# Unstructured data in Predictive Process Monitoring: lexicographic and semantic mapping to ICD-9-CM codes for the Home Hospitalization Service

Massimiliano Ronzani[3], Roger Ferrod[1], Chiara Di Francescomarino[3], Emilio Sulis[1], Roberto Aringhieri[1], Guido Boella[1], Enrico Brunetti[1 2], Luigi Di Caro[1], Mauro Dragoni[3], Chiara Ghidini[3], and Renata Marinello[2]

[1] University of Turin, Torino, Italy
{roberto.aringhieri,guido.boella,enrico.brunetti,
luigi.dicaro,roger.ferrod,emilio.sulis}@unito.it
[2] City of Health and Science, Torino, Italy
{ebrunetti,rmarinello}@cittadellasalute.to.it
[3] Fondazione Bruno Kessler, Trento, Italy
{dfmchiara,dragoni,ghidini,mronzani}@fbk.eu

**Abstract.** The large availability of hospital administrative and clinical data has encouraged the application of Process Mining techniques to the healthcare domain. Predictive Process Monitoring techniques can be used in order to learn from these data related to past historical executions and predict the future of incomplete cases. However, some of these data, possibly the most informative ones, are often available in natural language text, while structured information — extracted from these data — would be more beneficial for training predictive models.
In this paper we focus on the scenario of the Home Hospitalization Service, supporting the team in making decisions on the home hospitalization of a patient, by predicting whether it is likely that a new patient will successfully undergo home hospitalization. We aim at investigating whether, in this scenario, we can take advantage of mapping unstructured textual diagnoses, reported by the doctor in the Emergency Department, into structured information, as the standardized disease ICD-9-CM codes, to provide more accurate predictions. To this aim, we devise two different approaches involving respectively lexicographic and semantic distance for mapping textual diagnoses in ICD-9-CM codes and leverage the structured information for making predictions.

**Keywords:** Healthcare processes · Predictive Process Monitoring · Natural Language Processing· Home Hospitalization Service

## 1 Introduction

The improvement of healthcare processes and the support of clinical personnel in making decisions might have an impact on the efficiency of the healthcare services, as well as on the quality of the work of the clinical personnel, who sparing time in administrative tasks has more time available for taking care of patients, thus improving the patients' quality of life. Process Mining (PM) [1], which deals with the analysis of business processes based on their behaviour — observed and recorded in event logs — can be a useful instrument in this setting. PM deals with the analysis of business process event logs in different ways [3], including process discovery (i.e., extracting process models from an event log) [1], predictions of the future of ongoing cases [17] and process optimization [1]. PM techniques can be leveraged for the discovery and analysis of both clinical and administrative processes in healthcare. The application of PM techniques is further encouraged by the wide availability of administrative and clinical data in hospitals. These data could be leveraged for discovering (and improving) processes, as well as for supporting hospital teams in making decisions on clinical and administrative issues [4, 23]. It often happens that these data are collected in national standard forms and documents, shared among several hospitals on the national area. For instance, in Italy, one of these documents is the Hospital Discharge Form (HDF), which collects information related to the clinical history of a patient during his/her hospitalization. The data collected in the discharge form range from data (with temporal information) related to the hospital admission, discharge and examinations carried out during the hospitalization to data such as the number of days of hospitalization. Unfortunately, however, not all these data are structured. Some of them, possibly the most informative ones, are textual unstructured fields, as in the case of the patients' diagnoses reported by the doctor at the arrival of the patient at the Emergency Department.

In this paper we aim at investigating whether we can take advantage of mapping unstructured data into the structured information provided by the ICD-9-CM[4] taxonomy when making predictions in the scenario of the Home Hospitalization Service. We extend the work in [5], where we investigated a lexicographic distance for mapping textual diagnoses to ICD-9-CM codes, with a semantic distance. We first provide some preliminaries (Section 2) and introduce the Home Hospitalization scenario (Section 3). In Section 4 we report about the proposed approach that aims at (i) mapping unstructured data to ICD-9-CM codes via lexicographic or semantic match; and (ii) leveraging this structured information when making predictions. We report on the evaluations carried out in Section 5 and we finally conclude in Section 7.

## 2 Background

In this section we report the background concepts useful for understanding the remainder of the paper.

*Predictive Process Monitoring.* Predictive Process Monitoring (PPM) [17] is a relatively new branch of PM that aims at predicting at runtime and as early as possible the future development of an ongoing incomplete execution of a process.

[4] https://www.cdc.gov/nchs/icd/icd9cm.htm

Predictions related to the future of an incomplete process execution (as known as *case*) of state of-the-art approaches can be classified in macro-categories [10]: numeric predictions (e.g., time or cost predictions); categorical predictions (e.g., risk predictions or specific categorical outcome predictions such as the fulfillment or the violation of a certain property); as well as to next activities predictions (e.g, the sequence of future activities, possibly with their payloads).

Together with these techniques, few frameworks have also been recently developed implementing and collecting these techniques, such as for instance Nirdizati [21]. These frameworks take as input a set of past executions and use them to train predictive models to be used for providing users with predictions at runtime. They are usually characterized by two main modules: one for the case encoding, and one for the supervised learning. Each of them can be instantiated with different techniques.

*ICD-9-CM.* ICD-9-CM is the ninth edition of the *International Classification of Diseases*. It contains a structured standard codification of diseases and procedures that is used internationally both in the management of public health and for statistical and epidemiological purposes.

