Journal of Biomedical Informatics 136 (2022) 104242

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Original Research

Multisource and temporal variability in Portuguese hospital administrative datasets: Data quality implications

Júlio Souza [a,b,*], Ismael Caballero [c], João Vasco Santos [a,b,d], Mariana Lobo [a,b], Andreia Pinto [a,b], João Viana [a,b], Carlos Sáez [e], Fernando Lopes [a,b], Alberto Freitas [a,b]

[a] Department of Community Medicine, Information and Health Decision Sciences (MEDCIDS), Faculty of Medicine, University of Porto, Porto, Portugal
[b] Center for Health Technology and Services Research (CINTESIS), Faculty of Medicine, University of Porto, Porto, Portugal
[c] University of Castilla-La Mancha, Ciudad Real, Castilla-La Mancha, Spain
[d] Public Health Unit, ACES Grande Porto V - Porto Ocidental, ARS Norte, Portugal
[e] Biomedical Data Science Lab, Instituto Universitario de Tecnologías de la Información y Comunicaciones (ITACA), Universitat Politècnica de València (UPV), Spain

ARTICLE INFO

Keywords:
Data quality
Clinical coding
Data variability
Clinical classification software
International classification of diseases

ABSTRACT

Background: Unexpected variability across healthcare datasets may indicate data quality issues and thereby affect the credibility of these data for reutilization. No gold-standard reference dataset or methods for variability assessment are usually available for these datasets. In this study, we aim to describe the process of discovering data quality implications by applying a set of methods for assessing variability between sources and over time in a large hospital database.

Methods: We described and applied a set of multisource and temporal variability assessment methods in a large Portuguese hospitalization database, in which variation in condition-specific hospitalization ratios derived from clinically coded data were assessed between hospitals (sources) and over time. We identified condition-specific admissions using the Clinical Classification Software (CCS), developed by the Agency of Health Care Research and Quality. A Statistical Process Control (SPC) approach based on funnel plots of condition-specific standardized hospitalization ratios (SHR) was used to assess multisource variability, whereas temporal heat maps and Information-Geometric Temporal (IGT) plots were used to assess temporal variability by displaying temporal abrupt changes in data distributions. Results were presented for the 15 most common inpatient conditions (CCS) in Portugal.

Main findings: Funnel plot assessment allowed the detection of several outlying hospitals whose SHRs were much lower or higher than expected. Adjusting SHR for hospital characteristics, beyond age and sex, considerably affected the degree of multisource variability for most diseases. Overall, probability distributions changed over time for most diseases, although heterogeneously. Abrupt temporal changes in data distributions for acute myocardial infarction and congestive heart failure coincided with the periods comprising the transition to the International Classification of Diseases, 10th revision, Clinical Modification, whereas changes in the Diagnosis-Related Groups software seem to have driven changes in data distributions for both acute myocardial infarction and liveborn admissions. The analysis of heat maps also allowed the detection of several discontinuities at hospital level over time, in some cases also coinciding with the aforementioned factors.

Conclusions: This paper described the successful application of a set of reproducible, generalizable and systematic methods for variability assessment, including visualization tools that can be useful for detecting abnormal patterns in healthcare data, also addressing some limitations of common approaches. The presented method for multisource variability assessment is based on SPC, which is an advantage considering the lack of gold standard for such process. Properly controlling for hospital characteristics and differences in case-mix for estimating SHR is critical for isolating data quality-related variability among data sources. The use of IGT plots provides an advantage over common methods for temporal variability assessment due its suitability for multitype and multimodal data, which are common characteristics of healthcare data. The novelty of this work is the use of a set of methods to discover new data quality insights in healthcare data.

* Corresponding author at: Rua Dr. Plácido da Costa, 4200-450 Porto, Portugal.
E-mail address: juliobsouza@med.up.pt (J. Souza).

https://doi.org/10.1016/j.jbi.2022.104242
Received 22 July 2021; Received in revised form 18 August 2022; Accepted 6 November 2022
Available online 11 November 2022

1. Introduction

Healthcare administrative data are routinely collected during patient encounters in several settings, ranging from primary care and medical prescription to inpatient care [1]. This type of data is a major source for estimating indicators and outcomes for the assessment of the quality of care, also being reused for financing, research, epidemiological estimates and policy making. However, several data quality (DQ) issues have been reported within healthcare administrative datasets [2]. These datasets typically present a large among of coded information, for example based on the International Classification of Diseases (ICD), and the process of deriving clinical codes itself from health records is a very complex and naturally prone to errors task, often resulting in DQ issues.

Data variability among multiple sources (e.g., hospitals) may occur under natural circumstances and it is expectable at some degree, but it can also indicate DQ issues affecting data credibility. Variability in data distributions not only occur between different sources, but also over time, which in turn may lead to inaccurate, irreproducible or invalid conclusions if the data is used for research or decision-making [3–6]. Variability not related to natural circumstances may be attributed to several existing barriers, which are either systematic, such as lack or nonadherence to guidelines or data definitions, lack of standards in health care information systems and electronic health records (EHR), or due to random circumstances, -such as typing or transcription errors [7], or failures in coding diseases linked to the level of quality of inpatient documentation [8]. Thus, properly assessing and monitoring the stability of sources is essential for users to understand the data, to identify problems and biased sources, to discover patterns and to make decisions during the reutilization process [5].

Several studies have monitored variability in healthcare data, being generally performed using classical statistical methods [9–11], comparing populations' statistics [4,12], describing the distributions of individual variables [13] or using reference datasets [14]. The use of measurements such as the coefficient of variation, or other non-parametric equivalents, such as the quartile coefficient of dispersion, is also common in multisource variability assessment. However, such methods have some limitations, especially related to the loss of information as a consequence of summarizing data distributions into single scalar metrics, thereby affecting the capability of producing better insights. The coefficient of variation can be affected by the type or scale of the variables, whereas the quartile coefficient of dispersion may not reflect the shape of the variables' probability distribution function. Classical statistical tests are also used to detect differences among univariate data samples, but these may not be suited for multivariate, multimodal or multitype data, where other non-parametric information theory-based methods might be used instead [5].

Temporal variability can also manifest DQ issues that poses further challenges for the secondary use of data, particularly for research and machine learning [15–20]. For example, changes in coding systems, such as ICD, or modifications of protocols and clinical guidelines, often result in variable data representations across multiple diseases over time [21]. To monitor the aforementioned issues, authors have traditionally relied on statistical process control methods aimed at detecting the time-points at which changes occur. Common approaches for temporal variability assessment are also based on classical statistical methods, including Shewart charts, Levey–Jennings charts and Westgard rules [22,23]. Apart from not being suitable for multitype and multimodal data, these methods may not be adequate in the context of big data [24–26]. Autocorrelation or time series–based approaches have been applied to detect changes and periodicity within summary statistics obtained from longitudinal batches of data [4,9,12,27], but these methods tend to incur loss of information, especially when using coded data, namely categorical variables with a high number of labels (e.g., ICD-9-CM codes), or among multimodal distributions, which is common in biomedical data [21]. Finally, visual techniques or gold-standard methods for variability assessment are not usually provided. In the same line, typically gold-standard reference datasets are usually not available [17].

These limitations would represent a relevant drawback in settings using clinical and biomedical data, which is characterized by high heterogeneity in terms of sources, data types and distributions. Also, these methods are not suitable to deal with Big Data, which has been a reality in several healthcare organizations due to the various sources for big data, including hospital records, results of diagnostic examinations, procedures and treatments performed, nursing reports, discharge notes, Internet of Things (IoT) devices, as well as evidence produced by biomedical and public health research [28].

Guided by this motivation, in this paper we describe the process of detecting DQ implications in healthcare data from its multisource and temporal variability by applying two sets of existing methods: a Statistical Process Control (SPC) based on Funnel Plots and a previously validated probabilistic temporal data quality control approach. These sets of methods are suggested to constitute a data quality assessment framework, applicable to any multisource and temporal data in the health domain, addressing the limitations of the more common approaches, namely: (i) methods that are suitable for data with multiple health data domains, sources, types and distributions; (ii) add resources to assess current multisource variability beyond classical metrics (e.g., coefficient of variation and quartile coefficients of dispersion); (iii) and provide methods for temporal variability assessment suited to multimodal data, allowing a straightforward detection of temporal inconsistencies in the data. For the sake of demonstration, the methods were systematically applied to a large multisite data source, the Portuguese National Hospital Morbidity Database, to detect and assess abnormal variability regarding coding diseases between hospitals and over time.

2. Materials and methods

2.1. Data source

Data assessed in this study was extracted from CSV files exported from the National Hospital Morbidity Database, which holds data on inpatient and outpatient episodes occurred in all mainland hospitals within the Portuguese NHS [29]. Clinically coded data from all inpatient episodes with a discharge date between January 1st 2011 and December 31st 2017 were considered. Only inpatient episodes labelled in the database as statistically valid were included, that is, with a hospital stay of at least 24 h, or shorter than 24 h for patients who died, left against medical advice, or were transferred to another institution. Hospitals with missing data for an entire year were immediately excluded.

Variables at patient level that were considered relevant for assessing coding variability included age, sex, hospital, admission and discharge dates, as well as principal diagnosis (e.g., disease representing the cause of hospital admission) according to both ICD-9-CM, used until 2016 in Portuguese hospitals, and ICD-10-CM, used afterwards [30]. Furthermore, variables representing hospital characteristics, namely geographic region, hospital complexity category provided by the Central Authority for Health Services (ACSS, *Administração Central do Sistema de Saúde*) and teaching status were later added in order to account the effect of institutional characteristics on data variability. Hospitals that are not categorized according to the ACSS complexity category were also excluded from analysis. More detailed explanations on institutional characteristics variables are provided in section 2.4.

The data does not contain patient's identification as was previously anonymized, not requiring a complete review by the ethics council. Confirmation that the data is indeed anonymous is given by the Portuguese Central Authority for Health Services at the time of their assignment.

2.2. Outcomes for data variability assessment

Multisource variability was assessed based upon condition-specific Standardized Hospitalization Ratios (SHR), which is explained in more details below, whereas the primary outcomes to assess temporal variability were monthly relative frequencies of condition-specific hospitalizations. The Agency for Healthcare Research and Quality (AHRQ) Clinical Classification Software (CCS) was used to group each episode (hospital admission) into meaningful and mutually exclusive groups representing specific diseases or clinical conditions (e.g., liveborn), and thereby SHR were calculated by each condition available in the CCS. The CCS classification encompasses a set of 275 clinical condition categories, and the definitions to assign each episode into a CCS group according to ICD-9-CM or ICD-10-CM principal diagnosis codes are publicly available [31]. In this paper, we presented the analysis for the top 15 most commonly hospitalization conditions in Portugal to ensure that each hospital presented enough episodes to allow more robust statistical comparisons.

2.2.1. Standardized hospitalization ratio (SHR)

The SHR was calculated using the indirect standardization method, which provides individual rates by comparing a hospital with the reference population, which consists in the sum of all hospitals in the database. A logistic regression model including the patient's age group, sex and interaction between age and sex, and variables representing hospital characteristics was used to estimate the probability of hospitalization due to a specific condition (CCS). These models were estimated using the function *glm* of the R base package *"stats"* [32]. Subsets were created for each one of the 275 conditions, and a dummy (binary) variable indicating the occurrence of hospitalization due to a specific CCS, was included, which in turn was used as dependent variable for the hierarchical models. Stratified sampling preserving the same distribution of hospitals, years and outcome occurrences was performed for each condition subset to produce the logistic regression models. Individual probabilities were summed over the set of patients being admitted in a hospital to derive the expected number of hospitalizations in that hospital for that specific condition. The standardized hospitalization ratio was calculated as the ratio between the observed and the expected number of hospitalizations by a given condition, which in turn is multiplied by the overall standardized event (hospitalization) rate of the entire population (considering all hospitals) to obtain the SHR. We computed goodness-of-fit statistics for the logistic regression models, which included the Brier's score to measure the overall models' performance and C-statistics to assess model discrimination.

Since standardize hospital rate and standardized hospital ratio are inter-related measurements differentiated by factor, hence we used the term SHR interchangeably.

2.3. Multisource variability assessment

Multisource variability assessment was performed by constructing funnel plots of SHR according to the Spiegelhalter's method [33]. This method assumes that variation can be expected in any context and can be divided into two types: (1) common cause variation, when refers to an expected and stable level of variation; and (2) special cause variation, which refers to the unexpected variation, which is linked to systematic deviations and reflects out-of-control institutions. The SPC concept employed in the funnel plot assessments defines statistical boundaries to separate common-cause variation from special-cause variation.

The funnel plot is usually used to plot a given quality indicator against a measure of its precision (e.g., sample size), along with a horizontal line representing an internal summary (benchmark), such as the average across healthcare institutions, and control limits at 95 % (approximately-two standard deviations) and 99.8 % levels (approximately-three standard deviations) are also drawn. The control limits specify a range in which the values of the indicator would be statistically placed given the data distribution. If a given institution falls outside the control limits, its performance is said to be out of the expected, considering the benchmark value [33–36].

For multisource variability assessment, we assumed that funnel plots of condition-specific SHR would be useful tools to identify outlying data patterns that are potentially related to DQ issues, as these rates do not directly reflect performance and allow to adjust the event rate according to the hospitals' patient population and other contextual biases. As standardized ratios are usually computed as the ratio between the observed and expected number of events (i.e., number of hospitalizations due to a specific disease), the observed number of events is assumed to be an observation from a Poisson distribution [37]. Thus, in this paper, we constructed funnel plots using condition-specific SHR according to the definitions stated above, assuming a Poisson distribution to characterize the SHR distribution across the hospitals. The exact formula to draw Poisson control limits was used to minimize loss of robustness due to reduced sample size [33].

When assessing multisource variability, it is important to address overdispersion, a common phenomenon in health data which can be clearly displayed when using funnel plots [38]. In the context of this work, overdispersion occurs when there is true heterogeneity between hospitals and the mix of patients they treat, resulting in a variance that goes beyond that expected due to sampling variation [39]. There are several reasons for the occurrence of overdispersion, namely: (i) indicators estimated from a large number of cases, resulting in statistically significant differences with no practical importance; (ii) indicators that essentially determined by policy choices; and (iii) when hospitals admit patients with distinct characteristics for which the logistic regression model does not sufficiently corrects [40]. Furthermore, apparent overdispersion can occur if there are genuine major differences in DQ [40]. In this work, it was important to account for overdispersion attributable to poor risk adjustment (different between hospitals and case-mix). Thus, we considered method described by Spiegelhalter (2012) [40], which estimates a hierarchical model to draw overdispersed control limits in order to account for overdispersion. In this approach, it is assumed that each hospital has its own true underlying rate (the "random effects"), which themselves are distributed around the overall average with a "between" standard deviation that is added to the'within' standard deviation for the construction of overdispersed control limits [41].The R package *"FunnelPlotR"* [42], which implements the methods described by Spiegelhalter (2005) [33], was used to compute condition-specific SHR using the indirect method and considering the individual predicted probabilities obtained from the Logistic models, as well as to display the funnel plots with both the Poisson and the overdispersed control limits.

Adjustment for hospital characteristics in multisource variability assessment: Although some case-mix adjustment was addressed by using a logistic regression model accounting for age and sex, the initial models did not include hospital characteristics that could eventually explain data variability beyond DQ issues. Thus, to further minimize the bias introduced by hospital heterogeneity, we tested the association of condition-specific SHR with the following hospital characteristics:

(i) **Hospital group**: a discrete variable representing the cluster a hospital belongs according to an official categorization proposed by the ACSS for Portuguese NHS hospitals, which was based upon a clustering of institutions into five groups (Groups B, C, D, E and F) according to a hierarchical clustering method following the standardization of variables explaining hospital costs, thereby related to the case-mix and complexity of the institutions [43].
(ii) **Geographic region**: a discrete variable representing which geographic region the hospital is located on, as these regions may present relevant differences in terms of population density, income, and available resources, thereby influencing condition-specific hospitalization ratios. We considered the NUTS II

categorization developed by Eurostat, which defined five regions for mainland Portugal: Norte, Centro, Lisbon and Vale do Tejo, Alentejo and Algarve [44].

(iii) **Teaching status**: a discrete variable indicating whether the hospital provides medical education and training to future and current health care professionals. This variable is important in a sense that teaching hospitals tend to treat patients with more severe and complex diseases, thereby this status may influence higher SHR for some illnesses [45].

The Kruskal–Wallis test was used to examine associations between the SHR and the variables mentioned above. If a significant association between the indicator and hospital characteristics is found, it is advised to reconsider the funnel plot construction and perform case-mix correction improvement [39].

Descriptive analysis for multisource variability in condition-specificSHR: Apart from constructing and displaying funnel plots of condition-specific SHR, we reported the hospitalization ratio ratio of variance to mean for each condition-specific SHR in order to obtain a normalized measure of dispersion, assuming that SHRs follow a Poisson distribution.

2.4. Temporal variability assessment

To assess temporal variability, we compared probability distributions of relative frequencies in condition-specific hospitalizations over different periods of time using the temporal data quality control approach proposed by Sáez et al [6,21]. This approach makes use of the Jensen-Shannon distance (JSD) to measure similarity between two probability distributions. In databases with low variability, JSDs among distributions would be small, whereas different or anomalous data distributions would mean higher variability, resulting in higher JSD values. The main advantages of using JSD are: (1) its probabilistic interpretation, where 0 means equal distributions and 1 means non-overlapping distributions, and (2) that the measurements are not affected by large sample sizes, offering a robust alternative to classical statistical tests [17]. Furthermore, this approach provides a technique to explore variability among temporal batches of data, namely the Information-Geometric Temporal (IGT) plots and Data Temporal Heatmaps (DTHs), which help in uncovering temporal trends in the data, as well as to identify abrupt or recurrent changes in data distributions over time or time periods with similar data distributions. To compute JSD measurements and construct the IGT plots and DTHs, we used the R package *"EHRtemporalVariability"* [21]. To use this package, data was transformed into a matrix representing monthly relative frequency hospitalizations per hospital and condition. These methods and tools have been successfully applied to assess DQ temporal variability issues in US [6,21], UK [46] and Spanish [47] healthcare data.

All analyses were performed using R version 3.6.2 and RStudio version 1.2.1335 (RStudio Team, Boston, Massachusetts, United States).

Fig. 1 presents a step-by-step explanation of the processes to estimate and display visualization tools for multisource and temporal variability assessment in clinical coded data.

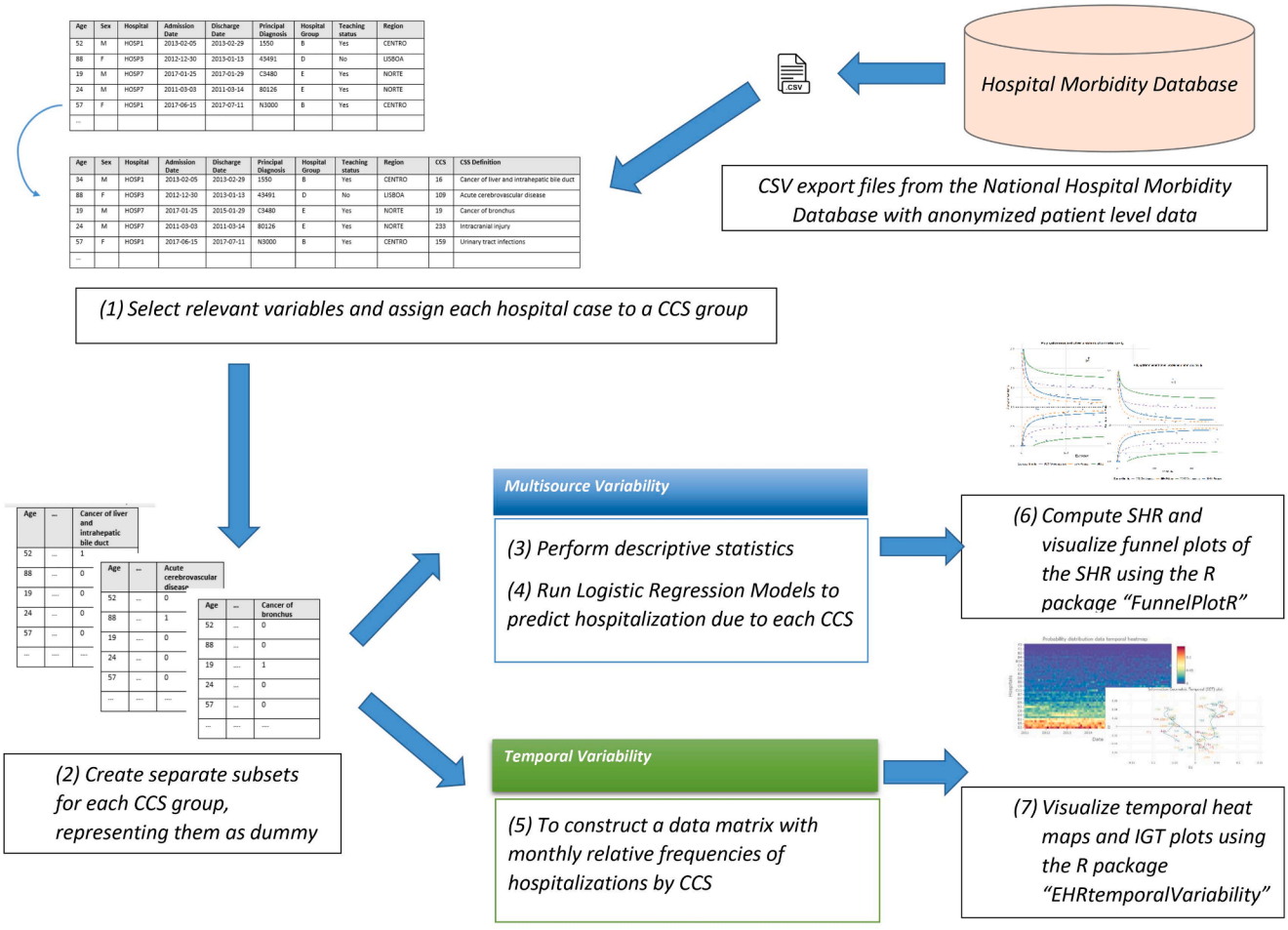

Fig. 1. Overview and steps considered in the methodological approach for multisource and temporal data variability in clinical coding.

3. Results

3.1. Data description

Data on 5,938,203 hospitalizations from 41 hospitals were included in the final sample. In Table 1, descriptive statistics of the sample considered for analysis is presented, including dispersion metrics of the crude hospitalization ratios (unadjusted) of the 15 most frequently admitted conditions, to provide basic insights regarding the magnitude of existing variation in coding condition-specific conditions.

The most frequently admitted condition in the period 2011–2017 was related to the "Liveborn" condition category (478,338 admissions, median crude ratio of 8,055 per 100,000 admissions). It is possible to observe, however, that the level of variation in data does not necessarily relate to the frequency, as conditions such as acute cerebrovascular disease presented the second-lowest interquartile range amongst top-15 conditions (668 per 100,000 admissions), despite being the third most frequently admitted condition, possibly indicating a lower degree of variability and more coding consistency across hospitals in comparison with other common conditions.

3.2. Multisource variability

To better quantify multisource variability that is not expected or explained by natural inter-hospital differences, condition-specific SHR were firstly computed using a logistic model adjusted for patient characteristics (age and sex). The Kruskall-Wallis tests results indicated that the association between the age-sex SHR with ACSS hospital group category was statistically significant ($p < 0.05$) for most of the diseases in the top-15 conditions, except for "Cardiac dysrhythmias" and "Biliary tract disease", whereas teaching status was statistically significant for "Acute bronchitis", "Biliary tract disease", "Complication of device", "Coronary atherosclerosis and other heart disease", "Fracture of neck of femur" and "Osteoarthritis". Amongst the top-15 conditions, the geographic region where the hospital is located was only significantly associated for "Acute myocardial infarction". The p-values obtained with the aforementioned tests can be found in Supplementary Table 1.

Thus, as statistically significant associations with the ACSS hospital group category and teaching status were verified for several conditions (represented by the CCS categories), we recomputed condition-specific SHRs by including these variables in the logistic models. Table 2 summarizes the overall variability in SHR for the 15 most frequently admitted conditions by means of the ratio of variance to mean, as well as goodness of fitness metrics (Brier's score and C-statistics) of the models employed for deriving the standardized rates.

The adjustment of SHR for hospital characteristics considerably reduced the ratio of variance to mean for all conditions, except for congestive heart failure, while it improved the overall discriminative ability (C-statistics) of the logistic models (Table 2). All logistic models presented a C-statistics above 0.68 (range: 0.68–0.97) following the adjustment for hospital characteristics. Moreover, Brier's score was close to 0 for all 15 conditions.

Fig. 2 shows the funnel plots of the SHR for three selected conditions before (Fig. 2A) and after (Fig. 2B) this adjustment. The benchmark (horizontal line) placed at 1 indicates fully control, where the expected and observed number of hospitalizations reported in a given hospital are equal. We highlighted the sources (hospitals) with the highest degree of deviation, which were those falling outside the 99.8 % overdispersed control limits.

Several hospitals presented a substantially higher or lower-than-expected SHR due to a specific disease. Some hospitals grouped into the same ACSS hospital group category, such as B3 and B4, were opposite outliers in coding acute myocardial infarction (Fig. 2), with the latter presenting a much lower-than-expected SHR, whereas its peer stood out in the opposite direction. Furthermore, substantial differences were observed in the funnel plots after adjusting for hospital

Table 1
Sample statistics in terms of crude hospitalization ratios for the 15 most frequently admitted conditions (CCS) in Portuguese public hospitals, 2011–2017.

CCS (Top-15 most frequently admitted)	Total number of hospitalizations	Overall Crude Ratio (per 100,000 hospitalizations)	Median Crude Ratio (per 100,000 hospitalizations)	Standard deviation Crude Ratio (per 100,000 hospitalizations)	Interquartile range (Q1 – Q3)
Liveborn	478,338	8,055.3	8,812.7	2,714.7	3,472.8 (6,857.1; 10,329.9)
Pneumonia (except that caused by tuberculosis or sexually transmitted disease)	291,639	4,911.2	4,978.4	2,193.6	2,783.6 (3,984.9; 6,768.5)
Acute cerebrovascular disease	177,225	2,984.5	3,159.3	1,060.4	667.9 (2,753.2; 3,421.0)
Biliary tract disease	162,354	2,734.1	2,968.3	1,157.4	1,450.3 (2,375.5; 3,825.8)
Urinary tract infections	142,344	2,397.1	2,384.1	1,055.2	1,307.1 (1,748.2; 3,055.3)
Congestive heart failure	120,642	2,031.6	2,102.4	1,087	1,252.8 (1,634.4; 2,887.3)
Acute bronchitis	96,737	1,629.1	1,482.5	981.2	823.2 (1,324.7; 2,147.8)
Abdominal hernia	94,995	1,599.7	1,631.5	853.4	1,026.3 (1,195.0; 2,221.3)
Fracture of neck of femur (hip)	87,893	1,480.1	1,744.8	620	856.0 (1,308.4; 2,164.4)
Acute myocardial infarction	85,701	1,443.2	1,255.7	806.2	882.0 (891.5; 1,773.5)
Osteoarthritis	82,681	1,392.4	1,534.1	864.4	1,148.0 (1,097.8; 2,245.8)
Coronary atherosclerosis and other heart disease	79,710	1,342.3	848.1	893	1,429.4 (346.3; 1,775.6)
Polyhydramnios and other problems of amniotic cavity	77,030	1,297.2	1,240.7	831.7	927.7 (830.6; 1,758.3)
Complication of device	70,609	1,189.1	871.9	469	532.7 (718.6; 1,251.3)
Cardiac dysrhythmias	70,271	1,183.4	1,181.1	573.5	843.7 (762.8; 1,606.5)

Table 2
Variability in condition-specific SHR for the 15 most frequently admitted conditions, 2011–2017.

Outcome	Ratio of variance to mean	Brier's Score	C-statistics	Ratio of variance to mean	Brier's Score	C-statistics
Liveborn	0.157	0.038	0.963	0.148	0.037	0.967
Pneumonia (except that caused by tuberculosis or sexually transmitted disease)	0.120	0.045	0.735	0.063	0.045	0.750
Acute cerebrovascular disease	0.122	0.028	0.738	0.066	0.028	0.750
Biliary tract disease	0.150	0.026	0.646	0.050	0.026	0.679
Urinary tract infections	0.139	0.023	0.684	0.087	0.023	0.692
Congestive heart failure	0.153	0.019	0.789	0.077	0.019	0.803
Acute bronchitis	0.285	0.016	0.733	0.152	0.016	0.755
Abdominal hernia	0.247	0.016	0.709	0.125	0.016	0.729
Fracture of neck of femur (hip)	0.145	0.014	0.838	0.114	0.014	0.850
Acute myocardial infarction	0.339	0.014	0.73	0.296	0.014	0.755
Osteoarthritis	0.341	0.014	0.772	0.242	0.013	0.798
Coronary atherosclerosis and other heart disease	0.481	0.013	0.772	0.586	0.013	0.802
Polyhydramnios and other problems of amniotic cavity	0.357	0.012	0.937	0.330	0.012	0.941
Complication of device	0.189	0.012	0.641	0.076	0.012	0.684
Cardiac dysrhythmias	0.228	0.012	0.702	0.148	0.012	0.717

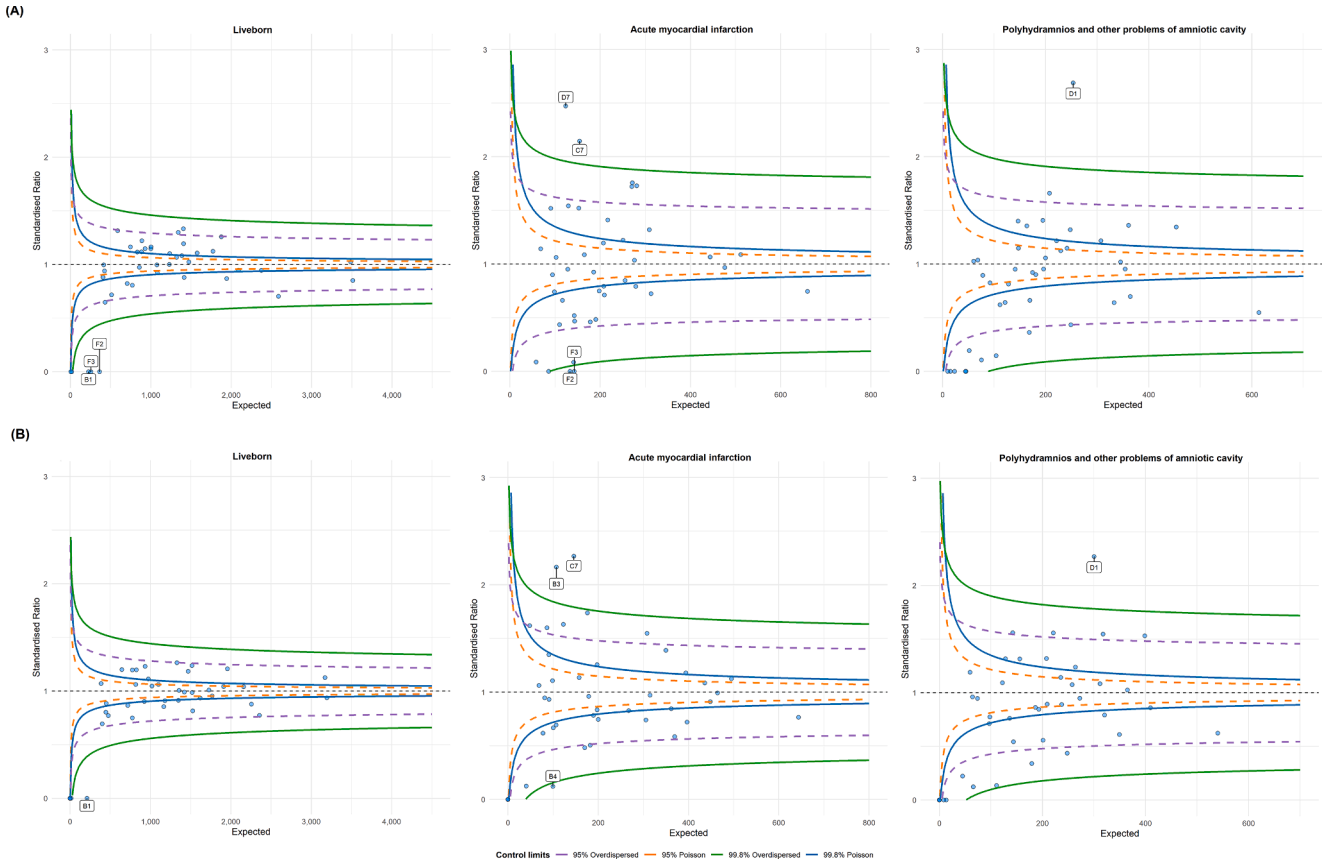

Fig. 2. Funnel plots of the SHR due to polyhydramnios and other problems of amniotic cavity, acute myocardial infarction and liveborn, before (A) and after (B) adjustment for hospital characteristics.

characteristics, either in shape or position of the outlying hospitals. For "Acute myocardial infarction" and "Liveborn", hospitals from group F are indicated as extreme outliers when adjusting SHR only for age and sex, even though such behavior is likely to be related to the level of specialization of these institutions, which are designated to cancer care. After adjusting for hospital characteristics, these institutions are no longer indicated as extreme outliers for these two conditions (Fig. 2B). Overall, the adjustment for hospital characteristics changed the set of outliers, except for "Polyhydramnios and other problems of amniotic cavity", in which hospital D1 remained as the only outlier, even after adjusting for hospital characteristics, with a much higher-than-expected SHR for that condition.

3.3. *Temporal variability*

To assess temporal variability, monthly relative frequencies of condition-specific hospitalizations (considering the discharge date) in each hospital were assessed. Fig. 3 presents the DTHs of monthly relative frequencies of hospitalizations for the studied period for four selected conditions within the top-15 ((A) Acute myocardial infarction; (B) Osteoarthritis; (C) Pneumonia; (D) Liveborn). The heat maps are ordered by relative frequency of condition-specific hospitalizations, where

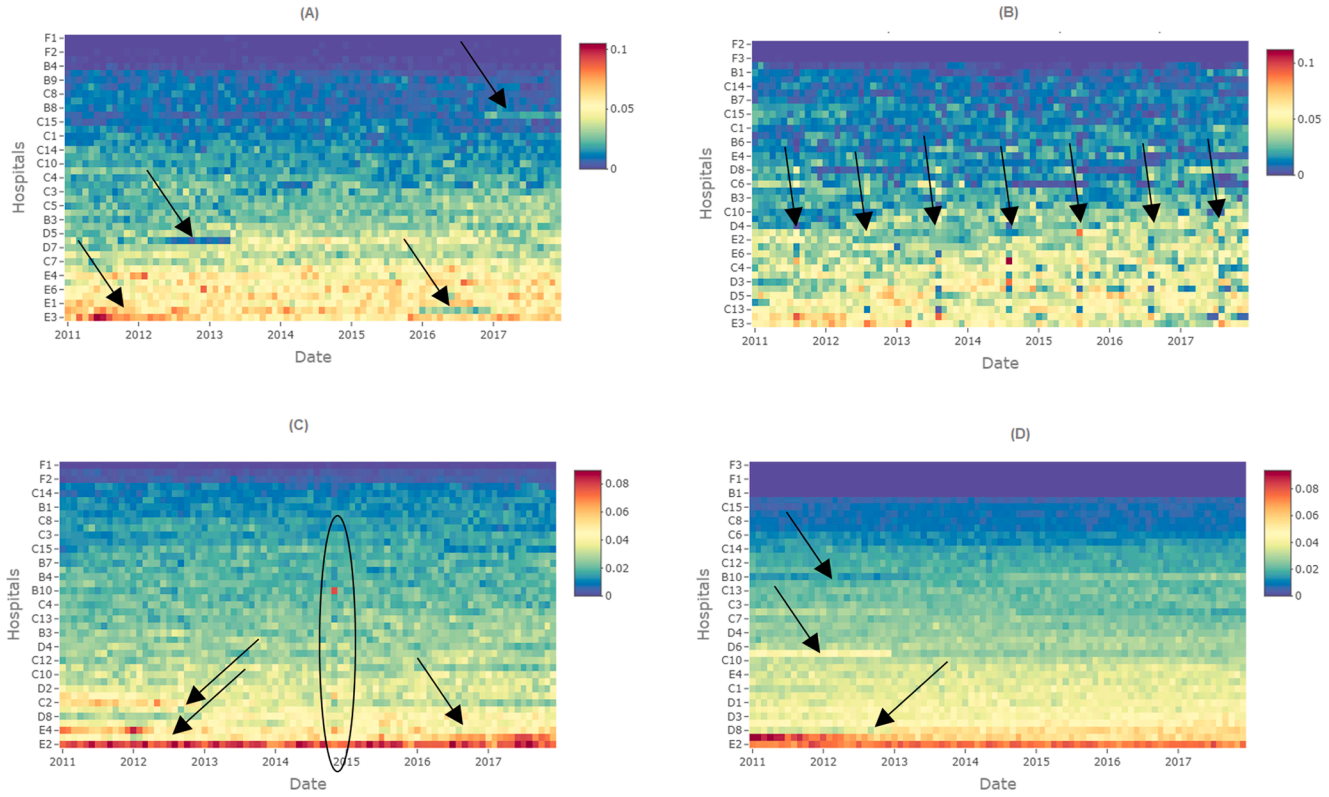

Fig. 3. Data Temporal Heatmaps of monthly relative frequencies of Portuguese hospitalizations due to Acute myocardial infarction (A), Osteoarthritis (B) Pneumonia (C) and Liveborn (D) 2011–2017.

hospitals located at the top are those presenting the lowest frequencies. The visual inspection of the temporal heat maps allowed the immediate detection of several affected distributions at institutional level over the years, highlighted by the arrows in Fig. 3. The "Liveborn" category (Fig. 3D) in general presented more stable relative frequencies during the studied period in comparison with the other presented conditions.

Regarding "Acute myocardial infarction", abrupt and prolonged decreases and increases in relative frequency are observed for several hospitals, as indicated by the arrows (Fig. 3A). Amongst the diseases in the top-15, "Osteoarthritis" was one of the conditions for which several discontinuities are observed, in which relative frequencies present sudden changes for the month of August, which is represented by the various isolated red, orange, or dark-blue dots that appear aligned in August of every year in the heat map, as indicated by the arrows (Fig. 3B).

The temporal evolution of "Pneumonia", the second most common admitted condition, also presented discontinuities in several hospitals, highlighting sudden decreases in hospitals E4 and C2 in the biennium 2012–2013, contrasting with a sudden increase in hospital D8 after 2013 (these cases are indicated by the arrows in Fig. 3C). Furthermore, it is also possible to observe outlier points aligned in the month of November 2014, as highlighted by the circle in Fig. 3C, with special emphasis for hospital B10 (Fig. 3C).

Although the relative frequencies of hospitalizations related to "Liveborn" were more homogeneous over time for most hospitals in comparison with other conditions, it is possible to detect some abrupt changes, namely for hospital C2 (line above C10, indicated by the arrow in Fig. 3D, where a prolonged yellow line is replaced by a prolonged green line until the end of the period), where a sudden decrease occurred after 2013, or hospital E3 (second line from the bottom, indicated by the arrow in Fig. 3D), which concentrated substantially higher relative frequencies (between 7 and 9 %) in the beginning of the period, but remained below 5 % until the end of the monitored period. An increase in relative frequencies of hospitalizations related to "Liveborn" was also detected for Hospital B10 after mid-2013 (as indicated by the arrow in Fig. 3D).

Furthermore, we constructed IGT plots to assess the overall temporal variability considering data from all hospitals. In Fig. 4, the IGT plot of the monthly relative frequencies of hospitalizations for four selected conditions among those within the top-15 ((A) Acute myocardial infarction; (B) Liveborn; (C) Congestive Heart Failure; (D) Osteoarthritis) can be seen for the entire period. Each point represents a data batch for a specific year and month, represented in the *yym* format, in which the first 2 numbers represent the year, and the last character indicates the month (Month abbreviations: {'J', 'F', 'M', 'A', 'm', 'j', 'x', 'a', 'S', 'O', 'N', 'D'}.). Colors indicate the annual seasons, with darker gradients indicating winter months and warmer gradients indicating summer months. The distance between monthly data points represents the JSD between their values (relative frequency of hospitalizations).

Overall, probability distributions changed over time for most diseases within the top-15, although with different intensities. In general, the plots suggest that the chronological order was a factor for increased similarity. Some of them presented accentuated shifts, such as "Acute myocardial infarction", where the flow of data points evolved irregularly over the studied period, and two abrupt changes were observed in 2013 and 2016, forming three temporal clusters (I - before 2013, II - between 2013 and 2016, and III - after 2016) (Fig. 4A). Other conditions with noticeable changes in temporal probability distributions included "Liveborn", in which an abrupt change can be observed after 2012 (Fig. 4B), and "Congestive heart failure", which presented a more distinct data distribution in 2017 in comparison with the remaining period (Fig. 4C).

Despite some discontinuities observed at hospital level in the DTHs, amongst the top-15 conditions, osteoarthritis was the one for which no clear abrupt change in data distributions was observed, forming an IGT plot composed of a single and more compacted cluster containing monthly data from the entire period (Fig. 4C). Nevertheless, outliers exclusively comprising data for the month of August (all years) were

Fig. 4. IGT plot of monthly relative frequency of Portuguese hospitalizations due to Acute myocardial infarction (A), Liveborn (B), Congestive heart failure (C) and Osteoarthritis (D), 2011–2017. Regarding Acute myocardial infarction (A), it is possible to observe two moments of sudden changes, one in 2013 (transition from cluster I to cluster II) and in 2016 (transition from cluster II to III). In terms of Liveborn admissions (B), there are clearly two clusters of similar temporal data, following an abrupt change in 2013. For Congestive heart failure (C), probability distributions started to differ in 2016, widening in 2017, whose points are placed distant from the remaining data; Osteoarthritis (D) was the condition with the least variation in probability distribution over time, forming a single and more compacted cluster composed of data of the entire period. Nevertheless, all data points for the month of August appear as outliers, possibly indicating potential DQ issues.

observed in the IGT plots for this condition.

4. Discussion

The purpose of this article was to describe the process of detecting DQ implications from applying a set of methods in the assessment of multisource and temporal variability in healthcare data. To this aim, we systematically applied the methods in a comprehensive multisite repository, the Portuguese Hospital Morbidity Database, comprising historical hospitalization data from all public hospitals in mainland Portugal. However, the methodological approaches used in this paper can be extended and reproduced for other healthcare data and clinical domains, addressing some gaps of common methods in the literature for variability assessment. Furthermore, in this paper, we sought to reinforce the importance of the systematic assessment of multisource and temporal variability as a key aspect of DQ among healthcare data, mainly when multisite data sources are considered.

The results obtained with the application of visualization tools such as IGT plots, clearly showed affected temporal data patterns that could be potentially related to the recent transition to ICD-10-CM in Portugal. The transition process to ICD-10-CM started in Portugal in August 2016, with three public hospitals being selected as pilots to implement the new coding system in October 2016, whereas the remaining hospitals would shift to ICD-10-CM by January 2017 [30]. Moreover, this new coding system adds more complexity into clinical coding tasks by offering increased granularity and a much higher number of codes to further specify the several diseases when compared to the previous version (ICD-9-CM) [30]. However, this impact seems to widely differ between conditions, as diseases such as acute myocardial infarction and congestive heart failure presented a much more accentuated change in probability distributions for 2016 and 2017 data in comparison with other conditions, such as osteoarthritis or liveborn.

It is important to point out that some comorbid conditions might occur in an inpatient episode as comorbidity or subsequent diagnosis following admissions for other causes. Nevertheless, chronic conditions are recorded inconsistently in hospital administrative datasets [48], and the usage of comorbidities as principal or secondary diagnoses may cause hospitals to present a substantially lower frequency of such conditions in the secondary diagnoses' fields, and vice-versa. For instance, in funnel plot assessment, hospitals B3 and B4 were opposite outliers in terms of SHR due to acute myocardial infarction, as some ICD codes related to this disease can be regarded as a comorbidity [49]. In this case, the latter presented a much lower-than-expected hospitalization ratios, whereas the first presented a much higher-than-expected (Fig. 2), despite belonging to the same hospital category and thereby no significant differences in their complexity level are supposed to occur. This aspect may severely affect measurements of quality of care, such as 30-day acute myocardial infarction in-hospital mortality, which often includes only patients with a principal diagnosis of acute myocardial infarction in its calculation [50]. A similar situation can occur for pneumonia, whose accurate coding may be affected by the variation in usage of sepsis or respiratory failure as principal diagnosis [51].

Changes in software usage, such as DRG grouper versions, may also be a potential source of data variability. For several conditions, namely acute myocardial infarction and liveborn, the analysis of IGT plots and temporal heat maps highlights substantial changes in data patterns coinciding with changes in the DRG grouper versions, namely after January 2013, when the AP (All-Patient)-DRG version 21 was replaced

by version 27 [52], and after January 2015, when the APR (All-Patient Refined)-DRG was adopted in Portugal [53], introducing a novel patient stratification concept whose financial structure relies more on secondary diagnoses coding, which may have driven different coding behaviors. Extreme temporal changes in DRG groups have also been found in the literature for the US National Hospital Discharge Survey (NHDS) data [54].

The use of the funnel plots provided several practical advantages, namely its usability for any clinical domain, capacity to uncover and easily display anomalous patterns between sources. Another advantage is to use the SPC present in the funnel plots to establish a reference for variability comparison, especially considering the lack of gold standard or reference datasets for healthcare data. Moreover, we advise the use of funnel plots with standardized rates related to the frequency of an outcome (i.e., hospitalizations, disease prevalence) for multisource variability assessment to minimize the effects of natural variability. In the application presented in this paper, we considered hospitalization ratios, but the indicator and its calculation should be determined by the reuse purpose. We reinforce the importance of properly controlling the indicators for other external factors, as it may be a critical aspect for isolating variability that could be related to DQ issues. Finally, we employed logistic regression models to compute individual probabilities of the target event (occurrence of a condition-specific hospitalization), but other models that are suitable for multimodal distributions and larger samples can also be explored for this adjustment, namely probabilistic machine learning models. The choice of the adjustment model will largely depend on the types, quantity, and distribution of the variables on the dataset, as well as the sample size.

Additionally, the use of the suggested visualization tools for temporal variability assessment to detect affected data batches provides a generic and reproducible method to be integrated into data quality monitoring frameworks, especially when such evaluations occur over time. The IGT plots should be regarded as a powerful and complementary tool to the DTHs, as it allows a global view of data distribution of the entire dataset and pattern changes that are not always clear in heat maps. Furthermore, the JSD metric displayed in the IGT plots can be applied to any type of variables or in transformed data (i.e., Principal Component Analysis), and it is comparable across studies [17]. These tools can also be useful to check data inconsistencies and support hospitals (and other healthcare settings) in monitoring the effects of changes in a given system, protocols, standards, software and even to evaluate the impact of new coding systems, such as the ICD-10-CM. This R-based open-source tool can be used on different sources, including raw Electronic Health Records (EHRs) and other healthcare datasets. Furthermore, unlike other methods mostly based upon classical statistical approaches, it is a suitable tool for assessing multimodal and highly coded data, which are common characteristics of healthcare/biomedical data. Finally, the JSD is complementary to DHTs in the sense that it provides a single metric to represent the entire repository in a given time point and allows the identification of sudden temporal deviations in an objective way through visual inspection. Additionally, the DHTs described in Fig. 3 provide a straightforward exploratory analysis for the details causing the change patterns, which in some cases is more difficult to find in simpler time series.

Some important limitations regarding both set of methods should be considered. Concerning the temporal variability assessment, the suggested set of methods rely exclusively on visual inspection. In this light, clustering analysis can be further applied on the projected months to find subgroups and outlying batches, enhancing temporal variability assessment. Regarding multisource variability assessment, we recommend testing for different data distributions is recommended to choose the formula for estimating the funnel plot control limits that better suits the data being analyzed. Although it is recommended in the literature that standardized assumes Poisson distribution, in this work, the ratio indicates SHR with binomial distribution (below 1), which thereby may have influenced the results. Additionally, multisource comparison was based upon standardized ratios computed by the indirect method, in which individual rates compare a hospital with the reference population representing the sum of all hospitals [56]. Therefore, ratios are not stable if a specific hospital presents a substantially different age distribution from the reference population [56]. Once unexpected multisource or temporal variability is observed, further explanation is always required to identify possible causes. Checking for possible DQ issues should ideally occur after ensuring that contextual predictors related to natural data variability are addressed, namely differences in the patient case-mix across hospitals, hospital characteristics, as well as external factors influencing natural temporal variability, i.e., increased prevalence of respiratory diseases during winter months. For instance, DTH for pneumonia showed an anomalous change in November 2014, which appears to coincide with an increased influenza activity during the 2014–2015 winter season in Europe [55], thereby minimizing the possibility of DQ issues. As future work, extensive research on the refinement of the predictive models used to compute the expected rates for multisource variability assessment, in which probabilistic machine learning models can be explored, is required. Furthermore, additional work should also focus on methods to explain root causes of unexpected variability and provide actionable insights. To this aim, a clinical coding process reference model, including specific DQ management process will be proposed as part of our future work.

5. Conclusions

In this article, we describe the process of detecting DQ implications by applying a set of methods based on SPC and probabilistic temporal data quality control approaches for monitoring multisource and temporal variability, using a nationwide Portuguese hospital database. The presented methods are generalizable, empirically driven and based on SPC, which constitutes and advantage considering the lack of gold standard for these datasets. For many diseases, relevant changes in temporal data distribution were observed and appear to coincide with some factors impacting the data generation process, such as the transition to ICD-10-CM and the adoption of different DRG grouper software versions. Those seem to be key factors impacting the observed variability and should be more carefully investigated. The use of funnel plots with standardized ratios, DTHs and IGT plots provides generalizable and reproducible tools based on previous literature that can be useful for discovering abnormal patterns in healthcare data, helping to detect potential random or systematic issues affecting the quality and reuse of the data. The main contribution of this work is demonstrating the successfulness of systematically applying a set of multisource and temporal variability methods for healthcare data for detecting potential DQ implications for their reuse. Given the characteristics of these types of datasets, and when multisite data repositories are considered, this work reinforces the importance of assessing variability in DQ checks processes, also highlighting novel possibilities, namely the capacity of monitoring changes impacting the data-generation processes, such as transition to new coding systems or DRG grouper software. The novelty of this work is the use of a set of methods to discover new DQ insights in healthcare data.

CRediT authorship contribution statement

Júlio Souza: Conceptualization, Methodology, Software, Formal analysis, Data curation, Writing – review & editing. **Ismael Caballero:** Methodology, Writing – review & editing. **João Vasco Santos:** Conceptualization, Methodology, Writing – review & editing, Supervision. **Mariana Lobo:** Methodology, Writing – review & editing. **Andreia Pinto:** Writing – review & editing. **João Viana:** Data curation, Writing – review & editing. **Carlos Sáez:** Conceptualization, Methodology, Software, Writing – review & editing. **Fernando Lopes:** Writing – review & editing. **Alberto Freitas:** Conceptualization, Methodology, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors would like to thank the Central Authority for Health Services, I.P. (ACSS) for providing access to the data. The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was financed by FEDER-Fundo Europeu de Desenvolvimento Regional funds through the COMPETE 2020-Operacional Programme for Competitiveness and Internationalisation (POCI) and by Portuguese funds through FCT-Fundação para a Ciência e a Tecnologia in the framework of the project POCI-01-0145-FEDER-030766 ("1st.IndiQare-Quality indicators in primary health care: validation and implementation of quality indicators as an assessment and comparison tool"). In addition, we would like to thank to projects GEMA(SBPLY/17/180501/000293)- Generation and Evaluation of Models for Data Quality, and ADAGIO (SBPLY/21/180501/000061) – Alarcos' Data Governance framework and systems generation, both funded by the Department of Education, Culture and Sports of the JCCM and FEDER; and to AETHER-UCLM: A smart data holistic approach for context-aware data analytics focused on Quality and Security project (Ministerio de Ciencia e Innovación, PID2020-112540RB-C42). CSS thanks the Universitat Politècnica de València contract no. UPV-SUB.2-1302 and FONDO SUPERA COVID-19 by CRUE-Santander Bank grant "Severity Subgroup Discovery and Classification on COVID-19 Real World Data through Machine Learning and Data Quality assessment (SUBCOVERWD-19)."

References

[1] C. Doktorchik, M. Lu, H. Quan, C. Ringham, C. Eastwood, qualitative evaluation of clinically coded data quality from health information manager perspectives, Health Inform. Manage. J. 49 (1) (2020) 19–27.
[2] M.F. Lobo, et al., Protocol for Analysis of Root Causes of Problems Affecting the Quality of the Diagnosis Related Group-Based Hospital Data: A Rapid Review and Delphi Process, in: Á. Rocha, H. Adeli, L. Reis, S. Costanzo, I. Orovic, F. Moreira (Eds.), Trends and Innovations in Information Systems and Technologies. WorldCIST 2020. Advances in Intelligent Systems and Computing, vol. 1159, Springer, Cham, 2020, pp. 93–103.
[3] A.J. McMurry, S.N. Murphy, D. MacFadden, G. Weber, W.W. Simons, J. Orechia, J. Bickel, N. Wattanasin, C. Gilbert, P. Trevvett, S. Churchill, I.S. Kohane, K. W. Carter, SHRINE: enabling nationally scalable multisite disease studies, PLoS ONE 8 (3) (2013) e55811.
[4] M.G. Kahn, M.A. Raebel, J.M. Glanz, K. Riedlinger, J.F. Steiner, A pragmatic framework for single-site and multisite data quality assessment in electronic health record-based clinical research, Med. Care 50 (2012), https://doi.org/10.1097/MLR.0b013e318257dd67.
[5] C. Sáez, M. Robles, J.M. García-Gómez, Stability metrics for multisource biomedical data based on simplicial projections from probability distribution distances, Stat. Methods Med. Res. 26 (1) (2017) 312–336.
[6] C. Saez, P.P. Rodrigues, J. Gama, M. Robles, J.M. Garcıa-Gomez, Probabilistic change detection and visualization methods for the assessment of temporal stability in biomedical data quality, Data Min Knowl. Discov. 29 (4) (2015) 950–975.
[7] R.J. Cruz-Correia, P. Rodrigues, A. Freitas, F.C. Almeida, R. Chen, A. Costa-Pereira. Data quality and integration issues in electronic health records. In: Information Discovery on Electronic Health Records, Chapman and Hall/CRC. pp. 55–95, 2009.
[8] P. Hay, K. Wilton, J. Barker, J. Mortley, M. Cumerlato, The importance of clinical documentation improvement for Australian hospitals, Health Inf. Manag. 49 (1) (2020) 69–73, https://doi.org/10.1177/1833358319854185.
[9] G. Svolba, P. Bauer, Statistical quality control in clinical trials, Control. Clin. Trials 20 (6) (1999) 519–530.
[10] J.J. Gassman, W.W. Owen, T.E. Kuntz, J.P. Martin, W.P. Amoroso, Data quality assurance, monitoring, and reporting, Control. Clin. Trials 16 (2) (1995) 104–136.
[11] G.L. Knatterud, Management and conduct of randomized controlled trials, Epidemiol. Rev. 24 (1) (2002) 12–25.
[12] F. Bray, D.M. Parkin, Evaluation of data quality in the cancer registry: Principles and methods. Part I: comparability, validity and timeliness, Eur. J. Cancer 45 (5) (2009) 747–755.
[13] K.L. Walker, O. Kirillova, S.E. Gillespie, et al., Using the CER Hub to ensure data quality in a multi-institution smoking cessation study, J. Am. Med. Inform. Assoc. 21 (6) (2014) 1129–1135.
[14] N.G. Weiskopf, C. Weng, Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research, J. Am. Med. Inform. Assoc. 20 (1) (2013) 144–151.
[15] D. Agniel, I.S. Kohane, G.M. Weber, Biases in electronic health record data due to processes within the healthcare system: retrospective observational study, BMJ 361 (2018), k1479.
[16] L. Knight, R. Halech, Cç Martin et al, 2011. Impact of changes in diabetes coding on Queensland hospital principal diagnosis morbidity data. Health Statistics Centre, Queensland Health, Brisbane, Queensland, Australia, 2011.https://www.health.qld.gov.au/hsu/tech report/techreport 9.pdf.
[17] C. Sáez, O. Zurriaga, J. Pérez-Panadés, et al., Applying probabilistic temporal and multisite data quality control methods to a public health mortality registry in Spain: a systematic approach to quality control of repositories, J. Am. Med. Inform. Assoc. 23 (2016) 1085–1095.
[18] A. Wright, J.S. Ash, S. Aaron, et al., Best practices for preventing malfunctions in rule-based clinical decision support alerts and reminders: results of a Delphi study, Int. J. Med. Inform. 118 (2018) 78–85.
[19] M. Sugiyama, N.D. Lawrence, A. Schwaighofer, et al., Dataset shift in machine learning, The MIT Press, Cambridge, Massachusetts, US, 2017.
[20] J.G. Moreno-Torres, T. Raeder, R. Alaiz-Rodríguez, et al., A unifying view on dataset shift in classification, Pattern Recogn. 45 (2012) 521–530.
[21] C. Sáez, A. Gutiérrez-Sacristán, I. Kohane, J. M. García-Gómez, P. Avillach, EHRtemporalVariability: delineating temporal data-set shifts in electronic health records, GigaScience, Volume 9, Issue 8, August 2020, giaa079, https://doi.org/10.1093/gigascience/giaa079.
[22] W.A. Shewhart, W.E. Deming, Statistical Method from the Viewpoint of Quality Control, Dover, New York, 1986.
[23] J.O. Westgard, Basic QC Practices: Training in Statistical Quality Control for Medical Laboratories, Westgard QC, Madison, WI, 2010.
[24] L.G. Halsey, D. Curran-Everett, S.L. Vowler, G.B. Drummond, The fickle P value generates irreproducible results, Nat. Methods 12 (3) (2015) 179–185.
[25] R. Nuzzo, Statistical errors, Nature 506 (13) (2014) 150–152.
[26] M. Lin, H.C. Lucas, G. Shmueli, Too Big to Fail: large samples and the p-value problem, Inform. Syst. Res. 24 (4) (2013) 906–917.
[27] G.E. Box, G.M. Jenkins, G.C. Reinsel, et al., Time Series Analysis: Forecasting and Control, John Wiley & Sons, Hoboken, New Jersey, US, 2015.
[28] S. Dash, S.K. Shakyawar, M. Sharma, et al., Big data in healthcare: management, analysis and future prospects, J. Big Data 6 (2019) 54, https://doi.org/10.1186/s40537-019-0217-0.
[29] Directorate-General of Health, National Hospital Morbidity Database. http://dis.dgs.pt/2010/08/23/base-de-dados-nacional-de-grupo-de-diagnostico-homogeneo-gdh/ (accessed 02 December 2021).
[30] J.V. Santos, R. Novo, J. Souza, F. Lopes, A. Freitas, Transition from ICD-9-CM to ICD-10-CM/PCS in Portugal: An heterogeneous implementation with potential data implications. Health Information Management Journal [epub ahead of print], 2021.
[31] Agency for Healthcare Research and Quality, Clinical Classification Software (CCS) for ICD-9-CM. http://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp (accessed 06 December 2021).
[32] R: A language and environment for statistical computing. http://www.R-project.org/ (accessed 06 December 2021).
[33] D.J. Spiegelhalter, Funnel plots for comparing institutional performance, Stat. Med. 24 (8) (2005) 1185–1202, https://doi.org/10.1002/sim.1970.
[34] O. Hirsch O, N. Donner-Banzhoff, M. Schulz, M. Erhart, 2018. Detecting and Visualizing Outliers in Provider Profiling Using Funnel Plots and Mixed Effects Models-An Example from Prescription Claims Data, Int. J. Environ. Res. Public Health. 15(9):2015. doi:10.3390/ijerph15092015.
[35] T. Rakow, R.J. Wright, D.J. Spiegelhalter, et al., The pros and cons of funnel plots as an aid to risk communication and patient decision making, Br. J. Psychol. 106 (2015) 327–348, https://doi.org/10.1111/bjop.12081.
[36] E.K. Mayer, A. Bottle, C. Rao, et al., Funnel plots and their emerging application in surgery, Ann. Surg. 249 (2009) 376–383.
[37] B.N. Manktelow, S.E. Seaton, Specifying the probability characteristics of funnel plot control limits: an investigation of three approaches, PLoS ONE 7 (9) (2021) e45723.
[38] D.C. Dover, D.P. Schopflocher, Using funnel plots in public health surveillance, Population Health Metrics 9 (1) (2011) 58, https://doi.org/10.1186/1478-7954-9-58.
[39] I.W. Verburg, R. Holman, N. Peek, A. Abu-Hanna, N.F. de Keizer, Guidelines on constructing funnel plots for quality indicators: A case study on mortality in intensive care unit patients, Stat. Methods Med. Res. 27 (11) (2018) 3350–3366, https://doi.org/10.1177/0962280217700169.
[40] D.J. Spiegelhalter, et al., Statistical methods for healthcare regulation: Rating, screening and surveillance, J. R. Stat Soc. A Stat. 175 (1) (2012) 1–47, https://doi.org/10.1111/j.1467-985X.2011.01010.x.
[41] D.J. Spiegelhalter, Handling over-dispersion of performance indicators, Quality & Safety in Health Care 14 (5) (2005) 347–351.
[42] Package "FunnelPlotR". Funnel Plots for Comparing Institutional Performance https://cran.r-project.org/web/packages/FunnelPlotR/FunnelPlotR.pdf (accessed 06 December 2021).
[43] Administração Central do Sistema de Saúde (ACSS). Abordagem Metodológica https://benchmarking-acss.min-saude.pt/BH_Enquadramento/AbordagemMetodologica (accessed 06 December 2021).
[44] Eurostat. Regions and cities – Overview. https://ec.europa.eu/eurostat/web/regions-and-cities/overview (accessed 06 December 2021).

[45] M. Ali, R. Salehnejad, M. Mansur, Hospital heterogeneity: what drives the quality of health care, Eur. J. Health Econ. 19 (3) (2018) 385–408, https://doi.org/10.1007/s10198-017-0891-9.
[46] P. Rockenschaub, V. Nguyen, R.W. Aldridge, et al, 2020. Data-driven discovery of changes in clinical code usage over time: a case-study on changes in cardiovascular disease recording in two English electronic health records databases (2001–2015) BMJ;10:e034396. doi: 10.1136/bmjopen-2019-034396.
[47] F.J. Pérez-Benito, C. Sáez, J.A. Pérez-Benito, S. Tortajada, B. Valdivieso, J. M. García-Gómez, Temporal variability analysis reveals biases in electronic health records due to hospital process reengineering interventions over seven years, PLoS ONE 14 (8) (2019) e0220369.
[48] H. Assareh, H.M. Achat, J.M. Stubbs, V.M. Guevarra, K. Hill, Incidence and Variation of Discrepancies in Recording Chronic Conditions in Australian Hospital Administrative Data, PLoS ONE 11 (1) (2016) e0147087.
[49] H. Quan, V. Sundararajan, P. Halfon, et al., Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data, Med. Care 43 (11) (2005) 1130–1139.
[50] P. Asaria, P. Elliott, M. Douglass, Z. Obermeyer, M. Soljak, A. Majeed, M. Ezzati, Acute myocardial infarction hospital admissions and deaths in England: a national follow-back and follow-forward record-linkage study, Lancet Public Health 2 (4) (2017) e191–e201.
[51] M.B. Rothberg, P.S. Pekow, A. Priya, P.K. Lindenauer, Variation in diagnostic coding of patients with pneumonia and its association with hospital risk-standardized mortality rates: a cross-sectional analysis, Ann. Intern. Med. 160 (6) (2014) 380–388, https://doi.org/10.7326/M13-1419.
[52] Directorate-General of Health, Portaria n.° 163/2013. https://data.dre.pt/eli/port/163/2013/04/24/p/dre/pt/html (accessed 14 July 2021).
[53] Directorate-General of Health, Portaria n.° 234/2015. https://data.dre.pt/eli/diario/1/153/2015/0/pt/html (accessed 14 July 2021).
[54] C. Sáez, J.M. García-Gómez, Kinematics of Big Biomedical Data to characterize temporal variability and seasonality of data repositories: Functional Data Analysis of data temporal evolution over non-parametric statistical manifolds, Int. J. Med. 119 (2018) 109–124, https://doi.org/10.1016/j.ijmedinf.2018.09.015.
[55] E. Broberg, R. Snacken, C. Adlhoch, J. Beauté, M. Galinska, D. Pereyaslov, C. Brown, P. Penttinen, WHO European Region and the European Influenza Surveillance Network. Start of the 2014/15 influenza season in Europe: drifted influenza A(H3N2) viruses circulate as dominant subtype, Euro. Surveill. 20(4): 21023 (2015), https://doi.org/10.2807/1560-7917.es2015.20.4.21023.
[56] J. Souza, I. Caballero, J. V. Santos, M. F. Lobo, A. Pinto, J. Viana, C. Saez, A. Freitas, 2021. “Chapter 19 Measuring Variability in Acute Myocardial Infarction Coding Using a Statistical Process Control and Probabilistic Temporal Data Quality Control Approaches”, Springer Science and Business Media LLC.