Forests and Naive Bayes (results not shown).

Metrics

To evaluate the experiments, micro-averaged F-measure is used (averaged over single codes). F-measure is the harmonic mean of precision and recall and is a good indicator for the overall predictive power of models. Our models return class probabilities, which allows for further tuning between optimizing precision and recall. This tuning was not carried out in this research, as the models were already optimized for F-measure. Micro-averaged F-measure was chosen for our interest in predicting correct codes for as many patients as possible, rather than ensuring good coverage of the different classes.[33]

Figure 4: Example of data integration for the ICD-9-CM code "430." This figure illustrates the difference between early and late data integration. (A) A pipeline for early data integration. (B) A pipeline for late data integration.

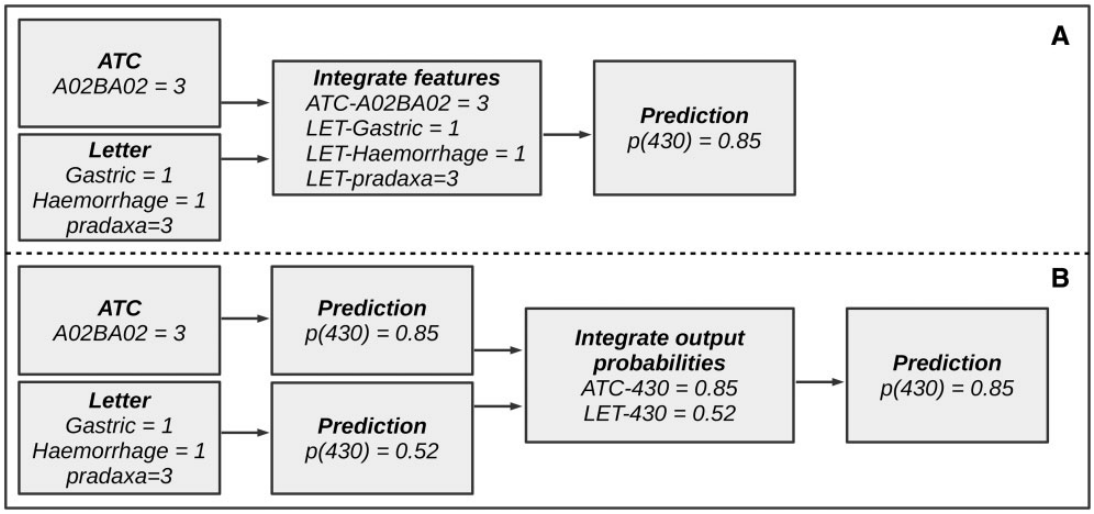


Table 1 shows the micro-averaged F-measure for the best performing algorithm for diagnostic ICD-9-CM codes and per combination of input data and medical specialty. Analogously, Table 2 represents the results for procedure codes. Complete performance results of both algorithms can be found in Supplementary Materials.

We compare these results to a frequency baseline, representing the achieved F-measure when classes present in >50% of the partials stays are assigned to each partial stay in that specific specialty. Data integration is evaluated for several interesting combinations of data sources. These combinations consist of the data types as grouped in the dataset section, with sometimes a specific input type added or removed.

In general, we notice that the F-measure varies significantly over the different specialties. Some sources seem to score better on specific specialties. For instance, both surgery reports and ATC-codes seem to be a good predictor for surgical specialties. The predictiveness of late data integration is superior to that of individual sources. For ICD-9-CM procedure codes, this is not the case, as RIZIV alone often beats late data integration. This can be explained by both RIZIV and ICD-9-CM procedure codes describing similar actions, often causing a one-to-one mapping between those codes. Disregarding the RIZIV data source, the performance increase by using late data integration is also visible for procedure codes.

Bag of Words text combination

The "all texts" combination is composed of a BoW of all different types of unstructured input associated with a stay. This can be considered an early form of data integration, with the difference that features representing the same word in different texts are merged instead of keeping them separated. In specialties such as pneumology, this combination will score better than the individual unstructured types, while for others (eg, gynecology), it will not quite reach the performance of the best text category. Overall, a smoothing effect of the individual unstructured types is seen, with a consistent improvement over the baseline.