The ICD-9-CM assigns specific codes (and associated descriptions) to both diseases and procedures. It is organized in the form of a taxonomy, so that each code corresponding to a specific disease variant (subprocedure) is classified as a disease (procedure), which, in turn, is classified as a category of diseases (procedures) and so on. In the case of the diagnoses, each code is composed of five digits: the first three digits represent a high level disease category, the fourth digit indicates the specific disease, while the last digit identifies the specific variant of the disease. In turn, the first three digits are further classified according to number interval ranges corresponding to families of diseases. For instance, the code **410.22** corresponding to the description *Acute myocardial infarction of inferolateral wall, subsequent episode of care* is a leaf of the hierarchy:

- **390–459**: *Diseases of The Circulatory System*
  - **410–414**: *Ischemic Heart Diseases*
    - **410**: *Acute myocardial infarction*
      - **410.2**: *Acute myocardial infarction of inferolateral wall*
        - **410.22**: *Acute myocardial infarction of inferolateral wall, subsequent episode of care*

This simple representation of the taxonomy allows us to select, for a given diagnosis code, the level of abstraction, i.e., the ancestor, among the low levels of the taxonomy, by truncating the last or the last two digits of the ICD-9-CM code.

## 3 The Home Hospitalization Service Scenario

The Home Hospitalization Service (HHS) of the City of Health and Science (CHS), which has been in operation for over 30 years, has proven to be a valid alternative to hospitalization for a variety of acute and chronic exacerbated diseases [22], such as uncomplicated ischemic stroke, congestive heart failure, exacerbations of chronic obstructive pulmonary disease, onco-hematological diseases with high transfusion requirements, dementia with behavioral disorders [14]. The HHS consists of a multidisciplinary team. The essential criteria for taking care of an acute patient at home are threefold: (i) clinical aspects, e.g., no need for continuous or invasive monitoring of vital parameters, as well as to perform invasive diagnostic-interventions; (ii) geographical aspects (residence in the area of competence of the HHS); (iii) social welfare (constant presence of one or more caregivers, formal or informal). Every year, the service manages about 500 admissions of patients coming in most cases from the same hospital and in small part upon direct request of the General Practitioner (GP). At the end of the treatment period, more than 80% of patients are discharged to the GP, 10.5% die during hospitalization and about 8% is moved to hospital. Over the past 8 years, the percentage of patients unable to continue care management at home has remained constant, despite the increase in clinical complexity and care burden of patients taken into care. In 2018, HHS patients were 492 with a high average age (about 84 years). The overall goal is supporting the HHS team in the timely identification and notification of the patients that can be managed through the HHS, as well as in the efficient management of the HHS processes.

*Data Description.* The administrative and clinical data available so far for the specific case study are related to Emergency Department Discharge Forms (EDDF) and to the Hospitalization Discharge Forms (HDF) of about 400 CHS patients benefitting from the HHS. The EDDF contains information collected at the Emergency Department (ED) such as: (i) date and time information related to the ED admission, triage, discharge, last and latest update of the anamnesis; (ii) structured information e.g., on the patient triage colour code; and (iii) textual notes e.g., on the diagnosis. The HDF contains instead information about the clinical history of the patient during the hospitalization, such as: (i) date and time information related to e.g., the hospital admission, discharge, main intervention; (ii) structured information related to e.g., patients' data (age, sex, civil status, etc.), number of visits; and (iii) textual information related to e.g., the hospitalization cause and the anamnesis.

## 4 Approach

In order to support the HHS team in making decisions on the home hospitalization of a patient, the overall idea is applying existing approaches of PPM to the data related to the administrative and clinical management of ED patients. To this aim, patient data need to be transformed into a trace describing the history of the patient and used as features to learn and provide predictions about the home hospitalization of the patient. Most of these data are structured bits of information, while others, equally or more informative, are collected as unstructured text, as for instance the diagnosis informally reported by the doctor when the patient reaches the ED. In order to be able to apply PPM approaches and properly leverage this information when making predictions, we devised the following pipeline:

- we preprocess data so as to generate an event log describing the patient histories (Section 4.1);
- we map the informal diagnosis descriptions into the standardized diagnosis codes of the ICD-9-CM taxonomy (Section 4.2);
- we leverage the mapped structured ICD-9-CM code or one of its ancestors as a structured feature to be used in making predictions (Section 4.3).

### 4.1 Data Preprocessing and Analysis

The dataset related to the HDFs extracted from the hospital information systems has first been cleaned by removing hospitalizations of few days or "routine" procedures and then joined with the dataset of the ED. The following steps have been then applied to the joined dataset:

- The dataset has been transformed into an event log. The hospital discharge id number has been used as *trace id*. For the HDF data, date and time fields related to the hospital admission, discharge, and to the interventions performed by the patient during the hospitalization have been used as timestamps for the activities `H_admission`, `H_discharge` and for the intervention activities (labelled with the corresponding ICD-9-CM code or with the procedure category they belong to in the ICD-9-CM procedures), respectively. Patient personal data and other structured data, such as the setting of referral, have been added as case attributes. Similarly, for EDDF data, date and time fields related to the ED admission, discharge, triage, anamnesis and diagnostic hypothesis have been used as timestamps for the `ED_admission`, `ED_discharge`, `ED_triage`, `ED_anamnesis`, `ED_diagnostic_hp`, respectively. Diagnosis and other few attributes have been instead used as case attributes. The resulting event log is composed of 413 cases with 270 different paths and 49 different activities.
- In order to be able to make predictions at the time of the discharge from the ED, each trace in the log has been truncated at the time of the activity `ED_discharge`, and the attributes that cannot be known at the time of the ED discharge have been removed, e.g. the attribute `H_number_of_days_in_the_facility`, which is known only at the end of the hospitalization.

