Applying probabilistic temporal and multi-site data quality control methods to a public health mortality registry in Spain: A systematic approach to quality control of repositories

Carlos Sáez[1, 2], Oscar Zurriaga[3, 4, 5], Jordi Pérez-Panadés[3], Inma Melchor[3], Montserrat Robles[1], Juan M García-Gómez[1, 6]

[1]Grupo de Informática Biomédica, Instituto de Tecnologías de la Información y Comunicaciones ITACA, Universitat Politècnica de València, Spain

[2]Centre for Health Technologies and Services Research, University of Porto, Portugal

[3]Dirección General de Salud Pública, Conselleria de Sanidad, Valencia, Spain

[4]FISABIO – Salud Pública, Consellería de Sanidad, Valencia, Spain

[5]CIBERESP, Madrid, Spain

[6]Unidad Mixta de Investigación en TICs aplicadas a la Reingeniería de Procesos Sociosanitarios (eRPSS), Instituto de Investigación Sanitaria del Hospital Universitario y Politécnico La Fe, Valencia, Spain

[*]**Corresponding author:**
Carlos Sáez
IBIME-ITACA, Universitat Politècnica de València, Building 8G, Access B,
Camino de Vera s/n, Valencia 46022, Spain.
Email: carsaesi@ibime.upv.es

Author version of manuscript published in Journal of the American Medical Informatics Association
(http://dx.doi.org/10.1093/jamia/ocw010)

ABSTRACT

Objective

To assess the variability in data distributions among data sources and over time through a case study of a large multi-site repository as a systematic approach to data quality (DQ).

Materials and methods

Novel probabilistic DQ control methods based on information theory and geometry are applied to the Public Health Mortality Registry of the Region of Valencia, Spain, with 512 143 entries from 2000 to 2012, disaggregated into 24 health departments. The methods provide DQ metrics and exploratory visualizations for (1) assessing the variability among multiple sources and (2) monitoring and exploring changes with time. The methods are suited to big data and multi-type, multivariate, and multi-modal data.

Results

The repository was partitioned into two probabilistically separated temporal subgroups following a change in the Spanish National Death Certificate in 2009. Punctual temporal anomalies were noticed, due to a punctual increment in the missing data, along with outlying and clustered health departments due to differences in populations or in practices.

Discussion

Changes in protocols, differences in populations, biased practices, or other systematic DQ problems affected data variability. Even if semantic and integration aspects are addressed in data sharing infrastructure, probabilistic variability may still be present. Solutions include fixing or excluding data and analyzing different sites or time periods separately. A systematic approach to assessing temporal and multi-site variability is proposed.

Conclusion

Multi-site and temporal variability in data distributions affects DQ, hindering data reuse, and an assessment of such variability should be a part of systematic DQ procedures.

Author version of manuscript published in Journal of the American Medical Informatics Association
(http://dx.doi.org/10.1093/jamia/ocw010)

BACKGROUND AND SIGNIFICANCE

Data sharing among multiple sites is gaining importance,[1] and multi-institutional data-sharing infrastructure has been successfully developed in many cases.[2-6] However, the value of such multi-site repositories depends greatly on the quality of their data.[6-14]

Data quality (DQ) in multi-site repositories continues to pose a challenge to health practitioners and researchers,[9] who demand increasing access to complete and accurate data as well as evaluation tools and metrics.[10] Therefore, multi-institutional platforms such as the US Clinical Effectiveness Research Hub now incorporate specific DQ assurance processes.[6] In fact, DQ is even considered one of the main components of any integrated data repository.[9]

Several systematic reviews have been carried out seeking agreement on different dimensions of DQ to be assessed in data repositories.[12-14] Despite the different approaches found in recent literature reviews[12,13], the methods and the dimensions aimed at similar fundamental DQ problems such as missing information, inconsistency among individual observations, and incorrect or outdated information. In the case of multi-site repositories, semantic and integration aspects are generally the first DQ problems to be addressed.[3,15]

However, two more DQ problems that can be particularly serious for large multi-site repositories have, in our opinion, received insufficient attention or lack appropriate methods for their assessment. These problems are caused by possible differences, or variability, in the probability distributions of data (1) among different sources of data (different sites or different practitioners, for example) and (2) with time.

Variability in data distributions can have several causes: differences in data acquisition methods, protocols or health care policies; systematic or random errors during data input and management (e.g., errors related to other intrinsic DQ dimensions such as changes in data completeness or consistency); geographic and demographic differences in populations;[16] or even falsified data.[17] These differences, if found among different data sources, constitute multi-source variability, and if found over time, either through a single source or multiple sources, constitute temporal variability.

Multi-source or temporal variability, if unmanaged, may lead to inaccurate or irreproducible results[3,18,19] or even to invalid results.[11] The reuse of data in multi-site repositories for population studies, clinical trials, or data mining rests on the assumption that the data distributions are to some degree concordant irrespective of the source of data or of the time over which the data have been collected and therefore allows generalizable conclusions to be drawn from the data. Differences in data distributions, by making the above assumption questionable, may hinder the reuse of repository data and may complicate data analyses, bias the results, or weaken the generalizations based on the data.

Common methods of assessing multi-source variability consist of comparing statistics of populations such as the mean[8,11], describing the distributions of variables[6], or comparing the data to a reference dataset.[12] Besides, methods for assessing temporal variability,

Author version of manuscript published in Journal of the American Medical Informatics Association
(http://dx.doi.org/10.1093/jamia/ocw010)

originally based on quality control of industrial processes, include statistical monitoring used in clinical contexts like Shewart charts[20] or, in laboratory systems, Levey–Jennings charts and Westgard rules.[21] Most of these methods are based on classical statistical approaches, which may face two main problems. First, classical statistical tests may not be suitable for multi-type data (e.g., numerical and categorical variables), multivariate data (several variables that change simultaneously), and multi-modal data (distributions generated by more than one component, e.g., data from several disease profiles)—the very characteristics of biomedical data.[18,22] Second, classical statistical methods may not prove adequate for big data.[23-25] Besides, specific DQ metrics or visual methods for variability are not generally provided. Finally, for data from multiple sources, a gold-standard reference dataset may not be available. These reasons support the current need for generalizable, systematic, empirically driven, statistics-based, and validated DQ assessment methods for the reuse of electronic health records.[12]

Meeting the requirements mentioned above, we developed two sets of methods for both multi-source and temporal variability assessment.[18,19] These methods allow variability to be measured and visually explored, by comparing the probability distributions of the data using information-theoretic metrics. These methods have been evaluated earlier using simulated problems as well as real registries, including the UCI Heart Disease public dataset[26] and the US NHDS data.[27]

In the present study, we systematically apply these methods to a large, multi-departmental, Public Health Mortality Registry, aiming to (1) emphasize the importance of systematic assessment of multi-source and temporal variability in multi-site biomedical data repositories, (2) propose that such variability be considered an aspect of DQ, and (3) highlight the novel possibilities opened up by these state-of-the-art methods.