Early data integration

For the early data integration approach, a slight decrease in predictiveness is often seen, compared to the best individual source in a certain combination. A combination existing of solely structured sources sometimes has slightly improved results, while a mixed combination with both structured and unstructured sources often shows a performance decrease. Early data integration is always outperformed by late data integration.

Late data integration

The results for late data integration with unstructured sources contain the same data sources as the "all texts" combination, but are based on a separate model for each individual source. This combination outperforms the "all texts" combination on all specialties, showing that the information is better captured when predictions are made individually. Adding predictions made on the "all texts" combination to the unstructured combination (ie, "unstruc + all texts"), we see a small increase in F-measure. Late data integration achieves superior results for other combinations as well.

The "all - RIZIV" combination encompasses all frequently seen sources in hospitals. The combination significantly improves results with respect only to the unstructured combination, showing that there is important and additional information present in structured data that is not captured by models using only unstructured data.

The "RIZIV" source, which is specific to Belgian hospitals, increases performance when added to the combinations for procedural ICD-9-CM codes. For diagnostic codes, they have a significantly smaller impact. An explanation for this is the one-to-one mapping between most RIZIV and ICD-9-CM procedure codes. Excluding RIZIV from the structured combination allows us to get a more objective view for predicting procedure codes.

In Figure 5, extra sources are iteratively added to the predictive model. In most cases, the average F-measure increases when adding sources. Each specialty seems to have different sources triggering a performance improvement.

DISCUSSION

The results shown above provide interesting insights into the combination of structured and unstructured data sources for automated clinical coding. First of all, individual data sources lead to major performance differences for the different specialties. From a perspective of portability towards other hospitals, countries and languages, the task of finding the most informative sources for each individual specialty would be arduous. Our proposed late data integration approach seems to be able to achieve this task in a scalable manner without the need to select a single best source.

