DIS: A New Natural Language Processing Inspired Methodology to Investigate Temporal Shifts (Drifts) in Healthcare Data

Bruno de Paiva, Marcos André Gonçalves, Leonardo Chaves Dutra da Rocha, Milena Soriano Marcolino, Fernanda Cristina Barbosa Lana, Maira Viana Rego Souza-Silva, Jussara M Almeida, Polianna Delfino Pereira, Claudio Moisés Valiense de Andrade, Angélica Gomides dos Reis Gomes, Maria Angélica Pires Ferreira, Frederico Bartolazzi, Manuela Furtado Sacioto, Ana Paula Boscato, Milton Henriques Guimarães-Júnior, Priscilla Pereira dos Reis, Felício Roberto Costa, Alzira de Oliveira Jorge, Laryssa Reis Coelho, Marcelo Carneiro, Thaís Lorenna Souza Sales, Silvia Ferreira Araújo, Daniel Vitório Silveira, Karen Brasil Ruschel, Fernanda Caldeira Veloso Santos, Evelin Paola de Almeida Cenci, Luanna Silva Monteiro Menezes, Fernando Anschau, Maria Aparecida Camargos Bicalho, Euler Roberto Fernandes Manenti, Renan Goulart Finger, Daniela Ponce, Filipe Carrilho de Aguiar, Luiza Margoto Marques, Luís César de Castro, Giovanna Grünewald Vietta, Mariana Frizzo de Godoy, Mariana do Nascimento Vilaça, Vivian Costa Morais

Submitted to: JMIR Medical Informatics
on: November 02, 2023

Table of Contents

DIS: A New Natural Language Processing Inspired Methodology to Investigate Temporal Shifts (Drifts) in Healthcare Data

Bruno de Paiva[1] MD; Marcos André Gonçalves[1]; Leonardo Chaves Dutra da Rocha[2]; Milena Soriano Marcolino[3]; Fernanda Cristina Barbosa Lana[3]; Maira Viana Rego Souza-Silva[3]; Jussara M Almeida[1]; Polianna Delfino Pereira[3]; Claudio Moisés Valiense de Andrade[1]; Angélica Gomides dos Reis Gomes[4]; Maria Angélica Pires Ferreira[5]; Frederico Bartolazzi[6]; Manuela Furtado Sacioto[7]; Ana Paula Boscato[8]; Milton Henriques Guimarães-Júnior[9]; Priscilla Pereira dos Reis[10]; Felício Roberto Costa[3]; Alzira de Oliveira Jorge[11]; Laryssa Reis Coelho[12]; Marcelo Carneiro[13]; Thaís Lorenna Souza Sales[1]; Silvia Ferreira Araújo[14]; Daniel Vitório Silveira[15]; Karen Brasil Ruschel[1]; Fernanda Caldeira Veloso Santos[16]; Evelin Paola de Almeida Cenci[17]; Luanna Silva Monteiro Menezes[1]; Fernando Anschau[18]; Maria Aparecida Camargos Bicalho[19]; Euler Roberto Fernandes Manenti[20]; Renan Goulart Finger[21]; Daniela Ponce[22]; Filipe Carrilho de Aguiar[23]; Luiza Margoto Marques[7]; Luís César de Castro[24]; Giovanna Grünewald Vietta[25]; Mariana Frizzo de Godoy[6]; Mariana do Nascimento Vilaça[26]; Vivian Costa Morais[7]

[1]Universidade Federal de Minas Gerais, Computer Science Department, Belo Horizonte, Brazil Belo Horizonte / MG BR
[2]Universidade Federal de São João del-Rei, Computer Science Department , Brazil São João del-Rei BR
[3]Universidade Federal de Minas Gerais, Faculdade de Medicina Belo Horizonte BR
[4]Hospitais da Rede Mater Dei Belo Horizonte BR
[5]Hospital de Clínicas de Porto Alegre Porto Alegre BR
[6]Hospital Santo Antônio Curvelo BR
[7]Faculdade Ciências Médicas de Minas Gerais Belo Horizonte BR
[8]Hospital Tacchini Bento Gonçalves BR
[9]Hospital Márcio Cunha Ipatinga BR
[10]Hospital Metropolitano Doutor Célio de Castro Belo Horizonte BR
[11]Hospital Risoleta Tolentino Neves Belo Horizonte BR
[12]Universidade Federal dos Vales do Jequitinhonha e Mucuri, Faculdade de Medicina Teófilo Otoni BR
[13]Hospital Santa Cruz Santa Cruz do Sul BR
[14]Hospital Semper Belo Horizonte BR
[15]Hospital Unimed BH Belo Horizonte BR
[16]Hospital Universitário de Santa Maria Santa Maria BR
[17]Hospital Moinhos de Vento Porto Alegre BR
[18]18Hospital Nossa Senhora da Conceição and Hospital Cristo Redentor Porto Alegre BR
[19]Fundação Hospitalar do Estado de Minas Gerais (FHEMIG) Belo Horizonte BR
[20]Hospital Mãe de Deus Porto Alegre BR
[21]Hospital Regional do Oeste Chapecó BR
[22]Universidade Estadual Paulista Júlio de Mesquita Filho, Faculdade de Medicina de Botucatu Botucatu BR
[23]Hospital das Clínicas, Universidade Federal de Pernambuco Recife BR
[24]Hospital Bruno Born Lajeado BR
[25]Hospital SOS Cárdio Florianópolis BR
[26]Hospital Metropolitano Odilon Behrens Belo Horizonte BR

Corresponding Author:
Bruno de Paiva MD
Universidade Federal de Minas Gerais, Computer Science Department, Belo Horizonte, Brazil
Av. Pres. Antônio Carlos, 6627 - Pampulha
Ap603
Belo Horizonte / MG
BR

Abstract

Background: Healthcare data is a valuable resource for improving patient's outcomes. If adequately treated and interpreted, it can enhance healthcare services and help to understand the impacts of new technologies and treatments. One important aspect of

healthcare data is that it is usually temporal, in the sense that it is collected over time and is susceptible to temporal shifts. For instance, COVID-19 vaccination dramatically changed the profile of hospitalizations and deaths, initially decreasing the mean age of the at-risk patients and then creating a large shift in the dying patient's characteristics. These temporal shifts may have significant impacts depending on the task one wishes to learn from and have particular relevance in understanding which factors (e.g., new technologies/treatments) affect patient outcomes.

Objective: We propose DIS(Detection, Initial Characterization, Semantic Characterization), a new methodology for analyzing the changes in health outcomes and variables over time while discovering outcome contextual changes in large volumes of data. DIS can help identifying patterns and trends, essential for decision-making and resource allocation, as well as improving the effectiveness of machine learning predictive algorithms applied to healthcare data.

Methods: The DIS methodology is based on a 3-step process that starts with drift detection, goes through an initial characterization that helps us direct the focus of the analysis, and evolves into a semantic characterization. By combining the outcomes from these three steps, our results can provide hints at specific factors, such as interventions and healthcare practice modifications, that drive the changes in patient outcomes.

Results: We applied the DIS methodology to two distinct datasets: the Brazilian COVID-19 multicenter cohort and the Medical Information Mart for Intensive Care, version IV (MIMIC-IV) dataset. In doing so, we were able to obtain insights hinting at the root causes for the drop in overall (all causes) mortality in the two datasets, such as the role of vaccination in the COVID-19 pandemic and the decrease in iatrogenic events and cancer related deaths in the MIMIC-IV dataset.

Conclusions: We successfully applied machine learning methods to detect, characterize and explain temporal shifts in healthcare data. Understanding these changes can improve patient outcomes, as well as healthcare resource allocation.

(JMIR Preprints 02/11/2023:54246)
DOI: https://doi.org/10.2196/preprints.54246

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

DIS: A New Natural Language Processing Inspired Methodology to Investigate Temporal Shifts (Drifts) in Healthcare Data

Bruno Barbosa Miranda de Paiva[1] MD, MSc (Paiva BBM; brunobarbosa.mpaiva@gmail.com; 0000-0001-8378-3650)
Leonardo Chaves da Rocha[2] MSc, PhD (Rocha LC; lcrocha@ufsj.edu.br; 0000-0002-4913-4902)
Jussara Marques de Almeida[1] MSc, PhD (Almeida JM; jussara@dcc.ufmg.br; 0000-0001-9142-2919)
Milena Soriano Marcolino[3-5] MD, MSc, PhD (Marcolino MS; milenamarc@gmail.com; 0000-0003-4278-3771)
Claudio Moisés Valiense de Andrade[1] MSc (Andrade CMV; claudiovaliense@gmail.com; 0000-0002-7366-2633)
Maira Viana Rego Souza-Silva[3] MD, MSc (Souza-Silva MVR; mairavsouza@gmail.com; 0000-0003-2079-7291)
Fernanda Cristina Barbosa Lana[3] MD (Lana FCB; fernandacblana@gmail.com; 0000-0002-8187-1152)
Polianna Delfino Pereira[3,4] MSc, PhD (Delfino-Pereira P; polidelfino@yahoo.com.br; 0000-0003-2406-6576)
Angélica Gomides dos Reis Gomes[6] MD, MSc (Gomes AGR; angelicagrgomes@gmail.com; 0000-0002-4568-0738)
Alzira de Oliveira Jorge[7] MD, MSc, PhD (Jorge AO; alzira.o.jorge@gmail.com; 0000-0003-1366-1732)
Ana Paula Boscato[8] MD, MSc (Boscato AP; apboscato@hotmail.com; 0009-0006-5894-8672)
Artur Vestena Rossato[9] (Rossato AV; arturrossato2003@gmail.com; 0009-0000-6727-4557)
Daniel Vitório Silveira[10] MSc (Silveira DV; danielvez@gmail.com; 0000-0002-7381-1651)
Daniela Ponce[11] MD, PhD (Ponce D; daniela.ponce@unesp.br; 0000-0002-6178-6938)
Euler Roberto Fernandes Manenti[12] MD, MSc, PhD (Manenti ER; eulermanenti@gmail.com; 0000-0003-1592-4727)
Evelin Paola de Almeida Cenci[13] BSc (Cenci EPA; assistencial6_hu@centrodepesquisaclinica.com.br; 0000-0001-8548-9279;)
Fernanda Caldeira Veloso Santos[14] MD (Santos FCV; fcvs.med@gmail.com; 0000-0002-1391-558X)
Frederico Bartolazzi[15] MD, MSc (Bartolazzi F; fredlazzi@hotmail.com; 0000-0002-9696-4685)
Giovanna Grünewald Vietta[16] MD, MSc, PhD (Vietta GG; ggvietta@gmail.com; 0000-0002-0756-3098)
Karen Brasil Ruschel[17] MSc, PhD (Ruschel KB; karenbruschel@gmail.com; 0000-0002-6362-1889)
Laryssa Reis Coelho[18] MD (Coelho LR; laryssa.coelho@ufvjm.edu.br; 0000-0001-9668-2349)
Luanna Silva Monteiro Menezes[19] MD, MSc (Monteiro LS; luannasmonteiro@gmail.com; 0000-0002-6621-3338)
Luís César de Castro[20] MSc, PhD (Castro LC; pharmlucamsc@gmail.com; 0000-0003-2379-0167)
Luiza Marinho Motta Santa Rosa[21] (Rosa LMM; luiza.motta26@gmail.com; 0000-0002-4741-4871)
Marcelo Carneiro[22] MD, PhD (Carneiro M; marceloc@unisc.br; 0000-0003-3603-1987)

Maria Angélica Pires Ferreira[23] MD, MSc, PhD (Ferreira MAP; mpiferreira@hcpa.edu.br; 0000-0003-0961-524X)
Maria Aparecida Camargos Bicalho[3,24] MD, MSc, PhD (Bicalho MAC; macbicalho@gmail.com; 0000-0001-6298-9377)
Mariana do Nascimento Vilaça[25] MD (Vilaça MN; mariananvilaca@gmail.com; 0000-0003-3950-476X)
Mariana Frizzo de Godoy[26] MD (Godoy MF; mfdegodoy@gmail.com; 0000-0002-6631-8826)
Milton Henriques Guimarães-Júnior[27] MD, MSc (Guimarães Júnior MH; miltonhenriques@yahoo.com.br; 0000-0002-2127-8015)
Priscilla Pereira dos Reis[28] MD (Reis PP; priscillamed@hotmail.com; 0000-0002-0340-9464),
Renan Goulart Finger[30] MD (Finger RG; renanfinger@yahoo.com.br; 0000-0002-5569-7787)
Thaís Lorenna Souza Sales[4,31] MSc, PhD (Sales TLS; thaislorennass30@yahoo.com.br; 0000-0002-1571-3850)
Marcos André Gonçalves[1] MSc, PhD (Gonçalves MA; mgoncalv@dcc.ufmg.br; 0000-0002-2075-3363)

[1]Computer Science Department, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil.
[2]Computer Science Department, Universidade Federal de São João del-Rei, São João del-Rei, Brazil.
[3]Medical School, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil.
[4]Institute for Health Technology Assessment (IATS/ CNPq), Porto Alegre, Brazil.
[5]Telehealth Center, University Hospital, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil.
[6]Hospitais da Rede Mater Dei, Belo Horizonte, Brazil.
[7]Hospital Risoleta Tolentino Neves, Belo Horizonte, Brazil.
[8]Hospital Tacchini, Bento Gonçalves, Brazil.
[9]Universidade Federal de Ciências da Saúde de Porto Alegre (UFCSPA), Porto Alegre, Brazil.
[10]Hospital Unimed BH, Belo Horizonte, Brazil.
[11]Medical School, Universidade Estadual Paulista Júlio de Mesquita Filho, Botucatu, Brazil.
[12]Hospital Mãe de Deus, Porto Alegre, Brazil.
[13]Hospital Universitário Canoas, Canoas, Brazil.
[14]Hospital Universitário de Santa Maria, Santa Maria, Brazil.
[15]Hospital Santo Antônio, Curvelo, Brazil.
[16]Hospital SOS Cardio, Florianópolis, Brazil.
[17]Hospital Mãe de Deus. R. José de Alencar, Porto Alegre, Brazil.
[18]Medical School, Universidade Federal dos Vales do Jequitinhonha e Mucuri, Teófilo Otoni, Brazil.
[19]Hospital Luxemburgo, Belo Horizonte, Brazil.
[20]Hospital Bruno Born, Lajeado, Brazil.
[21]Faculdade Ciências Médicas de Minas Gerais, Belo Horizonte, Brazil.
[22]Hospital Santa Cruz, Santa Cruz do Sul, Brazil.
[23]Hospital de Clínicas de Porto Alegre, Porto Alegre, Brazil.
[24]Fundação Hospitalar do Estado de Minas Gerais (FHEMIG), Belo Horizonte, Brazil.
[25]Instituto Orizonti, Belo Horizonte, Brazil.
[26]Hospital São Lucas PUCRS, Porto Alegre, Brazil.
[27]Hospital Márcio Cunha, Ipatinga, Brazil.
[28]Hospital Metropolitano Odilon Behrens, Belo Horizonte, Brazil.
[29]Hospital Universitário Professor Edgard Santos, Salvador, Brazil.
[30]Hospital Regional do Oeste, Chapecó, Brazil.

[31]Universidade Federal de São João del-Rei, Divinópolis, Brazil.

Corresponding Author

Bruno Barbosa Miranda de Paiva.
Computer Science Department. Avenida Antônio Carlos, nº 6627, Belo Horizonte/MG. CEP: 31.270-901.
Phone: +55 31 9971-0134
E-mail: brunobarbosa.mpaiva@gmail.com

Abstract

Background: Proper analysis and interpretation of healthcare data can significantly improve patient outcomes by enhancing services and revealing the impacts of new technologies and treatments. Understanding the substantial impact of temporal shifts in this data is crucial. For example, COVID-19 vaccination initially lowered the mean age of at-risk patients and later changed the characteristics of those who died. This highlights the importance of understanding these shifts for assessing factors that affect patient outcomes.

Objectives: To propose DIS (Detection, Initial Characterization, Semantic Characterization), a new methodology for analyzing changes in health outcomes and variables over time while discovering outcome contextual changes in large volumes of data.

Methods: The DIS methodology involves three steps: detection, initial characterization, and semantic characterization. Detection uses metrics like Jensen-Shannon divergence to identify significant data drifts. Initial characterization offers a global analysis of changes in data distribution and predictive feature significance over time. Semantic characterization employs NLP-inspired techniques to understand the local context of these changes, helping to identify factors driving changes in patient outcomes. By integrating the outcomes from these three steps, our results can provide hints at specific factors (e.g. interventions and modifications in healthcare practices), that drive changes in patient outcomes. DIS was applied to the Brazilian COVID-19 Registry and the MIMIC-IV datasets.

Results:Our approach allowed us to: (1) identify drifts effectively, especially using metrics such as the Jensen-Shannon Divergence, and (2) uncover reasons for the decline in overall mortality in both the COVID-19 and MIMIC-IV datasets, as well as changes in the co-occurrence between different diseases and this particular outcome. Factors such as vaccination during the COVID-19 pandemic and reduced iatrogenic events and cancer-related deaths in MIMIC-IV were highlighted. The methodology also pinpointed shifts in patient demographics and disease patterns, providing insights into the evolving healthcare landscape during the study period.

Conclusions: We developed a novel methodology combining machine learning (ML) and NLP techniques to detect, characterize, and understand temporal shifts in healthcare data. This understanding can enhance predictive algorithms, improve patient outcomes, and optimize healthcare resource allocation,ultimately improving the effectiveness of ML predictive algorithms applied to healthcare data. Our methodology can be applied to a variety of scenarios beyond those discussed in this article.

Keywords: Healthcare; Machine Learning; Data drifts; Temporal drifts.

Introduction

Healthcare data is a critical resource that can be used to improve patient outcomes and the financial performance of healthcare institutions [1,2]. By analyzing patient data, healthcare providers can gain insights into patient health status, identify trends, and make informed decisions about treatment plans. Properly collected, managed, treated, and interpreted healthcare data can help providers improve operational efficiency and reduce costs, thereby improving financial results [3].

One of the primary ways healthcare data can be used to enhance medical decisions and potentially improve patient outcomes is through predictive analysis. This technique utilizes historical data to identify patterns and predict future outcomes, hence enabling the recognition of high-risk patients, simulation of different therapeutic approaches, and the personalization of patient care. However, relying on historical data has its caveats, as the predictive capacity of different variables is not fixed over time. Ignoring these aspects of temporal data may lead to prediction errors and learning instabilities. These variations in performance are part of what is known as temporal data shifts [4-7].

A temporal data shift refers to a change in the statistical properties of a dataset over time, which can degrade model accuracy. In healthcare, this may occur due to various reasons, including changes in data collection practices, software updates or replacements, changes in patient behavior or lifestyle habits, and the introduction of new therapeutic technologies. These temporal events may lead to inconsistencies and discrepancies in the data, which may affect both the accuracy and the reliability of the data and models trained on it. The impacts can be significant [4,7], as they can lead to incorrect diagnoses, inappropriate treatment plans, and poor patient outcome predictions. This highlights the importance of managing, characterizing, and mitigating these temporal effects [8].

We are particularly interested in how temporal data drifts can be utilized to analyze the effectiveness of new patient treatment options. Changes in predictive capacity can provide insights into the impact of new treatments on patient outcomes. For instance, by comparing data collected before and after introducing a new treatment, we can identify any shifts that may indicate improved patient outcomes. If the data drift analysis indicates a positive impact of the new treatment, healthcare providers may choose to continue to monitor the data to ensure that the positive effects are sustained while maintaining the use of the new therapeutic option [9].

A notable example of a condition that experienced an important data drift over time is the human immunodeficiency virus infection. In the 1980s, HIV infection was a strong predictor of early death. However, it has now become more of a chronic condition, such as diabetes mellitus or systemic hypertension. In the same way, advancements in breast cancer treatment have significantly increased survivorship over the years [10].

Similarly, several infectious diseases, such as poliomyelitis or measles, have been nearly eradicated in most parts of the world, making them unlikely hypotheses for new diagnoses [11,12]. In the case of COVID-19, vaccination has dramatically changed the profile of hospitalizations and deaths [4,13], initially decreasing the mean age of patients at risk and creating a clear distinction between the periods before and after vaccination.

Our Main Contribution – The DIS Methodology

Building upon the idea of analyzing data drifts to obtain insights into how and if new technologies or treatments have impacted patient outcomes, this article proposes a novel 3-step healthcare temporal analysis methodology, called DIS – Detection, Initial characterization and Semantic characterization. The proposed DIS methodology is summarized in Figure 1. It consists of three main steps: (1) Detection, (2) Initial Characterization, and (3) Semantic Characterization, which are described in the following sections.

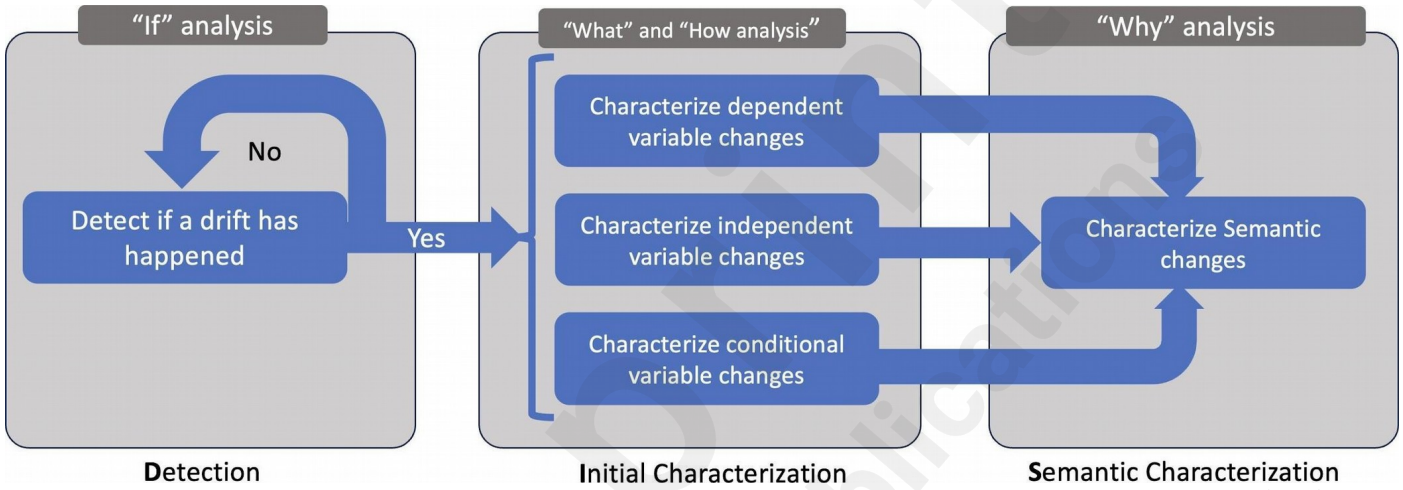


Figure 1. Overview of the Detection, Initial Characterization and Semantic Characterization (DIS) Methodology.

In summary, we exploit various drift detection metrics in the detection step to identify any significant instance of data drift. Some of the metrics we may explore in this step include Jensen-Shannon divergence [14], Autoencoder reconstruction error [15], and centroid distances [16]. If changes are detected, we proceed to the Initial characterization step, where we obtain a global (dataset-level) descriptive analysis of what has changed and how the discriminative/predictive power of each feature and the distribution of labels have evolved over time.

Additionally, we introduce the concept of temporal granularity in the data drift domain, which holds particular significance in healthcare data drifts and influences the instantiation of our third and final step. High temporal granularity refers to a dataset that allows the visualization of numerous events over time for individual patients, with a clear understanding of the chronological order among these events. Conversely, low temporal granularity describes a scenario where each patient is considered a singular event in time, lacking clarity regarding the precedence or sequence of different attributes.

Finally, guided by these principles, we proceed to the third Semantic Characterization step, which exploits concepts popularized in the Natural Language Processing (NLP) domain to provide a localized (instance level) perspective of why certain shifts occurred. To achieve this, we exploit vector embeddings derived from healthcare events, such as sequences of the International Classification of Diseases –(ICD) codes, vital data measurements, and consumption items. Each of these semantic units (ICD codes, measurements, consumed items, etc.) is treated as an "event" or, in NLP terminology, a "token". By employing NLP-inspired techniques to create semantic embeddings for these entities, we aim to uncover insights into the changing context and its impact

on the outcomes of interest over time.

Before delving into the details of each step in our methodology, it is crucial to emphasize that our DIS approach differs significantly from the common practices., While conventional methods usually involve an ad hoc combination of techniques for data collection, qualitative data processing and extraction, and data analysis, our DIS methodology offers a planned and structured procedure, as illustrated in Figure 1. This procedure delineates the required steps to understand data drift in healthcare data. As we will demonstrate and discuss, these steps can be tailored to various case studies by applying different techniques depending on specific data characteristics. We also offer guidance for selecting one particular approach given a specific scenario. Furthermore, we discuss how the results of each step can inform the execution of the following ones and how the combined results of all steps can support our understanding of the drift.

More broadly, to the best of our knowledge, this is the first study to examine data drifts in healthcare from a technology incorporation standpoint. Rather than solely focusing on enhancing the robustness of machine learning models, we delve into the underlying factors driving temporal shifts in patient outcomes. Our aim is to study the impact of emerging technologies such as new drugs, patient care policies, or vaccines. Over the following sections, we detail the steps of our DIS methodology and illustrate its application in two case studies with distinct characteristics in terms of temporal data shifts: (i) the Brazilian COVID-19 Registry dataset [17] and the Medical Information Mart for Intensive Care IV (MIMIC-IV) dataset [18]. By doing so, we illustrate how DIS can obtain insights into the reasons behind some real-life data drifts, as well as their potential impacts, both positive and negative, from a healthcare perspective.

In summary, the main contributions of this article are:

1. The proposal of a new data drift characterization and analysis methodology - DIS - that is flexible enough to work on different scenarios. DIS encapsulates and cohesively organizes a sequence of necessary steps for data drift analysis.
2. A new semantic analysis step based on NLP embeddings for temporal understanding which focuses on comprehending the context of relevant outcomes by examining changes in their embedding vector changes over time. By incorporating such semantic techniques, DIS provides deeper insights into the reasons behind temporal changes, especially when combined with domain-specific knowledge. This approach allows for a more nuanced analysis of data evolution over time, capturing complex patterns and relationships that may not be apparent with traditional methods such as Cluster Analysis.
3. The application of the DIS methodology to two different case studies with very different temporal granularity profiles illustrates some insightful analyses that can be obtained with it. We also offer guidelines to aid practitioners in making informed decisions about which methods to employ in each step of our methodology, based on particular characteristics of the data. This demonstrates the generalizability and applicability of DIS across different scenarios.

Methods

A Detailed Description of The DIS Methodology

Detection Step

In step 1 (Detection), the main focus is on assessing whether the data has relevant temporal variations. Monitoring and detecting such data drifts is crucial for upholding the accuracy and reliability of machine learning models and for identifying beneficial and detrimental changes in

healthcare caused by interventions, such as the introduction of new treatments or drugs. From the perspective of a healthcare service or company, this step identifies whether changes are occurring, potentially prompting further investigations that could enhance service efficiency over time.