MATERIALS AND METHODS

Methods

The methods used in the present study fall into two groups, namely those for assessing multi-source variability[18] and those for assessing temporal variability.[19] The methods are based on the comparison of probability distributions of the variables among different sources or over different periods of time. The comparisons are made by calculating the information-theoretic *probabilistic distances* between pairs of distributions, in concrete terms, we use the Jensen-Shannon distance (JSD), a symmetrized and smoothed version of the Kullback-Leibler divergence.[28,29]. The JSD permits measuring differences either in univariate and multivariate data including numerical data, such as ages, categorical data, such as ICD codes, or a combination of them. The assumption is that in a repository with low variability, JSDs among distributions would be small whereas different or anomalous data distributions would mean higher variability. Additionally, the JSD is bounded between zero and one, making it comparable among studies: a value of one indicates that the compared distributions are disjoint, i.e., they do not share common values. Further, the JSD measurements are not

affected by large sample sizes. These properties offer a robust alternative to classical statistical tests where they may not be appropriate.[22]

The multi-source variability methods include two metrics. The first measures the dissimilarity of a data source to a global central tendency of sources, namely the Source Probabilistic Outlyingness (SPO) metric. The second measures the global variability among all the data sources in a repository, namely the Global Probabilistic Deviation (GPD) metric. The metrics are complemented with an exploratory visualization of the variability among data sources, namely the Multi-Source Variability (MSV) plot. These methods serve to highlight anomalous behaviors in the data of specific sources, detect groups of sources with similar data, or provide an indicator of concordance among data sources.

The temporal variability is assessed by means of two methods. The first is an exploratory visualization for the variability among temporal batches of data, namely the Information-Geometric Temporal (IGT) plot. It helps in uncovering temporal trends in the data, abrupt or recurrent changes in distributions, conceptually-related time periods (periods with similar data distributions), and punctual anomalies in the data of batches. The second method is an automated Statistical Process Control (SPC) algorithm to monitor changes in data distributions, namely the PDF-SPC. It permits controlling the degree of current variability to a reference state supported by a control chart.

The present study also includes a new method for monitoring multi-source variability over time, which involves calculating the SPO and GPD metrics, or the MSV plot, through continuous temporal batches.

An extended description of the methods is provided in Table 1. Additionally, sections 1 and 2 of Appendix A provide basic, illustrative examples of the methods, and section 3 describes the main equations.

Table 1 Description of the methods for multi-source and temporal variability assessment applied in this study.

Note Meaning of the symbols: ➢ Purpose of the method ❖ How it works.

Multi-Source Variability	
Common basis	A geometric simplex the points of which represent data sources and the lengths of the lines that join the points represent the JSDs between the distributions of those sources. The centroid of the simplex represents the *latent* average distribution of the sources in the repository. The derived metrics are normalized by their maximum possible value given the number of sources to be [0-1]-bounded and consistent with the JSD.
Methods	• ***Source probabilistic outlyingness (SPO) metric*** ➢ Measures the dissimilarity of the distribution of a single data source to the global average distribution ❖ It is calculated as the distance between the point that represents a given source and the simplex centroid. • ***Global probabilistic deviation (GPD) metric***

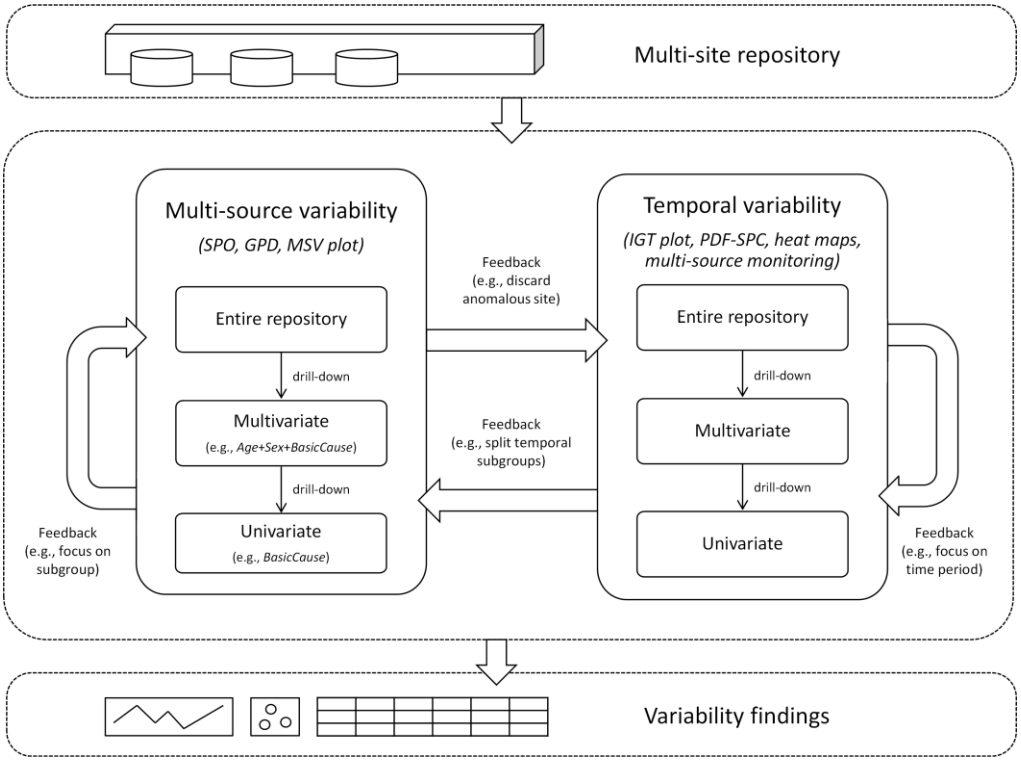
	➢ Measures the degree of global variability among the distributions of sources in a repository. ❖ It is calculated as the mean of the distances between each point that represents a source and the simplex centroid. • ***Multi-source variability (MSV) plot*** ➢ Visualizes the variability among data sources in a two-dimensional (2D) plot. ❖ The two components with largest variance of the simplex (which we named D1-simplex and D2-simplex) are projected using Multi-Dimensional Scaling (MDS)[30]. In the resultant 2D plot data sources are shown as circles in which the distance between two circles represents theJSD between their distributions, the radius of a given circle is proportional to the number of cases in the data source and its color indicates the SPO of the source.
	Temporal Variability
Common basis	Comparison of distributions through different batches of data in the repository, each batch representing a user-specified interval (weeks, months, years, etc.).
Methods	• ***Information geometric temporal (IGT) plot*** ➢ Visualizes the variability among time batches in a repository in a 2D plot. ❖ Time batches are positioned as points where the distance between them represents the JSD between their distributions, analogously to the MSV plot. To track the temporal evolution, temporal batches are labeled to show their date, supported by a smoothed timeline path, and colored according to their season • ***Probability distribution function statistical process control (PDF-SPC) algorithm*** ➢ Monitors changes in data distributions through an automated statistical process control (SPC), visualized in a control chart. ❖ Monitoring an upper confidence interval (e.g., one standard deviation) of the accumulated JSDs of time batches to a reference distribution (initially the first batch). The degree of change of the repository is classified into three states: in-control (distributions are stable), warning (distributions are changing), and out-of-control (recent distributions are significantly dissimilar to the reference). Warning states can be false alarms if the JSDs get closer to the reference once again, thus going back to the in-control state. However, when an out-of-control state is reached, a significant change is confirmed and the reference distribution is set to the current. • ***Temporal heat maps*** ➢ Facilitates a rapid and broad visualization of how the values of a variable evolve over time. ❖ These are 2D maps where the color of the pixel at a given (X,Y) position indicates the frequency (either absolute or relative) at which value Y was observed on date X.
	Combined methods
	➢ To monitor the multi-source variability over time ❖ Calculating the SPO and GPD metrics, or the MSV plot through continuous temporal batches.