Table 1: Micro-averaged F-measure for ICD-9-CM diagnostic codes

	ABDOMINAL SURGERY	NEUROSURGERY	ORTHOPEDICS	ONCOLOGY	OPTHALMOLOGY	CARDIOLOGY	GASTROLOGY	GYNEACOLOGY	HEMATOLOGY	OTORHINOLARYNGOLOGY	PAINCENTER	PNEUMOLOGY	THORAX AND VASCULAR	UROLOGY	AVERAGE
	2450	1750	4520	8700	4280	9550	4320	1950	4240	4150	3540	2200	1990	2950	4042
Frequency baseline	6.8	6.0	3.9	25.8	29.5	22.0	6.0	16.0	11.0	9.0	26.5	14.6	10.4	9.4	14.1
Single Source															
ATC	13.7	17.0	12.8	46.4	42.9	32.3	14.9	41.0	**37.8**	24.6	46.3	**32.3**	21.5	18.7	28.7
CODAP	12.9*	1.6	8.5	33.0*	5.6	24.1	14.4	24.9	14.0	13.5	25.7	22.2*	18.3	13.8*	16.6
LAB RESULT	13.6	14.5	10.3	47.2	36.2*	30.8	11.7	39.1	36.7	18.6	36.3	25.5	17.4	14.3	25.2
RIZIV	15.3	17.0	13.6	**52.4**	45.1	33.8	16.7	**43.5**	36.3	29.4	**50.3**	32.0	20.6	**22.3**	**30.6**
ATTESTATION	12.2*	10.0*	10.1	34.6*	47.6*	31.1*	9.9*	23.2*	18.4*	1.2*	39.7*	25.0*	16.4*	8.4*	20.6
LETTER	15.4*	**25.6***	**20.2***	43.8*	**48.1**	31.2*	**19.5***	37.7*	29.6*	30.3*	43.3*	31.2*	**29.2***	17.6*	30.2
NOTE	13.8	12.1	16.1*	42.0	44.5*	33.8*	14.0*	37.0*	29.9*	**32.5***	36.0*	27.6*	18.8	13.1	26.5
PROTOCOL	15.1*	20.6*	11.7*	34.5*	43.5*	**34.2**	18.6*	26.4*	18.1*	20.4*	35.8*	25.8*	22.7*	15.6*	24.5
OTHER TEXTS	3.3*	13.1*	1.3*	33.0*	47.5	10.5*	1.5*	0.0*	15.6*	1.6*	0.6	23.4*	3.6*	3.7	11.3
REQUEST	11.0	1.9	9.4	32.2*	4.0	30.3*	10.8*	6.8	12.7	19.6*	38.6*	8.8*	16.9*	2.8	14.7
SURGERY REPORT	**16.9***	16.0*	19.3*	34.6*	47.7*	28.6*	1.8*	27.8*	16.6*	29.5*	17.1*	4.3*	21.3*	16.6*	21.3
Bag of Words text combination															
ALL TEXTS	24.3*	25.9*	20.8*	46.6	46.3	35.1	20.7*	41.5	31.2	29.6	39.9	29.5	25.3	26.9*	31.7
Early data integration															
STRUCT – RIZIV	14.3	17.0	10.2	41.0	43.1	33.1	17.0	**45.4**	**36.1**	25.5	**46.4**	30.7	20.1	20.6	28.6
UNSTR	18.4	20.0	15.9	35.0	45.6	34.3	16.8	41.9	29.8	27.6	41.4	27.4	26.2	19.6	28.6
UNSTR + ALL TEXTS	**20.8***	**22.4**	17.3	36.1*	46.6	35.2	19.2	41.6	35.1	29.6*	41.1	29.4	**26.7***	21.8	30.2
ALL – RIZIV	**20.8***	20.5	16.0	42.1	46.6	35.1	17.0	44.1	35.4	29.7*	45.7	32.5	**26.7***	24.1	31.2
ALL	**20.8***	22.2	**17.9**	**44.5**	**46.9**	**36.3**	**19.3**	43.0	35.0	**29.8***	45.8	**36.2**	**26.7***	**26.0**	**32.2**
Late data integration															
STRUCT – RIZIV	16.8	19.6	14.0	53.4	46.1	36.2	19.5	44.7	43.9	27.8*	47.3	34.4	23.6	22.2	32.1
UNSTR	21.5*	26.5*	23.1*	48.0*	49.3*	37.7*	24.9*	44.6*	35.0*	34.6*	46.1*	33.7*	**29.9***	25.1*	34.3
UNSTR + ALL TEXTS	25.2*	27.4*	23.7*	51.1	49.1	37.0	24.9*	44.9*	35.3*	35.3*	45.6*	31.6	29.6	28.9*	35.0
ALL – RIZIV	**25.7***	27.7*	23.7*	59.5	49.2	38.4	**25.6***	48.4	47.2	35.6*	48.7	36.8	29.5	**29.8***	37.6
ALL	25.6*	**27.5***	**24.1***	**62.0**	**49.7**	**38.5**	25.3*	**49.6**	**48.9**	**35.9**	**51.4**	**38.9**	29.7	29.3*	**38.3**

Columns show different medical specialties, rows show the various input data sources and combinations. The best results for each group (namely, single source, data integration without meta-learning, data integration with meta-learning) are marked in bold. Only results for the best performing classifier are shown. A trailing asterisk indicates multinomial Naive Bayes, no mark indicates Random Forests. A darker background indicates better results.

The results show that integrating multiple data sources improves the classification of patient files with ICD-9-CM codes across all medical specialties in a consistent way. This strongly indicates that not all relevant information for assigning clinical codes is available in unstructured data and that adding structured data significantly improves performance. Since we use a relatively basic NLP pipeline, we hypothesize that improving the NLP pipeline will capture additional information from unstructured data that we are currently not able to use in our models. However, the performance impact of using additional structured information sources might be lower as a result of using improved NLP techniques.

Using an ensemble method makes data integration perform more consistently. This mostly shows that in order to get predictive information out of different data sources, deriving similar features and performing feature selection is not an effective approach. This suggests that more tailoring is needed towards selecting (and generalizing) the