For the *detection step*, we recommend splitting the data into temporal chunks and then comparing the data distributions in consecutive chunks. A drift is detected whenever the distributions of distinct chunks exhibit significant differences. Various metrics to compare empirical distributions are available in the literature. These metrics have different characteristics and underlying principles, which may lead to relevant differences to their effectiveness for detecting temporal data drifts. In this work, we consider the following metrics: Centroid Cosine Distance [16], Jensen-Shannon Divergence [14], Autoencoder Reconstruction Error [15], Classifier Error (in separating two-time chunks) [19] and Principal Component Analysis Reconstruction Error [20] metrics.

The Centroid Cosine Distance metric assesses changes in the central points of data clusters over time, and is sensitive to numeric outliers, particularly in heavy-tailed distributions where extremes can be multiple orders of magnitude larger than typical values. The Principal Component Analysis (PCA) Reconstruction Error captures variations in data structure by quantifying the difference between original and reconstructed data. Similarly, Autoencoder Reconstruction Error focuses on reconstruction accuracy. Both metrics measure the "novelty" of a data point and are sensitive to numerical outliers. On the other hand, the Classifier Error evaluates a model's ability to distinguish past from future data, providing insights into how drift affects predictive capabilities. Lastly, the Jensen-Shannon Divergence quantifies distributional changes, offering a broader perspective on underlying data distribution shifts over time. While reconstruction errors and centroids excel in detecting local outliers and structural changes, the Jensen-Shannon Divergence and Classifier Error provide a more comprehensive view of distributional shifts, making them valuable for modeling the impact of temporal drifts on data distributions.

As an example, our prior analysis of the Brazilian COVID-19 Registry [17] revealed a data drift that significantly impacted the death prediction task, suggesting that vaccination had a pivotal role in the profiles of hospitalized and deceased patients during the COVID-19 pandemic [4]. Although this is an interesting finding, the previous study did not present a proper structure to detect, monitor, and interpret such drifts generically, nor did it propose mechanisms to detect semantic information associated with specific outcomes.

The drift caused by vaccination can be initially hypothesized by comparing the data distributions of consecutive chunks (e.g., near future vs. recent past) using a classification approach. This involves monitoring the prediction model’s performance over time using metrics such as accuracy, precision, and recall. Alternatively, the distribution of different features over time can be tracked using metrics, such as Jensen-Shannon Divergence or Autoencoder Reconstruction Errors. If the model’s performance drops (or changes) significantly over time or if the differences between the metrics exceeds a certain threshold, it may indicate a data drift. A summary of this monitoring loop is illustrated in Figure 2.

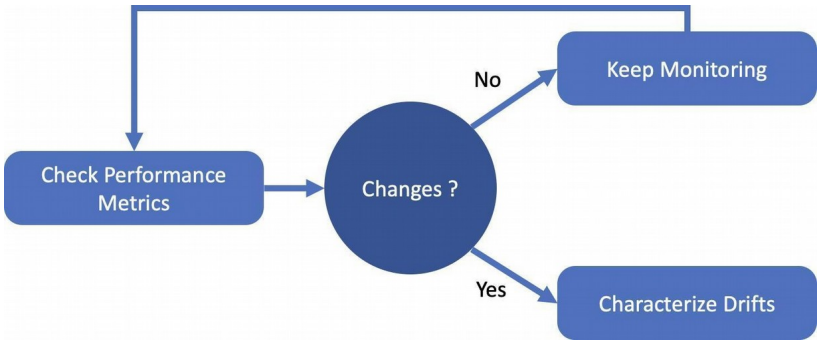


Figure 2. The temporal drift monitoring loop. We usually observe temporal shifts as important variations in model effectiveness over time.

Initial Characterization Step

Once a drift has been detected, we proceed to step 2 (Initial Characterization), where we begin to understand, from a global perspective (all data), *how* the data has changed (Table 1). This stage focuses on developing a general (global) comprehension of the *what´s* and the *how's* contributing to the changes observed in the data collection. Specifically, we are interested in characterizing variations in both *dependent* $P(y)$ and independent $P(x)$ variables, as well as the conditional probability of the dependent variables given the independent variables $P(y|x)$. To reach these goals, we examine how $P(y)$ has changed by plotting its frequency over time; the same is valid for $P(x)$. For $P(y|x)$, we can explore different complementary techniques that can help understanding the drifts globally. We can analyze how the different correlation metrics between the top independent variables and the dependent variable change over time – for instance, with Pearson [21] or Spearman [22] correlations – or analyze the feature Importance of tree-based learners or entropy-based, measures such as information gain or Chi-square over time [23]. Another possibility is to exploit explainability metrics based on game theory, such as SHapley Additive exPlanations (SHAP) values [24].

This type of analysis facilitates understanding how the relationship between predictive variables and the outcome of interest has evolved from a global perspective. Additionally, it is helpful to check the change rate for each selected outcome by using similarity metrics and comparing the different groups of patients over time. At this stage, it is feasible to answer valuable research and business questions. For instance, we may observe a decreased likelihood of a "death" outcome in a given population, such as COVID-19 or breast cancer patients. We may also spot changes in the profile of the patients who had adverse outcomes. Following these initial insights, the subsequent task is to understand *why* such changes happened – the goal of *step 3*.

Table 1. Drift types concerning the passing of time, in accordance to Moreno-Torres et al., (2012).“Sudden drift”, “incremental drift”, “gradual drift”, and “reoccurring drift”. “sudden drift” describes a situation where changes are abrupt and usually caused by a single event, such as a change in data collection practices, where an attribute stops being collected. “Incremental drift” describes gradual and directional changes in a data distribution, such as the observed increase in the overweight and obese population over the past years. “Gradual drift” is similar, but does not imply directional changes. Instead, it encompasses other gradual changes, such as the slow change in the hospital admission profile over many years. Finally, a “reoccurring drift” refers to a drift pattern that repeats over time, such as the seasonal increase in

emergency services admitting Influenza patients during predictable seasons of the year.

Data Drift Type	Description
Sudden Drift	Abrupt and unexpected changes in the data.
Incremental Drift	Gradual and continuous changes over time.
Gradual Drift	Slow and steady changes in the data distribution.
Reoccurring Drift	Periodic or repetitive shifts in the data.

Moreno-Torres JG, Raeder T, Alaiz-Rodríguez R, Chawla N V., Herrera F. A unifying view on dataset shift in classification. Pattern Recognit. 2012;45(1):521-530. doi:10.1016/j.patcog.2011.06.019.

Semantic Characterization Step

In step 3, the main focus is to learn *why* the changes we observed in step 2 happened. This step aggregates fundamental research and business value into our methodology, and is heavily dependent on the temporal granularity of the data under evaluation. To the best of our knowledge, this is the first study to look at data drifts in healthcare from a technology incorporation standpoint. For instance, as aforementioned, we may already have learned, as a result of step I, that a given disease or condition, such as COVID-19, had a decreased lethality over a specific time period. Given that information, what will add value to healthcare services is the discovery of which repeatable interventions within that time frame can be consistently beneficial .

We begin step 3 by proposing a novel NLP-inspired technique based on token embedding techniques, such as Word2Vec [25], to detect local or individual changes in outcome contexts over time. We opt for NLP-inspired techniques because they effectively model and comprehend "semantics" and "contexts". In this context, we treat each patient as a "document", and any temporally discrete healthcare event or information, such as disease codes or items used during a hospital stay, as a "token" (i.e., the equivalent of a "word" or a "subword" in NLP). For instance, the underlying premise is that a patient's semantics can be understood by examining their diseases and consumption history. Based on this representation, we characterize which entities or outcome groups have undergone the most significant changes regarding their defining characteristics in comparison to a baseline or initial time chunk. This assessment assumes a setting where we have an outcome y and a task of predicting this outcome using independent variables X. This characterization can be achieved by comparing the distance of each class's centroid to a reference centroid, where a "centroid" represents the arithmetic mean of each patient's features.

The procedure to compute each of these *centroids* is explained in Figure 3. In this Figure, we show a simplified view of two groups of patients in two dimensions and how the centroids are calculated to be at the spatial "center" of the groups by averaging their attributes. We can compare different centroids using either a cosine distance or cosine similarity (Equation 1). This type of analysis can guide our research towards a specific hypothesis, filtering down to the pattern changes in specific outcomes, such as death or the need for mechanical ventilation during a hospital stay.

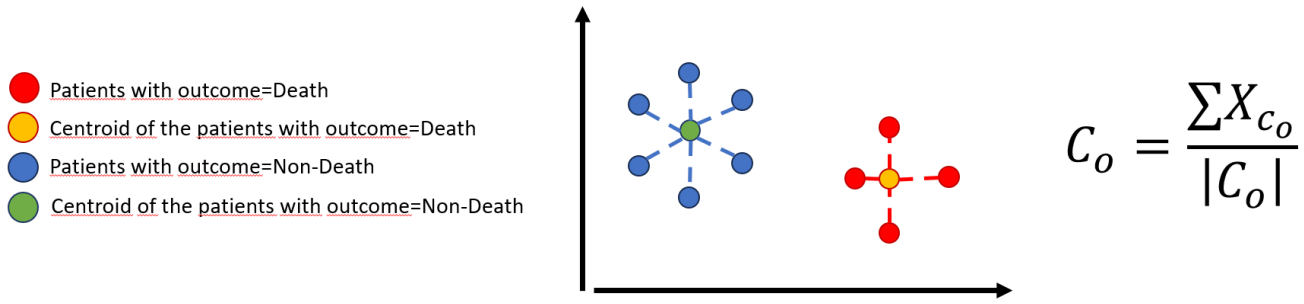


Figure 3. Modeling of the centroids as the arithmetic mean of the features in each outcome group. C_o is the centroid of cluster O, X_cO is the matrix of attributes including all patients in the outcome O and $|C_O|$ is the number of patients in the outcome group O.

$$\cos(\theta) = \frac{A \,.\, B}{¿ \vee A \vee ¿ \,.\, \vee ¿ B \vee ¿}$$

Equation 1. Cosine similarity. The cosine distance is simply 1 - cosine similarity.

The centroid of each class on the first (time) chunk will be analyzed over time, providing insight into which outcomes (e.g., death vs. non-death or hospitalization vs. non-hospitalization) underwent the most significant changes. From that observation, we can focus our analysis into the interest group. This approach, which will be further illustrated in our experiments, allows us to compute semantic distances among patients, between patients and outcomes, and among different outcomes.

To apply step 3 to a dataset, we need to remember that healthcare data comes in different temporal and semantic granularities. For instance, datasets like the Brazilian COVID-19 Registry (details presented in next section) treat each patient as a single data point, characterized by atomic temporal granularity, where temporal effects are only observed at a populational level. In datasets with such low temporal granularity, it is as if all events happened simultaneously on the patient level, and we only know the relationship between those events and the patients. In these cases, modeling the relationships between entities and their resulting semantic vectors may require techniques such as graph vectorization.

On the other extreme, datasets with high temporal granularity such as MIMIC-IV (details presented in next section) present patients existing within their own timelines, as well as on the populational level. Furthermore, MIMIC-IV has different levels of semantic detail, such as sequential disease codes that could be aggregated into broader groups, given by their chapters (e.g., both "prostate cancer" and "breast cancer" could be grouped under the "Neoplasms" disease code chapter).

In both cases, we would first refer to step 2 to identify suitable candidates for the NLP-inspired modeling. In the case of MIMIC-IV, as demonstrated later, the data has a gradual and trending shift over time, with -in-hospital mortality consistently decreasing over the years. Given this pattern and the granularity available on this data, we create sequences of discrete information tokens to elucidate t observed variations for each patient, such as ordinal disease codes or chapters, if a more compact set of possible semantic units is desired.

Finally, we can append "artificial tokens" at the appropriate positions on each patient's sequence, such as a "death" token at the end of the sequences of deceased patients or an "ICU" token when the patient is transferred to the intensive care unit, if applicable. With those sequences, we can obtain semantic vectors representing diseases, patients or outcomes. Following this process on discrete temporal chunks, such as years or months, we obtain distinct outcome tokens for each temporal chunk (e.g., "death 2020" and "death 2021", effectively separating the same outcome

over two years). With that, it is possible to compare the tokens, examining their relative distance and semantic similarity to each other and other tokens. This allows the identification of what has become more or less similar to the analyzed outcome over time.
Next, we will illustrate the application of our methodology to the two aforementioned case studies, with different temporal granularities. The two cases are very different in terms of their temporal granularity, volume, and nature of data, demonstrating the generalization capability of DIS.

DIS Instantiation

We illustrate the application of DIS to analyze temporal shifts by using the MIMIC-IV [18] and the Brazilian COVID-19 Registry datasets [17].
The MIMIC-IV dataset is a comprehensive, open-access, de-identified in-hospital patient record containing sequential diagnosis data, consumption items, vital data records, unstructured electronic health data (text data), and clinical notes for approximately 40,000 ICU patients from 2008 to 2019, designed for research in healthcare and medical science [18]. In this dataset, age is reported in age groups, as a requirement for de-identification.
The Brazilian COVID-19 Registry is a multicenter retrospective cohort of 10,897 patients with confirmed diagnosis of COVID-19, admitted between March 2020 and December 2021 from 41 different Brazilian hospitals. For the purpose of the present analysis variables collected at hospital presentation and at patient discharge were used. The dataset consists of over 200 features, including known comorbidities, patient's age and sex, laboratory tests (such as complete blood count, C-reactive protein and arterial blood gas analysis), vital signs at hospital presentation (i.e., arterial blood pressure, respiratory rate and heart rate) and clinical outcomes [17].
As aforementioned, we chose these two case studies as they illustrate scenarios where the available data has very different temporal granularity characteristics,meaning the patient's timeline can be reconstructed from the data at either a local (individual) or populational level.

Results

The MIMIC-IV dataset comprehended 299,712 patients (median age group 48 [interquartile range {–IQR} 29-65]), while the Brazilian COVID-19 Registry dataset comprehended 10,898 patients (median age 60 [interquartile range - IQR 48-71]).
Figure 4 illustrates how the DIS methodology is instantiated concerning the data's temporal granularity for each scenario. As explained, DIS consists of three steps (Detection, Initial Characterization, and Semantic Characterization). The temporal granularity of the available data affects specifically the last step (Semantic characterization). The figure also shows that several methods can be applied for the Detection step. n. In our experiments, we tested and compared five different methods regarding their capability of accurately identifying temporal drifts in the Detection Step. In the second step, different exploratory techniques that measure the relationship between the dependent P(y) and independent P(x) variables over time can be employed. We exploited multiple alternative techniques such as Feature Importance and Pearson Correlation. Finally, in the last step, our aim is to generate semantic embeddings for outcomes and other healthcare events over time and to derive insights from comparing these embeddings. We tested two different alternatives for producing such insights: (i) using our semantic embedding modeling and (ii) using traditional clustering techniques over the untreated (original data) without the semantic treatment. The goal of using these two techniques is to illustrate insights that can be

obtained with the semantic layer, which would be difficult to obtain otherwise.

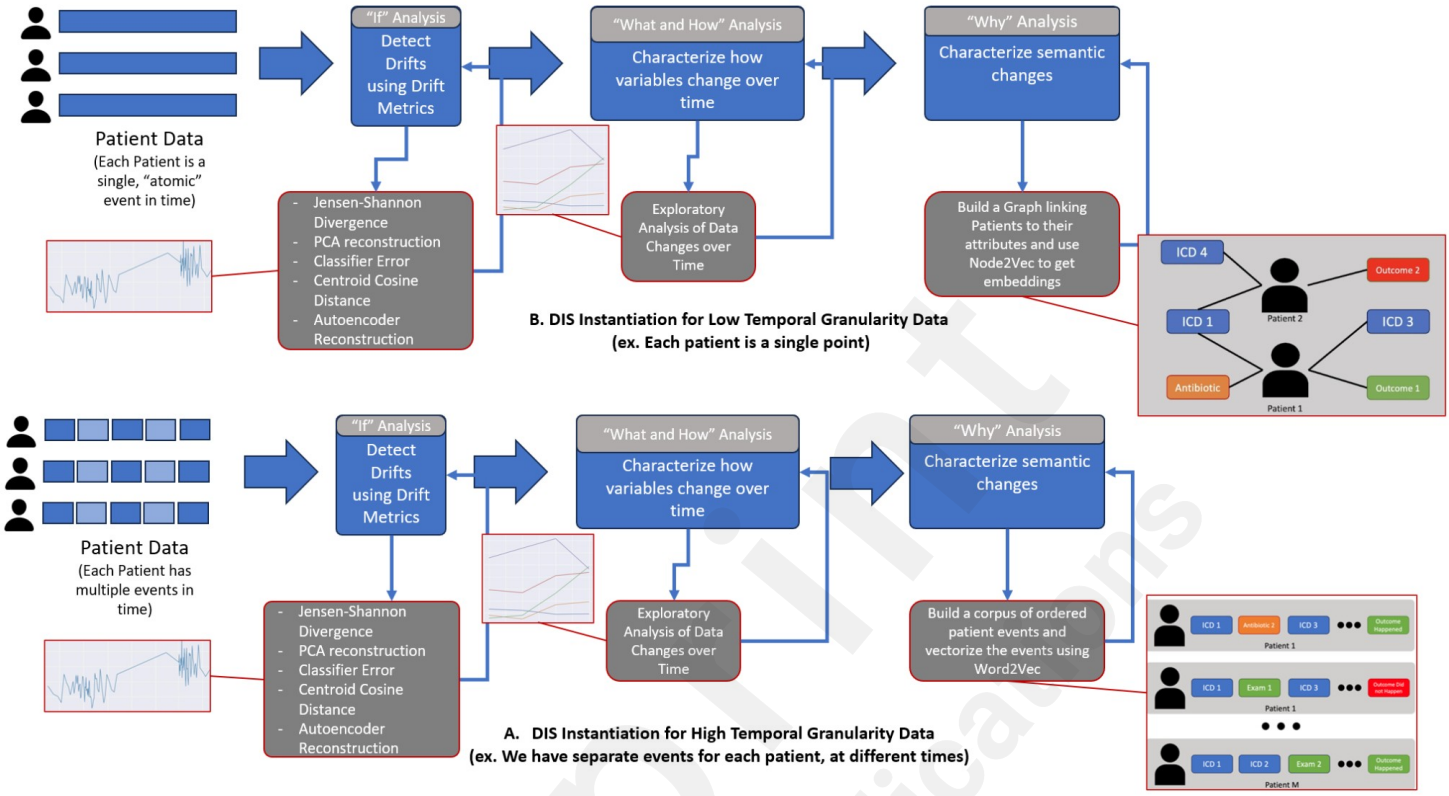


Figure 4. Overview of the Instantiation of the DIS Methodology to two scenarios with different temporal granularities (a) Medical Information Mart for Intensive Care, version IV (MIMIC-IV) DIS Instantiation and (b) Brazilian Covid Registry Instantiation.

DIS Instantiation for MIMIC-IV Dataset

A notable characteristic of this dataset is its high temporal granularity, enabling the tracking of time progression within each individual's hospital stay. High temporal granularity means we know the sequence of healthcare events at the individual level. This facilitates obtaining invaluable insights into the relationships between such events, much like it helps us learn about the semantics of words in NLP. It has been consistently shown that the order of precedence between words and how often they appear with other words is representative of that word's semantics [25]. We claim that the order of precedence and co-occurrence between healthcare events can also contain the "semantics" of those events. A distributed representation built from these relationships could cluster similar healthcare events, such as representing different types of diabetes or hypertension, and their associated complications, in close proximity in the space. Although all dates in the dataset are anonymized for privacy reasons, we can track each individual's sequence of events using a provided masking of dates.. This date-masking is consistent in a manner that allows for time tracking during each patient's hospital stay, and it contains a special attribute that allows us to associate patients with the yearly interval during which they were hospitalized. This yearly interval data allows us to compare how patients in each year group behaved as a group, meaning we can also measure temporal effects at the populational level. The period covered by this dataset ranges from 2008 to 2019.

In other words, the dataset offers temporal granularity at both the population and individual levels. However, breaking this dataset into arbitrary temporal chunks is challenging because the dates are masked. Despite this, the dataset contains a non-masked anchor year group that assigns each

patient to an actual year interval during which they were hospitalized. Figure 5 explains how this variable works. Essentially, a random time delta is fixed for each patient and added to all relevant dates, effectively masking them while preserving the relative time intervals for that patient. Consequently, direct comparison of dates between two different patients is not feasible, except for their "anchor_year_group" variables. For instance, a patient hospitalized in 2015 may have (through the added random time delta) dates that appear later than those of a patient hospitalized in 2020. We can only directly compare dates within the context of each patient. The real year interval during which each patient was hospitalized is preserved in their "anchor_year_group" variable, which we use in all chunking for this dataset henceforth.

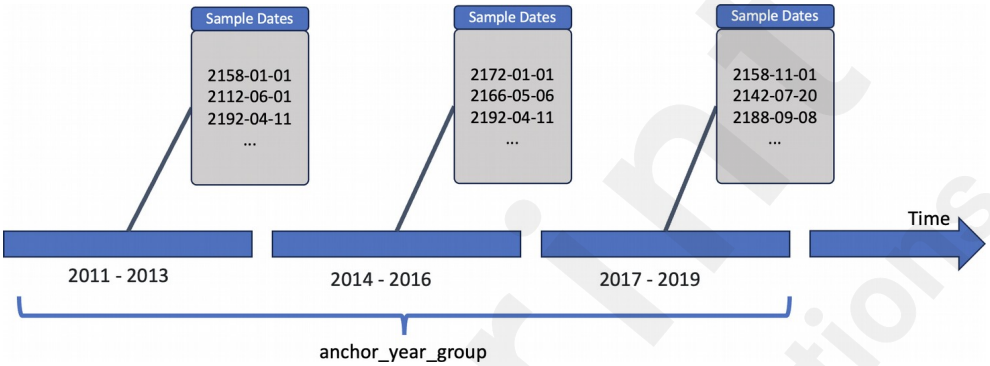


Figure 5. The "anchor_year_group" variable on the Medical Information Mart for Intensive Care, version IV (MIMIC-IV) dataset. Within each "anchor_year_group", the actual dates are masked, making it only possible to have a rough estimate of when the patient was at the hospital.

DIS- Detection Step (MIMIC-IV)

As described, the temporal chunks in MIMIC-IV are given by the "anchor_year_group" variable. We use this variable to separate patients into the four groups provided within the dataset.. We then employed alternative drift detection approaches, namely Jensen-Shannon divergence, Autoencoder reconstruction error, PCA reconstruction error, Centroid distances, and Classifier prediction error on separating time chunks plot for this dataset considering in-hospital ICD diagnosis. The Jensen-Shannon divergence formula is shown in Equation 1, where KL is the KL-divergence [26], P and Q are the two variables being compared.

We started step 1 of DIS with the *drift detection* sub-step. As previously described, the temporal chunks in MIMIC-IV are given by the "anchor_year_group" variable. We used that variable to separate patients into the four groups provided within the dataset. Figure 6 shows the Jensen-Shannon divergence plot for this dataset considering in-hospital ICD diagnosis. The Jensen-Shannon divergence is shown in Equation 2, in which *KL* is the KL-divergence, *P* and *Q* are the two variable distributions being compared, and $M/2^{*}(P + Q)$. To obtain the metric, we calculated the divergence for each feature, comparing a given temporal chunk's distribution to that of the reference chunk, and then average the result's overall attributes. This metric is tracked to evaluate whether the data distributions have changed over time, how fast they changed, and whether the data shift was temporary.

$$JSD(P||Q) = 1/2\,KL(P||M) + 1/2\,KL(Q||M)$$

Equation 2. The Jensen-Shannon divergence. Here, KL is the *Kullback-leibler* divergence, *M* is 1/(*P*

+ Q) and *P* and *Q* are the distributions of the variables we are comparing.

Figure 6 presents the results of our drift detection metrics, applied to the various "anchor_year_groups" in the MIMIC-IV dataset. The Figure depicts the normalized magnitude of the drift signal calculated per "anchor_year_group". The drift signals are normalized in the [0,1] range for visualization, as shown in Equation 2. The results for the Jensen-Shannon divergence, PCA reconstruction error, and centroid cosine distances reveal a trend towards increasing distance between the variable distributions over time, which does not revert to prior levels, suggesting a gradual temporal shift. As seen in Figure 5, this drift occurs gradually over several years, with a more pronounced change between the first two temporal chunks.

In contrast, when examining the autoencoder reconstruction error and classifier error metrics, a peak divergence is observed in the second time chunk (2011-2013), which gradually trends toward the baseline. As models with more parameters, these two drift metrics are sensitive to a combination of the data distribution, novel data points (i.e., rare diseases or diseases not present in the reference time slice), and numerical outliers, in the case of the autoencoder reconstruction error. For example, the disease codes appearing in the second chunk have the smallest intersection with the reference chunk, meaning they have the fewest diseases occurring concurrently in both chunks. This likely explains why the autoencoder reconstruction error and classifier error metrics exhibit their highest peaks in this slice.

In summary, the Jensen-Shannon divergence metric yielded more robust drift signals in our tests. It is important to note that the best metric depends on the most relevant type of drift for the data collection being analyzed. The Jensen-Shannon divergence is robust at detecting distribution changes, just as the classifier error metric. If we are interested in detecting the occurrence of outliers or novel samples not seen before, the reconstruction errors might result in better detection. The choice of metric must be informed by the characteristics of the metrics themselves, as well as the characteristics of the data stream being monitored.

$$NormalizedSignal = (X - min(X)) / (max(X) - min(X))$$

Equation 3. Normalization is used to calculate the normalized magnitude of the drift signal.