Systematic approach

Based on our experience of applying the described methods we propose a systematic approach to assessing multi-source and temporal variability in repositories (Figure 1). In a top-down approach, one starts by analyzing the temporal or multi-site variability of the complete data set and then, based on the results of the analysis and prior knowledge of the repository, drills down to specific variables or groups of variables. The process can be cyclic, similar to an On-Line Analytical Processing (OLAP) exploratory analysis, navigating through different levels of granularity; for example, a temporal change found in the complete repository could be caused by a sudden bias within a single site. Such an anomalous site may require a specific temporal analysis, and excluding it may facilitate the discovery of other patterns or sources of variability.

Author version of manuscript published in Journal of the American Medical Informatics Association
(http://dx.doi.org/10.1093/jamia/ocw010)

Figure 1: Proposed Systematic Approach to Assessing the Temporal, Multi-Site Variability of Repositories of Biomedical Data Using Probabilistic Data Quality Control Methods.



Materials

The above approach was applied to the Public Health Mortality Registry of the Region of Valencia (MRRV), an autonomous region of Spain. The repository comprises the records related to a total of 512,143 deaths that occurred between 2000 and 2012 (inclusive), disaggregated by 24 health departments covering 542 cities and towns with a total of 4.7 million inhabitants on average, representing 11% of the population of Spain. The repository includes the variables that make up the Spanish National Medical Death Certificate, an official paper document completed by a physician after the death of a person, according to the recommendations of the World Health Organization (WHO).[31] Any information that may disclose the identity of the person was removed before the analysis.

The studied variables are listed in Table 2. The initial, intermediate, immediate and contributive causes are the sequential causes leading to death, known as `multiple causes', for which up to three values are entered depending on the case. Further on, empty values up to the three possibilities will be labeled as `not applicable (NA)´. The basic cause of death is the official cause taken into account for national and international mortality statistics and is generally coded afterwards by specialist staff based on the multiple causes.

According to the WHO recommendations for facilitating statistical analysis and comparison of the present work with other international studies, the causes of death were re-coded using the WHO International Classification of Diseases (ICD) version 10 Mortality Condensed List,[32]

Author version of manuscript published in Journal of the American Medical Informatics Association
(http://dx.doi.org/10.1093/jamia/ocw010)

which condenses the full range of ICD three-character categories into 103 manageable items. Because this list brings together both the top-level ICD chapters and their subgroups of diseases, the chapter-level classifications were discarded to avoid duplication and to facilitate proper statistical distribution. Accordingly, a total of 92 unique causes of death (plus an additional category, namely NA) were used in this study (section 4, Appendix A). Deaths that occurred outside the Region of Valencia during this period (totaling 6,816) were excluded, leaving us finally with 505,327 entries. The CONSORT[33] diagram of the study and tables of sample sizes are included Appendix A.

Table 2: Studied Variables of the Public Health Mortality Registry of the Region of Valencia

Variable	Description	Type
Age	Age in years at the time of death	Numerical integer
Sex	Sex of the person	Discrete {Male, Female}
ImmediateCause[1,2,3]	Disease or condition directly leading to death (one to three options)	ICD-10 List 1 code
IntermediateCause[1,2,3]	Morbid conditions, if any, giving rise to the above cause (one to three options)	ICD-10 List 1 code
InitialCause[1,2,3]	Disease or lesion that initiated the process that eventually resulted in the death (one to three options)	ICD-10 List 1 code
ContributiveCause[1,2,3]	Other significant conditions contributing to the death but not related to the disease or condition that caused death (one to three options)	ICD-10 List 1 code
BasicCause	Basic cause of death	ICD-10 List 1 code
Health department	Health department the person was assigned to (associated with the city of residence)	Discrete code

RESULTS

The results of applying the proposed systematic approach to quality control of the MRRV repository led to the following four groups of main findings (other additional findings are described in Appendix A).

Temporal Anomalies

We first analyzed the temporal variability of the multivariate MRRV repository as a whole using IGT plots. To simplify the analysis, all the variables were combined using the principal component analysis (PCA) dimensionality reduction method. Figure 2 (a) shows the IGT plot for 2000–2012 giving the distributions of monthly temporal batches. The distributions from January to March 2000 (arrows a, b, and c) are located at anomalous positions with respect to

the distributions for other months and according to the time flow. This indicates anomalous behavior of the data for these three months. Drilling down to specific variables, the anomaly was found in all multiple causes as well. In concrete terms, we found a punctual increment on unfilled data for these months, reaching almost 100% in some variables, probably because the entries in the paper certificates were not electronically coded during those months.

To avoid a possible bias in the results pertaining to the year 2000, we proceeded excluding the entire year for subsequent analyses, given the difficulty in recovering all the missing data.

Figure 2: IGT Plots of the Multi-Variate Repository (all variables) on Monthly Basis. Each point represents one batch of the repository labeled with its date in 'YYM' format (YY: the last two digits of the year, M: the month as given in the list of abbreviations at the end), and the distances among them represent the dissimilarity in their distributions. a) The period 2000–2012, where the months January to March 2000 (arrows a, b, and c) are at anomalous positions according to the time flow. b) The period 2001–2012, after discarding the data for 2000. A gradual conceptual change is seen from the start until 2009 (arrow d), at which point the change is abrupt (arrow e), splitting the repository into two temporal subgroups. The cool (blues) and warm (yellows and reds) colors indicate winter and summer months, respectively, indicating a seasonal effect which is specially observed in the last temporal subgroup.

