Exploring Analogical Inference in Healthcare

Safa Alsaidi, Miguel Couceiro, Sophie Quennelle, Anita Burgun, Nicolas Garcelon, Adrien Coulet

Exploring Analogical Inference in Healthcare

Safa Alsaidi[1,2,*,†], Miguel Couceiro[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 $A : B :: C : D$ that are used to map similar relationships between two pairs of objects, A, B, and C, D. Analogies have long been a subject of research in the Natural Language Processing (NLP) community, where they have been applied to a variety of reasoning and classification tasks. Lately, machine and representation learning have shown to be useful for analogical reasoning. In this paper, we discuss the possibility of adapting the analogical framework to healthcare applications, in particular to medical decision support. We particularly hypothesize that as language representations help in analogical reasoning in NLP, patient representation learned from Electronic Health Records (EHRs) may help in healthcare. We define three different analogy based settings adapted to EHR data that we see as first steps to the development of analogical applications to this domain. We provide statistics on the first sets of analogies that we built from a publicly available dataset of EHRs, and report preliminary, but promising results to detect patient-stay analogies following our very first experimental setting.

Keywords

analogy classification, electronic health records, patient representation learning

1. Introduction and motivation

An analogical proportion, or an analogy, is a relation between four objects A, B, C, and D that is expressed as "A is to B as C is to D" and formally denoted as $A : B :: C : D$. There are two main tasks associated with analogical proportions: *analogy detection* and *analogy solving*. *Analogy detection* corresponds to the task of deciding whether a quadruple $\langle A, B, C, D\rangle$ is a valid analogy. *Analogy solving* corresponds to finding a fourth element x so that $A : B :: C : x$ is a valid analogy. This can be done either by retrieving x from a pool of candidates or by

IARML@IJCAI-ECAI'2022: Workshop on the Interactions between Analogical Reasoning and Machine Learning, at IJCAI-ECAI'2022, July, 2022, Vienna, Austria
*Corresponding author.
safa.alsaidi@inria.fr (S. Alsaidi); miguel.couceiro@loria.fr (M. Couceiro); 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); 0000-0002-4782-6737 (S. Quennelle); 0000-0001-6855-4366 (A. Burgun); 0000-0002-3326-2811 (N. Garcelon); 0000-0002-1466-062X (A. Coulet)

generating x. Analogies have been extensively studied and applied to various Natural Language Processing (NLP) tasks [1, 2, 3, 4]. Object representations called *embeddings* are low-dimensional representations of high-dimensional vectors, which have been used to improve deep learning methodologies. Some of these embeddings learn precise representations and are able to detect differences between objects. As a result they can discriminate between valid and invalid analogical proportions and solve analogical equations.

In this paper, we explore the possibility to leverage the analogy framework to solve tasks relevant to the healthcare domain. We particularly consider using Electronic Health Records (EHRs) to learn patient representations, *i.e.*, patient embeddings. We initiate the construction of sets of patient-based analogies using relationships existing between patient hospital stays from a publicly available set of EHRs. These health records consist of clinical and administrative data collected during patient hospital stays. Generally they are composed of structured (*e.g.*, diagnostic codes, lab tests) and unstructured data (*e.g.*, clinical notes, nursing reports, discharge summaries), either static (*e.g.*, patient demographics) or temporal (*e.g.*, vital signs).

EHRs have been secondary used to conduct epidemiological and observational studies. They have also been used as real word data to train predictive models [5]. In particular, deep learning methods have become increasingly popular in medical informatics for general tasks such as predicting mortality, in-hospital readmission, diagnoses, etc. A key element for such tasks is to effectively convert patient data from the raw EHR format to embeddings that can be further processed [6]. Representation learning thus consists of learning low-dimension feature representations from raw data. As EHR data are heterogeneous and complex, studies have shown that deep learning models are suited to encode complex EHR data to learn patient representations and that various architectures are suited to different biomedical tasks [7, 8, 9, 10, 11, 12]. For instance, Madhumita et al. [13] used a stacked denoised autoencoder and a paragraph vector model to learn generalized patient representations directly from clinical notes. Si and Roberts [14] utilized a three-level hierarchical attention-based recurrent neural network (HAN) with greedy segmentation to learn patient representation from clinical notes. Zhang et al. [15] proposed 2 multi-modal neural network architectures to enhance patient representation learning by combining sequential unstructured notes with structured data.

Analogies have only been sporadically applied to healthcare. Nonetheless, analogical reasoning has been applied in clinical practice by physicians for diagnosis and prognosis, as a way of linking visible signs and symptoms to possible causes. Indeed, medical reasoning relies on observations of previous patients with similar signs and symptoms, who happened to have a certain disease. Several studies have investigated analogies in healthcare by applying various machine learning methods. For instance, Rather et al. [16] used analogical proportions to identify hidden or unknown biomedical knowledge from free text resources. In their work, they defined analogies of the form "*acetaminophen* is a type of *drug* as *diabetes'* is a type of *disease*." Dynomant et al. [17] used analogical proportions to compare embedding methods trained on a corpus of French health-related documents. Each analogical proportion aimed to verify if $(\vec{Term}1 - \vec{Term}2) + \vec{Term}3 \approx \vec{Term}4$, allowing to check if the similarity between the first two terms is similar to the one between Term 3 and Term 4.

In this paper, we describe an ongoing work on analogical inference in healthcare. We introduce three analogy based settings, where each setting aims to investigate specific biomedical tasks, namely identity, predictive, and generative tasks. In comparison with previous studies

[14, 7, 8, 9, 10, 11, 12], we aim to build analogies based on patient-stay representations. One of the main contributions of our work is a framework to build sets of proportions for analogical inference in healthcare.

This paper is organized as follows. Section 2 provides a description of the MIMIC-III dataset. Section 3 defines our analogical settings and associated biomedical tasks, and justifies our task choices. Section 4 presents preliminary statistics of the analogical proportions built from MIMIC-III. Section 5 initiates a discussion addressing some analogical postulates that could be useful when generating our analogies. Section 6 illustrates the feasibility of our approach by providing preliminary results using one of our experimental settings. Section 7 discusses perspectives for future research.

2. Data description

We propose to use EHRs as a source of patient medical history data and aim to consider both its structured and unstructured data to define our analogies. In particular we experiment with a publicly available dataset of EHRs called MIMIC-III (Medical Information Mart for Intensive Care-III) [18]. MIMIC-III is a critical care database, developed by the Massachusetts Institute of Technology (MIT)'s Laboratory for Computational Physiology and distributed by PhysioNet [19]. It contains integrated, de-identifed health-related data in accordance with Health Insurance Portability and Accountability Act (HIPAA). It contains data associated with all patients admitted to the ICU (Intensive Care Unit) of Beth Israel Deaconess Medical Center between 2001 and 2012. It contains various data, such as patient demographics, vital signs, lab test results, medications, hospital length of stay, survival, clinical notes, imaging reports and more, structured into 26 tables. Each patient-stay is associated with diagnosis codes, motivating the stay and procedures performed during the stay. It encompasses data of more than 40,000 ICU patients and more than 60,000 ICU stays. Table 1 shows statistics for the subgroups of adult patients (aged 18 and above) with at least two stays, which is the subset that we consider in the rest of the article.

The database contains a combination of structured and unstructured data and is accessible to researchers under a data use agreement, where users are required to follow a HIPAA training course demanded by the National Institutes of Health (NIH).

	Statistics
Patients (total)	8,526
Gender, male (total)	4,818
Age (median, in years)	66.24
ICU stays (total)	23,345
Hospital stays (total)	19,709
ICU length of stay (median, in days)	2.33
Hospital length of stay (median, in days)	9.74
Clinical notes per stay (median)	18.0

Table 1
General statistics of the MIMIC-III EHR dataset, restricted to patients aged of 18 and above, with at least 2 stays.

3. Experimental settings and biomedical tasks

As we defined previously, an analogy is a 4-ary relation and is usually written as $A : B :: C : D$. In this paper, we define three analogy based settings and associated tasks that we are interested in investigating with EHR data. We name our three settings as follows: *(i)* Identity; *(ii)* Identity + Sequent; *(iii)* Identity + Directly Sequent. For these settings, we do not want to learn "full" patient representation, but *patient-stay representations* (*i.e.*, learn a numeric vector representation of EHR data that belong to a single hospital stay) which we hope to be simpler.

Identity In the first setting, we propose to build analogies of the form:

$$s_{t_1}^{i_1} : s_{t_2}^{i_1} :: s_{t_3}^{i_2} : s_{t_4}^{i_2}$$

where s_t^i 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, $s_{t_1}^{i_1}$ can happen before $s_{t_2}^{i_1}$ or the inverse. Note that i_1 and i_2 could be the same patient, and that t_1 and t_2, or t_3 and t_4, could represent the same time stamp. Furthermore, t_1 and t_3 or t_2 and t_4 could be the same when $i_1 = i_2$ (but not when $i_1 \neq i_2$). In this setting we aim at investigating identity tasks, *i.e.*, associating an unaffected sample of data to the patient it belongs. Note that this setting fits several data cleaning and data privacy related applications

Identity + Sequent For this setting, we add a temporal constraint to analogies, as we force

$$s_{t_1}^{i_1} \ll s_{t_2}^{i_1} \text{ and } s_{t_3}^{i_2} \ll s_{t_4}^{i_2}$$

where $\ll$ denotes temporal sequentiality between stays of a same patient, *i.e.*, $s_{t_2}^{i_1}$ takes place after $s_{t_1}^{i_1}$ and $s_{t_4}^{i_2}$ takes place after $s_{t_3}^{i_2}$ but not necessarily directly right after. We consider cases where $i_1 = i_2$. In this setting, we also define a relation named **diagnosis**, which forces $s_{t_1}^{i_1}$ and $s_{t_3}^{i_2}$ to have the same diagnosis. This relation provides more meaning to our analogies and gives us more medical insight into the relationship between our patients. For example, based on the different stays associated with a single patient we hope to see how a certain disease develops (similarly or differently) between two distinct patients.

Identity + Directly Sequent In this third setting, we make the temporal constraint more strict as we force the two stays of the same patient to be directly sequent (no other stay can exist in between). We note this constraint

$$s_{t_1}^{i_1} \prec s_{t_2}^{i_1} \text{ and } s_{t_3}^{i_2} \prec s_{t_4}^{i_2}$$

The **diagnosis** relation is kept between $s_{t_1}^{i_1}$ and $s_{t_3}^{i_2}$, and cases where $i_1 = i_2$ are also considered. With these three settings, we aim at investigating the applicability of two tasks: *analogy detection* and *analogy inference*. For instance, given an analogy of the form $A : B :: C : x$, we can either propose potential values for an unknown stay x (*i.e.*, predictive task) or generate stays which would enrich our dataset with synthetic stays (*i.e.*, generative task).

For the last two settings, we define additional settings by considering three levels of relaxation of the diagnosis constraint. It is satisfied either if both stays are associated with the very same primary diagnostic code (level 4) or in more relax settings, *i.e.*, if both codes belong to the same level-3 or level-2 branch of the hierarchy of the ICD-9-CM (International Classification of Diseases, Ninth Revision, Clinical Modification) [20]. Statistical details on the influence of the constraints is discussed in the next section.

4. Preliminary statistics on MIMIC-III

We computed some preliminary statistics on MIMIC-III dataset to check how many analogical proportions can be formed for each of the three analogical settings and based on the three level diagnosis constraint as shown in Table 3. To form the analogies, we built tuples of each of the two stays that belong to a single patient i. We kept only adult patients (aged 18 and above) that have at least two hospital stays. For our first setting, 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. Since we do not restrict the order of the stays for each of the pairs, our analogies were made of all the permutations of all the stays belonging to a patient.

For our second and third settings, we define a *valid analogy* to be a quadruple made 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 and $s_{t_1}^{i_1}$ and $s_{t_3}^{i_2}$ have the same diagnostic code. As an order constraint is introduced for these two settings, we had to make sure that $s_{t_1}^{i_1}$ takes place before $s_{t_2}^{i_1}$ and $s_{t_3}^{i_2}$ takes place before $s_{t_4}^{i_2}$. For the second setting, the stays do not necessarily happen directly right after, where other stays can exist in between. As for the third setting, one stay immediately follows the other, *i.e.*, there is no other stay in between them.

For the diagnosis constraint, we referred to the ICD-9-CM, which is the standard nomenclature for assigning diagnosis codes to each hospital stay. Indeed, each stay has a unique primary diagnosis code and a set of secondary codes. Diagnostic codes are organized hierarchically as follows: *(1) chapter*, *(2) block*, *(3) 3-digit category*, and *(4) full code*. As an example, the diagnosis code 767.4 and its hierarchy are presented Table 2.

Level	Level name	Example of code range
1	Chapter	760–779
2	Block	764–779
3	3-digit category code	767
4	Full code	767.4

Table 2
Hierarchy of the ICD9-CM (International Classification of Disease, Ninth Revision, Clinical Modifications), with examples for the code 767.4 ("Injury to spine and spinal cord due to birth trauma").

The MIMIC-III dataset associates each stay with an ICD-9 code (*i.e. full code*). For the second and third analogical settings, we preprocessed our diagnosis codes in the following manner. We kept only the primary diagnostic code associated with each patient-stay. We filtered ICD-9 codes that appeared only once. For the second level and third level diagnosis settings, we

performed the same preprocessing except that we had to first convert the ICD-9 codes into their corresponding *category* and *block* formats.

Table 3 provides the number of *valid* analogies that can be formed with the defined settings. As shown, the number of analogies is the highest for the *Identity* setting, which could be explained as a result of the absence of constraint on both the order of stays and diagnosis. The more strict the order constraint is, the less the amount of *valid* analogies that could be formed. The *diagnosis* constraint also influences the number of analogies that could be generated. The number of analogies is the lowest for *full code* (level 4) settings, where less patients share the very same primary diagnosis code. In comparison, more patients can share a single diagnosis code in the *category* (level 3) and *block* (level 2) settings, where we observe the highest number of analogies in the *block* setting for both the second and third analogical settings.

Setting	ICD Level	Analogies
Identity	N/A	$1,100,954,350$
Identity + Sequent	4	$648,169$
	3	$1,876,445$
	2	$3,243,699$
Identity+Directly Sequent	4	$545,892$
	3	$1,326,518$
	2	$2,780,507$

Table 3
Number of valid analogies that can be generated from MIMIC-III, depending on the different settings and on the level of flexibility allowed on patient ICD diagnosis.

5. Properties of analogies for data augmentation

As we are interested in exploring different deep learning models, we would need large amounts of data to train them. To enlarge the training datasets, one may use analogy properties to generate more analogies in a process called *data augmentation*. Training our model on different equivalent forms of the same analogy could help reduce overfitting. Previous works [21, 2, 22] have defined postulates that proportional analogy should obey; some of which include the following:

- reflexivity: $A : B :: A : B$
- inner reflexivity: $A : A :: C : C$
- determinism: $A : A :: A : D \rightarrow D = A$
- symmetry: $A : B :: C : D \rightarrow C : D :: A : B$
- inner symmetry: $A : B :: C : D \rightarrow B : A :: D : C$
- central permutation: $A : B :: C : D \rightarrow A : C :: B : D$.

However, not all these postulates hold for all our settings. Based on the current definitions of our analogical settings, we can apply *reflexivity* for all the three settings. *Inner reflexivity* can

only be applied for the *Identity* setting. Adding this postulate for the second and third settings would require to loose our order constraint, which is inconsistent with the temporal aspect of predictive modeling. *Determinism* holds for all the settings. We include this postulate even if it produces trivial analogies. *Central permutation* can be applied on our analogies for the first setting only and 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. Concerning the second and third analogy settings, *central permutation* cannot be applied as it violates the order constraint in most cases. Note that *central permutation* can be applied for these two settings for cases when $i_1 = i_2$, $t_2 \leq t_1$, $t_3 \leq t_4$, and the same diagnosis is associated to $s_{t_1}^{i_1}$ and $s_{t_3}^{i_2}$. *Inner symmetry* can be applied for the *Identity* setting, but it violates the order constraints for the other two settings. For all the three analogical settings, by applying *symmetry* to one valid analogy, we can increase the number of *valid* analogies as it does not violate any of the three constraints. In addition to valid forms, we can also consider *invalid* forms (*i.e.*, that contradict some of the setting constraints or that cannot be inferred from the base cases using the allowed postulates) for classification purposes.

6. Preliminary experiments: error analyses in the Identity setting

We set up a preliminary experiment on the analogy detection task, addressing our *Identity* setting. Inspired by [3, 23], we consider a CNN classifier adapted to patient-stay, to determine whether a given (A, B, C, D) constitutes a valid analogy. For the embedding model, we consider the Fusion CNN model developed by [15], which combines both structured and unstructured data to obtain patient-stay representations. For this very first experiment, we only consider structured data limited to demographics and admission-related information. For unstructured data, we group clinical notes associated with a hospital stay. We learn clinical note embeddings and concatenate them with static information following [15] to obtain our final patient-stay representations.

We considered hospital stays of 200 patients extracted randomly from MIMIC-III. 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. We do not define any order constraint for our stays; therefore, $s_{t_1}^{i_1}$ can happen before $s_{t_2}^{i_1}$. Quadruples where $i_1 = i_2$ are also included in the dataset. To generate other valid analogies, we make use of all postulates in Section 5, except for *central permutation* that is only applied in the case when $i_1 = i_2$. As *reflexivity* forces $i_1 = i_2$, it cannot be applied in the cases where $i_1 \neq i_2$. Accordingly, given a valid analogy $A : B :: C : D$, we generate 8 additional valid analogies, 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 2 invalid, namely, $D : A :: B : C$ and $A : C :: B : D$. When $i_1 = i_2$, we generate one more valid analogy of the form $A : B :: A : B$ and we consider invalid analogies as valid.

For training and evaluation, we split our dataset into 70% training set and 30% testing set, representing 939,638 analogies for training and 402,703 for testing. We randomly draw 50,000 analogies of each set (*i.e.*, training and testing) when loading the data. Following the data augmentation procedure introduced before, given a valid analogy, we generate 9 valid analogies

(*i.e.*, positive examples) and 2 invalid analogies (*i.e.*, negative examples) for cases when $i_1 \neq i_2$. In contrast, we generate 12 valid analogies and no invalid analogies for cases when $i_1 = i_2$. Based on this setting, we tend to generate more valid analogies than invalid ones. We trained our model on 10 epochs, with 3 random intializations to observe how the model behaves and how much it is able to learn. We only computed the accuracy and obtained 96.85 ± 1.75 for valid analogies and 70.31 ± 1.94 for invalid analogies. Our model performs the best for positive examples which can be explained as a result of the imbalance between valid and invalid examples in the training data. Nonetheless, these preliminary results seem to show that the model learns, to some extent, patient-stay identity relationships.

To gain a deeper insight on how our classification model works, we present four examples: one true positive, one false negative, one true negative, and one false positive. Patient ids in the examples below have been changed and dates have been shifted. We provide elements of interpretation to explain why our model correctly classifies some analogies and why in other cases it does not.

Analysis of a true positive example. We consider the stay $s_{t_1}^{i_1}$ of patient 1249, who is a female, with 83yo, suffers from Measles keratitis, admitted twice before, and with 12 Radiology reports documenting this stay. The second stay $s_{t_2}^{i_1}$ of the same patient 1249, but with 81yo, suffers from Pancreat cyst/pseudocyst, only admitted once before, and with 5 Radiology reports and 7 Nursing/Other reports. The stay $s_{t_3}^{i_2}$ belongs to patient 4695, who is a female, 21yo, with Acute venous embolism and thrombosis of superficial veins of upper extremity, admitted 5 times before, and with 5 Radiology reports and 2 Nursing/Other reports. The stay $s_{t_4}^{i_2}$ of the same patient 4695, but with 22yo, suffers from Hypertensive chronic kidney disease, admitted 8 times before, and with 5 Physician reports and 7 Nursing reports documenting this stay.

This example has been correctly classified as *valid* for all the 9 valid forms. As we do not introduce any order constraint for this setting, we can notice that for some forms like $A : B :: C : D$ and $A : B :: D : C$, $s_{t_2}^{i_1}$ would take place before $s_{t_1}^{i_1}$ in time. The model correctly classifies these analogy forms as valid.

Analysis of a false negative example. We consider the stays in this example to belong to the same patient 1109, who is a female. The stay $s_{t_1}^{i_1}$ of patient 1109, with 25yo, suffers from Malignant essential hypertension, admitted 7 times before, and with 1 Radiology report and 2 Nursing/Other reports documenting this stay. The second stay $s_{t_2}^{i_1}$ of patient 1109, but with 26yo, with Hypertensive chronic kidney disease, admitted 4 times before, and with 8 Physician reports and 4 Nursing reports. The third stay $s_{t_3}^{i_2}$ of patient 1109, but with 26yo, admitted once again for Hypertensive chronic kidney disease, admitted 3 times before, and with 3 Physician reports and 8 Nursing reports. The fourth stay $s_{t_4}^{i_2}$ of patient 1109, with 27yo, suffers from Vascular complications of medical care, admitted 5 times before, and with 3 Physician reports and 9 Nursing reports.

As we mentioned above, for cases where $i_1 = i_2$, applying central permutation would also give us valid analogies. In this example, our model incorrectly classified the form of $D : A :: B : C$ as invalid. As there were less analogies made of four stays that belong to the same patient included in our dataset, we noticed that our model is more likely to incorrectly classify these

analogies, particularly for invalid forms.

Analysis of a true negative example. We consider the stay $s_{t_1}^{i_1}$ of patient 553, who is a male, with 23yo, suffers from Hypertensive chronic kidney disease, admitted 4 times before, and with 4 Physician reports and 8 Nursing reports documenting this stay. The stay $s_{t_2}^{i_1}$ belongs to the same patient 553, but with 24yo, admitted once again for Hypertensive chronic kidney disease, admitted 6 times before, and with 6 Physician reports and 6 Nursing reports. The third stay $s_{t_3}^{i_2}$ belongs to patient 2387, who is a male, with 52yo, with Unspecified disease of pericardium, admitted 7 times before, and with 2 Radiology reports and 2 Nursing/Other reports. The stay $s_{t_4}^{i_2}$ belongs to the same patient 2387, but with 53yo, with Unspecified pleural effusion, admitted 9 times before, and with 8 Radiology reports and 4 Nursing/Other reports.

This analogy has been correctly classified as *invalid* for both invalid forms, $D : A :: B : C$ and $A : C :: B : D$.

Analysis of a false positive example. We consider the stay $s_{t_1}^{i_1}$ of patient 2771, who is a male, with 71yo, suffers from Subendocardial infarction, admitted 16 times before, and with 9 Nursing/Other reports. The second stay $s_{t_2}^{i_1}$ belongs to the same patient 2771, but with 70yo, suffers from Other pulmonary embolism and infarction, admitted 5 before, and with 1 Radiology report and 3 Nursing/Other reports. The stay $s_{t_3}^{i_2}$ of patient 2222, who is a female, with 69yo, with Diverticulosis of colon with hemorrhage, admitted 9 times before, and with 2 Physician reports and 10 Nursing reports. The stay $s_{t_4}^{i_2}$ belongs to the same patient 2222, but with 67yo, with Arterial embolism and thrombosis of lower extremity, admitted 5 times before, and with 3 Nursing/Other reports.

This analogy has been incorrectly classified as *valid* for the invalid form, $A : C :: B : D$. We noticed that when the category of the clinical notes is similar between two hospital stays and when our hospital stays do not include a lot of clinical notes, our model seems to struggle to distinguish between the two hospital stays. $s_{t_2}^{i_1}$ and $s_{t_4}^{i_2}$ have the same number of Nursing/Other reports. Thus these reports may not contain enough information to help our model differentiate between these two stays. As a result, the model incorrectly matches these stays to the same patient.

7. Conclusion

In this paper we discussed an exploratory approach to investigate analogical inference in healthcare. We started by briefly surveying some related work that address different applications of analogical reasoning in different domains. We defined three analogical settings for different healthcare tasks, and discussed the motivation behind our settings. We also presented preliminary statistics of the sets of analogical proportions that we built from MIMIC-III. The main contribution of our work is the formalization of settings that are meaningful in healthcare, and that guide the process of building sets of analogies in healthcare. We discuss the pertinence of certain widely used postulates in this healthcare context. Lastly, we also illustrated the *Identity* setting on which we addressed a preliminary experiment on the analogy detection task. These first results pave the way to conducting further experiments on the other two settings, and to an

in depth analysis of the potential of coupling representation learning and analogical reasoning in healthcare.

Acknowledgments

We thank IARML reviewers for their constructive and positive feedback. 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 named author is partially supported by TAILOR, a EU Horizon 2020 project (GA No 952215), and the Inria Project Lab "Hybrid Approaches for Interpretable AI" (HyAIAI).

References

[1] P. Murena, M. Al-Ghossein, J. Dessalles, A. Cornuéjols, Solving analogies on words based on minimal complexity transformation, in: Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI), 2020, pp. 1848–1854.

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

[3] 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.

[4] 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.

[5] P. B. Jensen, L. J. Jensen, S. Brunak, Mining electronic health records: towards better research applications and clinical care, Nature Reviews Genetics 13 (2012) 395–405.

[6] Y. Si, J. Du, Z. Li, X. Jiang, T. A. Miller, F. Wang, W. J. Zheng, K. Roberts, Deep representation learning of patient data from electronic health records (ehr): A systematic review, Journal of biomedical informatics (2020) 103671.

[7] Y. Li, S. Rao, J. R. A. Solares, A. Hassaine, D. Canoy, Y. Zhu, K. Rahimi, G. Salimi-Khorshidi, Behrt: Transformer for electronic health records, Scientific Reports 10 (2019) 1–12.

[8] I. Landi, B. Glicksberg, H.-C. Lee, S. Cherng, G. Landi, M. Danieletto, C. Furlanello, R. Miotto, Deep representation learning of electronic health records to unlock patient stratification at scale, npj Digital Medicine 3 (2020).

[9] R. Miotto, L. Li, B. A. Kidd, J. T. Dudley, Deep patient: an unsupervised representation to predict the future of patients from the electronic health records, Scientific reports 6 (2016) 1–10.

[10] Y. Huang, N. Wang, Z. Zhang, H. Liu, X. Fei, L. Wei, H. Chen, Patient representation from structured electronic medical records based on embedding technique: Development and validation study, JMIR Medical Informatics 9 (2021).

[11] T. Ruan, L. Lei, Y. Zhou, J. Zhai, L. Zhang, P. He, J. Gao, Representation learning for clinical

time series prediction tasks in electronic health records, BMC Medical Informatics and Decision Making 19-S (2019) 259.

[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] Y. Si, K. Roberts, Patient representation transfer learning from clinical notes based on hierarchical attention network, AMIA Summits on Translational Science Proceedings 2020 (2020) 597.

[15] 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.

[16] 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.

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

[18] 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).

[19] 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.

[20] Centers for Disease Control and Prevention, International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM), https://www.cdc.gov/nchs/icd/icd9cm.htm, 2015. Accessed: 2022-05-01.

[21] 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.

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

[23] 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.