Table 2: Micro-averaged F-measure for ICD-9-CM procedure codes

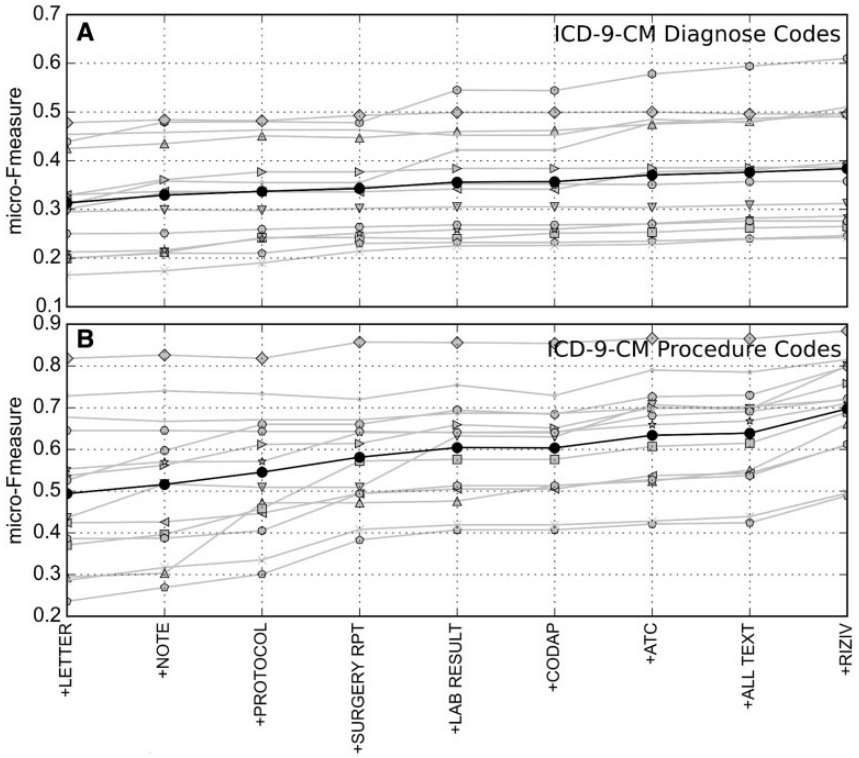
	ABDOMINAL SURGERY	NEUROSURGERY	ORTHOPEDICS	ONCOLOGY	OPTHALMOLOGY	CARDIOLOGY	GASTROLOGY	GYNEACOLOGY	HEMATOLOGY	OTORHINOLARYN-GOLOGY	PAINCENTER	PNEUMOLOGY	THORAX AND VASCULAR	UROLOGY	AVERAGE
Sample size	2450	1750	4520	8700	4280	9550	4320	1950	4240	4150	3540	2200	1990	2950	4042
Frequency baseline	5.0	6.0	3.6	62.1	60.4	44.6	14.1	31.3	27.1	60.5	31.5	25.9	10.1	11.5	28.1
	Single Source														
ATC	30.0	35.0	27.3	79.8	83.0	70.5	33.7	66.5	62.4	53.5	66.3	62.4	44.9	45.1	54.3
CODAP	14.9*	6.6*	2.5	71.5	49.0*	39.6	37.8	38.7*	25.1	8.3	36.9	20.2*	7.1*	19.2*	27.0
LAB RESULT	21.2	21.4	10.1	75.1	41.3	63.5	26.7	62.4	65.6	6.9	32.9*	57.3	27.3	20.5	38.0
RIZIV	**44.2**	**58.0**	**47.6**	**83.4**	**89.4**	**80.9**	**69.1**	**71.2**	**83.7**	**71.4**	**70.4**	**77.2**	**60.3**	**67.6**	**69.6**
ATTESTATION	10.8*	11.9*	19.3*	72.1	72.6	50.8*	17.5*	40.1*	32.0*	12.2*	33.4*	35.2*	14.4*	17.0*	31.4
LETTER	19.5*	39.0*	31.7	70.2*	80.9*	51.6*	28.2*	67.3*	41.7*	53.4	64.7*	52.9*	42.7*	26.8*	47.9
NOTE	20.0	21.3*	25.1*	73.0	78.7*	59.4*	20.8*	54.2	49.7	52.1	35.6*	51.7*	27.8*	18.1*	42.0
PROTOCOL	22.0*	32.5*	21.0*	68.9	71.0*	70.9	50.7	40.4*	35.5	26.9	33.4	51.6	38.2*	29.0*	42.3
OTHER TEXTS	12.9	12.8	4.2*	71.4	75.6	44.7*	7.5*	0.0*	32.3*	4.0*	0.0*	32.5*	11.3	10.9*	22.9
REQUEST	8.4*	11.6	15.8*	73.2*	73.5*	31.9*	37.7*	41.6*	24.5*	20.7*	34.8*	15.3*	12.8*	11.6*	29.5
SURGERY REPORT	34.7*	51.1*	41.8*	73.1*	85.8	52.8*	20.1*	40.9*	32.5*	63.0*	33.8*	13.4*	37.0*	48.0*	44.9
	Bag of Words text combination														
ALL TEXTS	33.9	41.2	32.7	72.9	80.6	67.5	45.3	62.1	52.9	58.1	57.6	60.2	44.5	45.7	53.9
	Early data integration														
STRUCT – RIZIV	28.1*	30.1	24.3	**76.2**	81.3	71.2	27.7	62.5	**70.2**	50.4	**66.1**	60.6	39.2*	30.8	51.3
UNSTR	33.1*	39.4	28.8	73.4	80.6	70.8	43.4	63.8	51.0	54.2	56.0	59.0	41.5	41.6	52.6
UNSTR + ALL TEXTS	33.1	41.0	30.9	73.6	79.6	70.2	43.2	62.5	53.0	56.6	52.2	59.5	43.5	45.9	53.2
ALL – RIZIV	33.0*	38.6	29.1	74.2	**82.2**	73.7	42.8	**66.0**	61.9	57.0	58.9	60.2	43.7	46.8	54.9
ALL	**33.5**	**42.4**	**31.7**	74.8	81.8	**74.9**	**45.2**	64.7	65.3	**59.6**	57.6	**63.5**	**46.3**	**48.2**	**56.4**
	Late data integration														
STRUCT – RIZIV	35.3	38.9	28.6	79.3	83.1	71.2	45.1	68.3	73.6	54.7	65.7	65.9	46.2	49.4	57.5
UNSTR	39.1*	49.5*	43.8	72.0*	85.6*	69.8	53.5*	67.1*	48.7*	64.2	63.9*	61.6*	51.0	57.6*	59.1
UNSTR + ALL TEXTS	40.0	50.8	44.9	70.4*	85.1	70.3	56.4	66.5*	54.6	64.3*	62.8*	61.7*	52.4	57.2*	59.8
ALL – RIZIV	42.3	53.7	45.7	77.5	85.5	74.0	57.1	70.8	71.3	67.5	68.9	69.5	55.6	61.1	64.3
ALL	**46.7**	**59.2**	**50.0**	**81.0**	**87.9**	**79.5**	**66.3**	**72.5**	**80.0**	**70.4**	**71.9**	**75.7**	**60.0**	**67.8**	**69.2**

Columns show different medical specialties, rows show the various input data sources and combinations. The best results for each group (namely, Single source, data integration without meta-learning, data integration with meta-learning) are marked in bold. Only results for the best performing classifier are shown. A trailing asterisk indicates multinomial naive Bayes, no mark indicates Random Forests. A darker background indicates better results.

individual features of different sources to improve early data integration.

When interpreting these results, we need to consider the concerns raised by Lawthers *et al* [34] and Romano *et al* [35] about the clinical validity of claims codes. Lawthers *et al* have shown that ICD-9-CM codes have limited value for the identification of medical complications because of the disconnect between discharge abstracts, which are collected for the sake of clinical processes, and ICD-9-CM codes, which are registered for billing and reimbursement purposes. While the disconnect cannot be denied, we designed a classification task that bears close resemblance to the actual task of clinical coders. When assigning clinical codes, coders are legally obliged to use the patient's medical record as the only source of information. A clinical coding application would use the same type of information. The implementation of ICD-10 coding systems only increases the need for solutions for the clinical coding task. The use of other measures with proven clinical

Figure 5: Micro-averaged F-measure for increasing number of data sources. From left to right on the *X* axis, additional data sources are added (using late data integration). (A) ICD-9-CM diagnostic codes, (B) ICD-9-CM procedural codes. Results for specific specialties are in gray, averaged results in black.



validity – such as outcome measures, process measures, or SNOMED codes – is a solution to the concerns raised, but they make for a different task.

A limitation of our study lies in the fact that the approaches have not been tested for portability to different hospitals, since the data originated from a single hospital. The factors missed here include local factors, such as hospital-specific aspects, country-specific regulations, goal-specific aspects (eg, clinical or administrative), and the difference between third-line and general hospitals. While using a relatively simple NLP-pipeline is beneficial to show portability across languages, a NLP-pipeline with more complex modules (eg, negation detection) has been shown to have superior performance.[12] Finally, the sole use of ICD-9-CM codes as the output space limits the portability to other coding systems.

