An analogy based framework for patient-stay identification in healthcare

Safa Alsaidi[1,2,*,†], Miguel Couceiro[3], Esteban Marquer[3], Sophie Quennelle[1,2,5], Anita Burgun[1,2,4,5], Nicolas Garcelon[1,2,4,5] and Adrien Coulet[1,2]

[1] *Inria Paris, F-75012 Paris, France*

[2] *Centre de Recherche des Cordeliers, Inserm, Université Paris Cité, Sorbonne Université, F-75006 Paris, France*

[3] *LORIA, CNRS, Universite de Lorraine, F-54000, France*

[4] *Imagine Institute, F-75015 Paris, France*

[5] *Service d'Informatique Biomédicale, Hôpital Necker-Enfants Malades, Assistance Publique - Hôpitaux de Paris, F-75015 Paris, France*

Abstract

Analogical proportions are statements of the form *"A is to B as C is to D"*. Analogies have been used in various reasoning and classification tasks, addressing different domains. Representation learning has enabled interesting progress in various analogy reasoning applications, where it focuses on the challenge of obtaining a vector representation of complex data. In the biomedical domain, representation learning has been adapted to patient data to solve various tasks such as predicting readmission, diagnosis, and length of stay. In this paper, we focus on the particular task of patient-stay identification, *i.e.*, does a hospital stay belong to a patient or not? This constitutes a building block for addressing key biomedical tasks such as patient matching and privacy preservation. We propose a prototypical architecture that combines patient-stay representation learning and the analogical reasoning framework. For evaluation, we constitute sets of analogies from real-word Electronic Health Records, where objects are patient-stay representations learned from the data. We enrich our analogies using analogical properties and use them to train a neural model to detect whether an analogy is valid. We define three first experimental setups to address our task, present our empirical results, and discuss further perspectives.

Keywords

analogy classification, patient matching, electronic health records, patient representation learning,

1. Introduction

An *analogical proportion*, or simply *analogy*, is a quaternary relation involving four objects A, B, C, and D that draws a parallel between the relation between A and B and the relation between C and D, and that supports analogical reasoning. There are two common tasks associated with

ICCBR Analogies'22: Workshop on Analogies: from Theory to Applications at ICCBR-2022, September, 2022, Nancy, France

[*] Corresponding author.

safa.alsaidi@inria.fr (S. Alsaidi); miguel.couceiro@loria.fr (M. Couceiro); esteban.marquer@loria.fr (E. Marquer); sophie.quennelle@inria.fr (S. Quennelle); anita.burgun@aphp.fr (A. Burgun); nicolas.garcelon@institutimagine.org (N. Garcelon); adrien.coulet@inria.fr (A. Coulet)

0000-0002-4132-1068 (S. Alsaidi); 0000-0003-2316-7623 (M. Couceiro); 0000-0003-2315-7732 (E. Marquer); 0000-0002-4782-6737 (S. Quennelle); 0000-0001-6855-4366 (A. Burgun); 0000-0002-3326-2811 (N. Garcelon); 0000-0002-1466-062X (A. Coulet)

analogies, namely, *analogy detection* and *analogy solving*. Analogy detection aims at deciding whether a quadruple $\langle A, B, C, D\rangle$ constitutes a valid analogy. Analogy solving aims at finding an x that makes $A : B :: C : x$ a valid analogy. Analogy reasoning has been applied to different Natural Language Processing (NLP) tasks such as mining paradigm tables in linguistics and image generation [1, 2].

Representation learning consists of learning low-dimension feature representations (*i.e.*, embeddings) from data. These embeddings, or *vector representations*, of objects (*i.e.*, words, images, characters, etc.) underpin much of modern machine learning and have demonstrated impressive performance on various downstream NLP tasks. For instance, Lim et al. [3] proposed a deep learning model to tackle analogies using semantic embeddings. Their architecture integrates the characteristics of analogies by design and relies heavily on pretrained GloVe embeddings [4]. These embeddings were not trained explicitly to find analogies; yet they were able to detect differences between objects. Hertzmann et al. [5] proposed an analogical framework to learn "image filters" between a pair of images to create an "analogous" filtered result on a third image. The generated image D should relate to C in the same way as B relates to A. Alsaidi et al. [6] developed a neural approach and used character-based embeddings to detect morphological analogies between words.