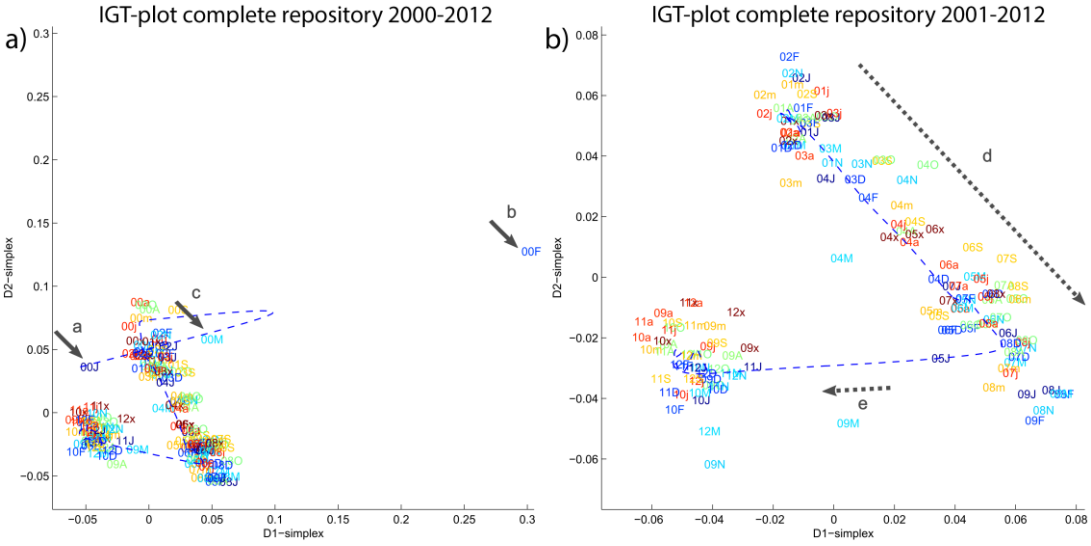
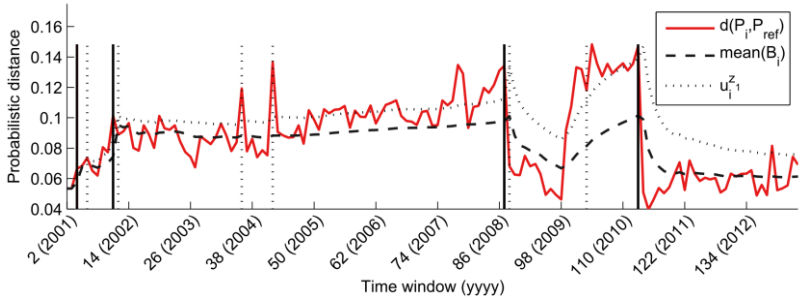


Figure 3: PDF-SPC Monitoring of the Variability of the Distribution of the Complete Multi-Variate Repository (all variables) on a Monthly Basis. The chart plots the current distance to the reference ($d(P_i,P_{ref})$), the mean accumulated distance ($mean(B_i)$), and the upper confidence interval being monitored (u^z_i) and indicates the warning and out-of-control states as broken or continuous vertical lines, respectively. After a transient state (2001), a gradual change is seen, alerting two warning states around 2004, until the threshold is reached in 2008, leading to an out-of-control state, which re-establishes the reference distribution. The abrupt change in 2009 is captured by the metric and confirmed afterward.



Temporal Subgroups

Figure 2 (b) shows the IGT plot of the multivariate MRRV repository in 2001–2012. The flow of points is continuous through the timeline (arrow d) until February 2009, indicating a gradual change in their distributions. An abrupt change in March 2009 (arrow e) then splits the repository into two temporal subgroups, i.e., conceptually-related time periods. Additionally, a yearly seasonal component can be observed, especially in the latter subgroup, based on the color temperature of the months.

Figure 3 shows the analogous PDF-SPC chart for 2001–2012. After a transient state (2001), the change is gradual, alerting two warning states around 2004 (broken vertical lines) until the accumulated threshold is reached in 2008 leading to an out-of-control state (solid vertical lines). The abrupt change in 2009 was detected by the method and confirmed afterward.

Drilling down to specific variables, we found that the change in 2009 was also present for most variables. For example, Figure 4 (a) shows the IGT plot of *IntermediateCause1*, where the change is observed in March 2009.

The corresponding temporal heat maps of the variables uncovered a major change in 2009 related to the number of causes specified in the certificate. However, even ignoring the NA category to check whether such an abrupt change was solely due to the number of specified causes, the change persisted (Figure 4, (b)), indicating that the frequencies of some causes of death changed abruptly as well (although to a small extent). Figure 4 (c) shows the temporal heat map of I*ntermediateCause1* without the NA category, where this finding is observed. It can be noted as well that in 2011 some of the affected frequencies were re-adjusted.

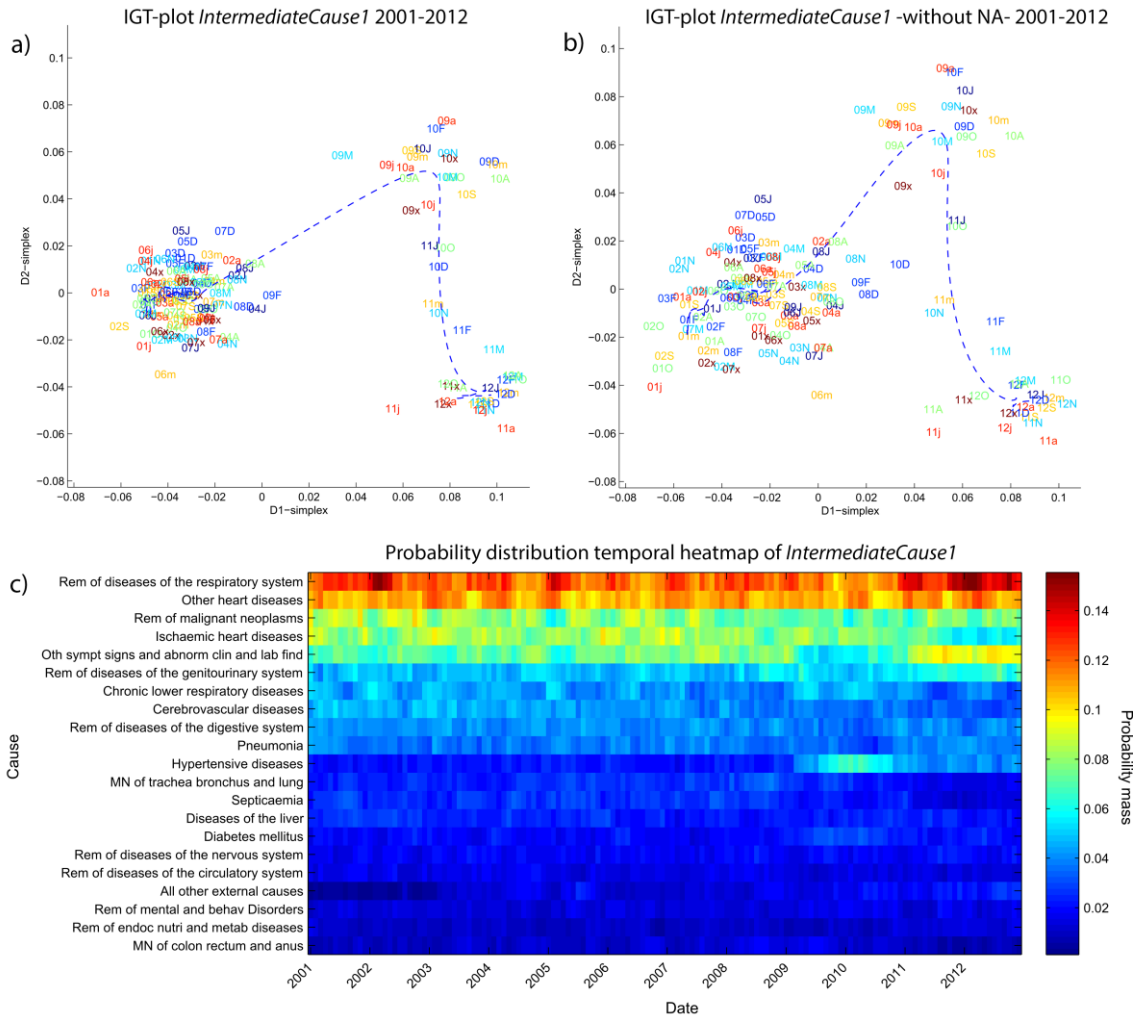
This abrupt change in 2009 is probably the most important finding from this study. This change coincides with the redesign of the National Certificate of Death in 2009. The new certificate was intended to meet the WHO recommendations to a greater extent. Two modifications to the certificate probably account for the abrupt change in 2009, namely (1) the use of a row of boxes, each to be filled with one letter, instead of blank lines that allowed continuous writing, and (2) renaming the field `Intermediate cause´ as `Antecedent cause´ and providing one more line for the entry. The first modification may have reduced the chances of filling more than one cause and encouraged filling at least one. The second modification probably increased the frequency of cases in which two intermediate causes were entered but, at the same time, limited the entries to only two causes—the option of entering a third cause was never used again. Additionally, the renaming caused some physicians to misunderstand `Antecedent cause´ as clinical antecedents; e.g., leading to the introduction of two prevalent chronic diseases such as hypertensive diseases and diabetes mellitus as antecedent causes, whereas introducing them as contributive causes would have been more appropriate. The Spanish National Statistics Institute warned the national Public Health institutions about this problem in 2011. To correct the situation, the term `Intermediate Cause´ was re-introduced. However, as seen in the results for *IntermediateCause1*, the practice was not abandoned entirely. Finally,