Finally, data have been labelled according to whether (i) the patient has been hospitalized at home and the hospitalization had a positive outcome (HH, i.e., Home Hospitalization); or (ii) she/he has been hospitalized in a different ward or the home hospitalization had a negative outcome (NO-HH). Out of the 413 cases, 368 (89%) were labeled with HH and 45 (11%) with NO-HH.

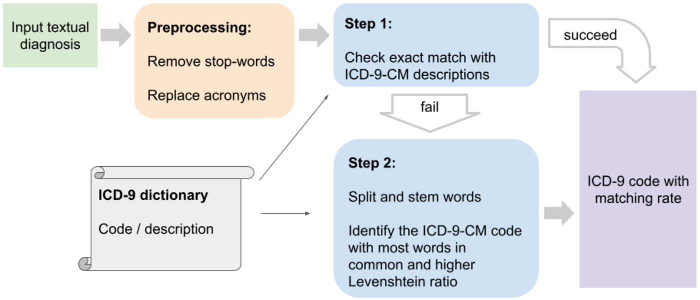

Input textual diagnosis
Preprocessing:
Remove stop-words
Replace acronyms
Step 1:
Check exact match with ICD-9-CM descriptions
succeed
fail
ICD-9 dictionary
Code / description
Step 2:
Split and stem words
Identify the ICD-9-CM code with most words in common and higher Levenshtein ratio
ICD-9 code with matching rate

Fig. 1: Overview of the ICD-9-CM mapping pipeline.

### 4.2 Mapping the Diagnosis field to the ICD-9-CM dictionary

In this section we briefly illustrate the Natural Language Processing (NLP) techniques applied to short unstructured textual diagnoses in order to map them to structured ICD-9-CM diagnosis codes. Since all the textual diagnoses we want to decode are in Italian, we refer to the Italian translation of the ICD-9-CM descriptions[5]. This is used to create a description/code dictionary of diseases, after the removal of those codes starting with letter 'E' (supplementary classification of external causes) and 'V' (supplementary classification of factors influencing health status and contact with health services). The technique developed is organized in three steps, which are illustrated in Fig. 1:

- **Preprocessing step:** the input textual diagnosis is preprocessed with the removal of stop-words and proper replacement of acronyms;
- **Step 1:** if the input diagnosis is already exactly matching one of the ICD-9-CM descriptions, then the corresponding code is taken from the dictionary;
- **Step 2:** if in the previous step there is no match, we try to identify among the ICD-9-CM descriptions the closest one to the input diagnosis. To this aim, we can follow a pure *lexicographic* or a *semantic* approach. We detail in the following the two alternative approaches for carrying out **Step 2**.

**Lexicographic approach.** In order to identify among the ICD-9-CM descriptions the one closest to the input diagnosis we can go through the following procedure:

- **Step 2 lexicographic:**
  - we stem words[6] in both input and ICD-9-CM diagnoses. Moreover we delete some undefined adjective (e.g. "non specificato" that means unspecified); this is done in order to prefer generic diagnoses to specialized ones.

[5] https://www.salute.gov.it/portale/documentazione/p6_2_2_1.jsp?lingua=italiano&id=2251

[6] We used snowball stemmer from nlkt package https://www.nltk.org/_modules/nltk/stem/snowball.html

- we identify the subset of ICD-9-CM diagnoses that share the maximum number of stems with the input diagnosis $D_{\text{input}}$
- among this subset we select the diagnosis $D_{\text{ICD9}}$ with the highest value of the metrics $g(D_{\text{input}}, D_{\text{ICD9}})$ defined as

$$g(D_1, D_2) = \frac{1}{len(D_1)len(D_2)} \sum_{\substack{stem_1 \in D_1 \\ stem_2 \in D_2}} lev.ratio(stem_1, stem_2) \tag{1}$$

  where $lev.ratio(s_1, s_2)$ is the Levenshtein ratio between two stems $s_1, s_2$ and $len(D)$ counts the number of stems composing the sentence $D$. The denominator normalizes the metrics: since the numerator grows with the number of words in the diagnoses, the metrics is a number between 0 and 1.

Once the input diagnosis is associated to an ICD-9-CM code, we assign a *Lexicographic score* ($SC_L$) from 0 to 100. This metrics aims at estimating the probability that the mapping is correct. If a match is found in **Step 1**, then $SC_L$ = 100; if the mapping comes in **Step 2**, then it is computed as follows:

$$SC_L = \min\left(\omega\, g(D_1, D_2)(1 + r(D_1, D_2)), 100\right) \tag{2}$$

where $\omega$ is a weight set to 50, $g$ is the metrics defined in (1) and $r$ is the number of stems in common between diagnoses $D_1$ and $D_2$. The quality of this choice for the metrics is investigated in Section 5.1.

The value of the *Lexicographic score* will be used to as a filter parameter: when its value is above a certain *Lexicographic score* threshold, we will use the associated ICD-9-CM code, otherwise we will assign a default code "0". The impact of the choice of the *Lexicographic score* threshold on the predictions is inspected in Section 5.2.

**Semantic approach.** The pipeline proposed above is based uniquely on the Levenshtein lexicographic distance. This means that diagnoses with the same semantics but with a different wording have a high lexicographic distance. For instance, *pyrexia* and *fever*, though having the same semantics, will result in a high lexicographic distance. In the semantic approach, instead, we want to take into account the semantic distance between words and for this we leverage a word embedding model.