Analogies have not been sufficiently exploited in healthcare, which thus motivates our work. However, practitioners unconsciously use analogical reasoning (*i.e.*, medical reasoning) in their daily clinical practice to understand the possible causes for a disease diagnosis and prognosis by linking visible signs and symptoms that have been observed among different patients. In addition, several machine learning methods were applied to investigate analogies in healthcare. For instance, Casteleiro et al. [7] utilized analogies to infer disease treatments from statements extracted from text. In their work, they try to extract biomedical facts by analogical reasoning from embeddings. Dynomant et al. [8] used analogical proportions to compare embedding methods trained on a corpus of French health-related documents (*i.e.*, discharge summary, procedure reports, and prescriptions). Analogical proportions were applied on the embeddings of medical documents to verify if $(\vec{A} - \vec{B}) + \vec{C} \approx \vec{D}$, thus allowing to check whether the similarity between A and B is similar to the one between C and D. An example of an analogical proportion they obtain is "(*cardiology* - *heart*) + *lung* ≈ *pneumology*." Rather et al. [9] used analogical proportions to identify hidden or unknown biomedical knowledge from free text resources. They proposed analogical proportions of the form "*acetaminophen* is a type of *drug* as *diabetes* is a type of *disease*."

In this paper, we aim to explore how the analogy framework can help in solving tasks relevant to the healthcare domain. We propose two models that learn patient-stay representations (*i.e.*, learn a vector representation of all the patient EHR data collected during a single stay) to detect analogies in healthcare. To do so, we define two crucial steps that are (1) the learning of embeddings adapted to patient data, and (2) the definition of a neural network dedicated to learn formal properties of analogy. As for the network, we use the same model that was proposed by Lim et al. [3] for word semantics, and later adapted by Alsaidi et al. [6] by incorporating character-based embeddings for morphological analogies. We argue that the framework itself has the potential to be applied in a wide range of domains, and we propose to use it here for healthcare applications, namely, for the patient identification task we introduce below.

Electronic Health Records (EHRs) are real world healthcare data that have been used to train predictive models (including neural network models) for different biomedical tasks, *e.g.*, predicting patient mortality, hospital readmission, length of stay, etc. These EHRs consist of clinical and administrative data collected during patient hospital stays in the form of both *structured* and *unstructured* data. Structured data generally includes diagnostic codes, lab tests, demographics, admission-related information, etc. It can be either static, *e.g.*, patient demographics, or temporal, *e.g.*, vital signs. Unstructured data includes various documents in natural language such as clinical notes, nursing reports, discharge summaries, lab reports, etc. For this work, we consider EHRs from the MIMIC-III (Medical Information Mart for Intensive Care, version 3) database [10] to learn patient representations (*i.e.*, patient embeddings) by converting patient data from the raw EHRs to embeddings that can be further processed. MIMIC-III is a free publicly available hospital database containing de-identified patient health data. This database has been widely used by researchers conducting data mining and machine learning studies applied to healthcare.

Several neural network architectures have been developed to represent biomedical data. For instance, Si et al. [11] adapted a multi-level CNN to learn patient representations from clinical notes through a multi-task learning framework to predict patient mortality and length of stay. Zhang et al. [12] used GRU-based RNN to capture relationships between clinical events and employed attention mechanism to learn a personalized representation to predict patient's future hospitalization using EHR data. Madhumita et al. [13] used a stacked denoising autoencoder and a paragraph vector model to learn generalized patient representations directly from clinical notes to predict patient mortality, primary diagnostic, procedural category, and patient gender. Zhang et al. [14] proposed two neural network architectures that enhance patient representation learning by combining sequential unstructured notes with structured data and evaluated these representations on 3 risk evaluation tasks (*i.e.*, in-hospital mortality, 30-day hospital readmission, and length of stay prediction). In our paper, we learn patient-stay representations and consider the task of patient-stay identification. We think that the tools that address this task will serve as building blocks for more complex and key biomedical tasks, such as patient matching and privacy preservation checking [15, 16].

In this paper, we particularly propose to tackle this task by relying on the detection of analogies in healthcare. In Section 2, we define the setting of analogy that we work on. The models we propose to detect analogies are described in Section 3, along with the procedures we use for data augmentation, training, and evaluation. In Section 4, we provide a description of the MIMIC-III dataset and detail how we build our experimental dataset. We present our experiments and report our results in Section 5. In Section 6, we discuss perspectives for future research.

The main contributions of this paper are the following:

- we propose an analogy based setting using patient-stay representations;
- we propose an embedding model to learn patient-stay representations;
- we display the performance of our classification model to detect analogies on patient-stay data.

2. Defining the task

As we defined previously, an analogy is a 4-ary relation written as $A : B :: C : D$ and expressed as "A is to B as C is to D". In this paper, we work on patient-stay analogies, *i.e.*, on analogies involving hospital stay. In our setting, A, B, C, and D represent patient-stay representations. We define an analogy based setting on patient-stay data that we refer to as *Identity setting*. For that, we consider patient-stay representations, which are vector representations of EHR data that belong to a single hospital stay. Based on the type of EHR data that we decide to include, our patient-stay representations can be made of a representation of either structured or unstructured data, or they can be made of the concatenation of both types of data. More details are provided in Section 5. For this setting, we propose to build analogies of the form:

$$s^{i_1}_{t_1} : s^{i_1}_{t_2} :: s^{i_2}_{t_3} : s^{i_2}_{t_4}$$

where s^i_t refers to the stay t of patient i. Here, pairs of the analogy quadruples are made of two random stays belonging to the same patient. Since there is no constraint on the order of stays, $s^{i_1}_{t_1}$ can happen before $s^{i_1}_{t_2}$ or the inverse. Note that i_1 and i_2 can be the same patient, and that t_1 and t_2, or t_3 and t_4, can represent the same time stamp. Furthermore, t_1 and t_3 or t_2 and t_4 can be the same when $i_1 = i_2$ (but not when $i_1 \neq i_2$). The *Identity* setting finds applications in several tasks relevant to biomedical informatics, including:

- data cleaning,
- data privacy related application,
- patient matching.

Data cleaning applications in the health domain involve repairing or removing patient health data that is inaccurate, incorrectly structured, duplicative, or incomplete. In data cleaning applications, we can associate an erroneously affected sample of data to the patient it belongs. Privacy related applications include verifying if patient data is de-identified, and whether it can be re-identified using different systems. Patient matching is defined as the identification and linking of one patient's data within and across health databases in order to obtain a comprehensive view of that patient's health care record [17]. In patient matching, we try to match patient-related information, either a single patient data (*e.g.*, a document) or full EHR data, that can coexist in one or several databases.

In this paper, we try to match *patient-stay representations* to the patient they belong to. We focus on the task of patient-stay identification, where we aim to determine if a particular hospital stay belongs to a certain patient. We address this task by learning a model to classify such quadruples into valid and invalid analogies. In this sense, we implement the task of analogy detection that aims to determine if a quadruple is a valid analogy. For our *Identity setting*, we define a *valid analogy* as a quadruple of four stays $(s^{i_1}_{t_1}, s^{i_1}_{t_2}, s^{i_2}_{t_3}, s^{i_2}_{t_4})$, where each pair of two stays belong to a single patient i_j; other forms of analogies are considered invalid.

3. Proposed Approach

Our model is made of two components: an embedding model and a classification model. The second takes as input patient-stay representations computed by the first (see Section 3.1).

Our embedding model is trained along with the classification model. We also detail the data augmentation procedure in Section 3.2, and describe the training and evaluation protocols that we followed in Section 3.3.

3.1. Embedding and Classification Models

The models described in this subsection are schematized in Figure 1.

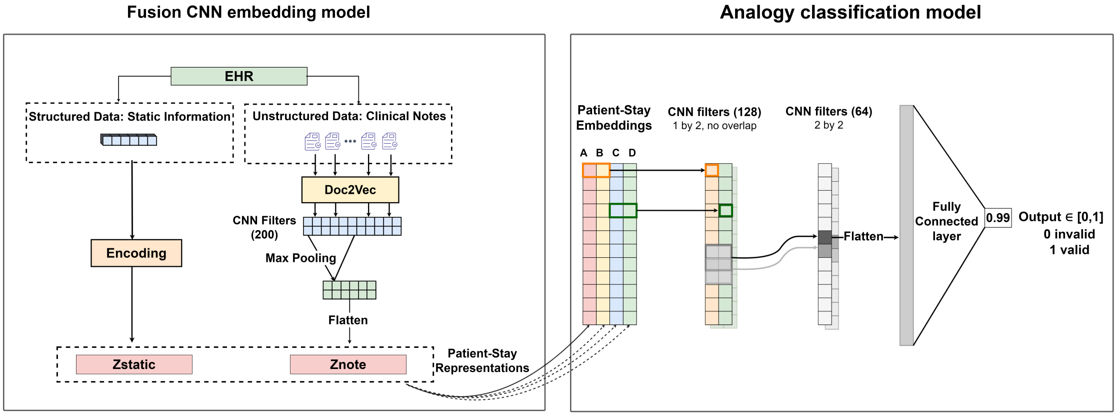

Figure 1: The Fusion CNN embedding model and the CNN classification model.

Embedding Model. As our embedding model, we adapt the Fusion CNN model that was developed by Zhang et al. [14] to obtain patient-stay representations. They proposed a neural network architecture that combines structured and unstructured data to obtain patient representations. The model consists of five parts: static information encoder, temporal signals embeddings, sequential notes representation, patient representation, and output layer that is used to predict three different clinical tasks.

In this work, we restricted structured data to static information, *i.e.*, demographics (Z_{demo}) and admission-related information (Z_{adm}), omitting deliberately vitals in this first attempt. Our model is thus made of static information encoder and sequential notes representation that are used to obtain patient-stay representations as illustrated by the left frame of Figure 1. The static categorical features are encoded as one-hot vectors through the static information encoder. The output of the encoder is $Z_{static} = [Z_{demo}; Z_{adm}]$, where $[Z_{demo}; Z_{adm}]$ is the concatenation of Z_{demo} and Z_{adm}. Z_{static} forms one part of the full patient-stay representation. As shown in Figure 1, the clinical notes representation part is made of a document embedding model, 2 convolutional layers, a max-pooling layer, and a flatten layer. To learn the document embeddings of the clinical notes, we use paragraph vectors (*i.e.*, Doc2Vec) [18]. The document embeddings are passed to the convolutional layers and max-pooling layers. The output of the max-pooling layer is then flattened into Z_{note}, the latent representation of the clinical notes. Based on the type of data that we decide to consider for our experiments, the final patient-stay representation can be made of the representation of only static information (*i.e.*, Z_{static}), the representation of only clinical notes (*i.e.*, Z_{note}), or the concatenation of the representation of clinical notes and static information (*i.e.*, $Z_{patient-stay} = [Z_{static}; Z_{note}]$). The final patient-stay representation is then passed to our classification model to detect analogies.

Classification Model. As in Alsaidi et al. [6], we adapt the neural architecture in Lim et al. [3] to our patient-stay setting. Our classification model determines if an analogy $A : B :: C : D$ is valid by verifying if A and B differ in the same way as C and D. The architecture of the classification model is a Convolutional Neural Network (CNN), which takes as input the embeddings of size n of four elements A, B, C, D. We stack them to get a matrix of size $n \times 4$. The CNN is made of three layers as depicted in the right frame of Figure 1. The first convolutional layer with 128 filters of 1 by 2 is applied on the embeddings, such that it analyses each pair separately without overlaps and measures how A and B, and how C and D differ for each component. The second convolutional layer with 64 filters of 2 by 2 is applied on the resulting matrix, after which the result is flattened into a $64 \times (n - 1)$ unidimensional vector and used as input of a fully connected dense layer that produces a single output. The second layer aims at checking if the difference between A and B is the same as the one between C and D. If A and B are different in the same way as C and D, then $A : B :: C : D$ is a valid analogy. The last layer aggregates this information using a sigmoid activation to get a result (*i.e.*, output of the classification model) between 0 (for invalid analogies) and 1 (for valid analogies). All layers, except the last one, use Regularized Linear Unit (ReLU) as activation function.

3.2. Data Augmentation

Deep neural network approaches require large amounts of data. Therefore we took advantage of properties of analogies to produce additional proportions based on our dataset in a process called *data augmentation*. Previous works [19, 20, 21] have proposed postulates that analogies should obey. For this study, we consider the following:

- $A : B :: A : B$ (reflexivity);
- $A : A :: C : C$ (inner reflexivity);
- $A : B :: C : D \to C : D :: A : B$ (symmetry);
- $A : B :: C : D \to B : A :: D : C$ (inner symmetry);
- $A : B :: C : D \to A : C :: B : D$ (central permutation).

Based on the definition of our analogical setting, we can apply all the above-mentioned postulates to generate valid analogical proportions except for central permutation, which can only be applied in the very particular case when $i_1 = i_2$. When $i_1 \neq i_2$, central permutation cannot be applied to increase our dataset as it would enable to associate stays of distinct patients, which is inconsistent with the aim of the *Identity* setting. Note that from reflexivity and central permutation we can deduce inner reflexivity. As reflexivity forces $i_1 = i_2$, applying it in cases where $i_1 \neq i_2$ would result in a case where $i_1 = i_2$.

For the cases where $i_1 \neq i_2$, given a valid analogy we can generate eight additional valid analogical proportions, namely

- $C : D :: A : B$,

- $D : C :: B : A$,
- $B : A :: D : C$,
- $A : A :: C : C$,
- $B : A :: C : D$,
- $A : B :: D : C$,
- $C : D :: B : A$,
- $D : C :: A : B$;

and two invalid analogical proportions, namely

- $D : A :: B : C$ and
- $A : C :: B : D$.

For cases where $i_1 = i_2$, we apply reflexivity to generate one more valid analogical proportion, namely $A : B :: A : B$. Note that for cases where $i_1 = i_2$, invalid analogical proportions would be considered valid.

3.3. Training and Evaluation

As mentioned, we define a *valid analogy* as a quadruple of four stays $(s_{t_1}^{i_1}, s_{t_2}^{i_1}, s_{t_3}^{i_2}, s_{t_4}^{i_2})$, where each pair of two stays belong to a single patient i_j. For each analogy in the dataset, we start by embedding the four stays. We augment the embeddings using the postulates that we recalled in Section 3.2. As a result, we generate 9 valid analogical proportions (*i.e.*, positive examples) and 2 invalid analogical proportions for cases where $i_1 \neq i_2$. For cases where $i_1 = i_2$, we obtain $10 + 2 = 12$ valid analogical proportions and no invalid analogical proportions. For optimization, we use the Binary Cross-Entropy (BCE) loss. To evaluate the classification model we use the same data augmentation process as for training, and we compute the accuracy and F1 score.

4. Dataset description

For our experiments, we used EHRs from the MIMIC-III [10] as a source of patient medical history data. MIMIC-III is a critical care database developed by the Massachusetts Institute of Technology (MIT)'s Laboratory for Computational Physiology and distributed by PhysioNet [22]. The database is publicly available, where it is accessible to researchers after finishing a HIPAA training course demanded by the National Institutes of Health (NIH). The database contains health-related information associated with all patients admitted to the ICU (Intensive Care Unit) of Beth Israel Deaconess Medical Center between the years 2001 and 2012. It encompasses data of more than 40,000 ICU patients with more than 60,000 ICU stays. All patients' data has been de-identifed in accordance with Health Insurance Portability and Accountability Act (HIPAA). The dataset contains various types of data such as patient demographics, vital signs, lab test results, medications, hospital length of stay, procedures, clinical notes, diagnosis codes (ICD-9), imaging reports, etc.

To build our dataset, we keep only adult patients (*i.e.*, patients aged 18 and above) with at least two admissions. As we do not define any order constraint, we obtain all the permutations of all the stays belonging to a patient. We organize our dataset in way where each pair of stay is associated to the patient it belongs to: $\langle S_1, S_2, PATIENT\_ID \rangle$, where S_1 corresponds to $s_{t_1}^{i_1}$, S_2 corresponds to $s_{t_2}^{i_1}$, and the associated $PATIENT\_ID$ that represents i_1. We obtain a dataset made of 46,986 triples, where for each two pairs of stays we produce an analogy. For our experiments, we use all hospital stays associated with randomly selected 200 patients. We use the data augmentation process to generate positive and negative examples. For training and evaluation, we perform a random split (using a fixed random seed) in a training set of 70% of the extracted analogies, the remaining 30% serving as the test set. We end up with 939,638 analogies for training and 402,703 for testing. To maintain reasonable training and evaluation time, we randomly selected 50,000 analogies from the training set and 50,000 analogies from the testing set.

5. Experiment Setup

We now present the three experiments that we conducted in the *Identity* setting. In Section 5.1, we describe the patient-stay features that we consider and the data preprocessing that we performed for structured and unstructured data. We describe the implementation details in Section 5.2. The results of our experiments are reported in Section 5.3 and discussed further in this section. The code used for our experiments is written in Python 3.9 and PyTorch and is available in the repository https://github.com/Safa-98/patient-stay-analogy.

5.1. Stay Features and Data Preprocessing

We consider both structured (*i.e.*, demographics and admission-related information) and unstructured data (*i.e.*, clinical notes) to define our analogies. In this subsection, we describe the patient-stay features that are utilized by our model and some data preprocessing details.

Static information. In our experiments, our static information consists of demographic information and admission-related information. For demographic information, we extract patient's age, gender, marital status, ethnicity, and insurance information. We keep only adult patients (*i.e.*, patients aged 18 and above). We split the age into 5 groups $[18, 25[$, $[25, 45[$, $[45, 65[$, $[65, 89[$ and $[89, +\infty[$. For admission-related information, we include admission type as features.

Clinical notes. Nursing, Nursing/Other, Physician, and Radiology notes make up the majority of clinical notes in MIMIC-III database. For each hospital stay, we only kept notes that belong to these 4 categories. We excluded notes that have an error tag and notes that lack a hospital admission id.

5.2. Implementation Details

To build our corresponding cohorts, we performed the preprocessing described in the previous section to obtain our patient-stay features. Patients without any records of clinical notes or with notes that do not belong to the 4 categories defined above were removed. We computed the median of notes per hospital admission to determine the number of clinical notes to extract

Table 1
Accuracy and F1 score (both in %) of 3 runs of the classification model. Embeddings used are concatenation of static information and clinical notes.

Epochs	Valid	Invalid	F1
40 epochs	98.41 ± 1.56	68.22 ± 1.94	95.79 ± 0.59
20 epochs	94.89 ± 1.74	72.08 ± 1.68	94.30 ± 0.80
10 epochs	96.85 ± 1.75	70.31 ± 1.94	95.20 ± 0.71

per hospital admission. Therefore, we kept the first 12 notes, and used padding (*i.e.*, completion with zeros) for hospital admissions with less than 12 notes.

For the unsupervised Doc2Vec model [18], we finetune it on the training set to obtain the document-level embeddings using the Gensim toolkit [23]. For the training algorithm, we use PV-DBOW (Paragraph vector-Distributed Bag of Words). We set the number of training epochs as 30, the initial learning rate as 0.025, the learning rate decay as 0.0002, and the dimension of vectors as 200 to train. The Fusion CNN model is trained with Adam optimizer with a learning rate of 0.0001 and ReLU as the activation function. The chosen batch size is 64.

In this paper we perform three experiments. In the first, we consider both structured and unstructured data. Therefore, we obtain our patient-stay representation by concatenating the representations of clinical notes along with static information. In this experiment, we verify if a particular hospital stay belongs to a patient by looking at both the structured and unstructured data associated with each stay. In the second, we only consider unstructured data, which means that our patient-stay representations are based solely on the representations of clinical notes. Therefore, by looking at clinical notes associated with a single hospital stay, we check if a particular hospital stay belongs to a patient. In the third, we only consider structured data, which means that our patient-stay representations are based solely on the representations of static information (*i.e.*, demographics and admission-related information). Therefore, we verify if a particular hospital stay belongs to a patient by looking at the static information that is associated with a hospital stay.

5.3. Results and Discussion

As mentioned previously, we conducted three experiments that mainly differ in what type of data was used to obtain our patient-stay representations. For all the experiments, we used 50,000 analogies for training and evaluation, and applied the same procedure for data augmentation. We report the accuracy and F1 score for each experiment. The F1 score gives a better measure of the incorrectly classified cases than the accuracy metric.

For the first experiment, we fed our embedding model with both structured (*i.e.*, demographics and admission-related information) and unstructured data (*i.e.*, clinical notes). Our patient-stay representations are thus made of the concatenation of static information and clinical notes. We chose the epochs where the training loss is at the local minimum. We trained our model for 10, 20, and 40 epochs, with 3 different random initializations in each case. Our results are detailed in Table 1. Our model performs the best for positive examples. For 40 epochs, the model gives the best result for valid analogies and performs best for invalid analogies for 20 epochs.

Table 2
Accuracy and F1 score (both in %) of 3 runs of the classification model. Embeddings used are based on only clinical notes.

Epochs	Valid	Invalid	F1
40 epochs	88.52 ± 6.91	77.20 ± 6.92	95.64 ± 1.27
20 epochs	97.71 ± 1.69	67.89 ± 2.43	94.74 ± 0.93
15 epochs	92.05 ± 6.42	74.45 ± 7.39	94.90 ± 1.25

Table 3
Accuracy and F1 score (both in %) of 3 runs of the classification model. Embeddings used are based on only static information.

Epochs	Valid	Invalid	F1
40 epochs	99.98 ± 0.001	66.14 ± 0.04	96.37 ± 0.01
20 epochs	99.98 ± 0.004	65.96 ± 0.31	96.35 ± 0.03
15 epochs	99.98 ± 0.002	66.12 ± 0.13	96.33 ± 0.05

For our second experiment, we used only unstructured data, *i.e.*, the Z_{note} part of the embedding for the patient-stay representations. Our patient-stay representations thus consisted of only the representation of clinical notes. The training loss was at the local minimum for 15, 20, and 40 epochs. Therefore, we trained our model for 15, 20, and 40 epochs, with 3 different random initializations in each case. As shown in Table 2, our model performs the best for positive examples when we train by 20 epochs.

For our third experiment, we used only structured data, *i.e.*, Z_{static}, to represent our patient-stay representations. Our training loss was at the local minimum for 15, 20, and 40 epochs. Therefore, we trained our model for 15, 20, and 40 epochs, with 3 different random initializations in each case. We report our results in Table 3. As seen, the accuracy for positive examples is high for all cases compared to negative examples where the accuracy drops.

In all our experiments, we can see that our model performs the best for positive examples regardless of whether we use $[Z_{static}; Z_{note}]$, only Z_{note}, or only Z_{static} for the patient-stay representations. This can be explained as a result of the imbalance between positive and negative examples in the training data. Balancing the data would be the next step as it proved to be a good solution for [6] to get similar results for positive and negative examples. The accuracy for valid analogies is the highest when our embedding model is fed with only static information. Between the first and the second experiment, the accuracy is the highest for valid analogies when the patient-stay representations are made of the concatenation in contrast to when our patient-stay representations are made of only clinical notes. This indicates that adding or using static information when learning patient-stay representations, as in the first and third experiment, improves the performance of our model, where it allows the model to better distinguish the stays and to match them to the patient they belong to. We also notice that the accuracy for invalid analogies is the highest when the embedding model is fed with only clinical notes. For all performed experiments, the F1 score is high, which indicates that our model is able to correctly classify analogies to the class they belong to (*i.e.*, valid or invalid).

To gain more insight into how our models perform, we conducted an error analysis where we noticed that most misclassifications were spotted in two cases.

1. **Cases where $i_1 = i_2$.**
   To recall, we do not generate invalid analogies for cases where $i_1 = i_2$; therefore, invalid analogy forms ($D : A :: B : C$ and $A : C :: B : D$) should be considered valid in these cases. In our error analysis, we noticed that when the four stays belong to the same patient, our model classifies the above-mentioned invalid analogy forms as invalid instead of valid. We believe that our model was not trained enough to distinguish these forms of analogies as there were less analogies with four stays belonging to the same patient generated in our dataset.
2. **Cases where representations are made of only clinical notes.**
   To recall, in our second experiment we only used the representations of clinical notes to obtain patient-stay representations. We noticed that when the category of the clinical notes is similar between two hospital stays or when two hospital stays have less than five clinical notes, our model struggles to distinguish between the two hospital stays. This indicates that in some cases using only clinical notes to learn patient-stay representations might not be sufficient as these notes might not contain enough information to help our model differentiate between two similar stays that belong to two distinct patients. As a result, the model would incorrectly match these two similar stays to the same patient.

In these experiments, we did not include temporal data, where we only used demographics and admission-related information as structured data. It would be interesting to also include temporal signals (*i.e.*, vital signs) along with demographics and admission-related information as structured data. Our patient-stay representations would be then made of the concatenation of the representations of static information and temporal signals as structured data and the representation of clinical notes as unstructured data.

6. Conclusion and Perspectives

We adapted the approach in [3, 6] from semantic and morphological analogies to patient-stay analogies. Our prototypical architecture has some limits, but seems promising for the task of patient identification. Our classification model is flexible in terms of the analogies that it classifies. Changing the way the data is augmented will change the way the model behaves. Our model can be adapted to different healthcare applications through dedicated embedding models [24]. Inspired by [14], we implemented a model to build patient-stay representations. As mentioned in Section 5.3, there are multiple plausible improvements to our approach, in terms of balancing valid and invalid analogies as well as including other types of data to build our patient-stay representations. As we limited ourselves to analogy detection, a future work would be to address analogy solving in the same setting that would allow the generation of synthetic patient-stays.

Acknowledgments

Experiments presented in this paper were carried out using computational clusters equipped with GPU from the Grid'5000 testbed (see https://www.grid5000.fr).

The research work of the second and third named authors is partially supported by TAILOR, a project funded by EU Horizon 2020 research and innovation program under GA No 952215, and the Inria Project Lab "Hybrid Approaches for Interpretable AI" (HyAIAI).

References

[1] R. Fam, Y. Lepage, Morphological predictability of unseen words using computational analogy, in: Workshops Proceedings for the Twenty-fourth International Conference on Case-Based Reasoning (ICCBR), volume 1815, 2016, pp. 51–60.

[2] S. E. Reed, Y. Zhang, Y. Zhang, H. Lee, Deep visual analogy-making, in: Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), 2015, pp. 1252–1260.

[3] S. Lim, H. Prade, G. Richard, Solving word analogies: A machine learning perspective, in: Proceedings of the Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU), volume 11726, 2019, pp. 238–250.

[4] J. Pennington, R. Socher, C. D. Manning, Glove: Global vectors for word representation, in: Proceedings of the Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1532–1543.

[5] A. Hertzmann, C. E. Jacobs, N. Oliver, B. Curless, D. Salesin, Image analogies, in: Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH), 2001, pp. 327–340.

[6] S. Alsaidi, A. Decker, P. Lay, E. Marquer, P.-A. Murena, M. Couceiro, A neural approach for detecting morphological analogies, in: Proceedings of the 8th IEEE International Conference on Data Science and Advanced Analytics (DSAA), 2021, pp. 1–10.

[7] M. A. Casteleiro, J. D. Diz, N. Maroto, M. J. F. Prieto, S. Peters, C. Wroe, C. S. Torrado, D. M. Fernandez, R. Stevens, Semantic deep learning: Prior knowledge and a type of four-term embedding analogy to acquire treatments for well-known diseases, JMIR Medical Informatics 8 (2020) 1–28.

[8] E. Dynomant, R. Lelong, B. Dahamna, C. Massonnaud, G. Kerdelhué, J. Grosjean, S. Canu, S. J. Darmoni, et al., Word embedding for the french natural language in health care: comparative study, JMIR medical informatics 7 (2019) 118–122.

[9] N. N. Rather, C. Patel, S. A. Khan, Using deep learning towards biomedical knowledge discovery, International Journal of Mathematical Sciences and Computing, (IJMSC) 3 (2017) 1–10.

[10] A. E. W. Johnson, T. J. Pollard, L. Shen, L. wei H. Lehman, M. Feng, M. M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, R. G. Mark, Mimic-iii, a freely accessible critical care database, Scientific Data 3 (2016).

[11] Y. Si, J. Du, Z. Li, X. Jiang, T. Miller, F. Wang, W. Jim Zheng, K. Roberts, Deep representation learning of patient data from electronic health records (ehr): A systematic review, Journal of Biomedical Informatics 115 (2021) 1–42.

[12] J. Zhang, K. Kowsari, J. H. Harrison, J. M. Lobo, L. E. Barnes, Patient2vec: A personalized interpretable deep representation of the longitudinal electronic health record, IEEE Access 6 (2018) 65333–65346.

[13] S. Madhumita, S. Simon, L. Kim, D. Walter, Patient representation learning and interpretable evaluation using clinical notes, Journal of biomedical informatics 84 (2018) 103–113.

[14] D. Zhang, C. Yin, J. Zeng, X. Yuan, P. Zhang, Combining structured and unstructured data for predictive models: a deep learning approach, BMC Medical Informatics and Decision Making 20 (2020) 280.

[15] P. Waruhari, A. Babic, L. Nderu, M. C. Were, A review of current patient matching techniques, in: Informatics Empowers Healthcare Transformation (ICIMTH), volume 238, 2017, pp. 205–208.

[16] F. N. Wirth, T. Meurers, M. Johns, F. Prasser, Privacy-preserving data sharing infrastructures for medical research: systematization and comparison, BMC Medical Informatics Decision Making 21 (2021) 242.

[17] B. H. Just, D. T. Marc, M. Munns, R. H. Sandefer, Why patient matching is a challenge: Research on master patient index (mpi) data discrepancies in key identifying fields., Perspectives in health information management 13 (2016) 1e.

[18] Q. V. Le, T. Mikolov, Distributed representations of sentences and documents, in: Proceedings of the 31th International Conference on Machine Learning (ICML), volume 32, 2014, pp. 1188–1196.

[19] L. Miclet, S. Bayoudh, A. Delhay, Analogical dissimilarity: Definition, algorithms and two experiments in machine learning, Journal of Artificial Intelligence Research 32 (2008) 793–824.

[20] Y. Lepage, De l'analogie rendant compte de la commutation en linguistique, Habilitation à diriger des recherches, Universit'e Joseph-Fourier - Grenoble I, 2003.

[21] C. Antic, Analogical proportions, ArXiv abs/2006.02854 (2020).

[22] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, H. E. Stanley, Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals., Circulation 101 23 (2000) E215–20.

[23] R. Řehůřek, P. Sojka, Software Framework for Topic Modelling with Large Corpora, in: Proceedings of the LREC Workshop on New Challenges for NLP Frameworks, 2010, pp. 45–50.

[24] S. Alsaidi, M. Couceiro, A. Burgun, N. Garcelon, A. Coulet, Exploring analogical inference in healthcare, in: Workshop on Interactions between Analogical Reasoning and Machine Learning (IARML), 2022 (to appear).