Author version of manuscript published in Journal of the American Medical Informatics Association
(http://dx.doi.org/10.1093/jamia/ocw010)

the several changes in multiple causes in 2009 carried the problem to the basic cause. The three versions of the certificates are shown in Appendix B.

The separation of the repository into two temporal subgroups, up to 2009 and from 2009 onward, gives the first hint that statistical analyses or models that treat the entire span as one may not be concordant, given the abrupt differences in their data distributions. Consequently, in some of the further steps, the two subgroups were analyzed separately.

Figure 4: IGT Plots (a, b) and Temporal Heat Map of Distribution (c) of *IntermediateCause1* for Men in 2001–2012 on Monthly Basis. In the IGT plots, each point represents one batch of the records labeled with its date in 'YYM' format (YY: last two digits of the year, M: the month as given in the list of abbreviations at the end), and the distances among them represent the dissimilarity in their distributions. The IGT plots were calculated considering (a) and discarding (b) unfilled values (NAs). The heat map shows the evolution of the probability distribution for 21 most prevalent causes after discarding the NA category, where the frequencies of `hypertensive diseases´, `chronic lower respiratory diseases´, and `diabetes mellitus´ increased, whereas those of `symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified´ decreased, among others. The three main temporal subgroups seen in both the IGT plots (split by months, namely 09M and 11J) are associated with the changes in the patterns of the frequencies of causes shown in the heat map for 2009 and 2011.



Author version of manuscript published in Journal of the American Medical Informatics Association (http://dx.doi.org/10.1093/jamia/ocw010)

Departmental anomalies

We next assessed the variability among different health departments. Figure 5 (a) shows the SPO monitoring of the multivariate MRRV repository on yearly basis. The health department of Requena showed a large SPO, indicating an outlying distribution. Besides, the health department of Torrevieja also increased its SPOs during 2005–2009.

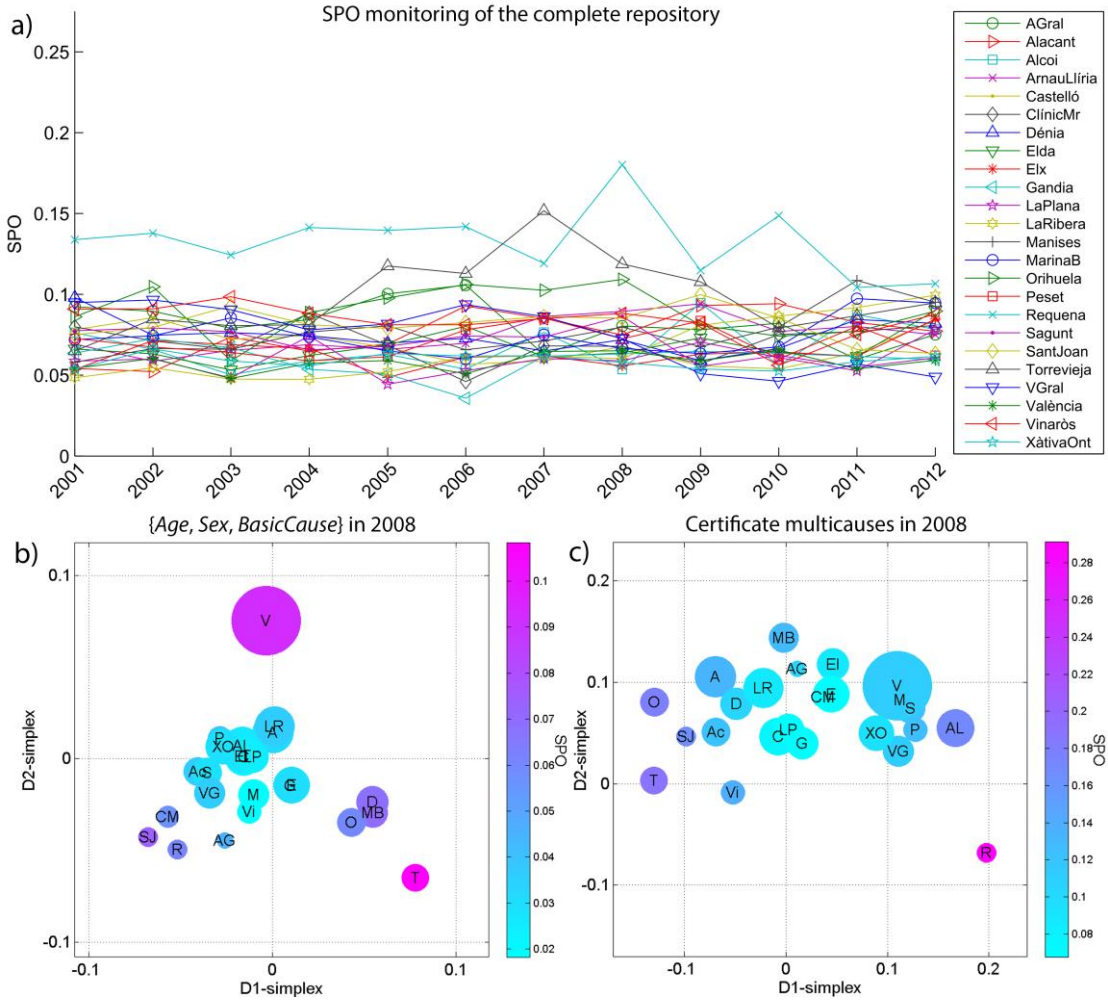
Further scrutiny led to the splitting of the set of variables into two subgroups: one classifying individual deaths by {*Age*, *Sex*, *BasicCause}* and other representing deaths as registered in the Certificate due to multiple causes. The latter subgroup behaved the same way as the entire group, with a predominant SPO in Requena, followed by Torrevieja and Orihuela. In contrast, in the former subgroup we found a predominant SPO in the departments of Torrevieja and Valencia. Figure 5 (b) and (c) show the MSV plots of the two subgroups in 2008, showing interdepartmental dissimilarities.