For the embedding model we relied on CODER [26], a multilingual model created specifically to deal with medical nomenclature, thanks to the integration of Knowledge Graph, such as UMLS, and mBERT [8]. Behind the functioning of word embeddings lies the principle of distributional semantics, according to which: *"linguistic items with similar distributions have similar meanings"*; therefore, vectors corresponding to similar words will appear close together in the embeddings space. For instance, the vectors of *pyrexia* and *fever* will be rather close in the embedding space. Moreover, we have observed how important it is, in this context, to extend this principle by integrating the information expressed in UMLS in order to correctly compute the similarity between medical terms. This similarity has been calculated on the basis of the cosine distance, which is defined as:

$$vect.distance(w_1, w_2) = 1 - cos(\theta) = 1 - \frac{w_1 \cdot w_2}{\|w_1\| \|w_2\|} \tag{3}$$

where $\theta$ is the angle between the vector representation of words $w_1$ and $w_2$.

In order to identify among the ICD-9-CM descriptions the one closest to the input diagnosis we can go through the following procedure:

- **Step 2 semantic:**
  - we split input and ICD-9-CM diagnoses in two lists of words and we delete some undefined adjectives (e.g. "non specificato" that means unspecified); this is done in order to prefer generic diagnoses to specialized ones.
  - we identify the subset of ICD-9-CM diagnoses that share with the input diagnosis $D_{\text{input}}$ the maximum number of *semantically similar* words. Given two sentences $D_1$ and $D_2$, two words $w_1 \in D_1, w_2 \in D_2$ are *semantically similar* when:

    $$vect.distance(w_1, w_2) \leq 0.15 \tag{4}$$

  - in this subset we select the diagnosis $D_{\text{ICD9}}$ with the highest value of the metrics $h(D_{\text{input}}, D_{\text{ICD9}})$ defined as:

    $$h(D_1, D_2) = (1 - Q(\mathbf{v}))(1 - D(\mathbf{v})) \tag{5}$$

    where $Q$ and $D$ are respectively the value of the first quartile and the value of the first decile computed on the population $\mathbf{v}$ of all the semantic distances between the words of the two diagnoses:

    $$\mathbf{v} = \{vect.distance(w_1, w_2), \forall\, w_1 \in D_1, w_2 \in D_2\}. \tag{6}$$

    The metrics is a number between 0 and 1.

Similarly to the lexicographic case (2), when the ICD-9-CM code is associated to the input diagnosis we assign a *Semantic score* ($SC_S$). This score is set to 100 if the diagnosis is matched during **Step 1**, otherwise it is computed with the following formula:

$$SC_S = \min\left(\omega\, h(D_1, D_2)(1 + s(D_1, D_2)), 100\right) \tag{7}$$

where $\omega$ is a weight set to 50, $h$ is Equation (5) and $s$ is the number of semantically similar words in $D_1$ and $D_2$, computed as described in Equation (4). The quality of this choice for the metrics is investigated in Section 5.1.

The two pipelines described in this section are used separately to associate the ICD-9-CM code to the input diagnosis; the respective results are then used in the predictive model and their performances are compared in Section 5.2.

### 4.3 Predicting the Home Hospitalization Outcome

The structured data, either extracted from the diagnosis textual fields or already stored in structured fields, can then be provided as input to PPM algorithms that use these features to learn a predictive model. At runtime, when the HHS team has to decide whether a new patient should undergo the home hospitalization, given the features of the new patient, the predictive model will predict whether it is likely that she/he will successfully undergo home hospitalization (HH) or whether it is better to proceed with the hospitalization in another ward (NO-HH). PPM algorithms, e.g., the ones available in Nirdizati [21], a PPM tool that collects a rich set of state-of-the-art approaches based on machine learning algorithms, can be used to train a predictive model able to learn the correlations between variables that describe the patient data and examinations he/she has carried out (features) and the hospitalization at home or in another hospital ward.

## 5 Evaluation

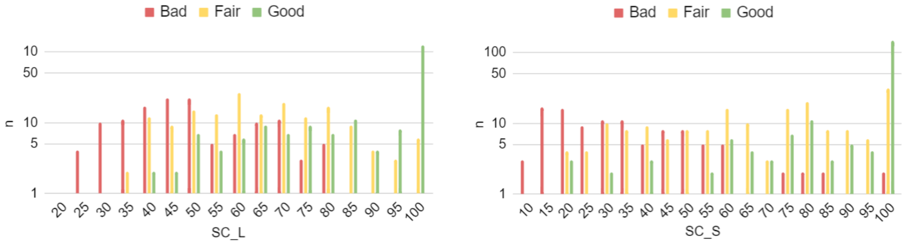
In this section we evaluate the proposed approach. In detail, we first evaluate the mapping of the textual fields to the ICD-9-CM disease codes (Section 5.1) and then the impact of the mapping to ICD-9-CM codes at different levels of abstraction of the ICD-9-CM taxonomy, when making predictions on the home hospitalization outcome (Section 5.2).

### 5.1 ICD-9-CM Mapping Evaluation

In this section we aim at evaluating: (i) the correctness of the ICD-9-CM mappings obtained using the two approaches presented in Section 4.2; (ii) whether the *Lexicographic score* and the *Semantic score* are good metrics to evaluate the quality of each ICD-9-CM mapping.

In order to evaluate their correctness, we analyzed the ICD-9-CM mappings given by the two approaches to 490 different textual diagnoses in the dataset. We then asked a domain expert to classify each mapping according to three categories:

- *Good* mapping: the assigned ICD-9-CM code correctly represents the semantics of the textual diagnosis, e.g. “anemia” (anemia) is mapped to code 599.0 corresponding to “altre e non specificate anemie” (other and unspecified anemias)
- *Fair* mapping: the assigned ICD-9-CM code represents only partially the semantics of the textual diagnosis, possibly it represents a superclass , e.g. “leucemia e polmonite” (leukemia and pneumonia) is mapped to code 208.9: “leucemia non specificata” (unspecified leukemia), so we miss the information about pneumonia
- *Bad* mapping: the assigned ICD-9-CM code represents a diagnosis that is uncorrelated to the textual one, e.g. “acufeni” (tinnitus) is mapped to code 706.1: “altre acni” (other acni)



(a)

(b)

Fig. 2: Number of diagnoses for *score* values. (a) Lexicographic approach. (b) Semantic approach.

| $T_L$ | **tot** | ***good*** | ***fair*** | ***bad*** |
|---|---|---|---|---|
| 0 | 100% | 41% | 33% | 26% |
| 53 | 70% | 38% | 23% | 9% |
| 70 | 45% | 33% | 10% | 2% |

(a)

| $T_S$ | **tot** | ***good*** | ***fair*** | ***bad*** |
|---|---|---|---|---|
| 0 | 100% | 42% | 36% | 22% |
| 50 | 70% | 40% | 26% | 4% |
| 79 | 45% | 33% | 11% | 1% |

(b)

Table 1: Percentage of ICD-9-CM diagnosis mappings with: (a) $SC_L$ value higher than $T_L$ for the lexicographic approach; (b) $SC_S$ value higher than $T_S$ for the semantic approach. The values of the thresholds $T_L$ and $T_S$ are chosen so to filter respectively 100%, 70% and 45% of all the diagnoses.

Based on the classification of the domain expert, we found that

- with the lexicographic approach: 41% of the mappings are *good*, 33% are *fair* and 26% are *bad*.
- with the semantic approach: 42% of the mappings are *good*, 36% are *fair* and 22% are *bad* mappings.

This represents a reasonable result. Indeed, by discarding the *bad* mappings we are able to fairly map 74% and 78% of the textual diagnoses for the lexicographic and the semantic approach, respectively. Moreover, we notice that the results returned by the semantic approach are overall better than the ones of the lexicographic approach.

In order to check whether the two *scores* are good metrics to evaluate the quality of the mappings, so as to use these metrics to discriminate the mappings we can trust as features for prediction tasks, we show in Fig. 2) the distributions of the three categories of diagnoses with respect to the relative *score* for each of the two approaches. The plot shows that in both cases most of the *bad* mappings have a low *score* value.

The metrics look reasonably good in separating *bad* mappings and hence, setting a *Lexicographic score* threshold value $T_L$ (respectively *Semantic score* threshold value $T_S$), they can be used to automatically exclude most of the bad mappings. Table 1 reports for different $T_L$ ($T_S$) values the percentage of diagnosis mappings that are above the threshold for each quality category.[7]

### 5.2 Home Hospitalization Outcome Prediction Evaluation

In this section we report about the accuracy of the predictions related to the HHS scenario. The accuracy of the predictions is evaluated using the *Matthews correlation coefficient metric* (MCC) [18] that is defined as follows:

$$\mathrm{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \qquad (8)$$

where TP, TN, FP, FN are respectively true positive, true negative, false positive and false negative predictions. The MCC metrics ranges from $-1$ to 1, where a perfect prediction measures 1, a random prediction measures 0 and a completely wrong prediction measures $-1$. In unbalanced datasets, like ours, where the number of positive and negative traces is very different (368 vs. 45), this metrics is more suitable than others like accuracy and F-measure for measuring the quality of the predictions [7].

In order to evaluate whether structured features, as the ICD-9-CM codes or its ancestors, rather than unstructured ones, as textual diagnoses, can be leveraged to get more accurate predictions, we analyzed and compared the results obtained with different sets of features:

- without the diagnosis (no_diag);
- with the textual diagnosis (text_diag);
- with the ICD-9-CM code assigned to the textual diagnosis or one of its ancestors via the lexicographic match (icd9_diag_lex(all)).
- with the ICD-9-CM code assigned to the textual diagnosis or one of its ancestors via the semantic match (icd9_diag_sem(all)).

These last two cases are further refined in different sub-cases based on two parameters: (i) the threshold values $T_L$, $T_S$; for each of them we consider two reference values: one that filters the 70% of the diagnoses ($T_L = 53$, $T_S = 50$) and one that filters the 45% of the diagnoses ($T_L = 70$, $T_S = 79$), see Table 1; and (ii) the level of abstraction of the ICD-9-CM classification, that corresponds to the number of digits that we trim from the right side of the ICD-9-CM codes (see Section 2): the higher the number of digits trimmed, the higher the abstraction level in the ICD-9-CM taxonomy. Here we consider two abstraction levels: the full ICD-9-CM code corresponding to the specific diagnosis, and the code with two digits trimmed corresponding to its ancestor diagnosis group.

As predictive model we used a Random Forest classifier on the incomplete traces properly preprocessed as described in Section 4.1. Moreover, we tested the predictions assuming we have observed only the first five activities at the ED.

[7] The percentages in Table 1 refer to the number of mappings per diagnosis. Note that these are in principle different from the number of mappings per trace in which the diagnosis appears, since the same diagnosis may appear in more than one trace.

| Diagnosis information | Description | avg(MCC) | $\sigma$(MCC) | max(MCC) |
|---|---|---|---|---|
| no_diag | without diagnosis | 0.51 | 0.09 | 0.65 |
| text_diag | textual diagnosis | 0.4 | 0.1 | 0.6 |
| icd9_diag_lex_70%-5 | ICD-9-CM, lexicographic, $T_L = 53$, 5 digits | 0.58 | 0.05 | 0.65 |
| icd9_diag_lex_70%-3 | ICD-9-CM, lexicographic, $T_L = 53$, 3 digits | 0.56 | 0.07 | 0.65 |
| icd9_diag_lex_45%-5 | ICD-9-CM, lexicographic, $T_L = 70$, 5 digits | 0.57 | 0.05 | 0.70 |
| icd9_diag_lex_45%-3 | ICD-9-CM, lexicographic, $T_L = 70$, 3 digits | 0.61 | 0.04 | 0.70 |
| icd9_diag_sem_70%-5 | ICD-9-CM, semantic, $T_S = 50$, 5 digits | 0.56 | 0.06 | 0.70 |
| icd9_diag_sem_70%-3 | ICD-9-CM, semantic, $T_S = 50$, 3 digits | 0.49 | 0.08 | 0.65 |
| icd9_diag_sem_45%-5 | ICD-9-CM, semantic, $T_S = 79$, 5 digits | 0.56 | 0.07 | 0.70 |
| icd9_diag_sem_45%-3 | ICD-9-CM, semantic match, $T_S = 79$, 3 digits | 0.55 | 0.07 | 0.65 |

Table 2: Prediction accuracy results obtained with different diagnosis information used in the encoding.

For the feature encoding we used the frequency-based encoding [16] enriched with trace attribute features. The classifier is trained with 70% of the traces; 10% of the traces is used to perform the hyper-parameter optimization on the MCC metrics (8); and finally the classifier is tested on the remaining 20% of the traces. Due to the non-deterministic trait of the prediction, each experiment is repeated 30 times, and the average value of MCC together with its standard deviation $\sigma$ are used as reference metrics.

The results are reported in Table 2. The first and the second columns of Table 2 show the diagnosis information used for the prediction and its description. The third and fourth columns contain respectively the mean and the standard deviation $\sigma$ of the MCC value computed in several (30) tests, while the fifth column contains the maximum values of MCC obtained during the (30) tests.

The worst performance is obtained when the textual diagnosis is used as feature (text_diag), while no_diag performs better than text_diag. This is possibly due to the high variability of the textual information, resulting in noise for the predictive model. On average, the best way of taking into account the diagnosis for the predictions seems to be via the mapped ICD-9-CM codes. Indeed, all the predictions obtained with mapped ICD-9-CM codes, except one (icd9_diag_sem_70%-3), provide better results than the no_diag prediction.

In order to further validate this analysis, we also checked the statistical significance of the identified differences:

$$\begin{aligned}
\textsf{no\_diag} &> \textsf{text\_diag} && \text{p-value} < 0.002 \\
\text{all } \textsf{icd9\_diag} \text{ except } \textsf{icd9\_diag\_sem\_70\%-3} &> \textsf{text\_diag} && \text{p-value} < 10^{-5} \\
\text{all } \textsf{icd9\_diag} \text{ except } \textsf{icd9\_diag\_sem\_70\%-3} &> \textsf{no\_diag} && \text{p-value} < 0.05 \\
\textsf{icd9\_diag\_sem\_70\%-3} &> \textsf{text\_diag} && \text{p-value} < 0.006
\end{aligned}$$

The results confirm that all the mappings based on ICD-9-CM codes — except icd9_diag_sem_70%-3 — are significantly higher than no_diag and text_diag, while icd9_diag_sem_70%-3 is only significantly better than text_diag.

We further analysed the results obtained with the mappings based on the ICD-9-CM codes by focusing on:

- the approach, i.e., the lexicographic or the semantic approach adopted;

- the threshold values used to filter bad ICD-9-CM mappings, i.e., $T_L$ and $T_S$: the higher the value of these thresholds, the lower the percentage of bad ICD-9-CM codes mapped;
- the number of digits used of the ICD-9-CM code, corresponding to the level of abstraction of the diagnoses: 5 digits represent detailed diagnosis codes, 3 digits represent groups of diagnoses.

Concerning the approach, the results obtained with the lexicographic approach provide slightly better results than the ones obtained with the semantic approach. However, when comparing the approaches with the same threshold value and number of digits, the difference is overall low and it is not statistically significant — except for icd9_diag_lex_45%-3 (lexicographic, $T_L = 70$, 3 digits) and icd9_diag_lex_70%-3 (lexicographic, $T_L = 50$, 3 digits) that perform better than their semantic-based counterpart (icd9_diag_sem_45%-3 and icd9_diag_sem_70%-3, respectively) with p-value $\leq 0.05$. This result is rather surprising considering the evaluation of the matching methods in terms of bad ICD-9-CM codes reported in Section 5.1.

Our understanding of this result is that a big part in the prediction performance is given by those ICD-9-CM codes which are classified in the *Fair* category. For example the diagnosis *pneumonia and cough* may be fairly mapped to both *486 pneumonia* and *786.2 cough*, but clearly the first one might be more important in the decision about home hospitalization than the second one. At the moment, however, we have not yet developed a method to select the most relevant sub-diagnosis when a diagnosis is composed of several sub-diagnosis.

Concerning the threshold values used to filter bad mappings, the results do not show any clear trends, although it seems that overall higher thresholds return very close or more accurate results than lower thresholds. This difference is however not statistically significant — except for the case of icd9_diag_lex_45%-3 that presents better results than icd9_diag_lex_70%-3 with a statistical significance.

Finally, concerning the level of abstraction of the ICD-9-CM mappings, we can observe that overall the accuracy obtained with more specific ICD-9-CM codes (5 digits) is higher than the accuracy obtained with more general ICD-9-CM codes (3 digits). This is however not true for icd9_diag_lex_45%-3 (3 digits) that has a significantly higher accuracy than icd9_diag_lex_45%-5 (5 digits).

In general, the statistical analysis shows that there are no significant differences between any of the ICD-9-CM results displayed in Table 2, except for two cases: icd9_diag_lex_45%-3 (lexicographic, $T_L = 70$, 3 digits) performs better than all the other ICD-9-CM mappings with p-value $\leq 0.005$ and icd9_diag_sem_70%-3 (semantic, $T_S = 50$, 3 digits) performs worse than all the other mappings with p-value $\leq 0.004$.

## 6 Related Work

The literature related to this work mainly pertains to two research areas: Predictive Process Monitoring (in particular with unstructured data) and the mapping of textual fields to the ICD.

Predictive Process Monitoring approaches can be classified based on the types of prediction they provide: (i) numeric predictions, (ii) outcome–based predictions, and (iii) next activity predictions. In this work we focus on outcome–based predictions, that is related to the fulfilment of a predicate on an ongoing trace, i.e., the outcome of the home hospitalization. Almost all the approaches in this field, rely on implicit models such as machine learning and statistical methods. Maggi et al. [17] report an approach that classifies the fulfilment of a predicate on an ongoing trace by exploiting both control flow and data flow. This work has then been extended in [9, 16, 25, 24]. Di Francescomarino et al. [9] extend the work adding clustering techniques on top of the previous approach. This results in training more classifiers with a smaller subsets of data. Leontjeva et al. [16] treat the execution traces as complex symbolic sequences, while Verenich et al. [25] combine these two approaches. Teinemaa et al. [24] exploit unstructured (textual) information contained in messages exchanged between process instances during execution in order to improve the accuracy of the predictions. Recently in [20] Pegoraro et al. apply natural process language techiniques and LSTM neural networks to integrate information from text documents written in natural language to the prediction model. In this work we borrow the idea of the works in PPM to extract structured information from textual data so as to improve the accuracy of outcome-based predictive models. However, to this aim, we leverage a mapping of textual diagnosis to ICD-9-CM diseases.

The mapping of free text to the ICD classification has been considered in several works. In [2] Akshara et al. provide an automated ICD-9-CM diagnosis prediction integrating structured patients' data together with unstructured clinical text notes. In [13] Gangavarapu et al. present a method for ICD-9-CM code group prediction from unstructured clinical nursing notes, using vector space and topic modeling approaches; in [12] this approach is integrated with a fuzzy similarity cleansing approach to merge anomalous and redundant data. In [19] machine learning and natural language processing approaches are used in the automatic mapping of ICD-10 codes from narrative text fields. In this work the performance of different classical machine learning classifiers are compared in terms of accuracy, precision and recall. In [15] and [11] machine learning techniques are used to map ICD-10 codes from textual death certificates. In [6] recurrent neural networks are used to map ICD-10 codes from Dutch cardiology discharge letters. Differently from all the above state-of-the-art approaches, we focus on Italian textual data and we defined an approach that is able to cope with the available NLP resources.

## 7 Conclusions

With the purpose of improving prediction accuracy by using structured rather than unstructured information in PPM, we have proposed a pipeline that leverages NLP methods and two different approaches — a lexicographic and a semantic one — for mapping textual fields to an existing dictionary, as in the case of textual fields mapped to ICD-9-CM codes. We have applied the proposed approach to a real-life healthcare scenario related to the HHS, and we have evaluated (i) the quality of the mappings; and (ii) the accuracy of the predictions without using the diagnosis information, using the textual diagnosis information, or using the structured information contained in ICD-9-CM codes. The results are overall reasonable and confirm that having structured rather than unstructured features improves the accuracy of the predictions.

We plan, as future work, to further refine the pipeline devised for mapping textual fields to the ICD-9-CM codes, e.g., by taking into account the fact that some textual descriptions are richer than a single ICD-9-CM code and can hence be mapped to more than one code.

## Acknowledgments

This research has been partially carried out within the “Circular Health for Industry” project, funded by “Compagnia San Paolo” under the call “Intelligenza Artificiale, uomo e società”.

## References

1. van der Aalst, W.M.P.: Process Mining - Data Science in Action, Second Edition. Springer (2016)
2. Akshara, P., Shidharth, S., Gokul S., K., Sowmya, K.: Integrating structured and unstructured patient data for icd9 disease code group prediction. In: 8th ACM IKDD CODS and 26th COMAD. p. 436. Association for Computing Machinery (2021)
3. van der Aalst W. M. P. et al.: Process mining manifesto. In: Daniel, F., Barkaoui, K., Dustdar, S. (eds.) BPM Workshops, Clermont-Ferrand, France, August 29, 2011, Revised Selected Papers, Part I. LNBI, vol. 99, pp. 169–194. Springer (2011)
4. Amantea, I.A., Sulis, E., Boella, G., Marinello, R., Bianca, D., Brunetti, E., Bo, M., Fernandez-Llatas, C.: A process mining application for the analysis of hospital-at-home admissions. Stud Health Technol Inform **270**, 522–526 (2020)
5. Aringhieri, R., Boella, G., Brunetti, E., Caro, L.D., Francescomarino, C.D., Dragoni, M., Ferrod, R., Ghidini, C., Marinello, R., Ronzani, M., Sulis, E.: Leveraging structured data in predictive process monitoring: the case of the ICD-9-CM in the scenario of the home hospitalization service. In: Proc. of the Workshop on Towards Smarter Health Care: Can Artificial Intelligence Help? co-located with AIxIA2021. CEUR Workshop Proceedings, vol. 3060, pp. 48–60. CEUR-WS.org (2021)
6. Bagheri, A., Sammani, A., Heijden, P.G., Asselbergs, F., Oberski, D.: Automatic icd-10 classification of diseases from dutch discharge letters. pp. 281–289 (01 2020)
7. Chicco, D., Jurman, G.: The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics **21**(1), 6 (2020)
8. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of, NAACL-HLT 2019. pp. 4171–4186. Association for Computational Linguistics (2019)
9. Di Francescomarino, C., Dumas, M., Maggi, F.M., Teinemaa, I.: Clustering-based predictive process monitoring. IEEE Trans. Serv. Comput. **12**(6), 896–909 (2019)

10. Di Francescomarino, C., Ghidini, C., Maggi, F.M., Milani, F.: Predictive Process Monitoring Methods: Which One Suits Me Best? In: BPM 2018, Proceedings. Lecture Notes in Computer Science, vol. 11080, pp. 462–479. Springer (2018)
11. Duarte, F., Martins, B., Pinto, C., Silva, M.: A deep learning method for icd-10 coding of free-text death certificates. pp. 137–149 (08 2017)
12. Gangavarapu, T., Jayasimha, A., Krishnan, G., Kamath S., S.: Predicting icd-9 code groups with fuzzy similarity based supervised multi-label classification of unstructured clinical nursing notes. Knowledge-Based Systems **190**, 105321 (2020)
13. Gangavarapu, T., Krishnan, G.S., Kamath S, S., Jeganathan, J.: Farsight: Long-term disease prediction using unstructured clinical nursing notes. IEEE Transactions on Emerging Topics in Computing **9**(3), 1151–1169 (2021)
14. Isaia, G., Bertone, P., Isaia, G.C., Ricauda, N.: Home care for patients with chronic obstructive pulmonary disease. Arch Phys Med Rehabil **100**, 664–5 (2010)
15. Koopman, B., Zuccon, G., Nguyen, A., Bergheim, A., Grayson, N.: Automatic icd-10 classification of cancers from free-text death certificates. International journal of medical informatics **84** (08 2015)
16. Leontjeva, A., Conforti, R., Di Francescomarino, C., Dumas, M., Maggi, F.M.: Complex symbolic sequence encodings for predictive monitoring of business processes. In: Business Process Management - 13th International Conference, BPM 2015, Innsbruck, Austria, Proceedings. vol. 9253, pp. 297–313. Springer (2015)
17. Maggi, F.M., Di Francescomarino, C., Dumas, M., Ghidini, C.: Predictive monitoring of business processes. In: Proceedings of CAiSE 2014. LNCS, vol. 8484, pp. 457–472. Springer (2014)
18. Matthews, B.: Comparison of the predicted and observed secondary structure of t4 phage lysozyme. Biochimica et Biophysica Acta (BBA) - Protein Structure **405**(2), 442–451 (1975)
19. Nkolele, R.: Mapping of narrative text fields to ICD-10 codes using natural language processing and machine learning. In: Proc. of the The Fourth Widening Natural Language Processing Workshop. pp. 131–135. Association for Computational Linguistics, Seattle, USA (Jul 2020)
20. Pegoraro, M., Uysal, M.S., Georgi, D., Aalst, W.: Text-aware predictive monitoring of business processes (04 2021)
21. Rizzi, W., Simonetto, L., Di Francescomarino, C., Ghidini, C., Kasekamp, T., Maggi, F.M.: Nirdizati 2.0: New Features and Redesigned Backend. In: Demonstration Track at BPM 2019. CEUR Workshop Proceedings, vol. 2420, pp. 154–158. CEUR-WS.org (2019)
22. Sulis, E., Amantea, I.A., Boella, G., Marinello, R., Bianca, D., Brunetti, E., Bo, M., Bianco, A., Cattel, F., Cena, C., et al.: Monitoring patients with fragilities in the context of de-hospitalization services: An ambient assisted living healthcare framework for e-health applications. In: 23rd ISCT. pp. 216–219. IEEE (2019)
23. Sulis, E., Terna, P., Di Leva, A., Boella, G., Boccuzzi, A.: Agent-oriented decision support system for business processes management with genetic algorithm optimization: an application in healthcare. J. Med. Syst. **44**(9), 1–7 (2020)
24. Teinemaa, I., Dumas, M., Maggi, F.M., Di Francescomarino, C.: Predictive business process monitoring with structured and unstructured data. In: BPM 2016. Proceedings. vol. 9850, pp. 401–417. Springer (2016)
25. Verenich, I., Dumas, M., La Rosa, M., Maggi, F.M., Di Francescomarino, C.: Complex symbolic sequence clustering and multiple classifiers for predictive process monitoring. In: Workshops - BPM 2015. vol. 256, pp. 218–229. Springer (2015)
26. Yuan, Z., Zhao, Z., Sun, H., Li, J., Wang, F., Yu, S.: Coder: Knowledge infused cross-lingual medical term embedding for term normalization (2021)