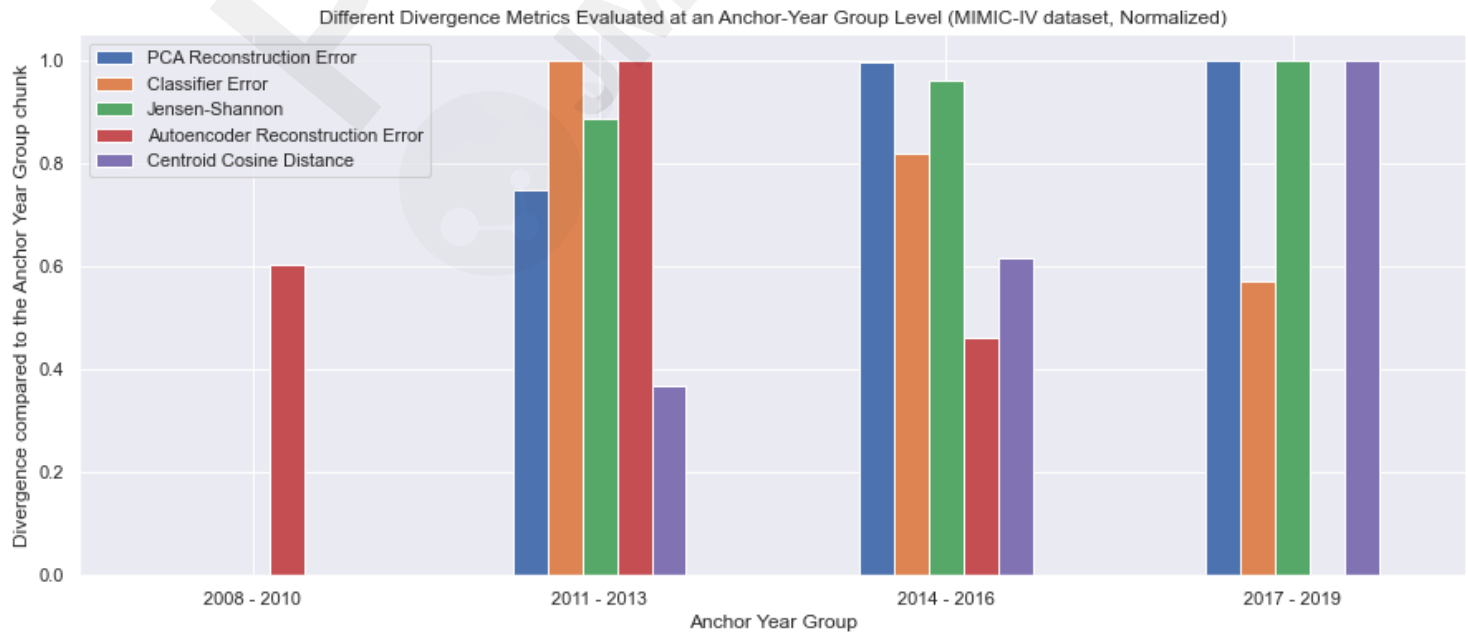


Figure 6. Different drift detection metrics over time on the Medical Information Mart for Intensive Care, version IV (MIMIC-IV) dataset, considering in-hospital ICD diagnosis.

DIS - Initial Characterization Step (MIMIC-IV)

After establishing that a drift has indeed occurred, especially based on the results of the most accurate method – the Jensen-Shannon divergence – we proceeded to *step 2*. In this step, we strived to understand how the independent variables ($P(X)$) affect the outcome, which is our *dependent variable* ($P(y|X)$), and how the relationship between dependent and independent variables change over time. This analysis can be accomplished using correlations and feature importance analysis over time, as well as by characterizing the distribution of different features over time. For instance, in Figure 7, we show how the relative distribution of the "death" outcome has changed over time in this dataset. This means that our data has a consistent trend towards in-hospital mortality reduction over time, which also means a change in the relative distribution of the two possible categories (deceased x not-deceased) for this outcome.

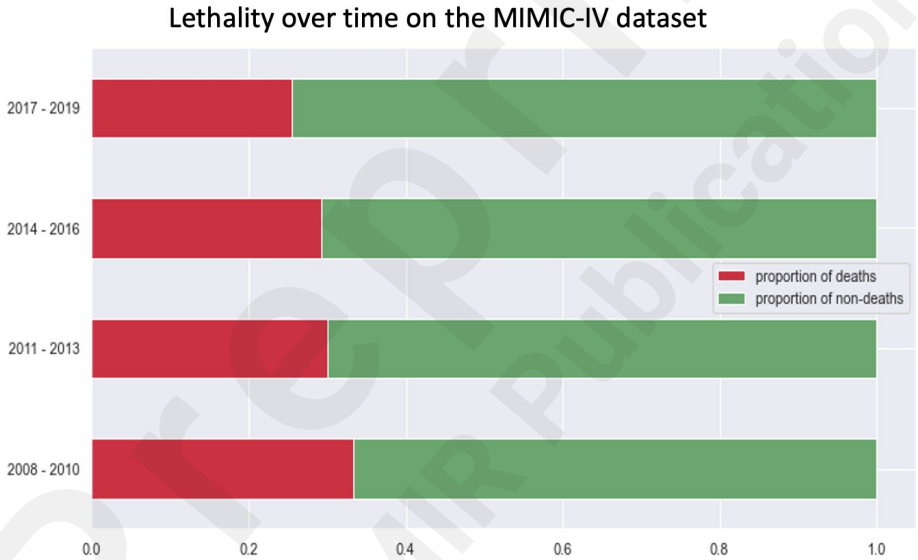


Figure 7. Lethality over time in the Medical Information Mart for Intensive Care, version IV (MIMIC-IV) dataset.

In Figure 8, we analyze the correlations and feature importance variations of the top five most correlated and the top five most predictive ~~international disease codes (~~ICD~~)~~ chapters (according to ICD-10) and the "death" outcome (according to the Feature Importance). For instance, we can see in Figure 8-A some expressive variations, such as how circulatory system diseases seem to grow more correlated with death over time, and, in Figure 8-B, how neoplasms seem to get less predictive of death over time.

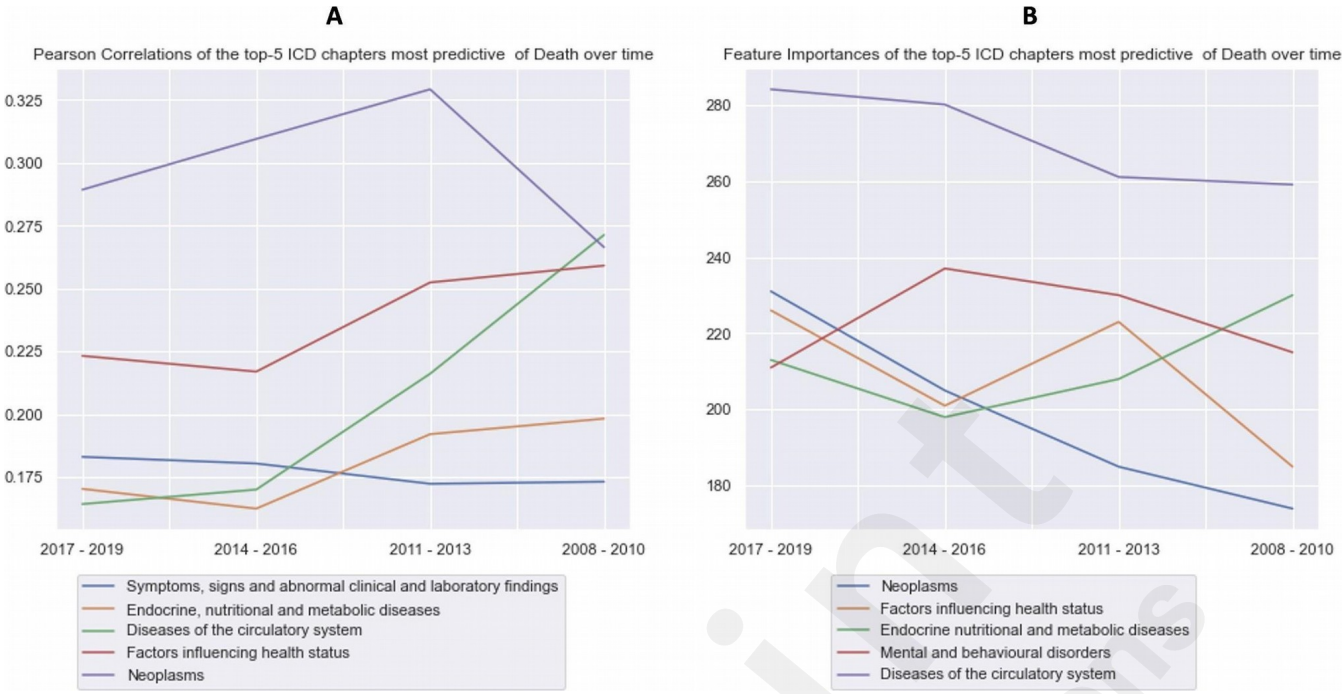


Figure 8. (A) Pearson correlations between the top-5 International Disease Codes (ICD) chapters (according to ICD-10) most correlated with the death outcome over time. (B) Feature importances between the top-5 ICD chapters most predictive of the death outcome over time.

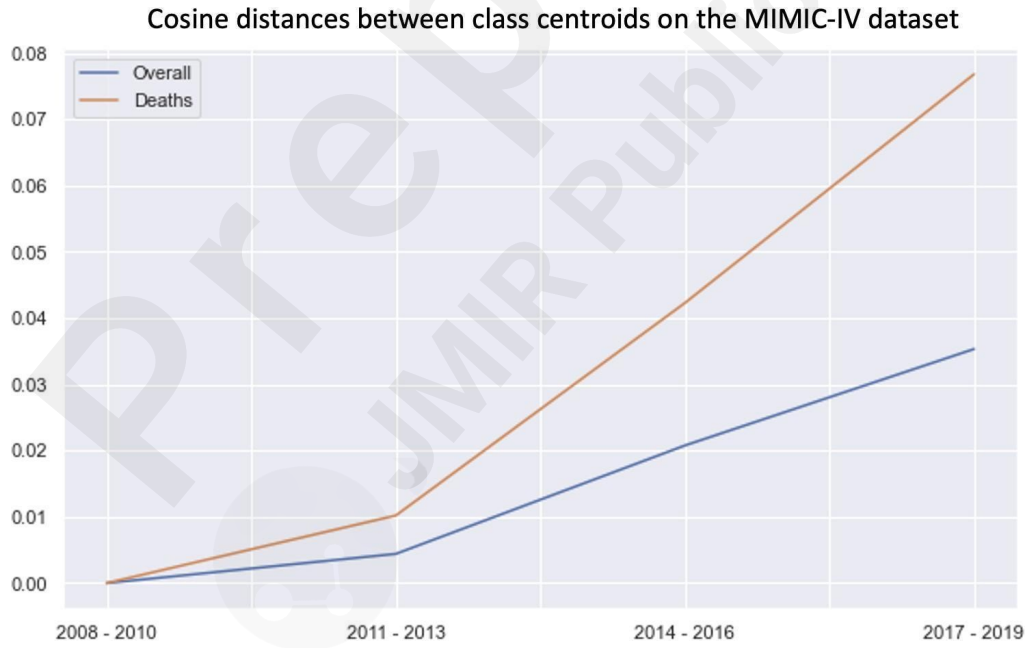


Figure 9. Drift of the arithmetic mean of each outcome class over time, as measured by cosine distances between each class"s means when compared to the mean of the first "anchor_year_group" on the Medical Information Mart for Intensive Care, version IV (MIMIC-IV) dataset.

In Figure 9, we showed how the different outcome groups behave over time concerning a given baseline, in particular the patterns of independent variables given the outcome categories $P(y|X)$ observed in the first temporal chunk. To obtain this result, we computed the arithmetic mean of

each class"s features on each "anchor_year_group" and calculated the cosine distance between these means over time, taking the first chunk as a reference to compare all other chunks against it. For this particular Figure, we represent each patient as a "corpus" containing all their healthcare events (such as diseases and medications used during the hospital stay), then encode each feature as a one-hot sparse matrix (each event can have values "0", in case it did not happen for a particular patient, or "1" if it did), and subsequently average these features. Note that this representation treats each patient as a "bag of healthcare events", disregarding the order of precedence between those events, unlike what we did in our semantic characterization step. In the specific case, we show how the "Death" outcome exhibits greater temporal drifts over the available time chunks in both datasets compared to the overall hospitalized patient population.

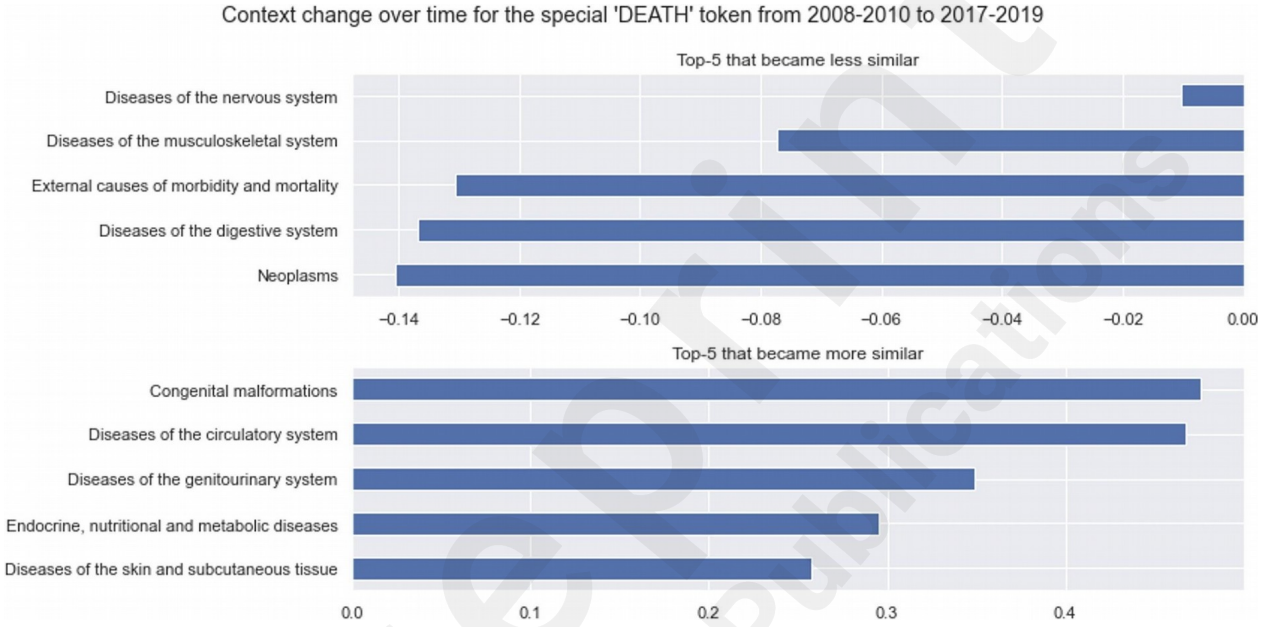


Figure 10. Evaluation of the drivers of lethality data drift on Medical Information Mart for Intensive Care, version IV (MIMIC-IV) dataset.

DIS - Semantic Characterization (MIMIC-IV)

In Figure 10, we show the top-five ICD-10 chapters that have become more and less similar to the "death" outcome over time. Notably, certain diseases , such as neoplasms, have become less similar, while others, such as malformations and circulatory system diseases, have become more similar. That is consistent with findings in step 2, and over the next paragraphs, we describe the procedure to obtain this similarity score. We explain the token level vectorization process for both - dependent and independent variables in Figure 11. First, we compile a temporally ordered list of patient data, consisting of discrete data points such as items consumed during hospital stay (e.g., antibiotics, anti-inflammatories, etc.), disease codes (using ICD), and procedures. At the end of each patient’s sequence, we append the outcome category for that patient. To classify the outcome, we divide binary outcomes into distinct tokens, such as "Deceased" and "Not-deceased", and use the corresponding token to generate our training corpus. Continuous outcomes and dependent variables can be binarized using a simple histogram binarization scheme, as demonstrated in the next analysis. Following the corpus generation, we use it to train token embeddings with Word2Vec [25]. This method produces embedding vectors for both dependent and independent variables,

allowing semantic comparisons between different entities, such as the "death" outcome and different disease codes. We create one outcome token for each outcome category and temporal chunk in our dataset. This allows us to evaluate how an outcome such as "Death" may have drifted closer to or farther from certain diseases or procedures over time.

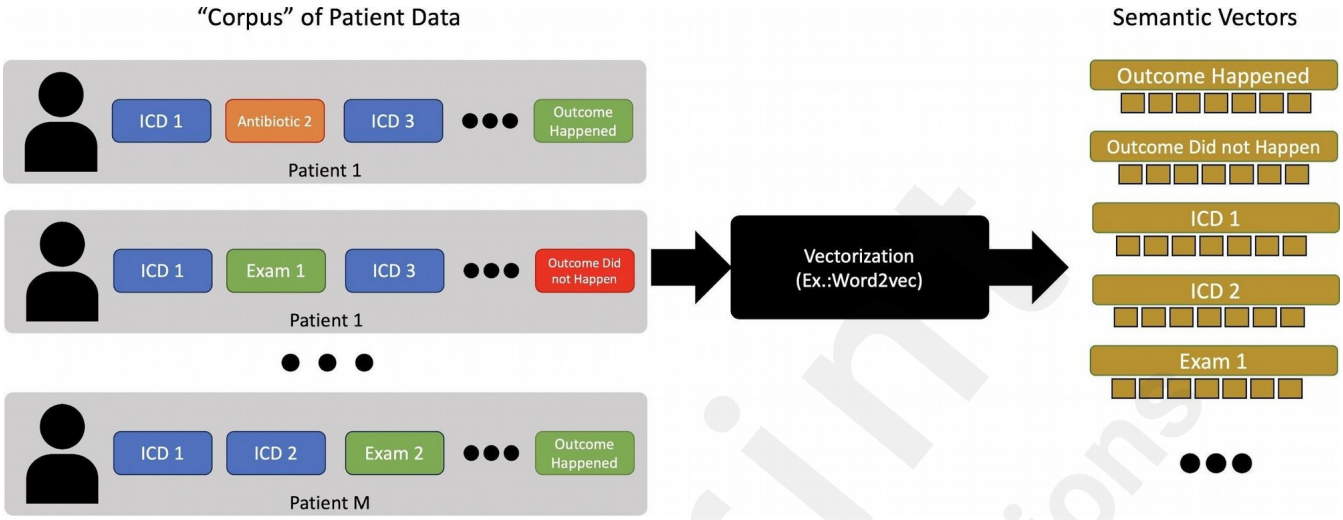


Figure 11. How to generate semantic vectors. We start by generating a corpus of temporally ordered patient discrete data points. Then, we vectorize the tokens of this corpus using Word2Vec to obtain semantic vectors for dependent and independent variables.

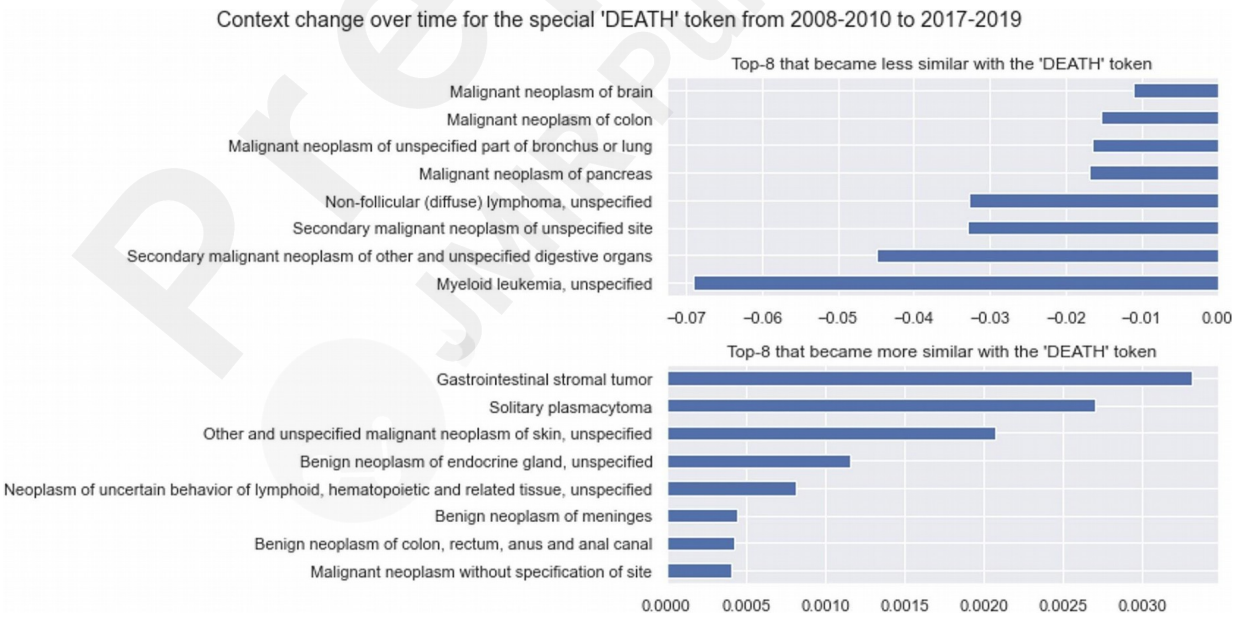


Figure 12. Evaluation of the drivers of lethality data drift on the Medical Information Mart for Intensive Care, version IV (MIMIC-IV) dataset.

In Figure 12, we show the top five conditions that became more similar to the "death" token and the top five that became less similar when comparing the first and the last time chunks. Since every entity is a "token", we can evaluate similarities between diseases, disease chapters, as well as

between patients and diseases they have not yet been diagnosed with, and between outcomes and diseases (Figure 12). In particular, in Figure 13, we demonstrate changes in similarity for the "dysphagia following stroke" ICD code within the MIMIC-IV dataset [18]. Our analysis reveals a rise in the simultaneous appearance of ICD codes related to obesity between the periods of 2011-2013 and 2017-2019. This trend aligns with broader observations indicating an uptick in obesity rates across the United States. Importantly, it is essential to recognize that this method does not permit us to establish causal relationships; rather, it emphasizes changes in correlation and co-occurrence.

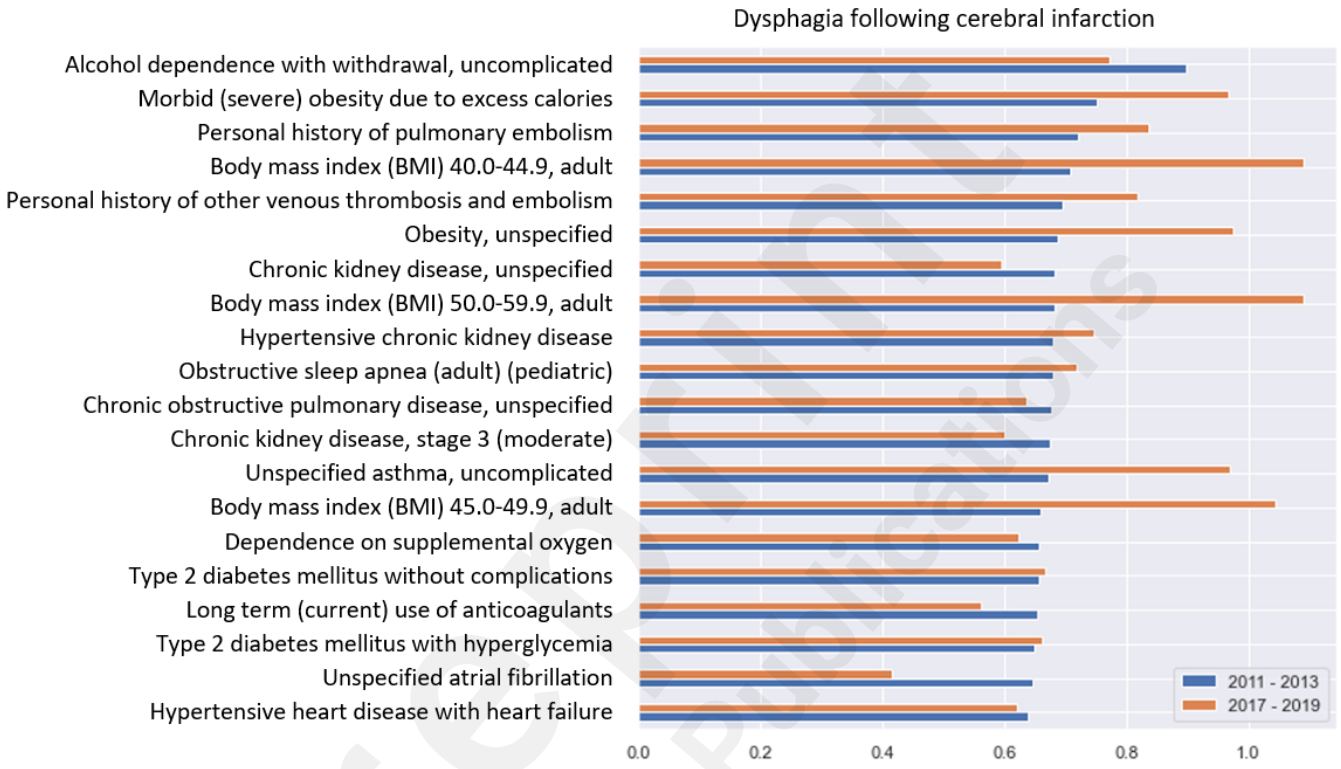


Figure 13. Changes in co-occurrence for the "dysphagia following stroke" International Disease Code (ICD) on the Medical Information Mart for Intensive Care, version IV (MIMIC-IV) dataset.

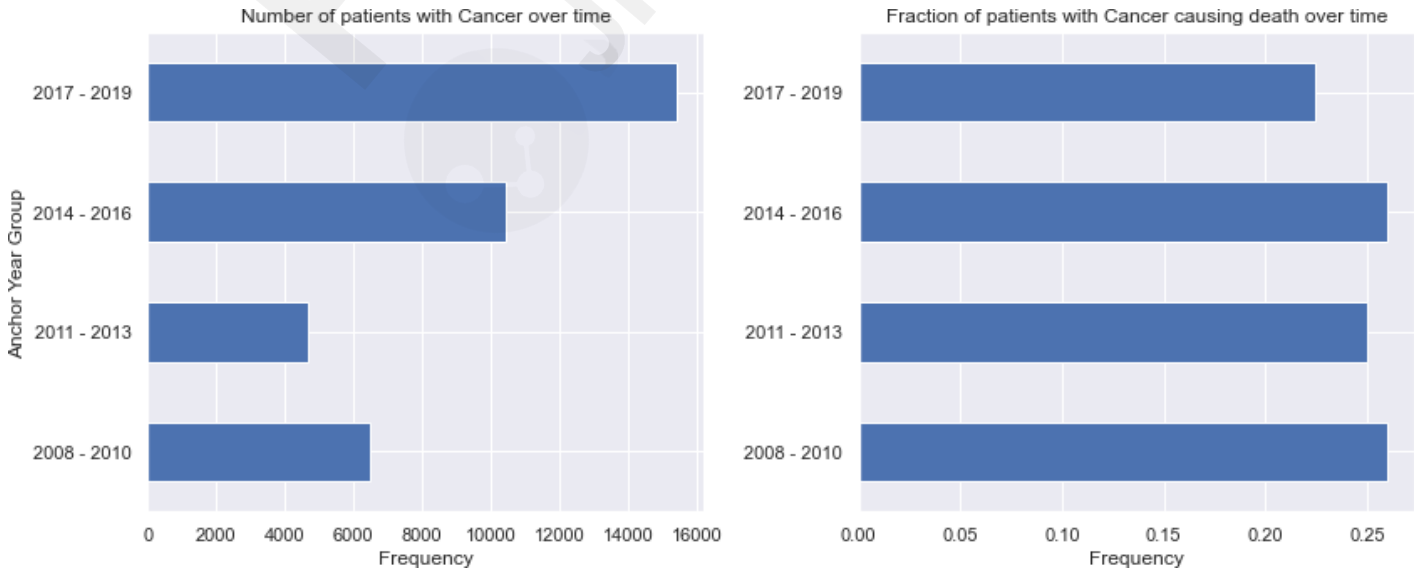


Figure 14. Validation of the data drift on cancer patients. On the left, we show the increase in

the absolute number of cancer patients, while on the right we show the overall lethality reduction for this disease group.

We can also conduct the *step 3* analysis at different levels of granularity to gain a deeper understanding of the observed changes. We know from step *2* that mortality has been decreasing and bears some relationship with particular disease groups. If we perform *step 3* at the disease code level, as in Figure 13, we can identify which chapters had considerable shifts in their similarity with the "death" outcome, either increasing or decreasing. For instance, notice how the findings confirm what we observed in Figure 7, where "cancer" shows a decreasing similarity with the outcome, while the variable "circulatory diseases" exhibits an increasing similarity with the outcome. This observation can be further supported by the results shown in Figure 14, where we can see an absolute increase in the number of cancer patients over time, associated with a relative decrease in in-hospital cancer-related deaths between 2008 and 2019.

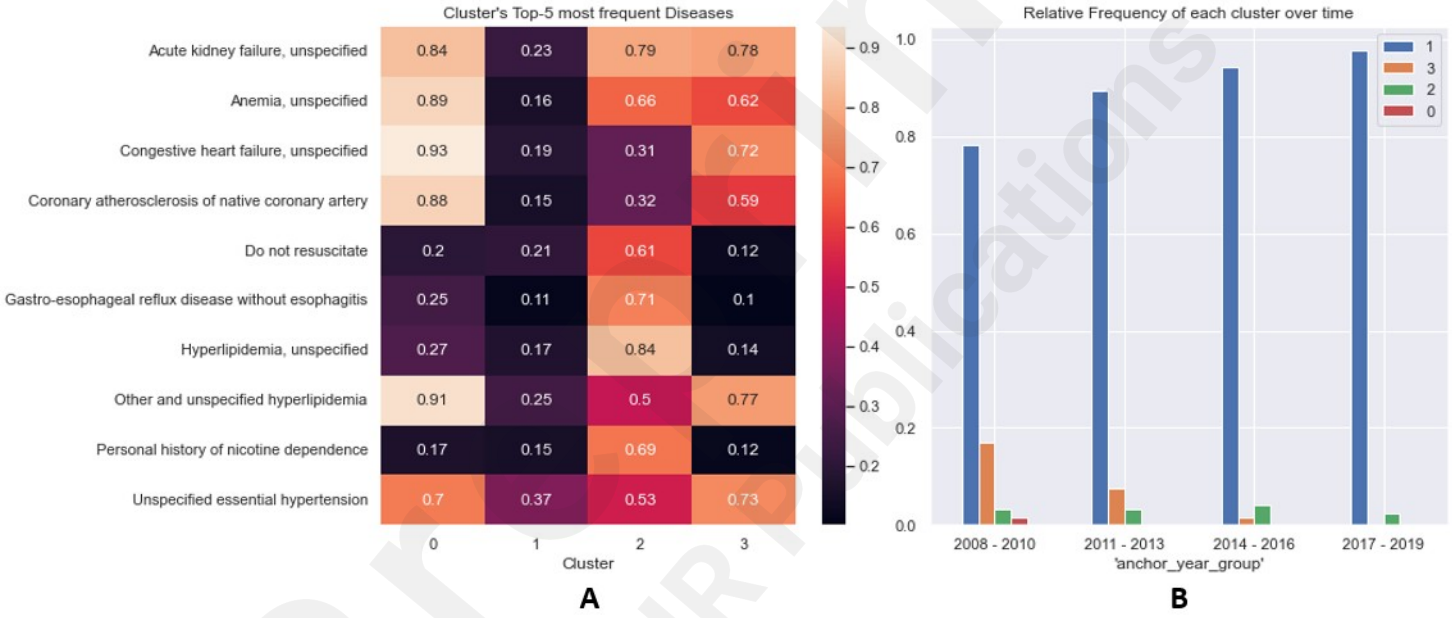


Figure 15. Cluster analysis of the Medical Information Mart for Intensive Care, version IV (MIMIC dataset). A: Top-5 highest-valued features per cluster. B: Relative frequency of each cluster over time.

To further illustrate how the proposed DIS Semantic Analysis based on embedding distances among entities of interest can help in better comprehending the reasons for the drifts, we contrast the previous analyses of our third step with a traditional clustering analysis for the MIMIC-IV dataset. This analysis uses a syntactically-oriented Term Frequency-Inverse Document Frequency (TF-IDF) [27] representation for the entities, built from the same corpus of clinical entities. In a TF-IDF representation, each dimension corresponds to a unique term (word) in the document corpus. The value in each dimension reflects the importance of that term in a specific document, calculated by multiplying the term's frequency in the document (TF) by the inverse frequency of the term across all documents (IDF). In our case, each "document" is a patient, and each "word" is a healthcare event, such as the identification of a novel disease. We apply a spectral clustering [28] procedure over the TF-IDF representation of the entities to create the clusters. The results are shown in Figure 15. To obtain the four clusters displayed in Figure 15, we used silhouette analysis ranging from 2 to

15 clusters.
Figure 15-A shows the top five most frequent diseases for each of the four clusters (y-axis). On the x-axis, we present the index of each cluster. In Figure 15-B, we show how the relative frequency of each cluster changed over each "anchor_year_group". A few points stand out from the clustering analysis in Figure 15. As it can be observed, the custer analysis using syntactically-oriented vectors makes it harder to interpret the drivers of a data drift when compared to DIS. For instance, some semantically similar diseases, such as "other and unspecified hyperlipidemia" and "hyperlipidemia, unspecified" may have very distinct profiles on different clusters, such as in clusters 0 and 2, each having a high concentration of patients with either one of these diseases. The main problem of this particular cluster analysis based on syntactically-oriented representation is the separation of semantically similar entities into distinct clusters. In DIS, similar entities will be represented similarly and thus be analyzed in conjunction.

DIS Instantiation for the Brazilian COVID-19 Registry dataset

The median age was 60 years old [interquartile range 48-71] and 46% were women (5,012 patients). In this dataset, 21% of registered patients died, yielding an unbalanced classification problem when predicting future deaths. The dataset has low temporal granularity, with only one data point per patient, which precludes time tracking during hospital stays. Consequently, we can only measure time at the populational level. In other words, unlike the previous case study, there is a single "snapshot" for each patient, with no temporal evolution at the individual level.

DIS- Detection Step (Brazilian COVID-19 Registry dataset)

As in the previous case study scenario, we evaluate the same five alternative techniques, namely: the PCA Reconstruction Error, Autoencoder Reconstruction Error, Classifier Error (In Separating Past vs. Future), and the Jensen-Shannon Divergence. All these metrics measure the drift compared to a reference temporal slice and do not require setting a specific outcome or using labeled data.
The outcomes of this procedure are illustrated in Figure 16, where the divergence sharply increases starting in the final quarter of 2020, based on the Jensen-Shannon divergence metric. Numerically, a drift is indicated in this interval as the divergence levels surpass a user-defined threshold, such as a fixed threshold of two standard deviations or a threshold informed by domain expertise. As depicted in the figure, the PCA reconstruction error, autoencoder reconstruction error, and centroid cosine distances indicate positive drift signals in the quarter starting in April 2020. During that semester, the Brazilian COVID-19 Registry dataset dataset exhibited a small number of numeric outliers, which were identified by these methods. Conversely, the Jensen-Shannon method signals a data drift in the quarter starting in October 2020, which aligns with the "official" start of the second wave in Brazil in November 2020. Meanwhile, the classifier-error method indicates a drift in July 2020, which falls between the identification of numeric outliers and the actual distribution change from the first to the second wave. Both the Jensen-Shannon method and the classifier-error method signal drift closer to known actual changes, while the other, more reconstruction-based methods are more sensitive to numeric shifts which are not necessarily associated with changes in the underlying distributions.

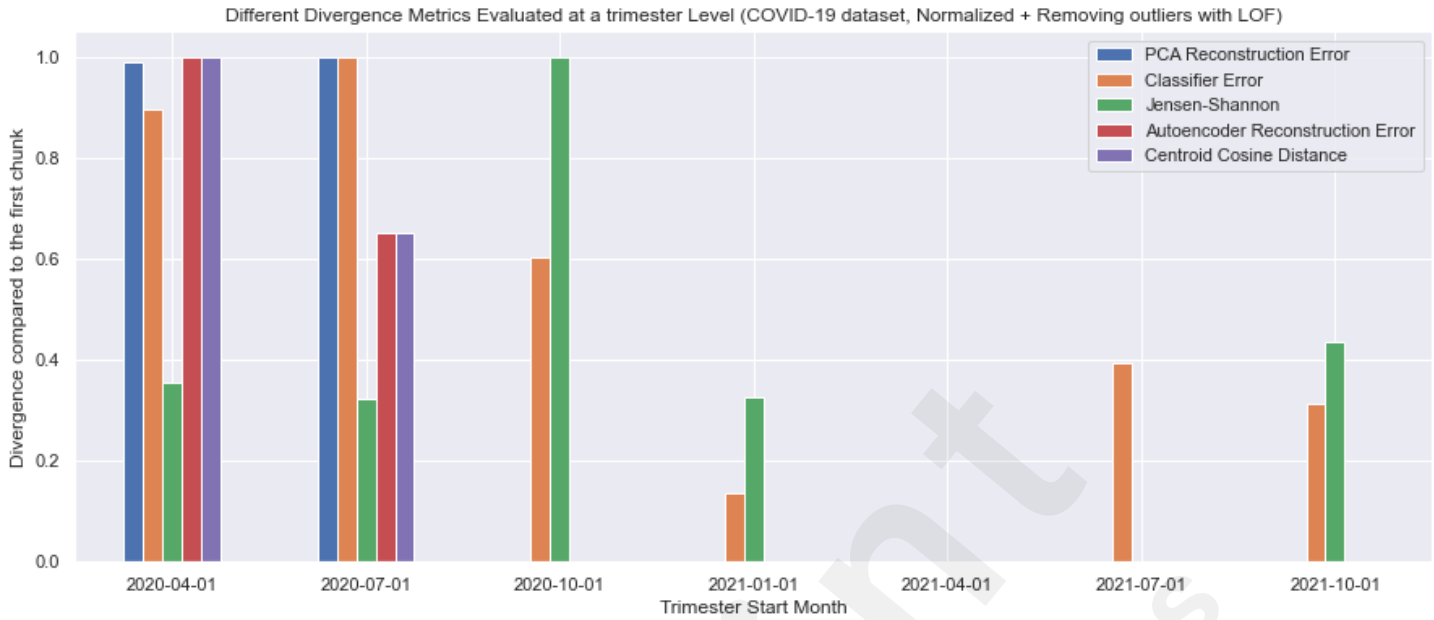


Figure 16. Different drift detection metrics over time on the Brazilian COVID-19 Registry dataset.

DIS- Initial Characterization Step (COVID-19)

Once a drift has been detected, we proceed with the second DIS step – Initial characterization. This step aims to understand the main drivers ("what") of changes during the considered period and "how" they affect the underlying outcomes. At a high level, we begin by characterizing the changes in the outcome (the independent variable) over time. In Figure 16, we illustrate this by evaluating the variation in COVID-19-related mortality in our dataset. This example displays a trend towards reduction in the death outcome over time. At the Initial characterization step, it is expedient to examining the distribution of the outcome of interest (e.g., death, ICU admissions, etc.) as well as those of the most predictive independent variables (e.g., those with the highest correlation with the desired outcomes or higher feature importance in a tree-based classifier).

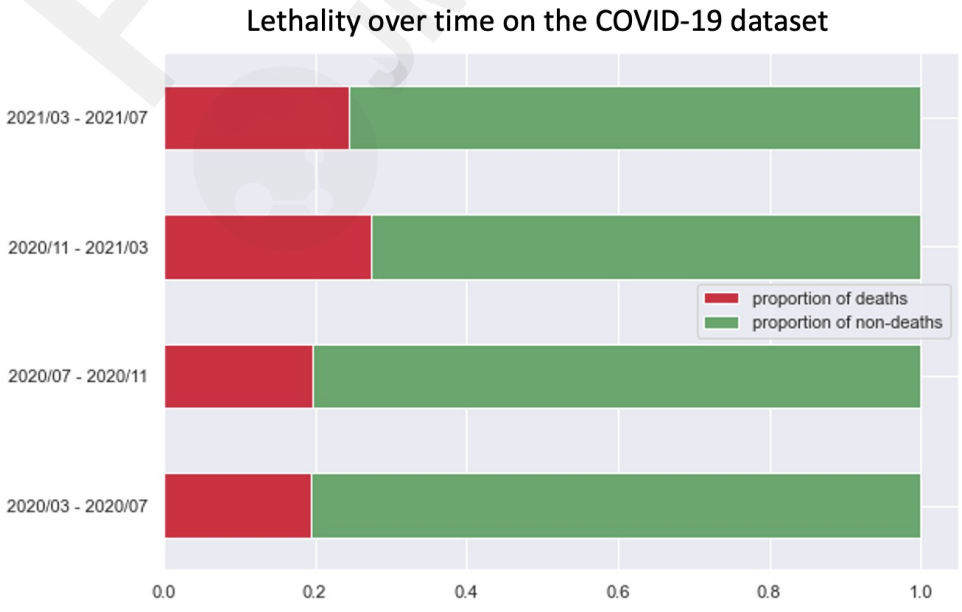


Figure 17. Lethality over time in the Brazilian COVID-19 Registry dataset.

To guide the next steps, it is helpful to check how much each outcome category's properties (such as the mean age of the deceased patient population or the prevalence of hypertension) have changed over time. In particular, focusing on which outcome(s) have changed the most helps to target specific subsets of our data that could better explain the observed phenomena. We show an example in Figure 18, where we analyze such variations in the Brazilian COVID-19 Registry dataset. To build the graphs in this figure, we split our dataset into time chunks. For each chunk, we separated all patients into classes according to their outcome (e.g., dividing the population into deceased and non-deceased, and then representing the chunk by averaging all of the patient's features in each category). For each subgroup of patients within the same time chunk and sharing the same outcome, we compute the centroid of that class (the arithmetic mean of all attributes). We then take the first chunk as a reference and compare each class's chunk arithmetic mean to the reference using a cosine distance. Figure 18 shows how much the deceased patients' characteristics changed more than the overall population during the same period.

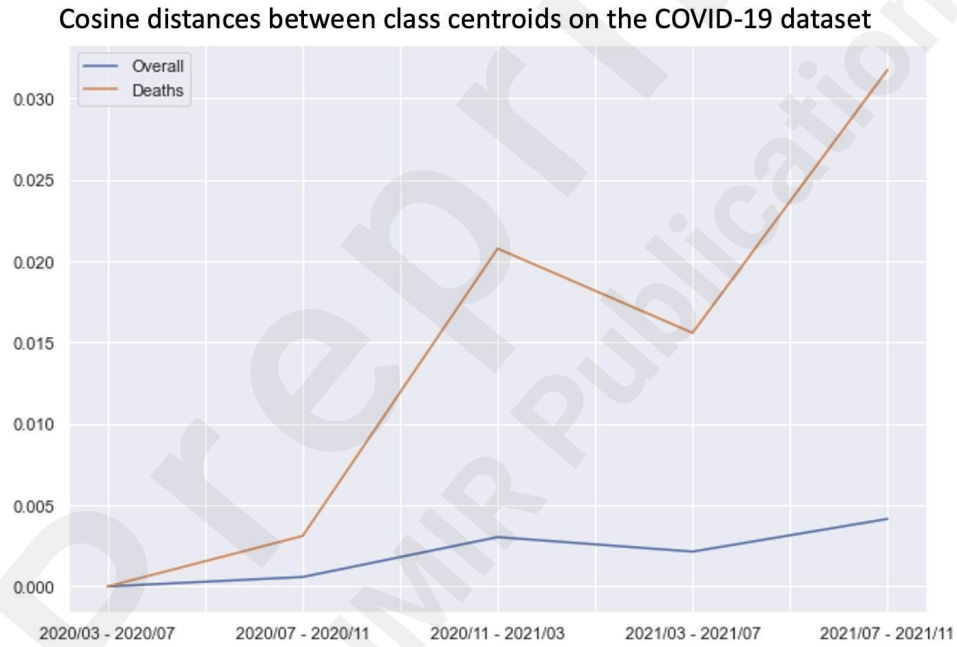


Figure 18. Drift of the arithmetic means of the dying patients versus the overall population over time, as measured by cosine distances between each class's means on each time chunk over time, on the Brazilian COVID-19 Registry dataset.

A better comprehension of the drift drivers during the COVID-19 pandemic emerges from Figure 19. In Figure 19-A, we observed how the overall best predictors of death changed over time through Pearson correlation analysis conducted each trimester on the dataset. At the beginning of the pandemic, age was the single best predictor of death, in trimesters 0 and 2. As the vaccination campaign started, the elderly were prioritized and received immunization first. This led to a progressive deterioration of the predictive value of age, as well as an overall decrease in mortality (Figure 17) culminating in the latest trimester where age was the worst predictor among the top five variables. In Figure 19-B, we can see the median age of the deceased patient population over time.

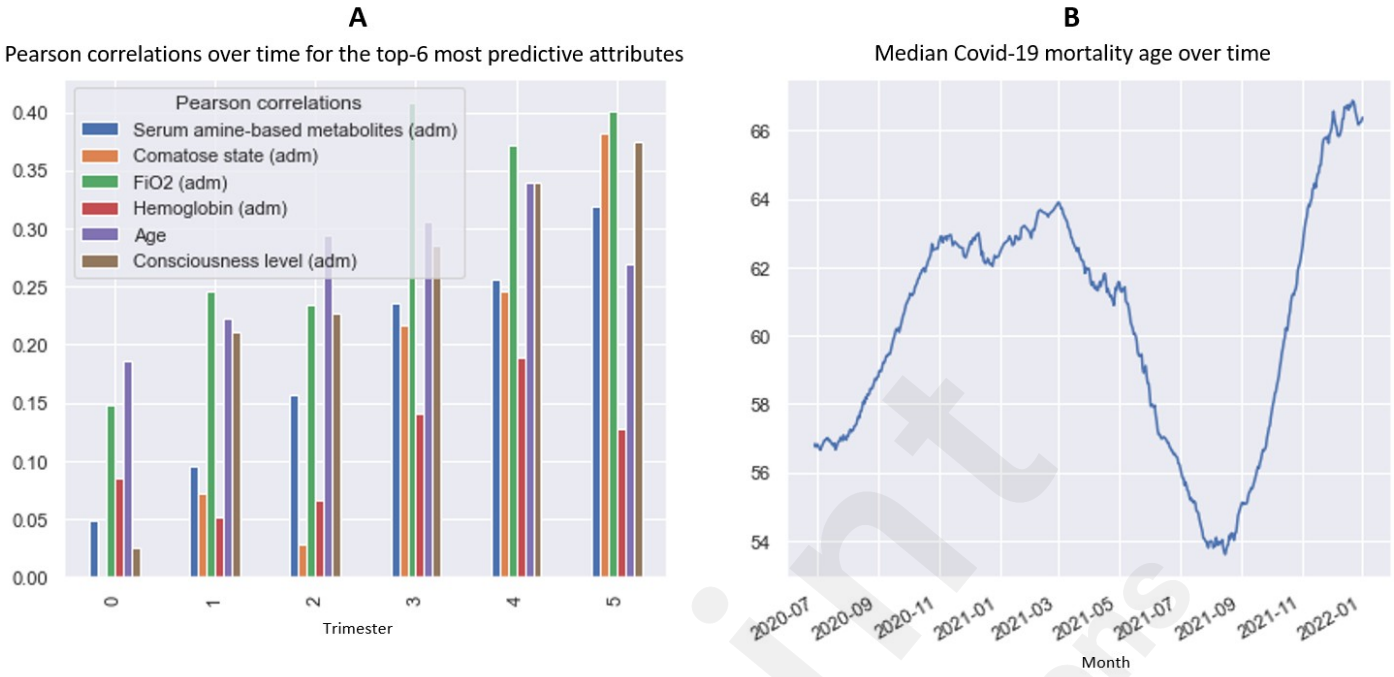


Figure 19. A: Pearson correlations over time for the overall top-6 most predictive variables on the Brazilian COVID-19 Registry dataset. B: Median age of the COVID-19's hospitalized dying patients.

In summary, the second step revealed that the COVID-19 data showed a progressive decrease in patient mortality (Figure 17), with a more pronounced change in the group of deceased patients (Figure 18). It was also possible to notice that the overall characteristics of these dying patients changed abruptly (Figure 19). From the remaining characterization steps in Figure 19, we can see that age lost its predictive capacity (Figure 19-A) over time, while clinical features such as the patients' fraction of inspired oxygen (FiO_2) became better predictors of death. Concurrently, there was a reduction in the median age of dying patients (Figure 19-B).

DIS- Semantic Characterization (COVID-19)

Following the conclusions of the previous step, we move further into the semantic characterization step. As The Brazilian COVID-19 Registry data has low temporal granularity and most of its features are continuous, what requires data categorization to enable the use of NLP techniques to treat words and other semantic units.

Subsequently, due to the low granularity at the individual level, we need to model relationships between these now discrete entities. In more detail, we assume that the temporal precedence between events imposes a relationship between them and that this relationship can be learned and embedded into a distributional representation. The issue with low temporal granularity data is that we do not know this order above precedence and, hence, are unable to model it directly. Due to that, we model all health events (from the perspective of a single individual) as if they happened simultaneously. Therefore, in this setting, we only model the passing of time from the perspective of the population and not the individual. That means we only know, for instance, that a given patient has events (such as new diseases or use of medications, etc.) 1, 2, and 3, but we do not know the order of precedence between those attributes, something that was explicit in the MIMIC-IV data, due to the high temporal granularity. We begin by discretizing the continuous features with a

Histogram Discretizer, that essentially breaks the data intervals into "equal width segments" and then assigns a "token "(i.e., a string or integer value that is unique to patients having that attribute in that specific range of values.

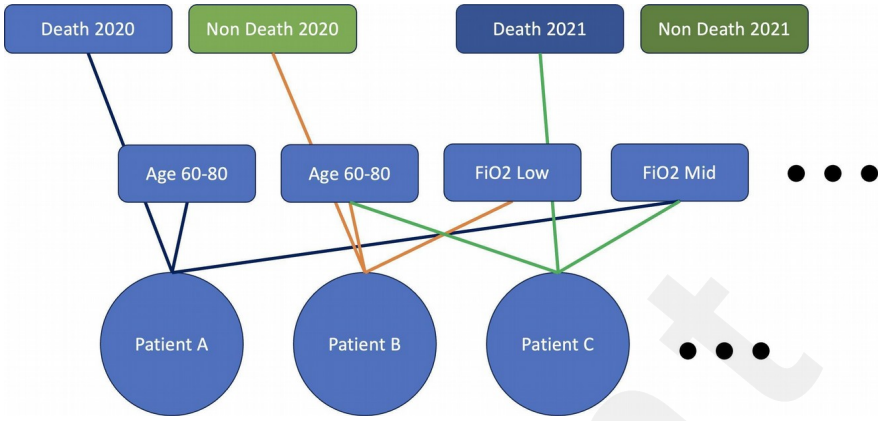


Figure 20. Example of how to build the patient graph with tokenized dependent variables and temporal outcome tokens.

After that, we build a graph with patients, discretized continuous attributes, discrete attributes, and outcomes like the one in Figure 20. To build this graph, we connect each patient to their attribute tokens and outcomes while creating one outcome token for each time chunk under analysis. Finally, we embed the graph using a node embedding algorithm such as Node2Vec [29]. We contrast this procedure with the one adopted to characterize the MIMIC-IV dataset in Figure 21. As discussed before, in MIMIC-IV, the temporal order is defined at the individual level, with entity relationships determined by the timeline. In contrast, the Brazilian COVID-19 Registry dataset dataset presents events as occurring "simultaneously" at the patient level, limiting our understanding of relationships to events and patients. In this case, to derive semantic vectors representing entity relationships, we approach it as a graph vectorization problem.

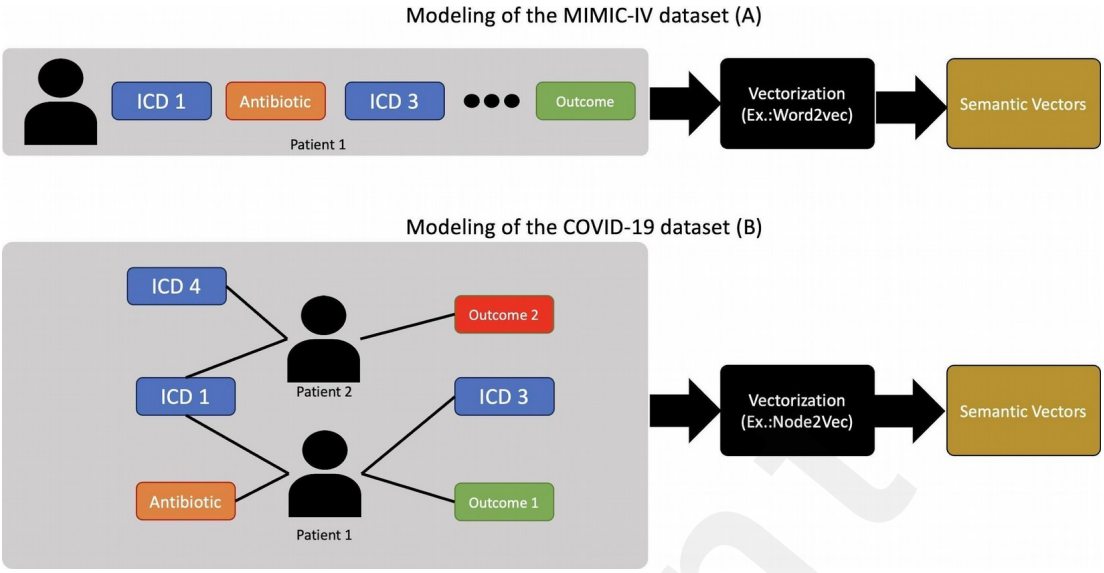


Figure 21. (A) Modeling of the Medical Information Mart for Intensive Care, version IV (MIMIC-IV) dataset as an ordered sequence of patient tokens. (B) Modeling of the Brazilian COVID-19 Registry dataset as a graph connecting multiple patients through their common token.

To analyze the resulting model, we compare the outcome embedding vectors to evaluate their similarity with each other and with other patient attributes. We show the results of this procedure in Figure 22. From that, it is evident that the 2021 death outcome token increased in similarity with lower age groups, such as ages 18-39 and 40-61, while decreasing in similarity with older age groups, such as ages 62-83 and 84-105. This observation further validates the previous findings and also introduces new elements not captured in earlier steps. We can also see an increase in similarity with lower admission heart rates and lower admission serum sodium values, as well as lower FiO_2 (fraction of inspired oxygen) at admission, showing a shift in disease severity markers over this time frame.

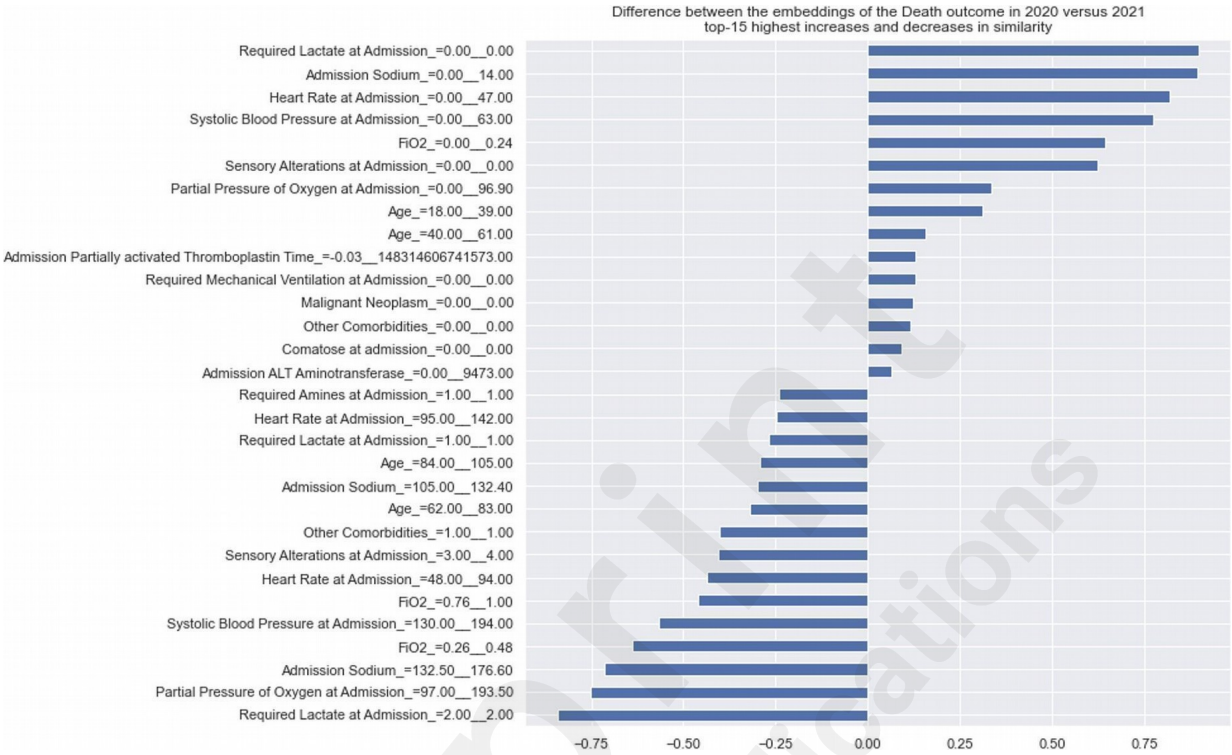


Figure 22. Top-15 largest increases and decreases in similarity between the "Death" tokens for 2021 and 202 on the Brazilian COVID-19 Registry dataset.

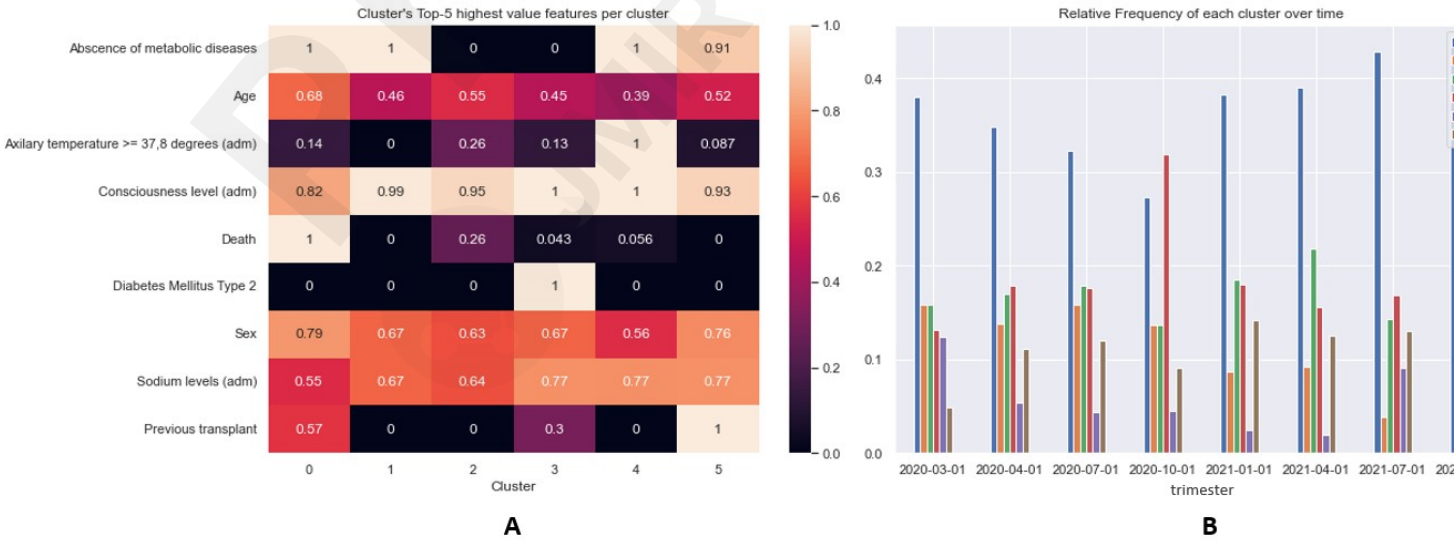


Figure 23. Cluster analysis of the Brazilian Brazilian COVID-19 Registry dataset. A: Top-5 highes valued features per cluster. B: Relative frequency of each cluster over time.

As aforementioned, for comparative purposes, to emphasize the semantic capabilities of the propose DIS procedures, we compare our semantic step results with those obtained with traditional clustering analys for the Brazilian COVID-19 Registry dataset in Figure 23. In this analysis, entities are represented by syntactically-oriented TF-IDF representation. In Figure 23-A, we show the top five highest value features fo each of the six clusters selected using the silhouette analysis, as was done for the MIMIC-IV case. In Figure 23 B, we show how the relative frequency of each cluster changed over time in each trimester.

Similar to the MIMIC-IV case, this clustering analysis of the COVID-19 data is not as straightforward t interpret as the DIS analysis when searching for the drivers of data drift. For example, we identify a cluster c "transplant" patients and another for "diabetes mellitus type 2", but driving conclusions for the reason why these particular clusters were selected and the reasons for the drifts are not easily derived by straightforward analysis of these syntactically-oriented clusters.

Discussion

Comparison with prior work

Multiple studies have analyzed variations observed over time in class distributions, model effectiveness, an their overall impacts. Studies such as Salles et al., (2016) and Mouro et al., (2008), for instance, perform detailed characterization of such effects in textual data sets of documents organized into topics. Healthcar data, however, is quite different from simple text data [30,31]. To begin with, this type of data is multimoda having tabular, sequential data in the form of vital data measurements, disease code diagnosis, and item consumed during a hospital stay, as well as, common text, images, wave forms, and sometimes even soun waves. Furthermore, it may experience sudden and specific drifts driven by new medications, vaccine surgeries, and public policies [9]. For example, an effective vaccine may cause the eradication of a disease resulting in a subsequent data drift I [32]. While most studies on healthcare data focus on either dri detection or drift adaptation [32,33], our work contrasts by focusing on drift detection, monitoring, an characterization. We advance the existing literature by leveraging these three steps to pursue explanation for healthcare data drifts.

Concerning terminology and problem-setting definitions, the authors of [34] define data changes as bein related to the distribution of the independent variables P(X), dependent variables P(y), or the condition probability of dependent variables given independent variables P(y|X). Works unified and consolidated som of the underlying terminologies [35,36]. As defined by Gama et al., (2019), data and concept drifts can b categorized based on how they behave over time, being: (i) sudden (i.e., one event permanently changes th "meaning" of a concept), (ii) incremental (i.e., one event incrementally generates gradual changes to th "meaning" of a concept); (iii) gradual (i.e., the concepts interchange gradually until the complete shift occurs or (iv) reoccurring (i.e., a transient concept drift) [36].

Approaches to detect and learn in the presence of concept drifts do exist. However, in most contexts, naive monitoring data drifts may be expensive, as it often requires data labeling. As an alternative approach, Haqu et al., (2017) utilize an ensemble of classifiers to report their prediction confidences and monitor changes i their confidence distribution to detect when concept drift occurred [37]. In the datasets used in this article however, deaths are readily available labeled data, which means that our main issue is related to learning the presence of data drift.

A common approach to drift detection is monitoring model outputs, as in Sahiner et al., (2023) [38]. Thes "model monitoring" approaches are not always possible or desirable, as argued in [39]. Tiwari and Agarwa for instance, argue that labels are a resource that is not always available and suggest exploring other option such as detecting drifts by monitoring changes in the underlying data distributions [39]. Following this ide we propose a *drift monitoring* procedure that is independent of labels and focuses on distribution change over time. Additionally, Tiwari and Agarwal (2022) provides a comprehensive review of useful healthcare dat type classification and data drift management strategies in data streaming scenarios [39]. The author propos a categorization of healthcare data into:

- *Clinical data*, such as the records in MIMIC-IV [18] and the Brazilian COVID-19 Registry [17]. This type data is desirable if the goal is to describe data drifts relating them to the impact of specific interventions, suc as the introduction of a new drug or therapy;
- *Self-administered data*, obtained from questionnaires, usually investigates lifestyle variables, such a smoking or alcohol consumption habits;
- *Biological data*, usually obtained from performing measurements on biological samples such as bloo and urine. This is often the result of a laboratory study;
- *Molecular data* are the kind of data encoded in protein databases such as UniProt [40], genom databases, or even drug-to-molecule interaction databases;
- *Exposure data* encode patients' exposure to given events, drugs, or intervention;
- *Modeling data*: data generated from models, including, for instance, estimated risks given the patient exposure.

In addition to the categorization mentioned above, Tiwari and Agarwal (2022) discusses the use of samplin in diverse forms to handle data streams and drifts. In healthcare data, it is common to encounter massiv datasets encompassing multiple years and thousands of patients [39]. For such cases, sampling may be viable option. Given the size and nature of our data sources, we opt to work with the full available datas instead of using sampling. The decision to use sampling should be evaluated depending on the type machine learning algorithm employed, the available computing capabilities, and the dataset size.

Drift detection has multiple beneficial impacts in healthcare. Once detected and treated, it can be used t help maintain and enhance model effectiveness. Additionally, it can be useful to detect whether a ne treatment is changing the outcomes of a disease in a meaningful way, or even understand population trends to derive health policies. A recent example is the COVID-19 pandemic. This topic was explored i [41,42], which showed differences in hospitalized patient profiles as new COVID-19 waves spread. Anothe study has explored how the death prediction task evolved throughout the pandemic, showing that facto such as vaccination changed the profile of severely ill patients [4]. These characterizations can help th detection of important pandemic events, such as the impacts of vaccination, emergence of new COVID-1 strains and the emergence of new viral strains resistant to current available therapies. In this context, w focus our characterization efforts on technology evaluation through the lens of data drifts in a healthcar setting.

Some solutions have been reported in the literature to address learning in the presence of data drifts, mo focused on sample selection and/or sample weighting, with variations on how they derive the final weightin and/or sampling. Klinkenberg (2004), for instance, tackles the problem by using Support Vector Machine (SVMs) for both sample selection and sample weighting, employing an iterative process that sequentiall trains SVMs to find the training instances that constitute the model's support vectors [43]. Kolter and Malo (2023) uses a special weighted ensemble to learn in the presence of such drifts [44]. Salles et al., (2010, 201 2017) use a temporal weighting function that can be automatically learned to select relevant samples fo

each training window [6,30,45]. Finally, Rocha et al., (2008) -tackles the problem using temporal contexts [7
The authors analyze document collections that evolve over time, and define a temporal context as portions
documents that minimize the temporal effects of class distribution, term distribution, and class similarity ove
time. This method is used to devise a greedy strategy to optimize the trade-off between undersampling an
temporal effects. We were inspired by this latter work in our methodology. Most of these approache
however, are not applied to the healthcare setting, focusing mostly on common text data.
Another relevant setting is detecting drifts in data streams. This is potentially relevant to some healthcar
data, especially sensor data, which are most commonly obtained in hospitalized patients, but also streame
from personal health devices such as smartwatches and heart rate sensors. Zliobaite et al., (2014), fo
instance, proposes a continuous loop of labeling new samples under a labeling budget and uses activ
learning to detect data drifts [46].
Class imbalance is another important aspect of detecting data drifts in healthcare data. Disease occurrence
naturally unbalanced, with common diseases such as diabetes or hypertension affecting between 5 to 30% o
the population [47,48]. Rare diseases, on the other hand, have a prevalence in the order of fewer than te
patients per 100,000 or 1,000,000 inhabitants, with a combined prevalence among all rare diseases –bein
estimated between 3.5% to 5% [49]. Most approaches to handle such class imbalances in the data dri
literature focus on oversampling, undersampling or a combination of both. Gao et al., (2008), for instanc
propose oversampling the minority class over multiple time slices, while undersampling the majority clas
using only the most recent slice [50]. Ditzler and Polikar (2013) on the other hand, focuses on usin
incremental learning combined with synthetic minority over-sampling technique (SMOTE) [51,52] to learn
classification ensemble that can deal with both the class imbalance and concept drifts in streamed data. I
particular, the combination of models and datasets used in our work was robust to such class imbalanc
issues and did not require using these types of techniques, as will be shown in the following sections.

Summary of the Main Results of Applying DIS to MIMIC-V

The instantiation of the drift detection step using several distribution comparison metrics showed th
flexibility of the methodology. It also demonstrated that, for the purpose of separating the temporal chunk
in this particular scenario, metrics such as the Jensen-Shannon divergence or the Classifier errors capture th
underlying distributions better than particular outliers or novel samples. Higher values in these metrics imp
more significant "populational" changes, such as a gradual shift in the composition of the in-hospita
population,"s disease burden.
As also shown in the drift detection step (Figure 6), there is a gradual but persistent pattern in MIMIC-IV
happening over several years. This gradual change may be caused by various factors, such as an increase
tendency for terminally ill patients to receive end-of-life care at home, and/or advancements in therapeut
techniques for certain diseases. The nature of the expected data change can be hypothesized based o
characteristics such as the suddenness or gradualness of the drift , its persistence, and its duration, along wit
the results from the next analytical steps in DIS. This difference becomes evident when comparing the MIMIC
IV and the Brazilian COVID-19 Registry datasets.
The Initial characterization step (shown in Figure 8), reveals a trend toward a decrease in overall mortalit
over time, and this is the "context" in which we will interpret subsequent findings . Additionally, Figure
indicates that the overall characteristics of the deceased patients changed more than the overall in-hospita
population over the observed time frame. This means thatthe reduction in overall mortality is due tochange

in the characteristics of the patients who died. Looking at the findings in Figure 8, we can inspect ho different diseases impacted mortality predictions over time. Figure 8 shows that two ICD-10 chapters "diseases of the circulatory system" and "cancer" - had important changes during this period. By associatin the findings of step 1 with those of *step 2*, we can begin to understand —the factors contributing to decreased mortality over time, but it does not provide the "full picture".

The DIS semantic characterization step, which measures how the contexts of the independent variables relat to those of dependent variables over time in a more semantic level, yields interesting results tha complement the previous ones. Figure 13 shows an example of such a result, showing changes in similarit for the "dysphagia following stroke" ICD-10 code within the MIMIC-IV dataset [18]. There has been a increase in the co-occurrence of many obesity-related ICD codes between the 2011-2013 and 2017-2019 tim slices. This is aligned with general observations of the increase in obesity prevalence in the overall U population. It is worth noting that this technique does not allow us to draw causal conclusions, but instea focus on the correlation and co-occurrence changes. The co-occurrence of death and "cancer", as well as wit "external causes" has decreased over the period, possibly indicating a reduction in iatrogenic events improved cancer treatment leading to lower lethality and/or that cancer patients are receiving more end-o life care at home. This may be an explanation as to why overall in-hospital mortality has decreased on th dataset.

As the overall mortality -decreases, -patterns affecting the similarity decrease of entities, such as cancer, ma be leaving other similarities, such as the lethality of circulatory diseases, unchanged. This means tha increases in similarity with the outcome may be simply due to the decrease in the lethality of other group To investigate this, we filtered the data only to cancer disease codes, as in Figure 12. The Figure reveale important decreases in mortality in mostly severe and hard-to-treat cancers, such as brain, colon, lung, an secondary (metastatic) tumors.

It is also possible that the observed patterns may be attributed to multiple factors at the same time. Fo instance, recent policy changes favoring home care for terminally ill patients may influence who dies in th hospital. If these patients are more likely to die at home, we might have a "survivorship bias", where most the ones who did not die received hospital care and the terminally ill were sent back home. Over this tim frame, we also had important advances in immuno-biological therapies for tumors, such as lung cancer, a well as early diagnostic techniques have made it possible to cure some early cases when the tumor is st resectable. Combining these factors yields a lower lethality ,which has decreased over time despite a increase in the total number of Neoplasm patients, as shown in Figure 7.

In summary, the application of the DIS methodology to the MIMIC-IV allowed us to determine importan trends that help to understand certain phenomena observed in the data. Moreover, it facilitates th formulation of interesting hypotheses which are harder to validate based only on the data itsel Nevertheless, in a real-world scenario, such hypotheses could be the subject of further investigation usin other data sources, such as official policy implementation records, country-wide demographic records or eve published literature.

Summary of Main Results of Applying DIS to the Brazilian COVID-19 Registr dataset

The *Drift Detection* step-, especially using the Jensen-Shannon Divergence, revealed important data drifts this dataset -shift, commencing about the same time interval as the vaccination rollout in Brazil, between lat

2020 and early 2021 [14]. The initial characterization revealed a trend towards decreasing mortality ov time, with the steepest decrease closely matching our drift detection. This means that thus far, we had a important variable distribution shift as well as a change in the distribution of the outcome itself.

We analyzed how the top-five highest Pearson correlation variables behaved over time (Figure 19). Figure 1 A shows how the relative ranking and correlation of the best predictors of death changed over the course o the pandemic, with features such as "age" being the strongest predictors at the early stages, and gradual becoming less predictive over time. Figure 19-A also shows how patient severity markers, such as "FiO_2" ar "altered level of consciousness" gradually became more important predictors over time, hinting at the chang from "older patients dying from COVID-19" to "patients who were severely ill at admission dying from COVID 19". From our analysis, the patient's age is shown to be a consistently robust predictor of COVID-19-relate hospitalization and death. In Figure 19-B, we show the median age of the COVID-19 deceased patients. Th shows how one of the most predictive features in this dataset has changed over time, with the median dyin age decreasing from about 63 years at its peak to about 55 in a time frame coinciding with the start of th vaccination campaign in Brazil [5]. However, the median age starts to rise again, possibly relating to anothe drift, such as the emergence of new viral strains that can disproportionately affect the elderly population. Th fluctuation in the median age of deceased patients leads to the aforementioned deterioration of th correlation scores. Furthermore, this pattern with the age variable decreasing over time is consistent wit how the vaccines were rolled out to the public, which prioritized older age groups for vaccination [46]. these groups received vaccines earlier and consequently reduced their probability of death, this would like reduce the median and mean deceased patients' age.

The main results of the semantic characterization step (Figures 18-23), where we compare the semant vectors for the "death" outcome in 2020 and 2021, validate several findings from the Initial characterizatio step and introduce new findings. For instance, it shows a decrease in similarity between the outcome an older groups (e.g., -"84-105" vs. "62-83" age groups) with an increase in similarity with younger groups. Th validates the findings in Figure 19-A, where median age declines steadily up until roughly September 202 Figure 19-A also shows how the "death" outcome had an increase in similarity with several disease severit markers, such as lower admission serum sodium, lower admission arterial blood pressure,fewe comorbidities, and lower FiO_2. This potentially indicates that, when compared to 2020, patients who died i 2021 were more severely ill at admission, had fewer comorbidities and were younger (presumab unvaccinated). This is a significant pattern change, especially compared to the bulk of deceased patients i the initial chunk, which were mostly elderly patients with lower severity at admission. This change in patter implies that, at the analyzed time frame, young and severely ill at admission patients were now a mor common pattern of dying patients. However, that should be analyzed in conjunction with the previou findings from the other steps. For instance, we know that the overall mortality has decreased, and th patient profile (young and severely ill at admission) could also be present in the first temporal chunk. Wha possibly happened was the removal of a significant portion of elderly dying patients from the population b events such as vaccination, evidenced by the reduced mortality and diminished predictive power of age.

To conclude, the DIS analysis hints at the central role of vaccination in the COVID-19 pandemic, whic reduced the odds of elderly patients dying from the disease following the rollout of the vaccines. Th hypothesis was raised by the alignment between the detected data drift and mortality reduction during th vaccination period. Additionally, the observed decrease in median age of the dying patients corresponded t the age-stratified vaccination strategy. Furthermore, the shift of mortality burden to the young and severe ill patients upon admission, who were likely unvaccinated, demonstrates how they possibly kept dying whil this process unfolded.

Limitations

We have proposed a methodology to discover and interpret temporal shifts in healthcare data. While ou approach provides valuable insights by uncovering many correlations and semantic connections, DIS st cannot establish causal relationships outcomes and semantic units. The causal part is only hypothesized an inferred, but the methodology does not go so far as to return causal links for arbitrary outcome Furthermore, we have not applied the methodology to certain relevant healthcare domains, such as image (e.g., x-rays, computed tomography or ultrasound) and wavelets (e.g., electrocardiograms o electroencephalograms).

That said, here we offer some insights on how we could apply DIS to handle temporal shifts in no quantitative data or raw magnetic resonance imaging (MRI) data. For this, we would first need to obtain distributed representation of the data in such a manner that samples from similar patients have simila embedding vectors. For instance, we could use DINOv2 embeddings or Contrastive Language-Image Pre training (CLIP) embeddings in images. This type of pre-trained neural network exists for multiple data type which facilitates the application to multiple domains. From the embeddings, we can apply the first step of ou methodology as performed to tabular data, computing Jensen-Shannon divergence (or Auto-encoder error classifier errors, etc.) to detect whether a drift exists on that data or not. Exploring this data for the secon step presents some challenges, as it might involve exploring both the embedding and raw data spaces. Fo instance, we can use clustering and centroid analysis (applied to the embeddings) to find samples where th drift is particularly pronounced. Then, we can go back to the raw data and analyze the samples to check fo patterns. In essence, the third step remains similar in nature. The idea is to train a neural network model suc that the embeddings of the samples closely resemble the embeddings of the outcomes experienced by thos patients over time. One such way to obtain these embeddings, starting from pre-trained ones, is to us losses like the triplet loss to approximate patient sample embeddings from outcome embeddings. Th interpretation of the triplet loss, as we presented in our work, will change according to the tempora granularity of the samples. If the data has high temporal granularity, the positive pairs (which the loss w learn to represent more closely in space) will obey an ordered sequence of events. For instance, two MI tests will be proximate if they belong to the same patient and happen close to each other in time -and if the are visually and semantically similar. Conversely, if the data has a low temporal granularity, the embedding should be learned to align patient samples to their outcome embeddings. Then, for the analysis of suc embeddings, we would have to analyze the raw data samples closer to the outcome embeddings. If one split the time, say, in two years and is working with the "death" outcome, one would be expected to have on such outcome for each year. Then, and this might require some domain expertise, analyzing the sample closer to each of the outcome embeddings should help build an understanding of the relevant changes in more generalized setting. We intend to explore these ideas in future work.

Finally, we cannot claim that our three steps (encompassing the “if”, “what” and “why” a data dri happened) are a comprehensive list of all possible steps to analyze a temporal shift. Instead, we believe ou steps to be a minimum required subset. While it is possible that these steps might not cover all possibl situations, they allowed us to obtain interesting insights from the two datasets presented in our work, a discussed above. We and other researchers plan to continue to study, extend, and adapt this methodology i future work to test the limits of our approach and whether new steps or a refinement of the ones propose in a fiber granularity level is necessary.

As part of a future work, we intend to propose exploring methods for enhancing models resilience to dat drifts, as well as exploring different healthcare-relevant domains, such as images and wavelets, as well a multimodal data.

Conclusions

We have proposed DIS, a temporal data drift methodology for analyzing the changes in health outcomes an variables over time while discovering outcome contextual changes in large volumes of data. We applied DIS t two very different case studies and demonstrated how it can provide valuable insights about changin patterns in the data and the underlying reasons driving such changes.

The DIS methodology goes beyond simple detection; it comprehensively characterizes temporal data drift By analyzing the underlying causes, patterns, and magnitudes of drifts, healthcare stakeholders can gain deeper understanding of the factors influencing data changes over time. This deeper understanding ha practical implications for healthcare organizations, allowing them to improve patient care, optimize resourc allocation, and enhance operational efficiency by leveraging the insights gained from monitoring an characterizing temporal data drifts.

The practical implications of our methodology are far-reaching. Early detection of data drifts can trigge timely interventions, enabling proactive adjustments to treatment plans, healthcare policies, and qualit improvement initiatives. Our methodology empowers healthcare practitioners and data analysts to effective monitor and manage temporal data drifts, ultimately leading to better healthcare outcomes and informe decision-making processes.

Acknowledgements

We would like to thank the hospitals and staff for their support in this project.

Funding

This study was supported by Minas Gerais State Agency for Research and Development (*Fundação de Ampar à Pesquisa do Estado de Minas Gerais* - FAPEMIG) [grants number APQ-01154-21; APQ-00262-22], Nation Institute of Science and Technology for Health Technology Assessment (Instituto de Avaliação de Tecnologia em Saúde – IATS)/National Council for Scientific and Technological Development (Conselho Nacional d Desenvolvimento Científico e Tecnológico – CNPq) [grant number 465518/2014-1]. National Council fo Scientific and Technological Development [grants number 421773/2022-7; 403184/2021-5; 401898/2022-9] This study was also partially financed with resources from the Center for Innovation and Artificial Intelligenc for Health (CI-IA Saúde), in part with resources from the São Paulo State Research Support Foundatio (FAPESP) Process n° 2020/09866-4, from the Foundation of Minas Gerais Research Support (FAPEMIG Process No. PPE-00030-21 and UNIMED Belo Horizonte. MSM was supported in part by CNPq [grant numbe 310561/2021-3]. MAG was supported in part by CNPq [310538/2020-3]. The funding bodies played no role i the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Authors' contribution

Substantial contributions to the conception or design of the work: BBMP, LCR, JMA
MSM, CMVA, FCBL, MVRSS, PDP, MAG.

Substantial contributions to the acquisition, analysis, or interpretation of data for th
work: all authors.

Drafted the work: BBMP, LCR, JMA, MSM, CMVA, FCBL, MVRSS, PDP, MAG.

Revised the manuscript critically for important intellectual content: all authors.

Final approval of the version to be published: all authors.

Conflict of Interest

The authors declare no conflict of interest.

Abbreviations

CLIP: Contrastive Language-Image Pre-training

DIS: Detection, Initial Characterization, Semantic Characterization

FiO_2: Fraction of Inspired Oxygen
ICD: International Classification of Diseases
ICU: Intensive Care Unit
MIMIC-IV: Medical Information Mart for Intensive Care IV
MRI: Magnetic Resonance Imaging
NLP: Natural language processing
PCA: Principal Component Analysis
SHAP: SHapley Additive exPlanations
SMOTE: Synthetic minority over-sampling technique
SVMs: Support Vector Machines
TF-IDF: Term Frequency-Inverse Document Frequency

References

1. Vayena E, Dzenowagis J, Brownstein JS, Sheikh A. Policy implications of big data in the healt
sector. Bull World Health Organ. 2018;96(1):66-68. doi:10.2471/BLT.17.197426

2. Pastorino R, De Vito C, Migliara G, et al. Benefits and challenges of Big Data in healthcare: A overview of the European initiatives. Eur J Public Health. 2019;29:23-27. doi:10.1093/eurpub/ckz168

3. Roski J, Bo-Linn GW, Andrews TA. Creating value in health care through big data: opportunitie and policy implications. Health Aff (Millwood). 2014;33(7):1115-1122. doi:10.1377/hlthaff.2014.0147

4. Barbosa B, De Paiva M, Delfino-Pereira P, et al. Characterizing and Understanding Tempor Effects in COVID-19 Data. Proceedings of the 1st Workshop on Healthcare AI and COVID-19, ICML. 202 PMLR 184:33-40

5. Moura EC, Cortez-Escalante J, Cavalcante FV, Barreto IC de HC, Sanchez MN, Santos LMP. COVID 19: temporal evolution and immunization in the three epidemiological waves, Brazil, 2020–2022. *Re saúde pública*. 2022;56:105. doi:10.11606/s1518-8787.2022056004907

6. Salles T, Rocha L, Pappa GL, Mourã F, Meira W, Gonçalves M. Temporally-aware algorithms fo Document Classification. SIGIR 2010 Proceedings - 33rd Annual International ACM SIGIR Conference o Research and Development in Information Retrieval. Published online 2010:307-314 doi:10.1145/1835449.1835502

7. Rocha L, Mourão F, Pereira A, Gonçalves MA, Meira W. Exploiting temporal contexts in tex classification. In: International Conference on Information and Knowledge Management, Proceedings. 2008:243-252. doi:10.1145/1458082.1458117

8. Abdul AR, C. R N, B. R S, Lahza H, Lahza HFM. A survey on detecting healthcare concept drift i AI/ML models from a finance perspective. Front Artif Intell. 2022;5. doi:10.3389/frai.2022.955314

9. McLean C, Capurro D. Concept Drift Detection to Assess the Diffusion of Process Innovations i Healthcare. AMIA Annu Symp Proc. 2023;2022:746-755. Published 2023 Apr 29. PMID: 37128394

10. Sundquist M, Brudin L, Tejler G. Improved survival in metastatic breast cancer 1985-2016. Breas 2017;31:46-50. doi:10.1016/j.breast.2016.10.005

11. Lima ES, Romero EC, Granato CFH. Current polio status in the world. J Bras Patol Med Lab 2021;57:1-6. doi:10.5935/1676-2444.20210022

12. Dabbagh A, Patel MK, Dumolard L, et al. Progress Toward Regional Measles Elimination Worldwide, 2000-2016. MMWR Morb Mortal Wkly Rep. 2017;66(42):1148-1153. Published 2017 Oct 27 doi:10.15585/mmwr.mm6642a6

13. Ghosn L;, Evrenoglou T;, Jarde A;, et al. Efficacy and safety of COVID-19 vaccines. Cochran Database of Systematic Reviews. 2022;(12):2022. doi:10.1002/14651858.CD015477

14. Menéndez ML, Pardo JA, Pardo L, Pardo MC. The Jensen-Shannon divergence. J Franklin Ins 1997;334(2):307-318. doi:10.1016/S0016-0032(96)00063-4

15. Menon AG, Gressel G. Concept Drift Detection in Phishing Using Autoencoders. Communication in Computer and Information Science. 2021;1366:208-220. doi:10.1007/978-981-16-0419-5_17

16. Deng Z, Li C, Song R, Liu X, Qian R, Chen X. Centroid-Guided Domain Incremental Learning for EEG Based Seizure Prediction. IEEE Trans Instrum Meas. 2024;73:1-13. doi:10.1109/TIM.2023.3334330

17. Marcolino MS, Pires MC, Ramos LEF, et al. ABC2-SPH risk score for in-hospital mortality in COVID 19 patients: development, external validation and comparison with other available scores. International Journal of Infectious Diseases. 2021;110:281-308. doi:10.1016/j.ijid.2021.07.049

18. MIMIC-IV v1.0. https://physionet.org/content/mimiciv/1.0/ [accessed Oct 18, 2023].

19. Wang H, Fan W, Yu PS, Han J. Mining concept-drifting data streams using ensemble classifier Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining Published online 2003:226-235. doi:10.1145/956750.956778

20. Parmar HS, Nutter B, Mitra SD, Long R, Antani SK. Automated signal drift and global fluctuatio

removal from 4D fMRI data based on principal component analysis as a major preprocessing step fo fMRI data analysis. *SPIE*. 2019;10953:109531E. doi:10.1117/12.2512968

21. Benesty J, Chen J, Huang Y, Cohen I. Pearson Correlation Coefficient. Springer Topics in Signa Processing. 2009;2:1-4. doi:10.1007/978-3-642-00296-0_5

22. Myers L, Sirois MJ. Spearman Correlation Coefficients, Differences between. Encyclopedia o Statistical Sciences. Published online July 15, 2004. doi:10.1002/0471667196.ESS5050

23. Kazemitabar SJ, Amini AA, Bloniarz A, Talwalkar A. Variable Importance Using Decision Trees. Ad Neural Inf Process Syst. 2017;30.

24. Lundberg SM, Lee SI. A Unified Approach to Interpreting Model Predictions. Neural Informatio Processing Systems. Published online 2017. 25. Mnih A, Kavukcuoglu K. Learning word embedding efficiently with noise-contrastive estimation. Adv Neural Inf Process Syst. 2013;26

26. Kullback–Leibler divergence. Available from:. https://en.wikipedia.org/wiki/Kullbac %E2%80%93Leibler_divergence [accessed Oct 18, 2023].

27. Jones KS. A statistical interpretation of term specificity and its application in retrieval. Journal o Documentation. 1972;28(1):11-21. doi:10.1108/EB026526/FULL/XML

28. Hamad D, Biela P. Introduction to spectral clustering. 2008 3rd International Conference o Information and Communication Technologies: From Theory to Applications, ICTTA. Published onlin 2008. doi:10.1109/ICTTA.2008.4529994

29. Grover A, Leskovec J. Node2vec: Scalable feature learning for networks. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016;13-17-Augus 2016:855-864. doi:10.1145/2939672.2939754

30. Salles T, Rocha L, Gonçalves MA, et al. A quantitative analysis of the temporal effects o automatic text classification. *J Assoc Inf Sci Technol*. 2016;67(7):1639-1667. doi:10.1002/ASI.23452

31. Mouro F, Rocha L, Arajo R, Couto T, Gonçalves M, Meira W. Understanding temporal aspects i document classification. *WSDM'08 - Proceedings of the 2008 International Conference on Web Searc and Data Mining*. Published online 2008:159-169. doi:10.1145/1341531.1341554

32. Rotalinti Y, Tucker A, Lonergan M, Myles P, Branson R. Detecting Drift in Healthcare AI Mode Based on Data Availability. Communications in Computer and Information Science. 2023;1753 CCIS:243 258. doi:10.1007/978-3-031-23633-4_17

33. Abdul AR, Nirmala CR, Aljohani M, Sreenivasa BR. A novel technique for detecting sudde concept drift in healthcare data using multi-linear artificial intelligence techniques. Front Artif Inte 2022;5:950659. doi:10.3389/FRAI.2022.950659/BIBTEX

34. Gama J, Zliobaite I, Bifet A, Pechenizkiy M, Bouchachia A. A survey on concept drift adaptation ACM Computing Surveys (CSUR). 2014;46(4). doi:10.1145/2523813

35. Moreno-Torres JG, Raeder T, Alaiz-Rodríguez R, Chawla N V., Herrera F. A unifying view o dataset shift in classification. *Pattern Recognit*. 2012;45(1):521-530. doi:10.1016/j.patcog.2011.06.019

36. Lu J, Liu A, Dong F, Gu F, Gama J, Zhang G. Learning under Concept Drift: A Review. *IEEE Tran Knowl Data Eng*. 2019;31(12):2346-2363. doi:10.1109/TKDE.2018.2876857

37. Haque A, Chandra S, Khan L, Hamlen K, Aggarwal C. Efficient multistream classification usin direct density ratio estimation. *Proc Int Conf Data Eng*. Published online May 16, 2017:155-158 doi:10.1109/ICDE.2017.63

38. Sahiner B, Chen W, Samala RK, Petrick N. Data drift in medical machine learning: implications an potential remedies. Br J Radiol. 2023;96(1150):20220878. doi:10.1259/bjr.20220878

39. Tiwari S, Agarwal S. Data Stream Management for CPS-based Healthcare: A Contemporar

Review. IETE Technical Review. 2022;39(5):987-1010. doi:10.1080/02564602.2021.1950578

40. UniProt Consortium. UniProt: a hub for protein information. *Nucleic Acids Res*. 2015;43(Databas issue):D204-D212. doi:10.1093/nar/gku989

41. Jung C, Excoffier JB, Raphaël-Rousseau M, Salaün-Penquer N, Ortala M, Chouaid C. Evolution c hospitalized patient characteristics through the first three COVID-19 waves in Paris area using machin learning analysis. *PLoS One*. 2022;17(2):e0263266. Published 2022 Feb 2 doi:10.1371/journal.pone.0263266

42. Jassat W, Mudara C, Ozougwu L, et al. Difference in mortality among individuals admitted t hospital with COVID-19 during the first and second waves in South Africa: a cohort study. *Lancet Glo Health*. 2021;9(9):e1216-e1225. doi:10.1016/S2214-109X(21)00289-8

43. Klinkenberg R. Learning drifting concepts: Example selection vs. example weighting. *Intelliger Data Analysis*. 2004;8(3):281-300. doi:10.3233/IDA-2004-8305

44. Kolter JZ, Maloof MA. Dynamic Weighted Majority: An Ensemble Method for Drifting Concept *Journal of Machine Learning Research*. 2007;8(91):2755-2790. *45.* Salles T, Rocha L, Mourão Gonçalves M, Viegas F, Meira W. A Two-Stage Machine learning approach for temporally-robust tex classification. *Inf Syst*. 2017;69:40-58. doi:10.1016/J.IS.2017.04.004

46. Zliobaite I, Bifet A, Pfahringer B, Holmes G. Active learning with drifting streaming data. *IEE Trans Neural Netw Learn Syst*. 2014;25(1):27-39. doi:10.1109/TNNLS.2012.2236570

47. World Health Organization. Hypertensio https://www.who.int/news-room/fact-sheets/detail/hypertension [accessed June 11, 2023].

48. American Diabetes Association. About Diabetes. https://diabetes.org/about-diabetes [accessed June 11, 2023].

49. Nguengang Wakap S, Lambert DM, Olry A, et al. Estimating cumulative point prevalence of rar diseases: analysis of the Orphanet database. *Eur J Hum Genet*. 2020;28(2):165-173. doi:10.1038/s4143 019-0508-0

50. Gao J, Ding B, Fan W, Han J, Yu PS. Classifying data streams with skewed class distributions an concept drifts. *IEEE Internet Comput*. 2008;12(6):37-49. doi:10.1109/MIC.2008.119

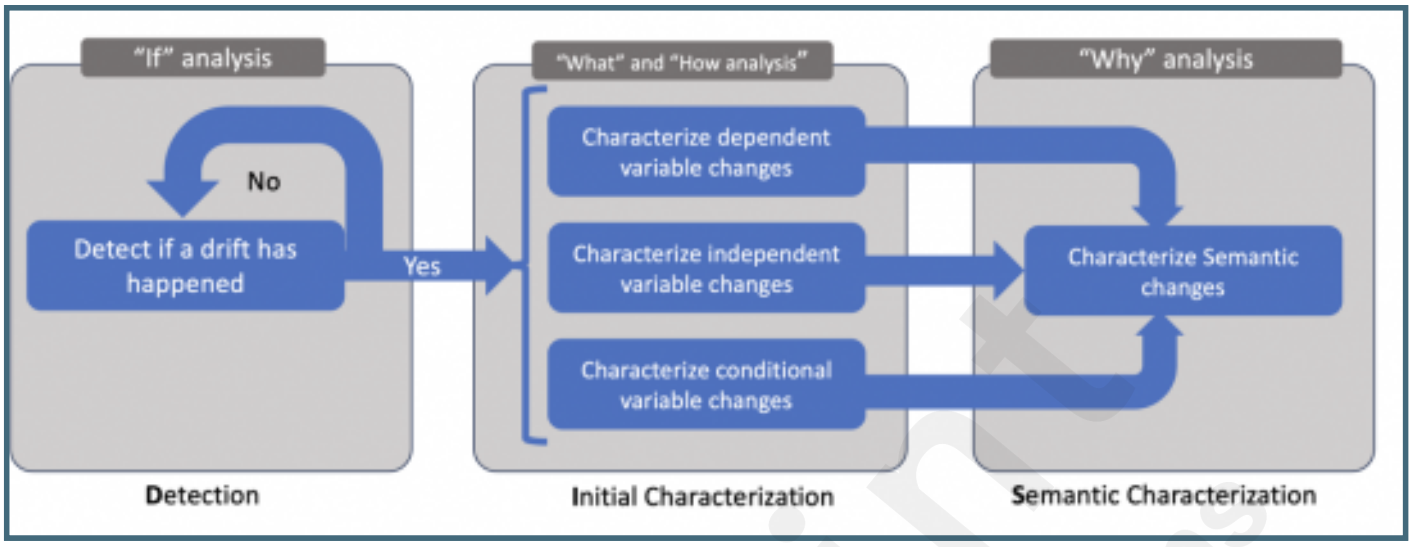
51. Ditzler G, Polikar R. Incremental learning of concept drift from streaming imbalanced data. *IEE Trans Knowl Data Eng*. 2013;25(10):2283-2301. doi:10.1109/TKDE.2012.136

52. Chawla N V., Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-samplin Technique. *Journal Of Artificial Intelligence Research*. 2011;16:321-357. doi:10.1613/jair.953

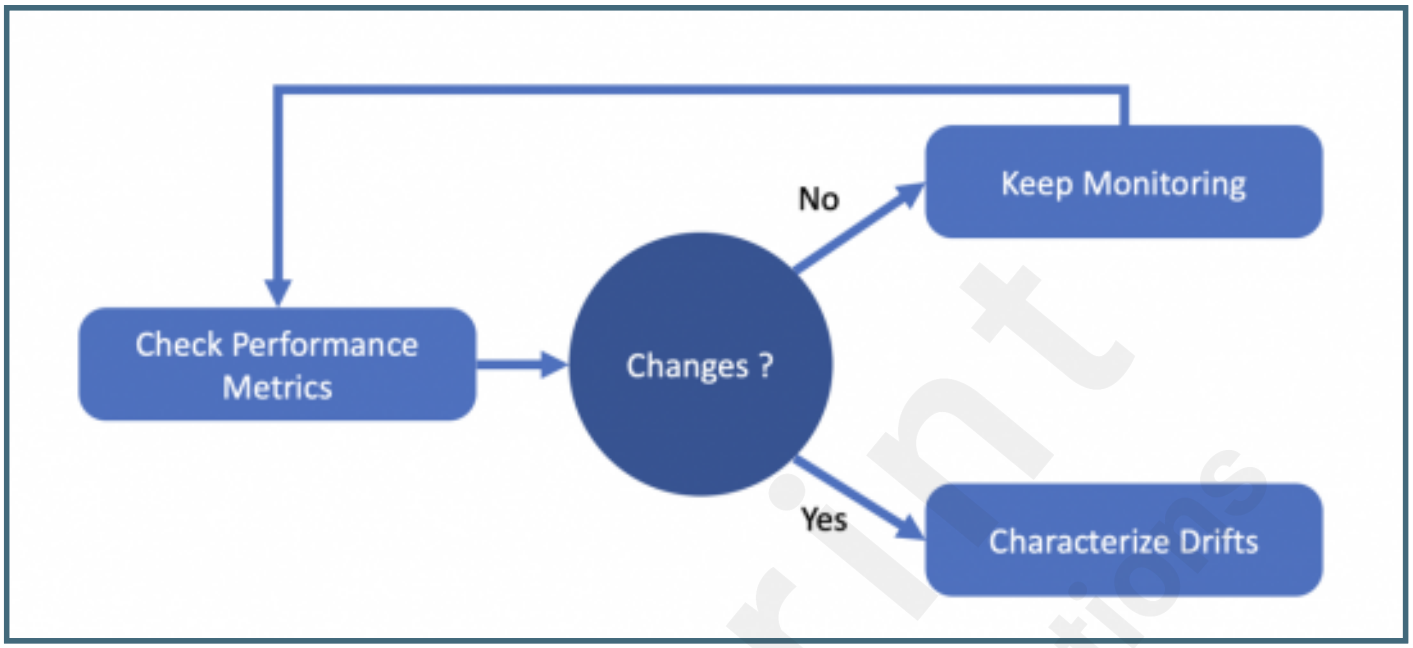
Supplementary Files

Figures

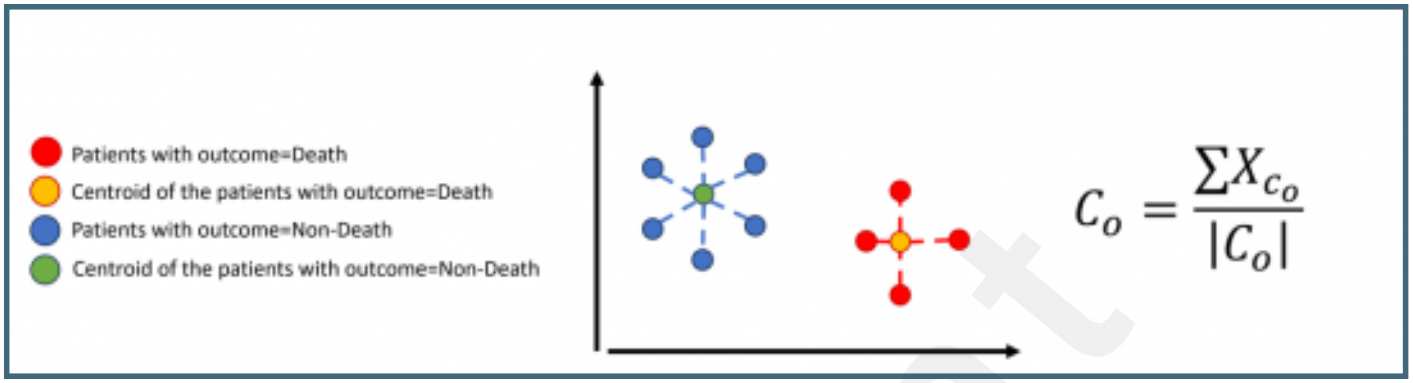
Overview of the detection, initial characterization and semantic characterization framework.



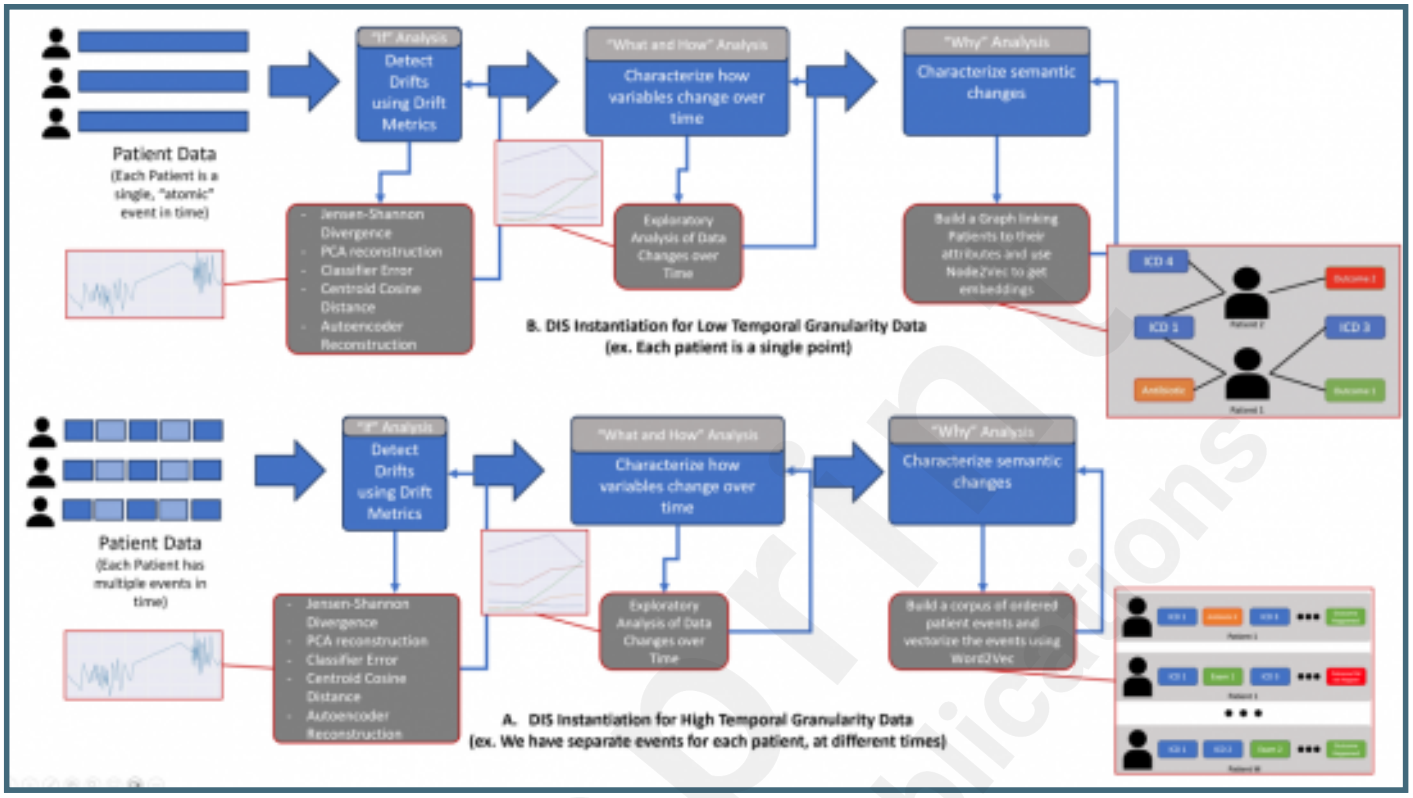
The temporal drift monitoring loop. We usually observe temporal shifts as important variations in model effectiveness through time.



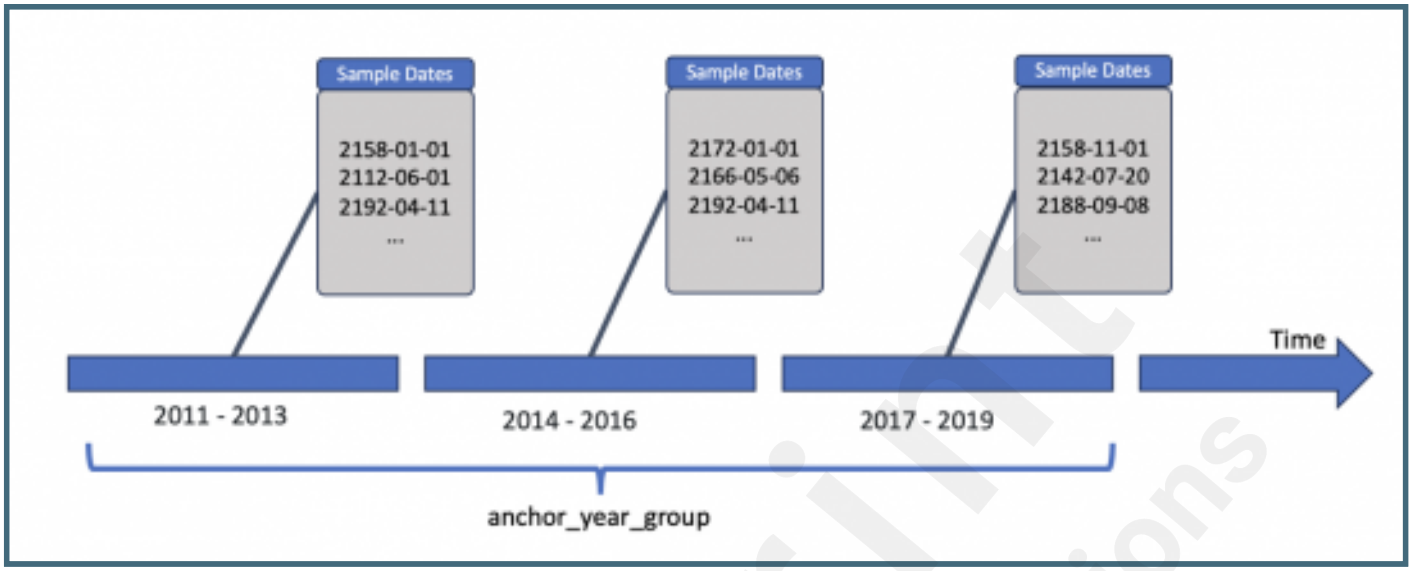
Modeling of the centroids as the arithmetic mean of the features in each outcome group. Co is the centroid of cluster O, XCO is the matrix of attributes including all patients in the outcome O and |CO| is the number of patients in the outcome group O.



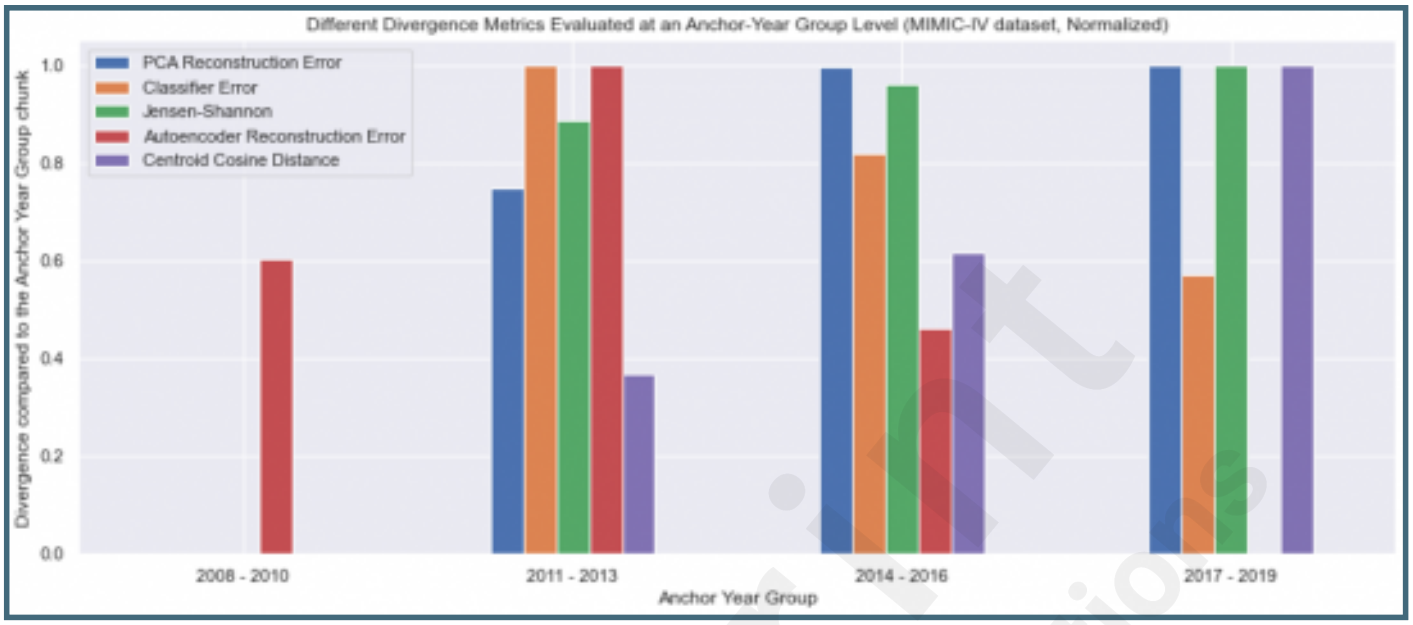
Overview of the Instantiation of the DIS Methodology to two scenarios with different temporal granularities (a) Medical Information Mart for Intensive Care, version IV (MIMIC-IV) DIS Instantiation and (b) Brazilian Covid Registry Instantiation.



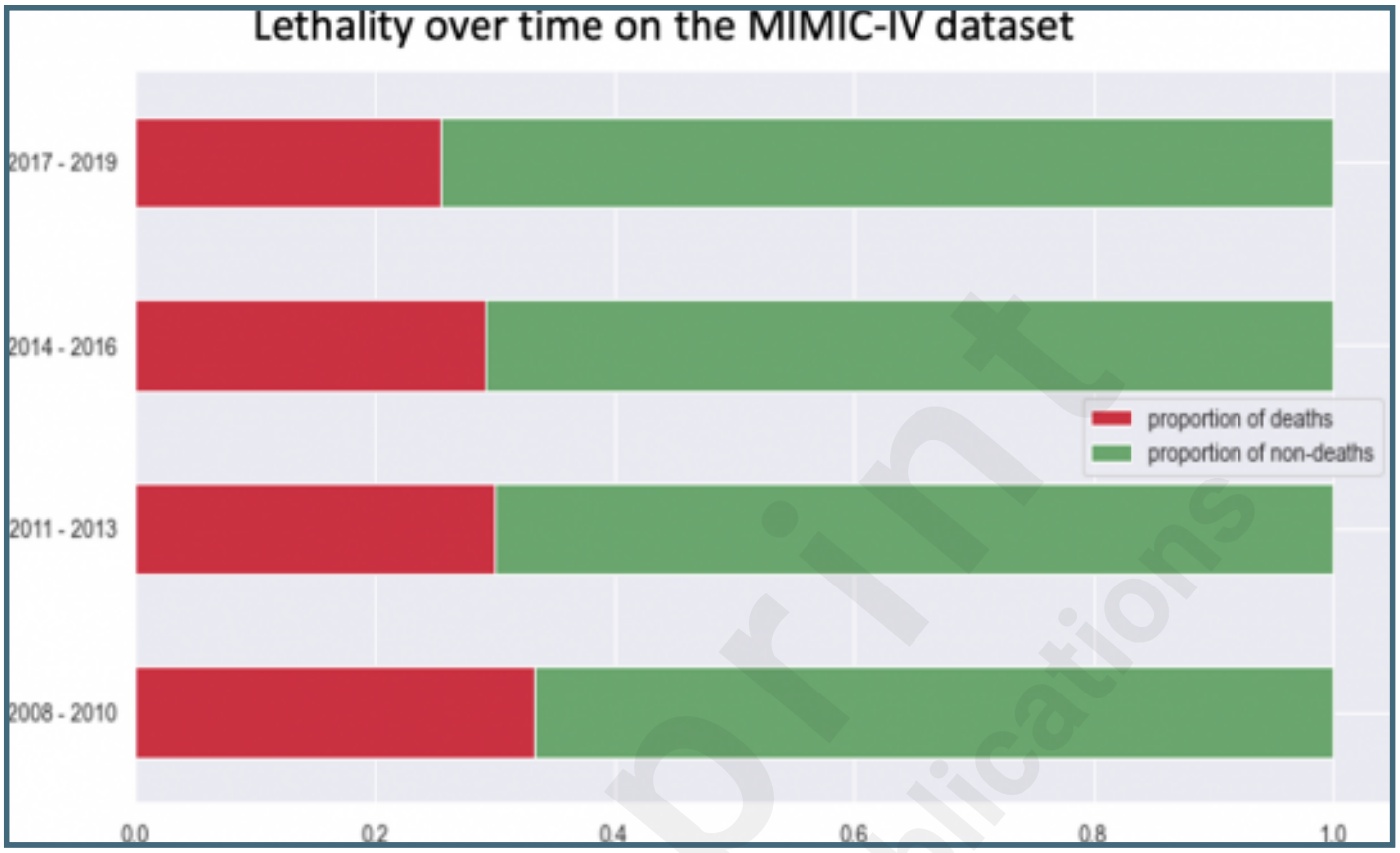
The "anchor_year_group" variable on the Medical Information Mart for Intensive Care, version IV (MIMIC-IV) dataset. Within each "anchor_year_group", the actual dates are masked, making it only possible to have a rough estimate of when the patient was at the hospital.



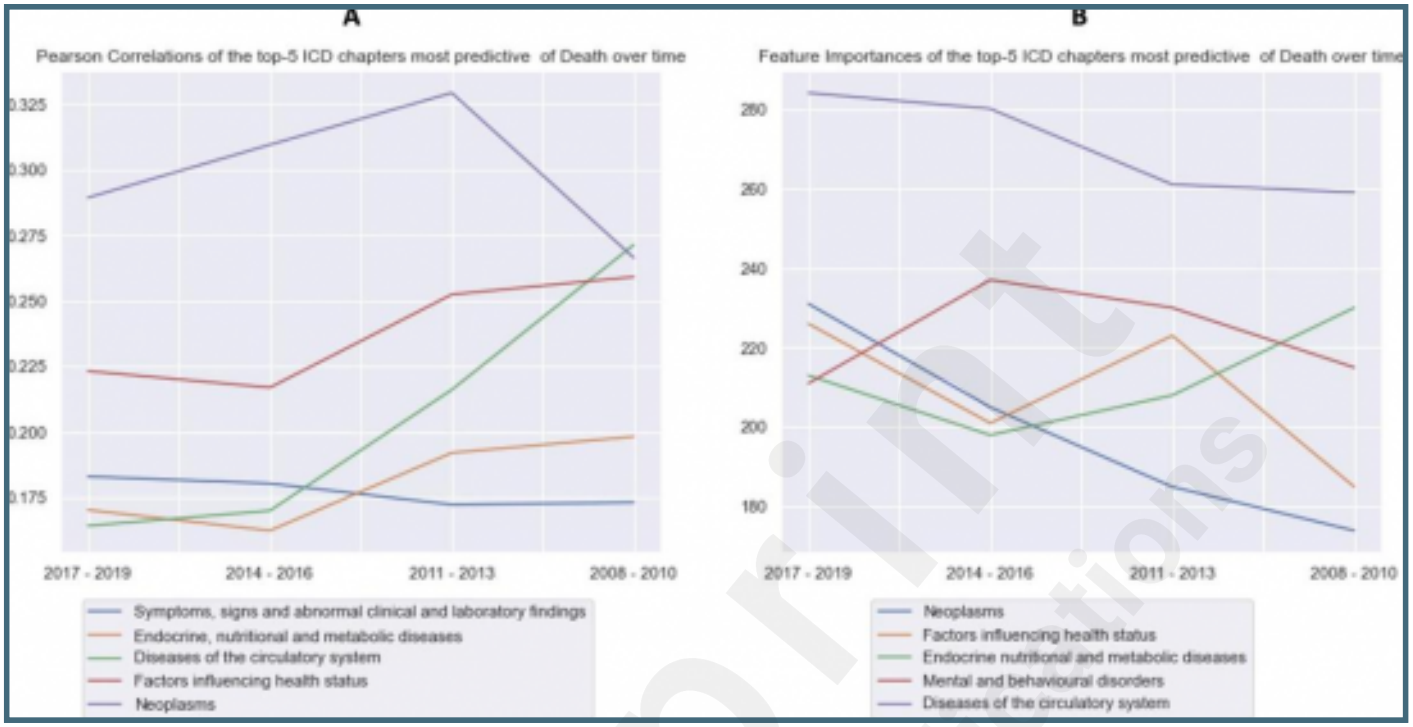
Different drift detection metrics over time on the Medical Information Mart for Intensive Care, version IV (MIMIC-IV) dataset, considering in-hospital ICD diagnosis.



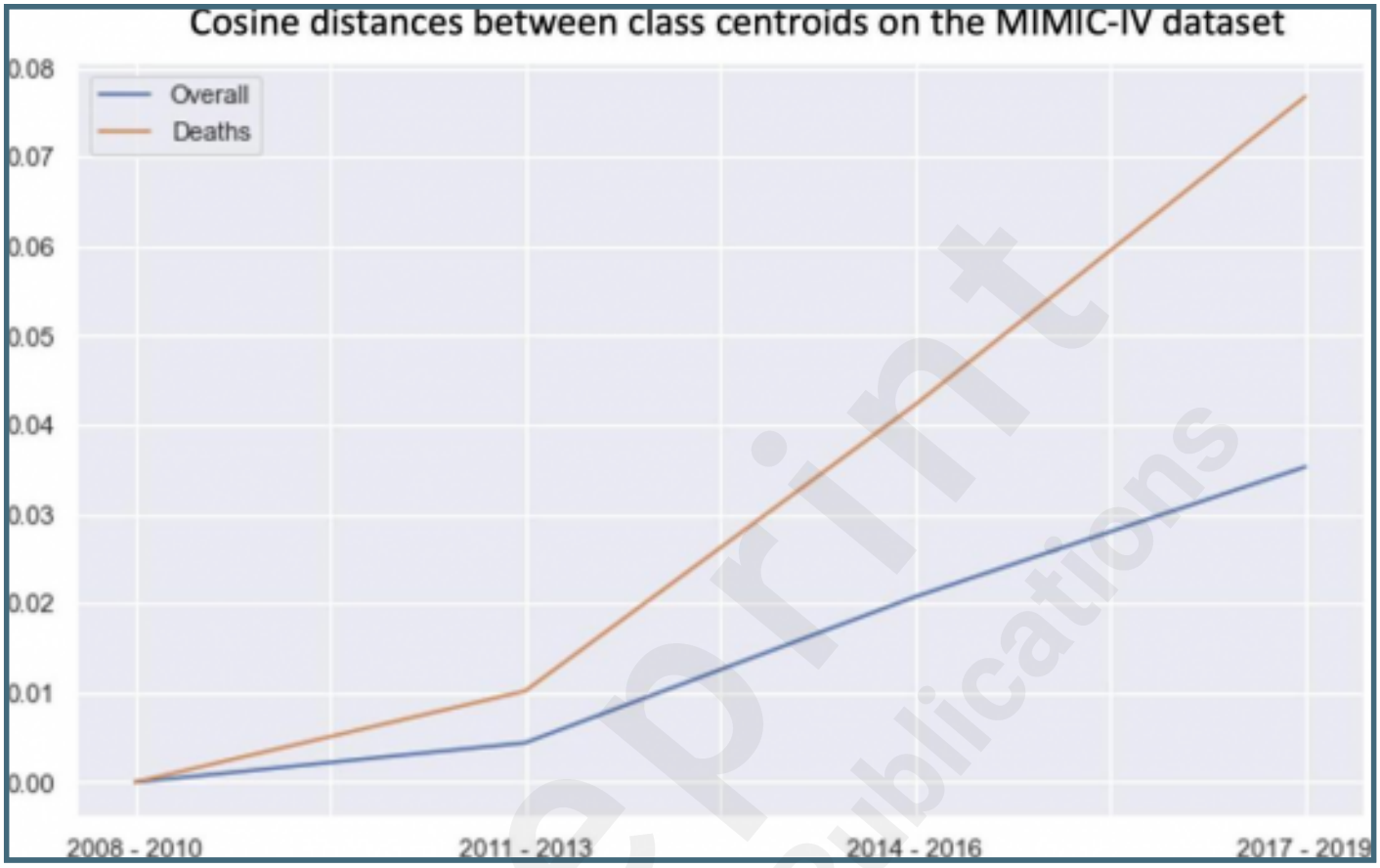
Lethality over time in the Medical Information Mart for Intensive Care, version IV (MIMIC-IV) dataset.



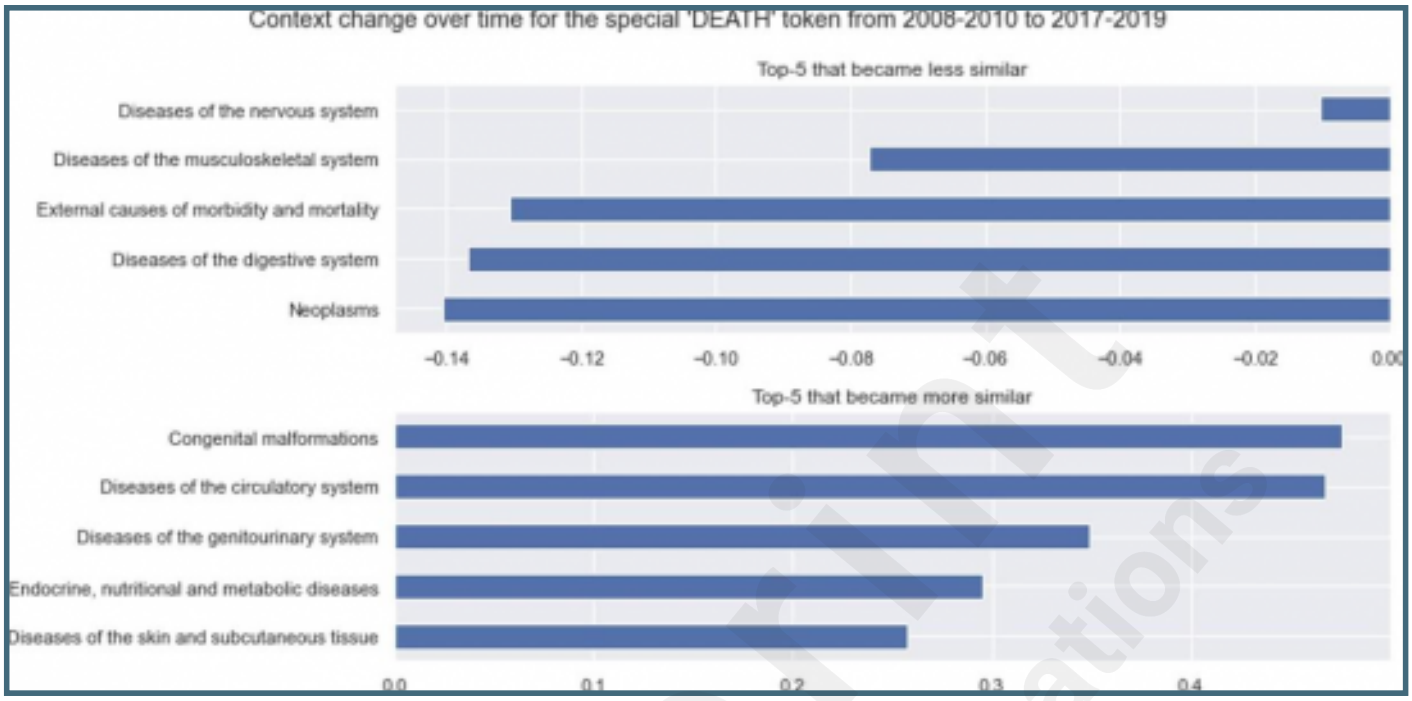
. (A) Pearson correlations between the top-5 International Disease Codes (ICD) chapters (according to ICD-10) most correlated with the death outcome over time. (B) Feature importances between the top-5 ICD chapters most predictive of the death outcome over time.



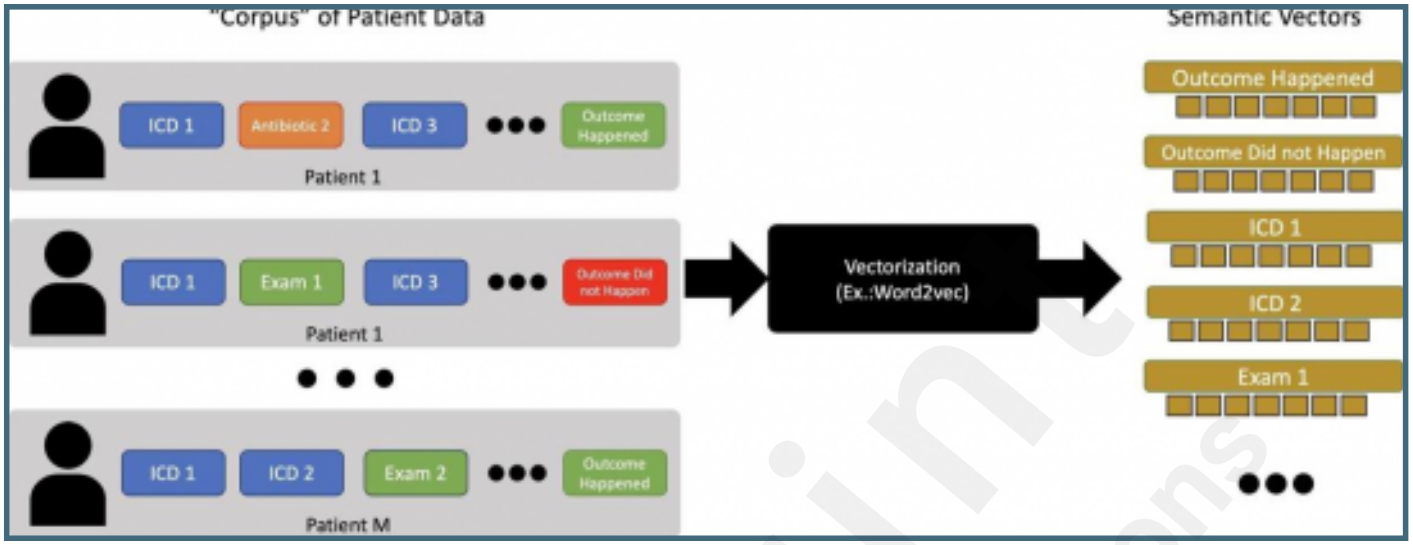
Drift of the arithmetic mean of each outcome class over time, as measured by cosine distances between each class"s means when compared to the mean of the first "anchor_year_group" on the Medical Information Mart for Intensive Care, version IV (MIMIC-IV) dataset.



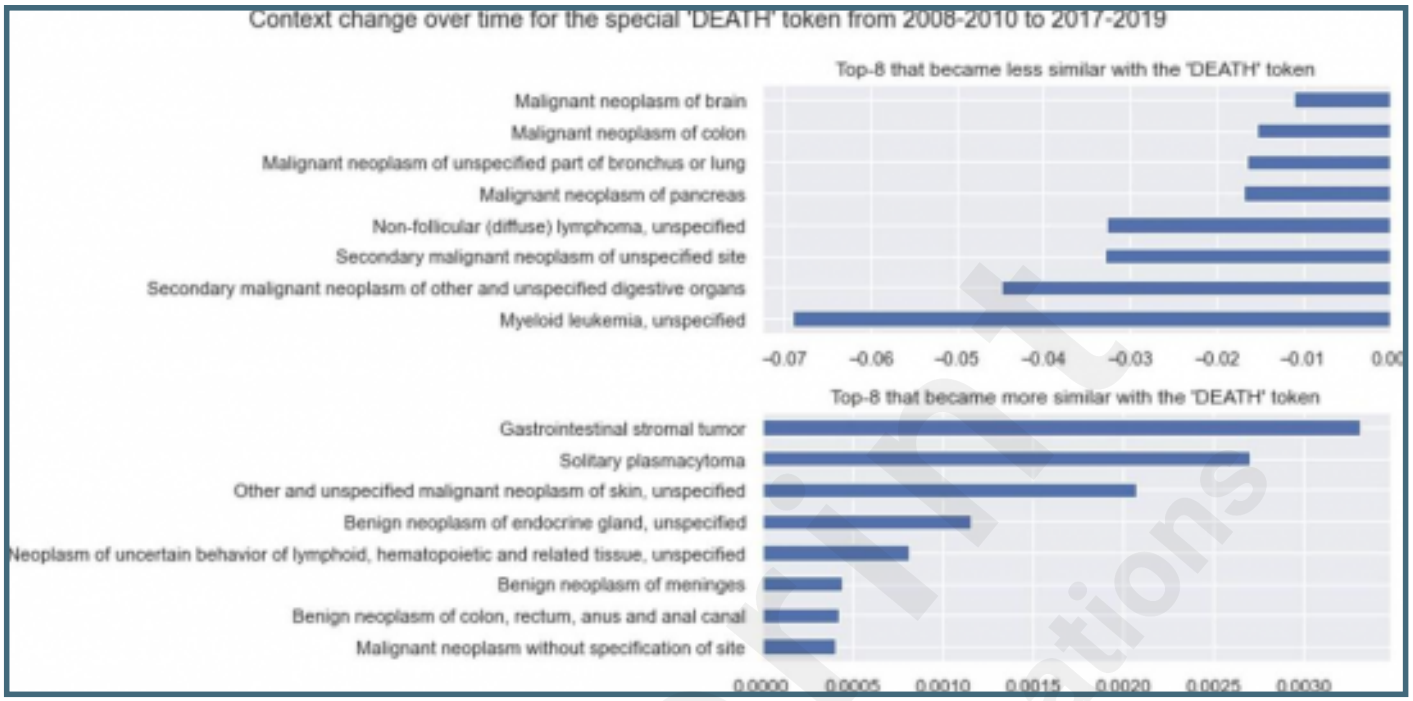
Evaluation of the drivers of lethality data drift on Medical Information Mart for Intensive Care, version IV (MIMIC-IV) dataset.



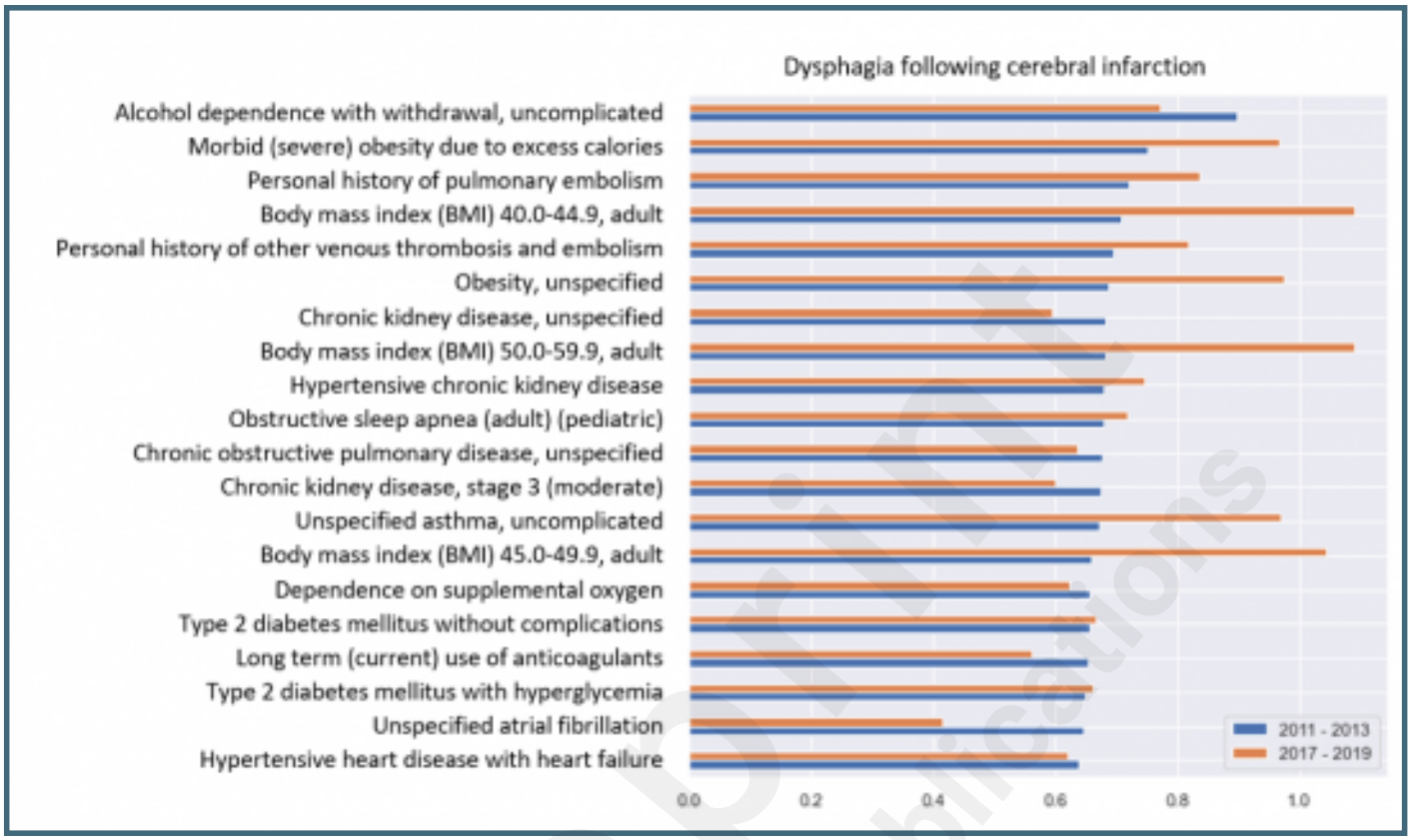
How to generate semantic vectors. We start by generating a corpus of temporally ordered patient discrete data points. Then, we vectorize the tokens of this corpus using Word2Vec to obtain semantic vectors for dependent and independent variables.



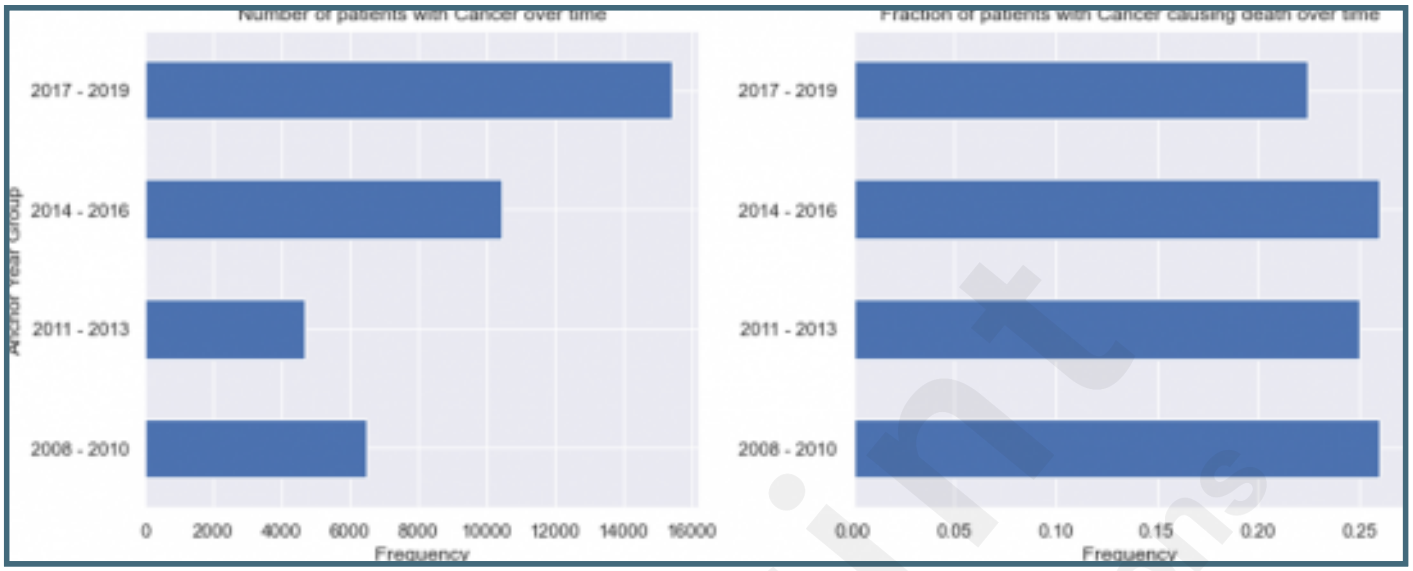
Evaluation of the drivers of lethality data drift on the Medical Information Mart for Intensive Care, version IV (MIMIC-IV) dataset.



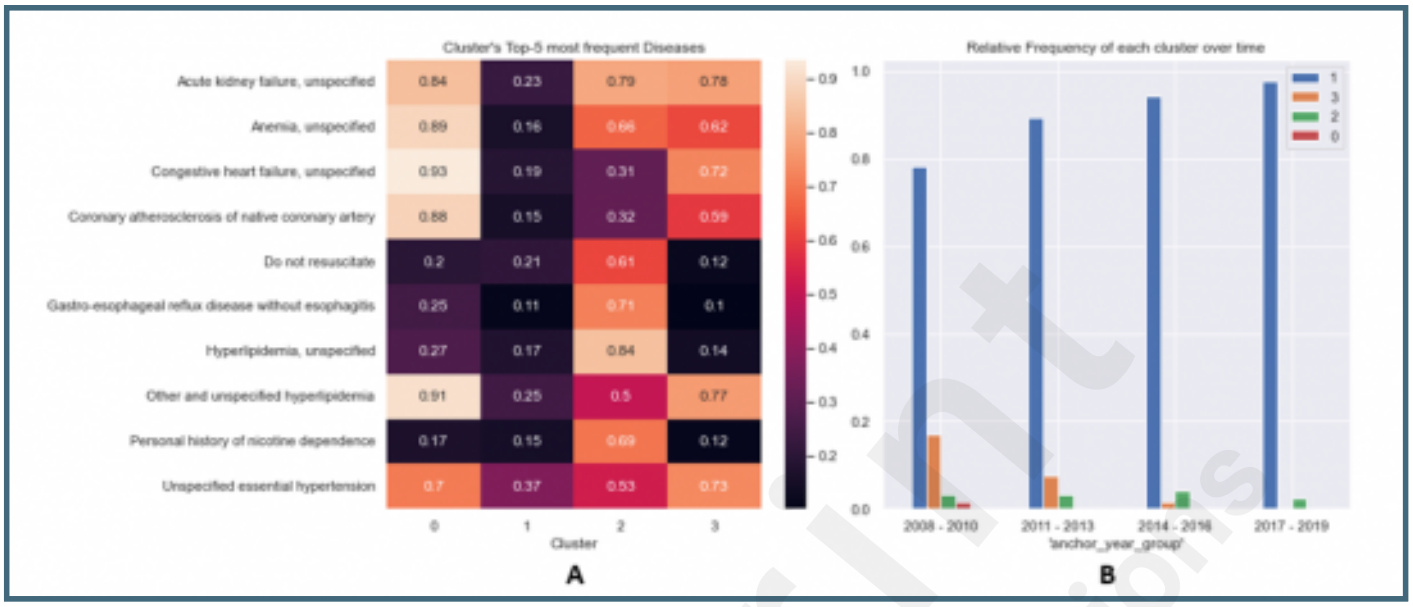
Changes in co-occurrence for the "dysphagia following stroke" International Disease Code (ICD) on the Medical Information Mart for Intensive Care, version IV (MIMIC-IV) dataset.



Validation of the data drift on cancer patients. On the left, we show the increase in the absolute number of cancer patients, while on the right we show the overall lethality reduction for this disease group.

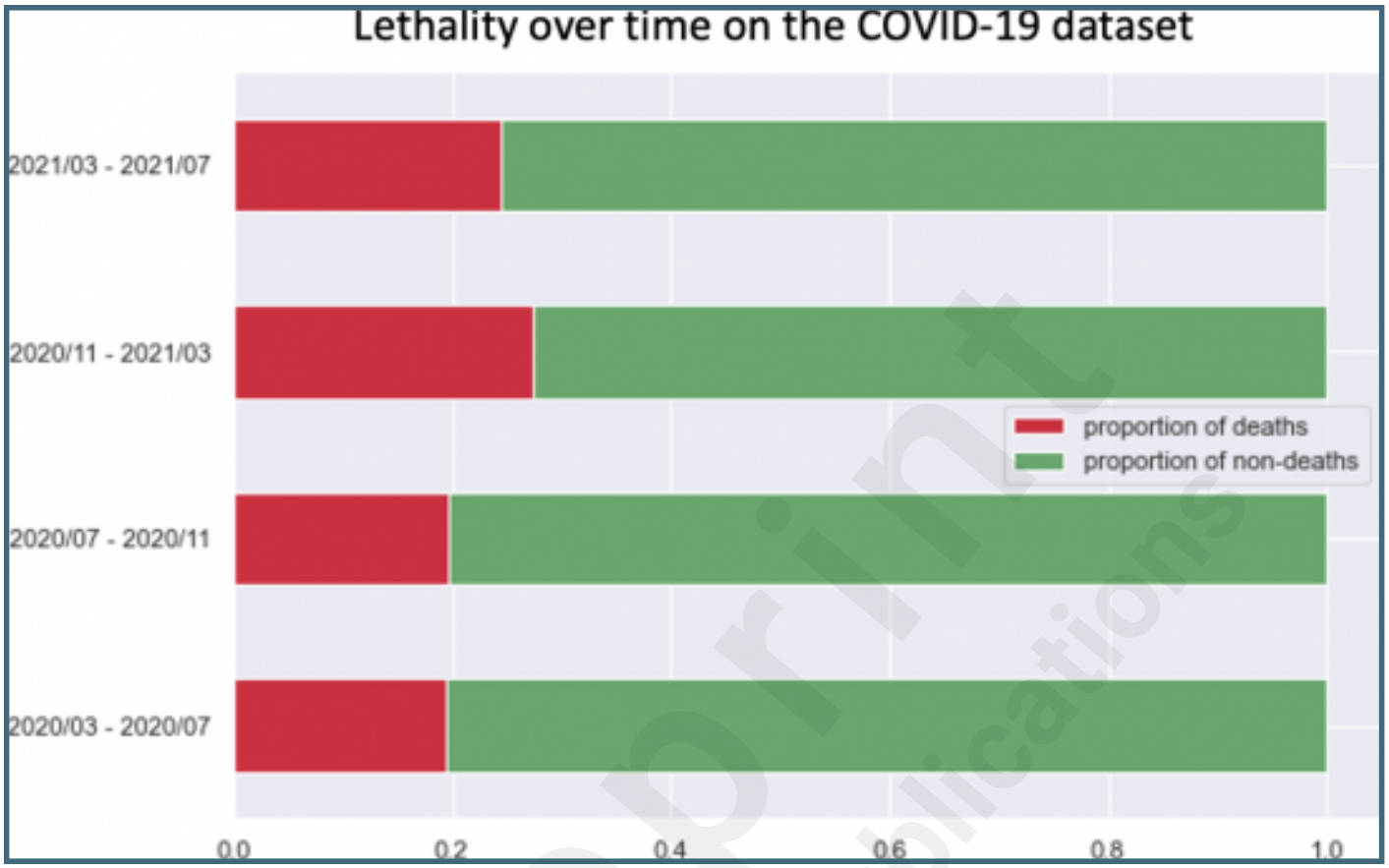


Cluster analysis of the Medical Information Mart for Intensive Care, version IV (MIMIC dataset). A: Top-5 highest-valued features per cluster. B: Relative frequency of each cluster over time.

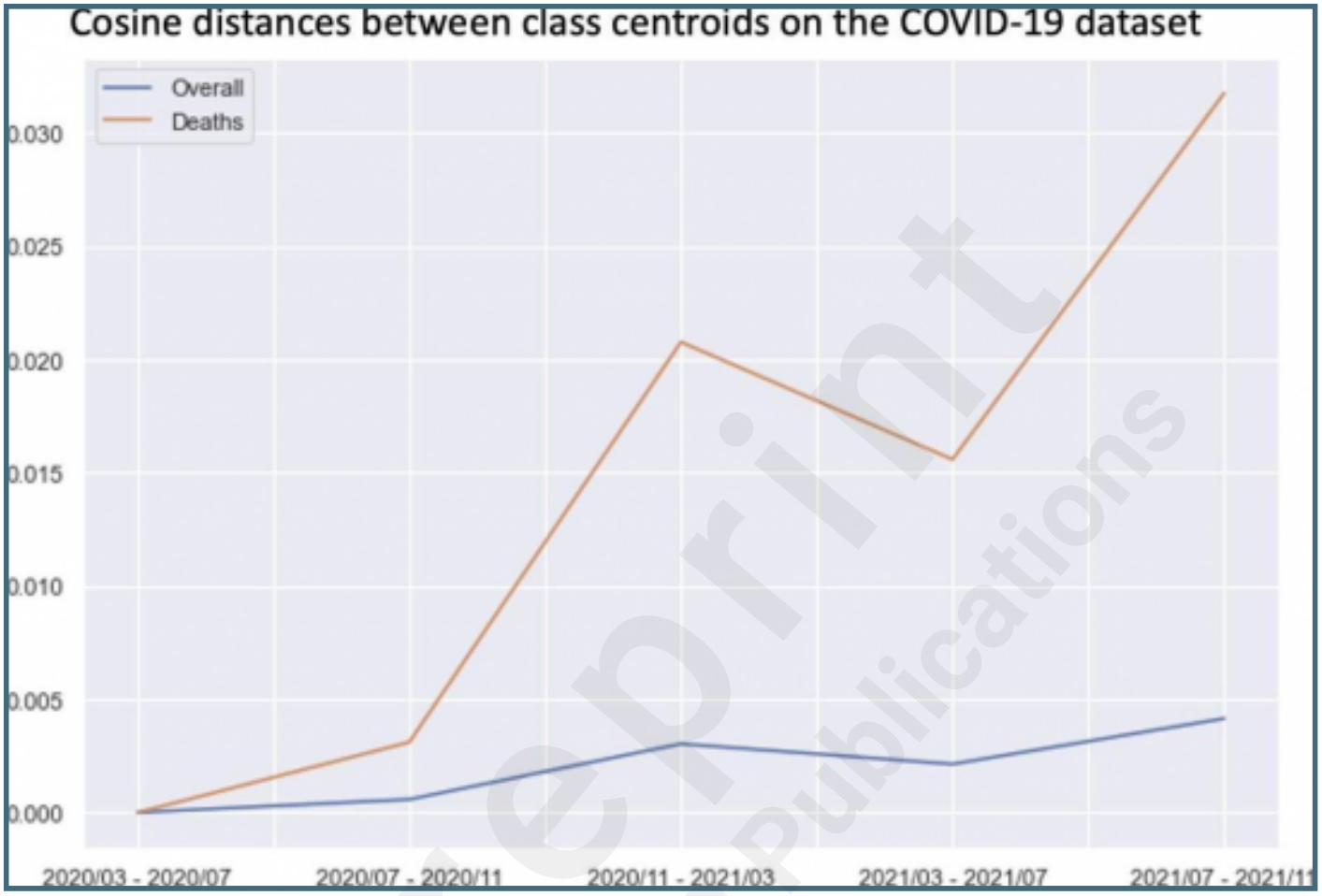


Different drift detection metrics over time on the Brazilian COVID-19 Registry dataset.

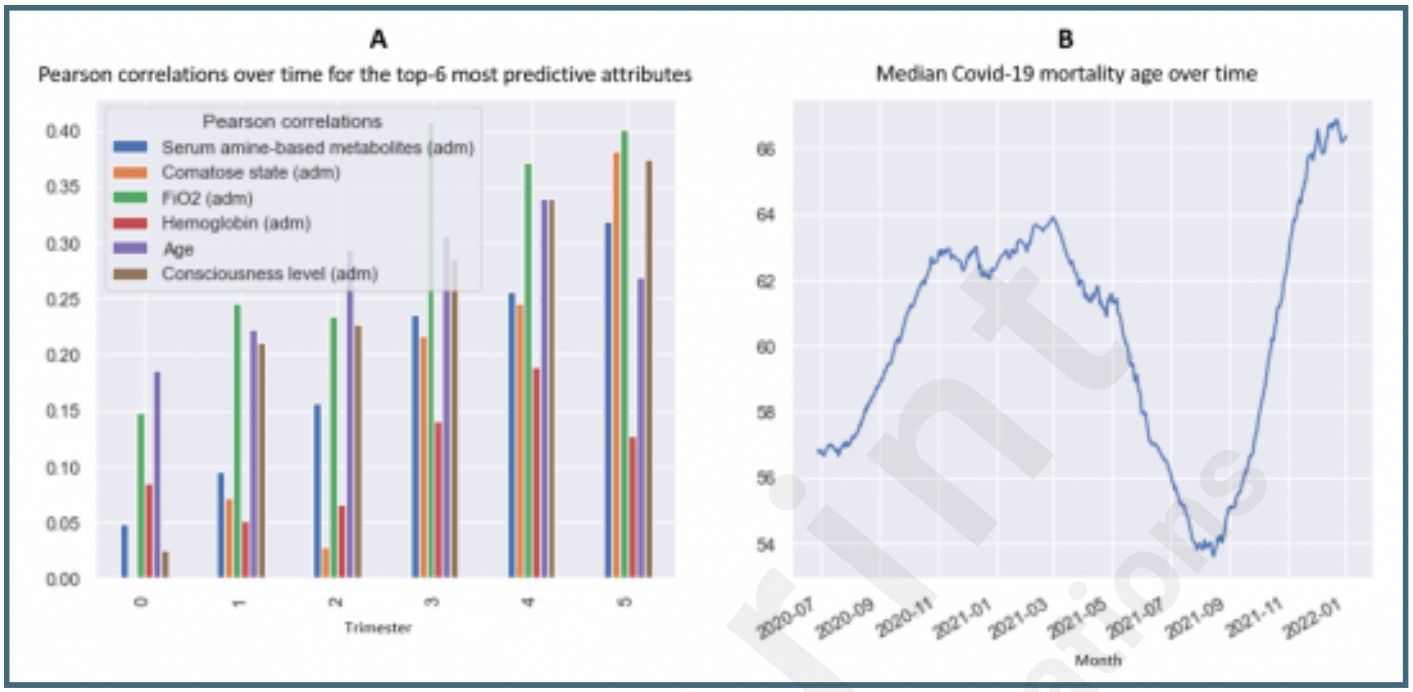
Lethality over time in the Brazilian COVID-19 Registry dataset.



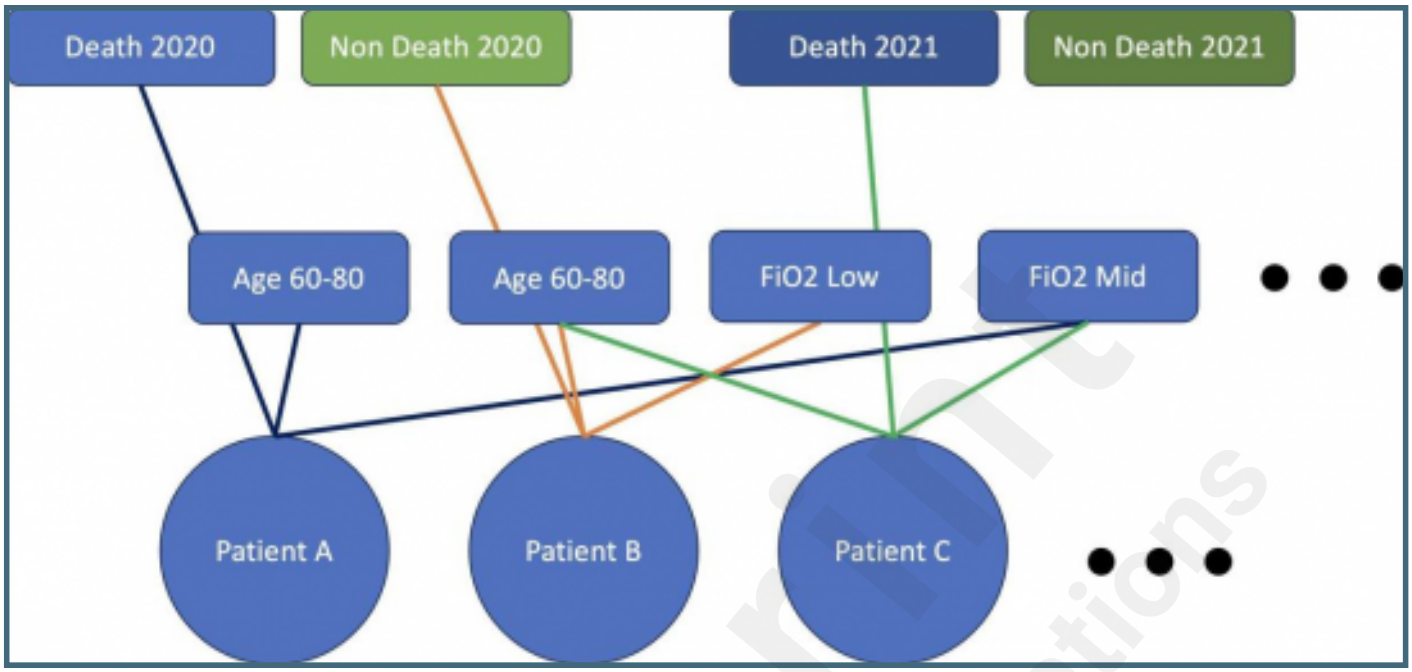
Drift of the arithmetic means of the dying patients versus the overall population over time, as measured by cosine distances between each class's means on each time chunk over time, on the Brazilian COVID-19 Registry dataset.



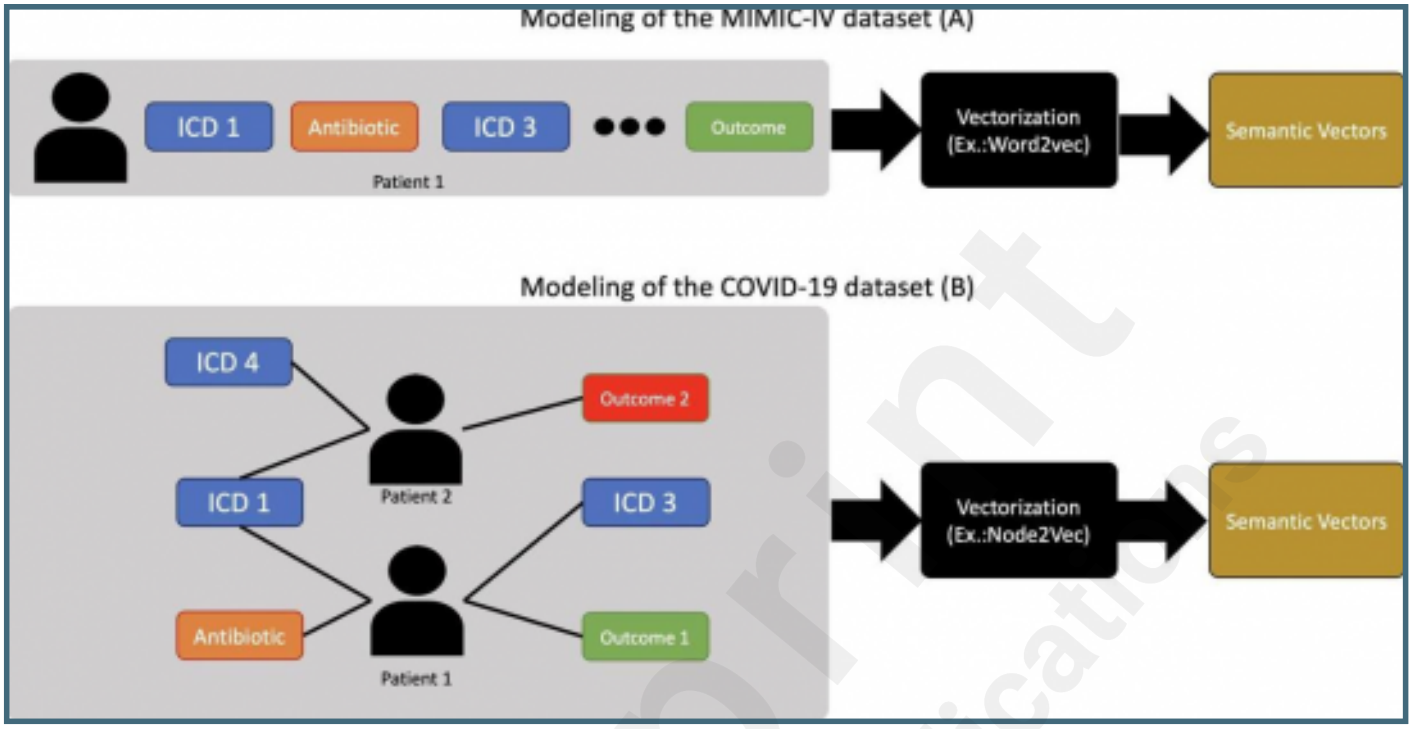
A: Pearson correlations over time for the overall top-6 most predictive variables on the Brazilian COVID-19 Registry dataset.
B: Median age of the COVID-19's hospitalized dying patients.



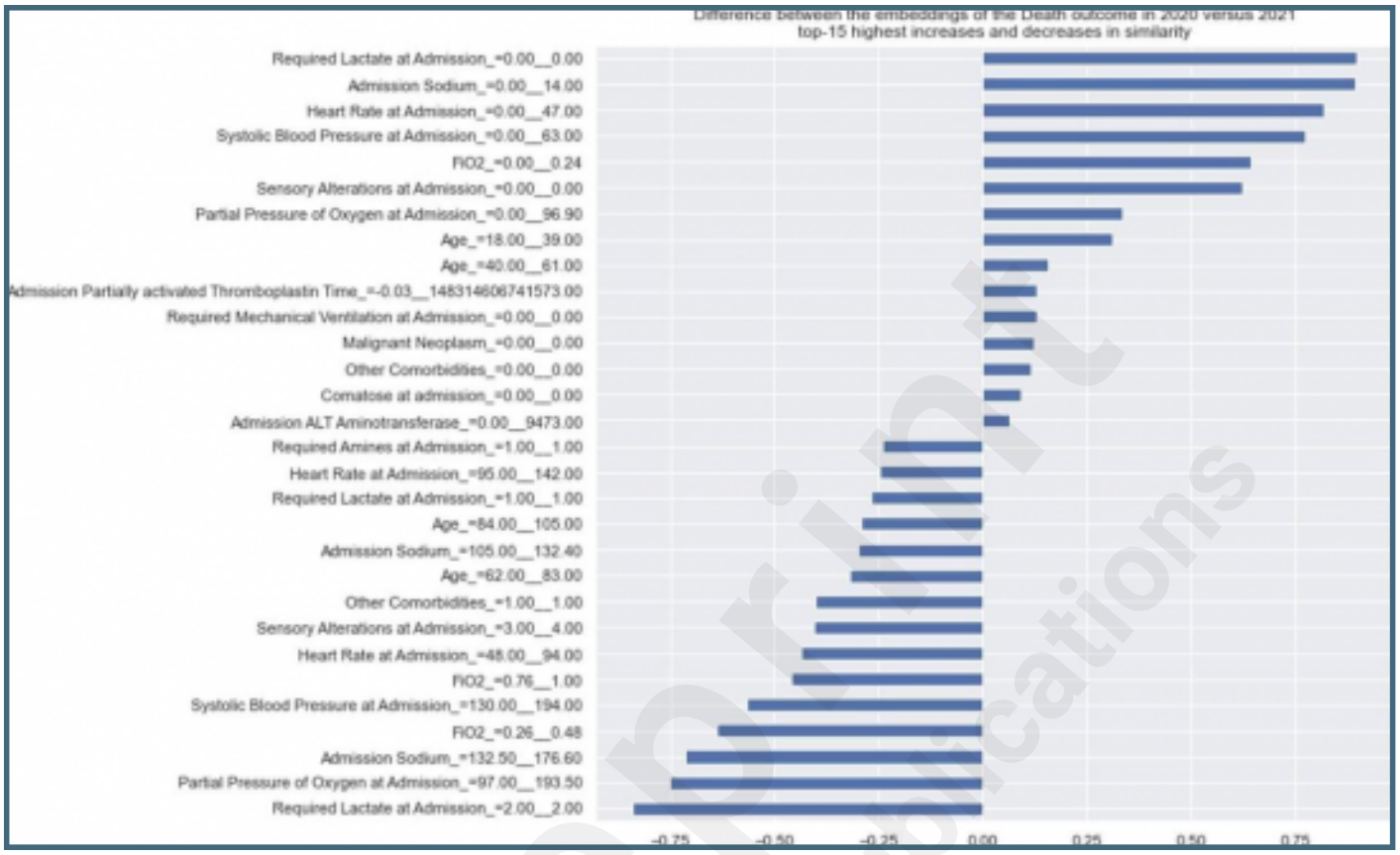
Example of how to build the patient graph with tokenized dependent variables and temporal outcome tokens.



(A) Modeling of the Medical Information Mart for Intensive Care, version IV (MIMIC-IV) dataset as an ordered sequence of patient tokens. (B) Modeling of the Brazilian COVID-19 Registry dataset as a graph connecting multiple patients through their common token.



Top-15 largest increases and decreases in similarity between the "Death" tokens for 2021 and 2020 on the Brazilian COVID-19 Registry dataset.



Cluster analysis of the Brazilian Brazilian COVID-19 Registry dataset. A: Top-5 highest-valued features per cluster. B: Relative frequency of each cluster over time.