When applying the algorithms to a computer-assisted clinical coding environment, the most important factor is improving the throughput time while maintaining the coding quality. When suggesting a code to a clinical coder, a model should return codes with a high recall, while limiting the amount of codes returned. When completely automating (a part of) the clinical code assignment, precision is a more important factor, as automatically assigned codes need to be completely correct.

CONCLUSIONS

Current state-of-the-art algorithms for prediction of ICD-9-CM codes are typically built on discharge summaries or radiology reports and are often tailored towards specific medical specialties. We evaluated the effect of integrating additional information sources, both structured and unstructured, and compared early and late data integration approaches for different machine learning algorithms (naive Bayes, Random Forests) and across a wide range of medical specialties. We show that the late data integration approach significantly improves the performance of these algorithms across all investigated medical specialties. All available data sources, independent of their (un) structured origin, can be added to the model without loss of predictive power for each of the different medical specialties. Evaluations have been performed on (Dutch-language) EHR data of a single hospital, but our approach was specifically designed to be portable to different contexts such as medical specialties, hospitals, specific coding systems, and languages.

FUNDING

This work was supported by the Agency for Innovation by Science and Technology in Flanders (IWT) grant number 131137.

COMPETING INTERESTS

The authors report no potential conflicts of interest.

CONTRIBUTORS

All the authors contributed to the design of the study. E.S. collected and analyzed the datasets and performed the experiments. All authors contributed in writing the manuscript. K.L., L.L., W.D., and T.V.dB. provided guidance and contributed from their respective areas of expertise. W.D. and T.V.dB. academically supervised the project. All authors approved the final version of the paper.

SUPPLEMENTARY MATERIAL

Supplementary material is available online at http://jamia.oxfordjournals.org/.

REFERENCES

1. Hsiao C-J, Hing E. *Use and Characteristics of Electronic Health Record Systems Among Office-Based Physician Practices, United States, 2001-2012*. US Department of Health; Human Services, Centers for Disease Control; Prevention, National Center for Health Statistics, United States.
2. Cimino, JJ. Improving the electronic health record—are clinicians getting what they wished for? *JAMA*. 2013;309(10):991–992.
3. WHO. *International Classification of Diseases*. http://www.who.int/classifications/icd/en/. Accessed 25 March 2015.
4. WHO. *International Classification of Primary Care*. 2nd edn. 2003. http://www.who.int/classifications/icd/adaptations/icpc2/en/. Accessed 25 March 2015.
5. WHO. *International Classification of Diseases, Clinical Modification (Ninth Revision)*. http://www.cdc.gov/nchs/icd/icd9cm.htm. Accessed 25 March 2015.
6. Ramanathan R, Leavell P, Stockslager G, *et al*. Validity of International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) Screening for Sepsis in Surgical Mortalities *Surg Infect*. 2014;15 (5):513–516.
7. Carroll RJ, Bastarache L, Denny JC. R PheWAS: data analysis and plotting tools for phenome-wide association studies in the R environment. *Bioinformatics*. 2014;30(16):2375–2376.
8. Goldstein I, Arzumtsyan A, Uzuner Ö. Three Approaches to Automatic Assignment of ICD-9-CM Codes to Radiology Reports. *AMIA Ann Symp Proc*. 2007:279–283.
9. Kevers L, Medori J. Symbolic classification methods for patient discharge summaries encoding into ICD. In *Advances in Natural Language Processing*. Springer; 2010:197–208.
10. Pakhomov SVS, Buntrock JD, Chute CG. Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques. *JAMIA*. 2006;13 (5):516–525.
11. Farkas R, Szarvas G. Automatic construction of rule-based ICD-9-CM coding systems. *BMC Bioinformatics*. 2008;9(3):S10.
12. Pestian JP, Brew C, Matykiewicz P, *et al*. A shared task involving multi-label classification of clinical free text. *Assoc Computational Linguistics, In Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*. 2007;97–104.
13. Stanfill MH, Williams M, Fenton SH, *et al*. A systematic literature review of automated clinical coding and classification systems. *JAMIA*. 2010;17(6):646–651.
14. Perotte A, Pivovarov R, Natarajan K, *et al*. Diagnosis code assignment: models and evaluation metrics. *JAMIA*. 2014;21(2):231–237.
15. Saeed M, Villarroel M, Reisner AT, *et al*. Multiparameter intelligent monitoring in intensive care II (MIMIC-II): a public-access intensive care unit database. *Crit Care Med*. 2011;39(5):952.
16. Abhyankar S, Demner-Fushman D, Callaghan FM, McDonald CJ. Combining structured and unstructured data to identify a cohort of ICU patients who received dialysis. *JAMIA*. 2014;21(5):801–807.
17. Pathak J, Bailey KR, Beebe CE, *et al*. Normalization and standardization of electronic health records for high-throughput phenotyping: the SHARPn consortium. *JAMIA*. 2013;20(e2):e341–e348.
18. Pletscher-Frankild S, Pallejá A, Tsafou K, *et al*. DISEASES: Text mining and data integration of disease–gene associations. *Methods*. Elsevier; 2014;74: 83–89.
19. Daemen A, Gevaert O, Ojeda F, *et al*. A kernel-based integration of genome-wide data for clinical decision support. *Genome Med*. 2009;1(4):39.
20. Brookes BC. The derivation and application of the Bradford-Zipf distribution. *J Document*. 1968;24(4):247–265.
21. WHO. *Anatomical Therapeutic Chemical (ATC) Classification System*. 2015. http://www.whocc.no/atc/structure_and_principles/. Accessed 25 March 2015.
22. BDSP. Cahier d Observation Descriptif de L Activité Palliative (CODAP). http://www.bdsp.ehesp.fr/Base/327741/. Accessed 25 March 2015.
23. RIZIV. *Rijksintituut Voor Ziekte- En Invaliditeitsuitkeringen Nomenclature*. http://www.riziv.fgov.be/NL/nomenclatuur/Paginas/default.aspx#.VOX1TzU2x0x. Accessed 25 March 2015.
24. Nouraei SAR, O'Hanlon S, Butler CR, *et al*. A multidisciplinary audit of clinical coding accuracy in otolaryngology: financial, managerial and clinical governance considerations under payment-by-results. *Clin Otolaryngol*. 2009;34 (1):43–51.
25. Bosch AVd, Busser B, Canisius S, Daelemans W. An efficient memory-based morphosyntactic Tagger and Parser for Dutch. *LOT Occasional Series*. 2007;7:191–206.
26. McCallum A, Nigam K. A comparison of event models for Naive Bayes text classification. *In AAAI-98 Workshop on 'Learning for Text Categorization'* . 1998.
27. Breiman L. Random forests. *Mach Learn*. 2001;45 (1):5–32.
28. Hall M, Frank E, Holmes G, *et al*. The WEKA Data Mining Software: an Update. *SIGKDD Explorations*. 2009;11(1):10–18.
29. Salton G, Buckley C. Term weighting approaches in automatic text retrieval. *Inform Process Manag*. 1988;24(5):513–523.
30. Peng H, Long F, Ding C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 2005;27(8):1226–1238.
31. Hamid JS, Hu P, Roslin NM, *et al*. Data integration in genetics and genomics: methods and challenges. *Hum Genom Proteomics*. 2009;1:1
32. Cooper G, Herskovits E. A Bayesian method for the induction of probabilistic networks from data. *Mach Learn*. 1992;9(4):309–347.
33. Jackson P, Moulinier I. *Natural Language Processing for Online Applications: Text Retrieval, Extraction and Categorization*. John Benjamins Publishing; 2007:5.
34. Lawthers AG, McArtney RG, Davis RB. Identification of in-hospital complications from claims data: is it valid? *Med Care*. 2000;38:785–795.
35. Romano PS, Chan BK, Schembri M, Rainwater J. Can administrative data be used to compare postoperative complication rates across hospitals? *Med Care*. 2002;40:856–867.

AUTHOR AFFILIATIONS

[1]ADReM (Advanced Database Research and Modelling), Biomedical Informatics Research Center Antwerp (biomina), University of Antwerp, Antwerp, Belgium

[2]Department of Medical Information, Antwerp University Hospital, Antwerp, Belgium

[3]Computational Linguistics and Psycholinguistics (CLiPS) Research Center, University of Antwerp, Antwerp, Belgium

[4]Biomedical Informatics Research Center Antwerp (biomina), University of Antwerp - Antwerp University Hospital, Belgium; ADReM (Advanced Database Research and Modelling), University of Antwerp, Antwerp, Belgium