Drilling down to individual variables, we found Requena to be the outlier with respect to *ContributiveCauses1–3*. However, the anomaly disappeared after discarding the category NA. We therefore analyzed the number of filled causes by the departments and found that Requena was the department that had filled the maximum number of contributive causes. This may reflect an isolated practice in a small department composed of an older population.

We also found that Torrevieja was the outlier with respect to the age at death, being the opposite of Requena. This difference may be due to the large number of deaths of young men in Torrevieja, which additionally counts with large settlements of immigrants from Eastern Europe and Russia. Other studies have noted the much greater incidence of cancer in Torrevieja and other places close to it probably related to immigration.[34] Lastly, the dissimilarity between Valencia and other departments is mainly due to its lowest proportion of deaths of men in Valencia.

Author version of manuscript published in Journal of the American Medical Informatics Association
(http://dx.doi.org/10.1093/jamia/ocw010)

Figure 5: Monitoring of the Departmental Anomalies Based on the Distribution of All Variables in the Repository During 2001-2012 using SPO monitoring (a). MSV plots for visualizing the variability among the distributions of the health departments in 2008 are shown for the multivariate combinations {*Age*, *Sex*, *BasicCause*} (b), and multiple causes {*InitialCause[1,2,3], IntermediateCause[1,2,3], ImmediateCause[1,2,3], ContributiveCause[1,2,3]*} (c). Circles represent the health departments (see the key to the names at the end), their color represents the source SPO, and their size reflects the sample size.



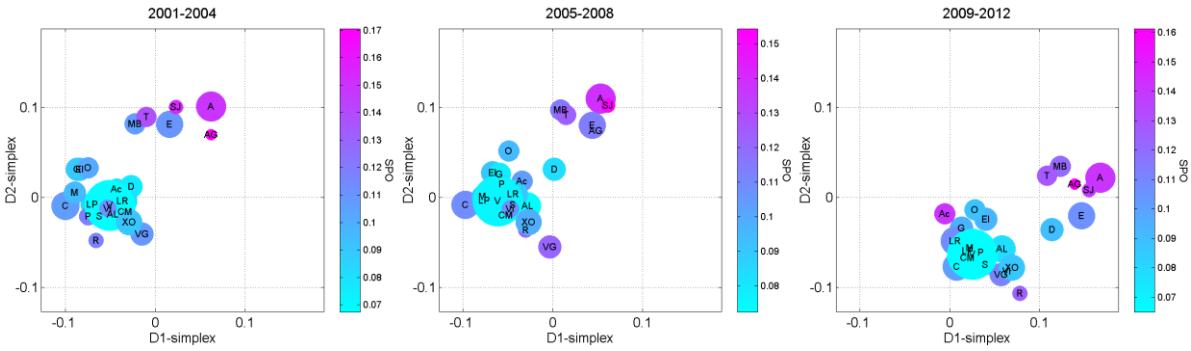
Departmental Subgroups

The existence of source subgroups, i.e., groups of sources with similar probability distributions, was addressed next. The MSV plots uncovered a multi-site subgroup formed by most departments in the province of Alicante, mainly found in *ImmediateCause1*, *IntermediateCause1* (Figure 6), and *InitialCause1*. Discarding the category NA, the subgroup was not present in *InitialCause1*; however, it still was present in *ImmediateCause1* and *IntermediateCause1*. This indicates a local variation of such departments both in terms of the number of filled causes and the causes of death, which may reflect an isolated practice in death certification (for example, we found that 27% of the records were left unfilled with respect to *InitialCause1* in the subgroup of the province of Alicante whereas for the rest, the proportion was 12%). The subgroups were empirically confirmed using clustering algorithms

based on the dissimilarity matrix of interdepartmental distribution distances obtained from the method.

Additionally, the change to the death certification in 2009 can be seen in Figure 6 as a global change affecting all data points in the last batch (2009-2012), but not necessarily equally.

Figure 6: Variability of *IntermediateCause1* Among the Distributions of the Health Departments over Time (4-Month Batches using MSV plot Monitoring). Circles represent the health departments (see the key to the names at the end), their color represents the source SPO, and their size reflects the sample size. A subgroup formed by most departments in the province of Alicante is at upper right part throughout. Besides, the change to the death certification in 2009 can be seen as a global change affecting all data points in the last batch (2009-2012), but not necessarily equally.



DISCUSSION

Table 3 summarizes the main findings and their causes. Such a table may constitute a form of feedback for the management of variability in repositories, either to avoid any problem or bias in the data to be reused, or to improve the processes of data acquisition and repository maintenance, as well as to prevent future problems related to DQ.

It is important to note that in some cases, variability may be inherent to environmental or population differences. However, in other cases, variability may be undesired, e.g., that due to faulty acquisition processes, biased actuations or variations in protocols. Regardless of whether variability is inherent or undesired, variability findings suggest investigating such a lack of concordance for a proper data reuse.

Hence, before reusing the data in the MRRV repository, users should consider the problems the above-mentioned findings may cause. A selection of them is described in Table 4, in which we attempt to provide a generic list of findings related to multi-source or temporal variability in repositories along with their possible causes, problems in reusing the data, and solutions.

Table 3: Variability in the Mortality Registry and its Causes.

Note Observable causes are those intrinsic to the data and found during the assessment process. The possible original causes are the external factors that cause the variability. The causes are linked to generic findings in Table 4 using the codes given in the first column.

Finding (generic code in Table 4)	Observable Cause	Possible Original Cause	Detected by

Temporal anomaly from January to March 2000 (F1)	A great deal of missing data in temporal batches	Lack of electronic coding of paper certificate	- IGT plot (Figure 2 (a)) - Temporal heat map (section 8, Appendix A)
Gradual change through the period of study (F2)	Gradual shifts in probability mass of causes of death through time	Increase of life expectancy, social and clinical changes in practice	- IGT plot (Figure 2 (b)) - PDF-SPC (Figure 3)
Abrupt change in March 2009 dividing the repository into two temporal subgroups (F3)	Abrupt change in probability of NAs for most variables, and to a small extent in other specific causes of death including the basic cause	Change in the national certificate of death	- IGT plot (Figure 2 (b), Figure 4 (a,b)) - PDF-SPC (Figure 3) - Temporal heat map (Figure 4 (c)) - MSV plot monitoring (Figure 6)
Other minor abrupt changes in 2005, 2009, and 2011 (F3)	Abrupt changes in probability mass of specific causes of death	National programs for control and prevention of diseases, redesign of certificate, change of disease patterns	- IGT plot (Figure 4 (a,b)) - Temporal heat map (Figure 4 (c))
Seasonal variations in causes of death (F4)	Seasonality of diseases, mainly winter-specific respiratory diseases and greater incidence of heart diseases in summer	Normal environmental and social effects	- IGT plot (Figure 2 (b))
Department of Requena as an outlier (F5)	Requena provides more number of causes, specially contributive causes	Isolated certificate filling practice in the small department with older population	- SPO monitoring (Figure 5 (a)) - MSV plot (Figure 5 (b))
Anomalous Department of Torrevieja (F5)	Anomalous population, with more deaths of young men	Different population due to immigration	- SPO monitoring (Figure 5 (a)) - MSV plot (Figure 5 (c))
Subgroup composed of departments in the province of Alicante (F6)	More intermediate and initial causes filled but fewer filled with immediate causes. Other differences in incidence of causes.	Isolated certificate filling practices	- MSV plot (Figure 6)

The problems listed in Table 4 are associated with basic research uses of data, namely for empirical derivation of hypotheses or statistical models. The proposed solutions vary with the sites or time affected and include fixing or excluding data or applying specialized data analysis methods. For example, for statistical modeling, an abrupt temporal change may reduce the model's effectiveness when using the data for the entire period, where a model with a good further generalization would be one giving more importance to latest data Besides, a probabilistically isolated site or group of sites may bias the results of a global analysis.[35] Excluding biased sites would improve the global results and in the case of multi-site subgroups, a good solution would be to analyze them separately or using a mixed-model approach.[36] An alternative solution which may reduce user involvement could be using *incremental learning* approaches, which rank the data in terms of importance by their age[37] or provenance.[38]

Fixing problematic data may also be considered when variability is associated with intrinsic problems with DQ such as changes in completeness or consistency of data.

Table 4: Generic Temporal and Multi-site Variability Findings and Possible Causes, Problems, and Solutions

Generic Finding (code)	Generic Possible Cause	Possible Data Reuse Problems	Possible Solutions
Punctual temporal anomaly (F1)	Biased temporal batch	Biased container time period (a year given a biased month), inaccurate research hypotheses or statistical models	Fix temporal batch; remove container time period
Gradual change (F2)	Normal evolution of population or clinical practice	Outdated statistical models	Incremental learning of models
Abrupt change causing temporal subgroups (F3)	Change of protocols, systematic errors, environmental or social effects	Inaccurate research hypotheses or statistical models: results that are not concordant before or after	Separate analyses, incremental learning of models
Seasonality (F4)	Normal environmental or social effects	Inaccurate statistical models	Season-specific models, mixed models
Anomalous sites (F5)	Anomalous population, biased clinical practice or systematic errors	Biased research hypotheses or statistical models: incompatible decisions or models among sources	Separate analyses or separate models for outlying sites, mixed models
Multi-site subgroups (F6)			Separate analyses or separate models for subgroups, mixed models

The proposed approach does not attempt to be a general DQ assurance approach for multi-site repositories, as shown by Kahn et al.[11] or Walker et al.[6], but to provide a comprehensive approach for the specific problems of multi-source and temporal variability. Nevertheless, we are not the first to suggest that variability needs to be managed before reusing data. Controlling the variability of data and outcomes to some extent is common in clinical trials that use classical statistical methods.[39-41] The variability among sites—mainly related to semantic or integration aspects—is controlled as well[6, 11, 41, 42]. However, semantic interoperability does not ensure that the aforementioned variability problems are properly managed; unfortunately, these problems will be reflected in probability distributions of data. Therefore, we advocate a `probabilistic interoperability´ assessment. In fact, Walker et al.[6] remark that statistical summaries of multi-site data are important for the shareability of their datasets under a centralized quality assurance assessment. Probabilistic methods such as those applied in the present study permit simultaneous managing of most of these biases, intrinsic DQ problems, and population differences in a metric and visual way. Further, in big data environments classic statistical methods may be inadequate. For example, analysis of variance (ANOVA) is aimed at testing for differences among Gaussian homoscedastic groups of data; however, the test is greatly affected by large sample sizes and not suited to multi-modal distributions. As an alternative, the GPD and SPO metrics are independent of sample size, useful with non-parametric continuous and categorical variables and even with a mix containing multiple variables of both types.

The proposed generalizable approach may be adopted in controlling data variability in research projects or multi-site data-sharing infrastructure. Additionally, ensuring DQ requires

specific areas of research and investment in public health.[8,14] For example, the WHO recommends conducting regular checks to validate death certification in hospitals as well as investigating new technologies to understand large data sets,[30] where the multi-source and temporal methods presented here may prove particularly useful.

Limitations

Due to the high number of possible combinations of variables, the efficiency of the approach may be improved through automated procedures or a guided Graphical User Interface. Besides, although the methods permit quantitative and qualitative descriptions of variability, it is the duty of the investigator to look for external original causes of variability, based on the insights provided by these methods.

Although the methods permit the analysis of multivariate joint distributions, we used the PCA dimensionality reduction method because it was simple and enabled us to find the most relevant problems. However, other non-linear methods may be better suited to multi-type and multi-modal data. We also found that the PDF-SPC algorithm may require a calibration of its thresholds in some situations, instead of using the classical three-sigma rule used in the present study.

Future Work

In functional terms, the next item of work is to incorporate the systematic approach into a general DQ assessment procedure as well to find ways of automatic monitoring and navigation through increasingly granular variables, data sources, and time periods.

In terms of research, we need to focus more on multivariate interactions; e.g., removing individual variable effects using mutual information.[43,44] Finally, the applied metrics need to be characterized with respect to alternative statistical methods such as effect size analysis to facilitate practical interpretation of the metrics.

CONCLUSION

Undesired variability in data distributions among sites or over time can be considered a DQ problem, which may lead to inaccurate or irreproducible results when the data are reused. The present study shows that the applied probabilistic methods may be useful as a systematic and generalizable approach to detect and characterize multi-site and temporal variability in large multi-site data that need to be reused. We suggest that, in addition to integration and semantic aspects, the temporal and multi-site probabilistic variability of data be incorporated in systematic procedures of assessing DQ to help ensure that valid conclusions are drawn when such data are reused.

Acknowledgments

This work was funded by the Spanish Ministry of Economy and Competitivity through the Retos-Colaboración 2013 Program (RTC-2014-1530-1) and the project `Caracterización de firmas biológicas de glioblastomas mediante modelos no-supervisados de predicción estructurada basados en biomarcadores de imagen´ (TIN2013-43457-R), and by the Universitat Politècnica de València through the `*Prueba de Concepto 2015´* project: *`Servicio de evaluación*

de la estabilidad espacio temporal de repositorios de datos biomédicos´ (SP20141432). The authors thank Carmen Alberich from the Dirección General de Salud Pública, Conselleria de Sanidad, Valencia, Spain, for the comments and support in the case study.

Acronyms for months

J: January, F: February, M: March, A: April, m: May, j: June, x: July, a: August, S: September, O: October, N: November, D: December

Acronyms for Health Departments

AG: AGral, A: Alacant, Ac: Alcoi, AL: ArnauLLíria, C: Castelló, CM: ClínicMR, D: Dénia, El: Elda, E: Elx, G: Gandía, LP: LaPlana, LR: LaRibera, M: Manises, MB: MarinaBaixa, O: Orihuela, P: Peset, R: Requena, S: Sagunt, SJ: SantJoan, T: Torrevieja, VG: VGral, V: València, Vi: Vinaròs, XO: XàtivaOnt

REFERENCES

1. Toubiana L, Cuggia M. Big Data and Smart Health Strategies: Findings from the Health Information Systems Perspective: IMIA Yearb. 2014;9(1):125–7.

2. Weber GM, Murphy SN, McMurry AJ, MacFadden D, Nigrin DJ, Churchill S, et al. The Shared Health Research Information Network (SHRINE): A Prototype Federated Query Tool for Clinical Data Repositories. J Am Med Inform Assoc. 2009 Sep 1;16(5):624–30.

3. McMurry AJ, Murphy SN, MacFadden D, Weber G, Simons WW, Orechia J, et al. SHRINE: Enabling Nationally Scalable Multi-Site Disease Studies. Carter KW, editor. PLoS ONE. 2013 Mar 7;8(3):e55811.

4. Natter MD, Quan J, Ortiz DM, Bousvaros A, Ilowite NT, Inman CJ, et al. An i2b2-based, generalizable, open source, self-scaling chronic disease registry. J Am Med Inform Assoc. 2013 Jan 1;20(1):172–9.

5. Weber GM, Barnett W, Conlon M, Eichmann D, Kibbe W, Falk-Krzesinski H, et al. Direct2Experts: a pilot national network to demonstrate interoperability among research-networking platforms. J Am Med Inform Assoc. 2011;18:157–60.

6. Walker KL, Kirillova O, Gillespie SE, Hsiao D, Pishchalenko V, Pai AK, et al. Using the CER Hub to ensure data quality in a multi-institution smoking cessation study. J Am Med Inform Assoc. 2014 Nov 1;21(6):1129–35.

7. Kuula A, Borg S. Open access to and reuse of research data - The state of the art in Finland. Finnish Social Science Data Archive. 2008;7.

8. Bray F, Parkin DM. Evaluation of data quality in the cancer registry: Principles and methods. Part I: Comparability, validity and timeliness. Eur J Cancer. 2009 Mar;45(5):747–55.

9. MacKenzie SL, Wyatt MC, Schuff R, Tenenbaum JD, Anderson N. Practices and perspectives on building integrated data repositories: results from a 2010 CTSA survey. J Am Med Inform Assoc. 2012 Jun 1;19:119–24.

10. Massoudi BL, Goodman KW, Gotham IJ, Holmes JH, Lang L, Miner K, et al. An informatics agenda for public health: summarized recommendations from the 2011 AMIA PHI Conference. J Am Med Inform Assoc. 2012 Sep 1;19(5):688–95.

11. Kahn MG, Raebel MA, Glanz JM, Riedlinger K, Steiner JF. A Pragmatic Framework for Single-site and Multisite Data Quality Assessment in Electronic Health Record-based Clinical Research: Med Care. 2012 Jul;50:S21–9.

12. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J Am Med Inform Assoc. 2013 Jan 1;20(1):144–51.

13. Liaw ST, Rahimi A, Ray P, Taggart J, Dennis S, de Lusignan S, et al. Towards an ontology for data quality in integrated chronic disease management: A realist review of the literature. Int J Med Inf. 2013 Jan;82(1):10–24.

14. Chen H, Hailey D, Wang N, Yu P. A Review of Data Quality Assessment Methods for Public Health Information Systems. Int J Environ Res Public Health. 2014 May 14;11(5):5170–207.

15. Cruz-Correia RJ, Rodrigues P, Freitas A, Almeida FC, Chen R, Costa-Pereira A. Data quality and integration issues in electronic health records. Inf Discov Electron Health Rec. 2009;55–95.

16. Galea S, Ahern J, Karpati A. A model of underlying socioeconomic vulnerability in human populations: evidence from variability in population health and implications for public health. Soc Sci Med. 2005 Jun;60(11):2417–30.

17. Knatterud GL, Rockhold FW, George SL, Barton FB, Davis CE, Fairweather WR, et al. Guidelines for quality assurance in multicenter trials: a position paper. Control Clin Trials. 1998;19(5):477–93.

18. Sáez C, Robles M, Garcia-Gomez JM. Stability metrics for multi-source biomedical data based on simplicial projections from probability distribution distances. Stat Methods Med Res. 2014 Aug 4;Published Online First [In Press].

19. Sáez C, Rodrigues PP, Gama J, Robles M, García-Gómez JM. Probabilistic change detection and visualization methods for the assessment of temporal stability in biomedical data quality. Data Min Knowl Discov. 2015 Jul;29(4):950–75.

20. Shewhart WA, Deming WE. Statistical method from the viewpoint of quality control. New York: Dover; 1986.

21. Westgard JO, Barry. Basic QC practices: training in statistical quality control for medical laboratories. Madison, WI: Westgard QC; 2010.

22. Sáez C, Robles M, Garcia-Gomez JM. Comparative study of probability distribution distances to define a metric for the stability of multi-source biomedical research data. In Osaka: IEEE; 2013. p. 3226–9.

23. Halsey LG, Curran-Everett D, Vowler SL, Drummond GB. The fickle P value generates irreproducible results. Nat Methods. 2015;12(3):179–85.

24. Nuzzo R. Statistical errors. Nature. 2014;506(13):150–2.

25. Lin M, Lucas HC, Shmueli G. Too Big to Fail: Large Samples and the *p* -Value Problem. Inf Syst Res. 2013 Dec;24(4):906–17.

26. Asunción A ND. UCI Machine Learning Repository; University of California, Irvine, School of Information and Computer Sciences [Internet]. [cited 2015 May 19]. Available from: http:// archive.ics.uci.edu/ml/

27. National Center for Health Statistics. National Hospital Discharge Survey (NHDS) data. US Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics [Internet]. [cited 2015 May 19]. Available from: http://www.cdc.gov/nchs/nhds.htm

28. Lin J. Divergence measures based on the Shannon entropy. IEEE Trans Inf Theory. 1991 Enero;37(1):145–51.

29. Cover TM, Thomas JA. Elements of information theory. 2nd ed. Hoboken, N.J: Wiley-Interscience; 2006.

30. Borg, Ingwer, and Groenen, P.J.. Modern multidimensional scaling: Theory and applications. Springer Science & Business Media, 2005.

31. World Health Organization. Strengthening civil registration and vital statistics for births, deaths and causes of death: resource kit. WHO Press; 2012.

32. World Health Organization. International statistical classification of diseases and related health problems. - 10th revision, 2008 edition. Geneva: WHO Press; 2009.

33. Moher, D., Hopewell, S., Schulz, K. F., Montori, V., Gøtzsche, P. C., Devereaux, P. J., Elbourne, D., Egger, M., Altman, D. G. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials.Journal of clinical epidemiology, 63(8), e1-e37, 2010.

34. Zurriaga O, Vanaclocha H, Martinez-Beneito MA, Botella-Rocamora P. Spatio-temporal evolution of female lung cancer mortality in a region of Spain, is it worth taking migration into account? BMC Cancer. 2008;8(1):35.

35. García-Gómez JM, Luts J, Julià-Sapé M, Krooshof P, Tortajada S, Robledo JV, et al. Multiproject–multicenter evaluation of automatic brain tumor classification by magnetic resonance spectroscopy. Magn Reson Mater Phys Biol Med. 2009 Feb;22(1):5–18.

36. Cnaan, A., Laird, N. M., & Slasor, P. (1997). Tutorial in biostatistics: using the general linear mixed model to analyse unbalanced repeated measures and longitudinal data. Stat Med, 16, 2349-2380.

37. Gama J, Gaber MM. Learning from Data Streams: Processing Techniques in Sensor Networks. