Bag-of-Attributes Representation: a Vector Space Model for Electronic Health Records Analysis in OMOP

José M. Clementino Jr*, Christian C. Bones*, Bruno S. Faiçal*, Oscar C. Linares*
Daniel M. Lima*, Marco A. Gutierrez†, Caetano Traina Jr.* and Agma J. M. Traina*
*Institute of Mathematics and Computer Sciences, University of São Paulo (USP)
São Carlos, SP, BR 13566-590
Email: juniorclementino@usp.br
†Heart Institute (InCor) Clinical Hospital, Faculty of Medicine, University of São Paulo (HCFMUSP)
São Paulo, SP, BR 05403-900

***Abstract*—Several studies have been performed worldwide to improve health services using data generated by digital medical systems. The increasing volume of data generated by these systems is making the use of knowledge discovery and data analysis techniques essential to improve the quality of the health services, which are offered by the medical facilities. However, it is possible to observe a gap, in the literature, about generic and flexible vector space models (VSM) that are well adapted to handle electronic health records (EHR), requiring that each knowledge discovery effort develop their own VSM or other representation model. This restriction can turn a knowledge discovery task over clinical pathways nonviable for comparative evaluations among different methods. Targeting such scenario, we propose the Bag-of-Attributes Representation (BOAR). BOAR represents an EHR as an n-dimensional vector space. Since BOAR takes advantage of the OMOP (Observational Medical Outcomes Partnership) standard, BOAR is able to represent records retrieved from different data models. The experimental results show that BOAR is flexible and robust to representing EHR from several sources, and allows the execution and evaluation of several clustering algorithms.**

***Index Terms*—clinical pathway, OMOP, Vector Space Model, VSM, electronic health records**

I. Introduction

Analyzing the healthcare interventions received by a patient admitted in a hospital, the so-called "clinical pathways", has gained more attention from researchers around the world [1, 2, 3]. The main reason is that extracting useful information from Electronic Health Records (EHR) aids to improve the quality of the health services. Following well-defined clinical pathways during the analysis of a patient's symptoms can bring important benefits, such as early diagnosis, which increases the likelihood of recovery and reduction of operating costs [2]. Clinical pathways are composed of several pieces of information: medical procedures, drug exposures, patient complaints, and procedure occurrences.

The authors would like to thank "Brazilian Coordination of Superior Level Staff Improvement' (CAPES), grant PROEX-11357281/M; the "Sao Paulo Research Foundation" (FAPESP), grants 2016/17078-0, 2018/06228-7, 2019/04660-1, 2018/06074-0, 2020/07200-9; and the "National Council for Scientific and Technological Development" (CNPq).

The amount and variability of data in the EHR are expanding rapidly. This is due to the development of new drugs, procedures, and protocols for analyzing medical practices, which arise with medical advances. In addition, EHR databases and platforms are widely used, and they are expected to continue to improve the diversity of clinical systems [4].

Extracting useful information from EHR in order to define a patient healthcare pathway is a challenging task. Several techniques are involved: big-data mining, high dimensional data representation, complex data visualization, and explicability of the data mining algorithms results [5]. Several methods to extract clinical pathways from EHR data can be found in the literature. However, they present important limitations, such as a limited variety of scenarios in which they may be applicable, and low flexibility for using in different databases. Besides, it is still complex to use them to compare new methods [3, 2, 6, 1].

In this paper we propose a baseline representation for EHR data, which we call the *Bag-of-Attributes* representation (BOAR) model. It is an efficient and flexible data model that allows to design a single representation for different databases and clustering methods. Finding clusters of patients and correlated procedures is a central task for data analytics. Our method allows performing comparative studies among different clustering methods and databases, contributing to lessen the difficulty of comparing methods on distinct databases. Fig. 1 shows the motivation for our method. Scenario (a) illustrates the difficulty of directly comparing two databases with different data models. In scenario (b) the databases are standardized, by using the OMOP (Observational Medical Outcomes Partnership) model [7], but their data models still remain different, keeping the comparison among analytical methods over the database hardsome. Scenario (c) presents a data representation using our Bag-of-attributes model, and how it allows a swift comparison between different analytical methods.

The remaining of this paper is organized as follows. Section II presents the related concepts and related work. The BOAR

2372-9198/20/$31.00 ©2020 IEEE
DOI 10.1109/CBMS49503.2020.00045

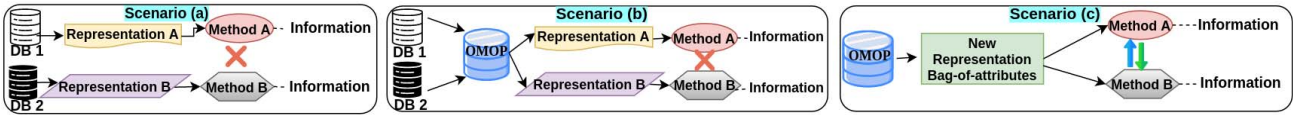


Fig. 1: Three scenarios that motivate our proposal. Scenario (a): the difficulty to compare data mining methods using different databases and representations. Scenario (b): in spite of the databases being standardized by OMOP, different representations make it hard to compare methods. Scenario (c): a new data representation using Bag-of-Attributes Representation that allows comparing different methods, since they are in the same settings.

method is described in Section III. Experiments and implementation details are presented in Section IV. Results are given in Section V. Finally, Section VI presents the conclusions of our work and possible future research.

II. Background and Related Work

Considering groups or clusters of related patients and procedures support on the organization of clinical pathways. Therefore, it is important to evaluate clustering techniques well suited to this task. There are many clustering algorithms, each one using their own structure to represent data [8, 9, 10]. In [8] the authors propose the Hierarchical Agglomerative Clustering (HAC), which uses a merge strategy of average linkage. It minimizes the average of the distances between all pairs of observations in each cluster. The SK-Means algorithm [9], uses the mean of distances between each new element and the clusters' centroids. Their difference is that whereas SK-Means update the cluster for each element during its execution, HAC doesn't do that. The cosine similarity is used as the distance function for both HAC and SK-Means. However, it is not trivial to execute such comparison, as they use different data representations – a limitation that our BOAR method overcomes.

A method to represent data stored in EHR was introduced in [11]. The authors conducted the study on 19 datasets, all of them obtained from *the Stockholm EPR Corpus* [12]. However, this method is not flexible enough to support different data models, being the main problem of this work.

A Vector Space Model (VSM) was used in [13, 14] to represent free text and other complex objects, organizing them into an n-dimensional vector. It creates a dictionary based on the entire set of texts available, where each word corresponds to a dimension. Subsequently, each text is represented as a copy of the dictionary (the "bag"), where the frequency of each word occurring in the text is the value of the corresponding dimension.

The VSM methods produce different feature vectors when the source texts have different sizes, which is not desirable, since they may have similar semantics [14]. To reduce the impact of these differences, the Term Frequency–Inverse Document Frequency (TF-IDF) method has been used to assess the importance of each term in each document, considering the occurrences of the term in the whole set of documents [15, 14]. Thus, when texts have similar semantics, they are placed in the same cluster and the text size contributes little to differentiate them. Several studies have achieved good results combining the VSM and TF-IDF methods [14].

Another initiative aimed at better representing, organizing, and storing medical data is the Observational Medical Outcomes Partnership (OMOP) with the Common Data Model (CDM). OMOP is a data standardization model that allows a systematic analysis of disjoint EHR observational databases [16]. The OMOP data model is being widely adopted, as it allows to uncomplicate Extract-Transform-Load (ETL) processes, using the tools provided by the Observational Health Data Sciences and Informatics (OHDSI) [7].

OMOP was developed for the purposes of analysis, research, and business using health data. In the scientific environment, OMOP allows high portability of methods and interchangeability of EHRs among different studies. On the other hand, in the business environment, OMOP supports data analysis processes to evaluate methods proposed in the scientific literature to optimize their private clinical paths.

The convergence of scientific and business environments leads to fast advances in the optimization of clinical paths. In practice, clinical paths are valuable: (i) for patients, they improve both the efficiency of medical care and quality of life during treatment; (ii) for healthcare professionals, they provide more accurate and early diagnosis; (iii) and for healthcare institutions, they optimize the clinical pathway processes and help reducing operating expenses.

Table I provides a comparison of representative related works, highlighting their desirable characteristics. Most methods are evaluated using records with more than one disease, but using a single database. The lack of methods that use multiple databases can be understood by the difficulty in creating and/or adapting methods that retrieve data from selected patients (cohorts) from databases with distinct organizational models. Different models may present distinct relations between tables and their names, given that, as there are no standards, the attributes may differ completely.

Table I shows that the low flexibility of methods supporting different databases is reflected in the existence of a few methods able to compare their approaches. To compare two methods, it is necessary to adapt both of them to handle a common database, which is a very complex task, due to the high time/cost required to understand and implement the translation of the respective structures to the common model. Therefore, implementing an algorithm able to compare two methods using databases with different data models may be unfeasible.

TABLE I: Related methods comparison.

Methods	Goal	Different scenarios or aplications	Compared with related methods	Scalable to other database
[3]	Pattern discovery by modeling clinical pathway	Yes	No	No
[2]	Discover and cluster clinical pathways	Yes	No	No
[6]	Guide healthcare instances in applying data analytics	No	No	No
[1]	Clinical pathways mining with added temporal pieces of information	Yes	Yes	No
BOAR	**EHR Bag-of-Attributes data representation**	**Yes**	**Yes**	**Yes**

III. Bag-of-Attributes Representation

In this section we describe the proposed Bag-of-Attributes Representation (BOAR) based method. BOAR consists of two main steps, *Cohort selection* and *Builder*. These steps are detailed in Algorithm 1 and visualized in Fig. 2.

Algorithm 1 Bag-of-Attributes Representation, Cohort selection and building steps

Input: : A EHR DataBase: EHR_DB
Input: : Query method: $Query$
Input: : Attributes of interest $AttributesOfInterest$
Output: : Representation $LVector$

$Cohort \leftarrow SelectCohort(EHR\_DB, Query)$
for each $Attribute \in AttributesOfInterest$ **do**
  for each $Column \in Cohort$ **do**
    if $Attribute = Column.Name$ **then**
      $Vfeatures.AddAsRow(Column.values)$
    end if
  end for
end for
$Vfeatures.RemoveSpecialSymbols()$
$Vfeatures.ToLowerCase()$
$Vfeatures.RemoveSpaces()$
$Vfeatures.AttributeToWord()$

$RVector \leftarrow ComputeAttrFrequency(Vfeatures)$
$RVector.ComputeTFIAF()$

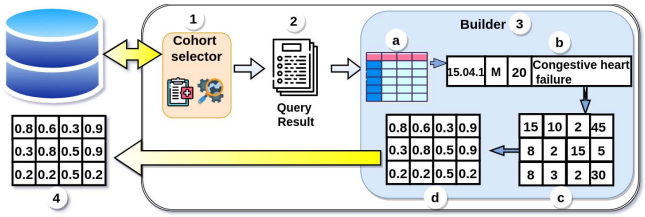


Fig. 2: Bag-of-Attributes Representation (BOAR) workflow. Cohort selection step in (1) and (2), Building step in (3), and the resulting representation vector in (4).

A. The cohort selection step

In this step, the Cohort selector module accesses the database and retrieves the relevant information to be exported to the builder module. This module can be adapted to use databases with different data models, according to the user preference and demand. This process is visualized in Fig 2 item 1. As the generated cohort can be stored in a file, it is possible to use BOAR without doing step 1. We give more details in Section IV.

B. The builder step

The builder module is responsible for creating a table using the information exported by the Cohort selector. In which, the *attributes* of interest, already setup by the user are selected. Each row of the table contains specific information, e.g., regarding the admission of patients into a hospital. Then, a dictionary is built using the selected *attributes* values (see Fig. 3). We assume that the value of each attribute is a unique word, and the Bag-of-Words technique does not allow duplicate values.

In this step, the frequency of attributes for each record is also computed and stored as an n-dimensional vector. We adapted the Term Frequency–inverse Document Frequency (TF-IDF) transformation to handle attributes instead of documents, calling it the *Term Frequency–inverse Attribute Frequency* (TF-IAF) transform. We apply TF-IAF to weight (penalize) the attributes in the vector that appear more frequently. The main difference is that we consider attribute values as words, because OMOP data is structured, using succinct terms previously inserted the vocabulary. The output of the builder module is a matrix composed of the TF-IAF vector. This module was implemented to use the data either retrieved by the cohort selector module or stored in a file.

Id	Concept	Duration (Days)	Gender
1	Essential (primary) hypertension	35	M
2	Congestive heart failure	35	F
3	Essential (primary) hypertension	20	F

Dictionary

essential_primary_hypertension	congestive_heart_failure	35	20	m	f
2	1	2	1	1	2

Fig. 3: Example of a dictionary created by the Builder module.

It is important to highlight that, by representing the data in this way, including the assumption that the entire content of each attribute is a single word, the BOAR method allows the comparison of any number of different algorithms with distinct databases, data models and settings. Moreover, this data representation allows BOAR to generate compact dictionaries and to avoid the need for some widespread steps (such as token

removal) that are needed to handle free text, without degrading the quality of the analysis results.It is important to highlight that some problems, such as misspellings and semantically similar words are solved directly in the OMOP structure.

IV. Experiments

A vector space model should highlight the features of the represented objects that contribute to generate well-defined clusters i.e., dense clusters of objects with well-defined separations. In our case, similar symptoms or diseases may lead the decision maker specialist to choose the clinical procedures, which is given by the clusters of patients that follow similar clinical pathways. That clusters yield useful information to improve health services, making it possible to support medical procedures with more confidence. In special, the largest cluster presents preponderant features to characterize the cohort analyzed, because this cluster is built on the majority of cohort records.

We highlight that we use the term *symptom* to refer to the patient condition retrieved from the database. In OMOP both symptoms and diseases are stored in the same database table for the same attribute.

We apply a cross-validation technique to evaluate the proposed method. We perform the cross-validation technique with $KFold = 10$. To estimate k, at each iteration, we selected 20% of the BOAR representation vector, and to evaluate the clustering we selected 80%. During the estimate step, we divided the 750 symptoms into 10 subsets ($KEstim$) of size 75 each. We compute the silhouette coefficient [8] to select the number of clusters. During the clustering evaluation step (with $KClust$), we evaluate the quality of clustering using the k estimated. This process is detailed in Algorithm 2.

Algorithm 2 Cross-Validation setup

Input: : BOAR representation vector $BVector$

for each $KFold \in [1, 10]$ **do** ▷ **Estimating step**
  $LSilhouette \leftarrow Zeros(10, 2)$ ▷ 10×2 matrix
  for each $KEstim \in [75, 150, ..., 750]$ **do**
    $Labels \leftarrow Clustering(BVector, KClust, 20\%)$
    $Silhouette \leftarrow ComputeSilhouette(Labels)$
    $LSilhouette.Add(Silhouette, KClust)$
  end for
  ▷ **Clustering step**
  $K \leftarrow MaxSilhouette(LSilhouette).KClust$
  $Labels \leftarrow Clustering(BVector, K, 80\%)$
  $WriteLabels(Labels, KFold)$
end for

Once the BOAR vector was labeled ($WriteLabels$ function in Algorithm 2), we selected the most frequent label of each $KFold$ to retrieve its 10 most frequent respective symptoms.

A. Electronic health records dataset

We carried out the experiments on a small anonymized sample [16] of an Electronic Health Records (EHR) dataset provided by the *Instituto do Coração* (InCor) - The Heart Institute - of the University of Sao Paulo, Brazil. This study was performed complying with the applicable data protection laws. This research was approved by the InCor Review Board (IRB) under number $CAEE17146019.0.0000.0068$.

That data include electronic records of $94,603$ patients treated at InCor during the period from $08/04/1998$ to $09/23/2018$, represented in the OMOP common model. The source code is publicly available[1].

B. Experiments setup

The *Cohort selector* module allows the selection of two different cohorts referring to hospital admissions in a time interval. The first cohort has all data records within a defined period of time, and the second cohort has only records with at least one symptom from a provided list, in the same period of time. From now on, we refer to these two cohorts as *cohort-1* and *cohort-2*, respectively. The implementation of the cohorts can be found in the repository[2].

Our evaluation used the interval from $05/29/2007$ to $05/31/2008$. If a patient has more than one admission within this period, the data for each admission are treated independently. The result is stored in a file using the JavaScript Object Notation (JSON) format, so that multiple experiments can be performed using the same selected cohort. Thus, cohort-1 retrieved $45,720$ admissions with 861 attributes each and cohort-2 retrieved $5,530$ with 543 attributes. Therefore, the created table has the date of the procedures and the admission code identifier. The identifier is used only in the analyses to map the representative vector to the normal records to the corresponding records stored in the database.

Based on the cohort selected, the *Builder* constructs a dictionary with n dimensions, corresponding to the max quantity of attributes retrieved in the cohort. Then, using our TF-IAF approach, the frequency of each attribute is inserted in the corresponding dimension of the representative vector. Finally, the table with weighted values of frequency is sent to the clustering algorithm for pattern recognition.

V. Results and discussion

In this section, we show the results of using BOAR applied to two-well known clustering algorithms, SK-Means and Hierarchical Agglomerative Clustering (HAC). Notice that our BOAR method can be used with other clustering algorithms as well. The results obtained in both clustering algorithms using the cohort-1 were compatible. In both algorithms the largest cluster was mainly formed with admissions without a symptom associated. Additionally, the second largest cluster created by each algorithm using the cohort-1 is compatible with the largest cluster of cohort-2 for the respective algorithms. This result evidence that BOAR is robust to highlight characteristics of the admissions, regardless the clustering algorithm used.

[1]https://github.com/clementinojr/Bag-of-Attribute-Representation-BOAR-
[2]https://github.com/clementinojr/Bag-of-Attribute-Representation-BOAR-/tree/master/SelectCohort

These results are also available in our experiment repository at github[3], for further analysis.

From now on, we will discuss the results obtained by cohort-2, due to space limitations and because cohort-2 brings more analytical results to discuss. For cohort-2, Table II summarizes the useful information about the clustering results and the largest clusters generated by each algorithm, considering the 3 most frequent symptoms. Both Table II and Fig. 4 depict the largest cluster obtained with the algorithms HAC and SK-Means.

TABLE II: Data highlights of the cohort-2 in the period from $05/29/2007$ to $05/31/2008$ and about the clustering made by the SK-Means and Hierarchical Agglomerative Clustering algorithms.

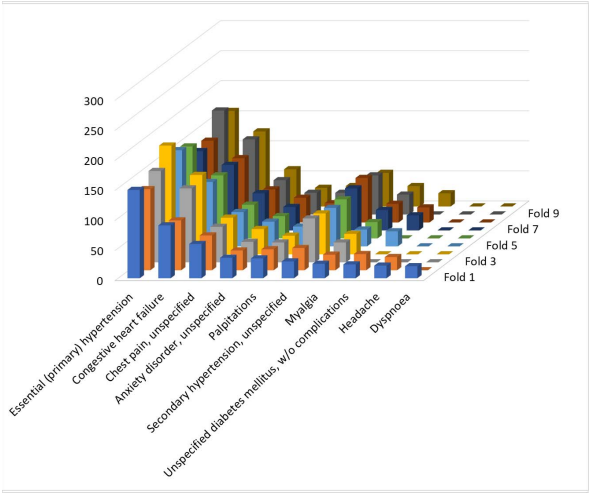
Overall information		**Quantities**
	Hospital admissions	$5,530$
	Total symptoms recorded	$12,684$
	Symptoms	755
SK-Means		**Means of symptoms in cluster**
	Biggest cluster	$1,020 \pm 96$
Most frequent symptoms		
	Essential (primary) hypertension	153 ± 16
	Congestive heart failure	110 ± 16
	Chest pain, unspecified	66 ± 6
Hierarchical Agglomerative Clustering		**Means of symptoms in cluster**
	Biggest cluster	$1,885 \pm 98$
Most frequent symptoms		
	Essential (primary) hypertension	255 ± 19
	Congestive heart failure	161 ± 15
	Chest pain, unspecified	88 ± 11

Fig. 4 shows the frequency (axis y) of symptoms that compose the largest cluster (axis x) in each cross-validation (axis z) cycle for both algorithms. The 10 most frequent symptoms were selected based on the first cross-validation cycle (Fold 1).

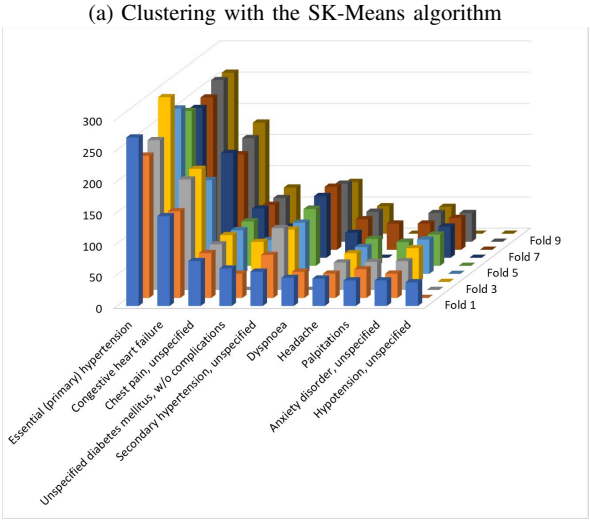
Most symptoms in the records from the largest cluster are related to heart diseases. The frequency of the first 3 symptoms occur in the same order in both clusters (Fig. 4a and 4b). Note that 90% of symptoms appeared in both graphs. That is, 18 out of 20 symptoms are present in both results, as shown in the graphs.

Despite the differences in absolute values shown in Fig. 4a and 4b, it is possible to see a pattern in the amount of records clustered considering the symptoms "Essential (primary) hypertension", "Congestive heart failure" and "Chest pain, unspecified" – the three most frequent symptoms. In addition to being in the same frequency order when comparing both graphs, the same pattern can also be seen when the cycle of cross-validation happens in the other graph. This

[3]https://github.com/clementinojr/Bag-of-Attribute-Representation-BOAR-/tree/master/Experiments/



(a) Clustering with the SK-Means algorithm



(b) Clustering with the Hierarchical Agglomerative algorithm

Fig. 4: Top 10 symptoms obtained by each clustering method, considering only occurrences with symptoms.

similar behavior is also seen in records of other symptoms that are not in the same order, such as: "Secondary hypertension, unspecified" and "Anxiety disorder, unspecified".

It is interesting to note that the numbers of different records from the same symptoms and the patterns of presence follow a similar behavior. The variation in the number of records is a consequence of the cluster algorithm used, because SK-Means uses the distance of each element to a centroid to define its cluster, and HAC uses the mean distance of each element to all other elements of the base to define its cluster. On the other hand, the resembling cluster formation pattern indicates consistency and robust performance in both cases (BOAR+HAC and BOAR+SK-Means).

Considering the behavior of both clustering algorithms, the graphs in Fig. 4 indicate that the SK-Means algorithm presents a more stable performance, since the absent symptoms are

the less frequent ones. Although in certain cycles of cross-validation the HAC algorithm shows some absent records, in other cycles it shows a larger frequency than other symptoms (e.g. records with symptoms "Dyspnoea"). Considering this scenario, we highlight the advantage of enabling the comparison between different algorithms with the same database and settings. In addition to the progress made in the scientific literature about the representativeness of EHR, our proposed BOAR method enables healthcare professionals to evaluate the impact of semantic knowledge discovery using different algorithms, taking advantage of a simple but effective comparison approach.

VI. Conclusion

In this paper, we introduced a Bag-of-Attributes Representation (BOAR) baseline representation of Electronic Health Records (EHR). Our BOAR method provides a representation that any algorithm using the Vector Space Model (VSM) as input can benefit from.

We evaluated and validated our proposal on two representative databases, which were created by two cohorts and two clustering algorithms. Experimental results show that BOAR effectively produced a robust and adaptable common representation of EHR in a vector space, where similar patterns were revealed.

The main advantage of BOAR is that it provides a common model to evaluate, compare, and discover similar information in EHR, regardless of different cohorts and/or clustering algorithms. In addition, it also provides an environment well-suited to evaluate the trade-off between computational cost and quality of data semantic of distinct cohorts. Prior to our baseline representation, performing this task was costly to implement due to the inherent complexity in understanding and matching different models.

As future work, we intend to enhance our method by including measurements to compare semantic information.

References

[1] Z. Huang, X. Lu, and H. Duan, "On mining clinical pathway patterns from medical behaviors," *Artificial Intelligence in Medicine*, vol. 56, no. 1, pp. 35 – 50, 2012.

[2] A. A. Funkner, A. N. Yakovlev, and S. V. Kovalchuk, "Towards evolutionary discovery of typical clinical pathways in electronic health records," *Procedia Computer Science*, vol. 119, pp. 234 – 244, 2017, 6th Intl. Young Scientist Conference on Computational Science, YSC 2017, 01-03 November 2017, Kotka, Finland.

[3] Z. Huang, W. Dong, L. Ji, C. Gan, X. Lu, and H. Duan, "Discovery of clinical pathway patterns from event logs using probabilistic topic models," *J. of Biomedical Informatics*, vol. 47, pp. 39 – 57, 2014.

[4] P. Galetsi, K. Katsaliaki, and S. Kumar, "Big data analytics in health sector: Theoretical framework, techniques and prospects," *Intl. J. of Information Management*, vol. 50, pp. 206–216, 2020.

[5] V. N. Gudivada, R. Baeza-Yates, and V. V. Raghavan, "Big data: Promises and problems," *J.Computer*, vol. 48, no. 3, pp. 20–23, Mar 2015.

[6] J. Lismont, A.-S. Janssens, I. Odnoletkova, S. vanden Broucke, F. Caron, and J. Vanthienen, "A guide for the application of analytics on healthcare processes: A dynamic view on patient pathways," *Computers in Biology and Medicine*, vol. 77, pp. 125 – 134, 2016.

[7] O. H. D. Sciences and Informatics, *The Book of OHDSI*. Independently published, 2019. [Online]. Available: https://www.ohdsi.org/data-standardization/the-common-data-model/

[8] R. Gupta, A. Singhal, and A. Sai Sabitha, "Comparative study of clustering algorithms by conducting a district level analysis of malnutrition," in *2018 8th Intl. Conference on Cloud Computing, Data Science Engineering (Confluence)*, Jan 2018, pp. 280–286.

[9] W. Usino, A. S. Prabuwono, K. H. S. Allehaibi, A. Bramantoro, H. A, and W. Amaldi, "Document similarity detection using k-means and cosine distance," *Intl. J. on Advanced Computer Science and Applications*, vol. 10, no. 2, pp. 165–170, 2019.

[10] S. Ramasamy and K. Nirmala, "Disease prediction in data mining using association rule mining and keyword based clustering algorithms," *Intl. J. on Computers and Applications*, vol. 42, no. 1, pp. 1–8, 2020.

[11] J. Zhao, P. Papapetrou, L. Asker, and H. Boström, "Learning from heterogeneous temporal data in electronic health records," *J. of Biomedical Informatics*, vol. 65, pp. 105 – 119, 2017.

[12] H. Dalianis, M. Hassel, A. Henriksson, and M. Skeppstedt, "Stockholm epr corpus: A clinical database used to improve health care," in *Swedish Language Technology Conference*, 2012, pp. 17–18.

[13] G. Salton and M. E. Lesk, "Computer evaluation of indexing and text processing," *J. ACM*, vol. 15, no. 1, p. 8–36, Jan. 1968.

[14] N. Garcelon, A. Neuraz, V. Benoit, R. Salomon, S. Kracker, F. Suarez, N. Bahi-Buisson, S. Hadj-Rabia, A. Fischer, A. Munnich *et al.*, "Finding patients using similarity measures in a rare diseases-oriented clinical data warehouse: Dr. warehouse and the needle in the needle stack," *J. of biomedical informatics*, vol. 73, pp. 51–61, 2017.

[15] J. Yang, Y.-G. Jiang, A. G. Hauptmann, and C.-W. Ngo, "Evaluating bag-of-visual-words representations in scene classification," in *Proceedings of the Intl. Workshop on Workshop on Multimedia Information Retrieval*, ser. MIR '07. New York, NY, USA: ACM, 2007, pp. 197–206.

[16] D. M. Lima, J. F. Rodrigues-Jr, A. J. Traina, F. A. Pires, and M. A. Gutierrez, "Transforming two decades of epr data to omop cdm for clinical research," *Stud Health Technol Inform*, vol. 264, pp. 233–237, 2019.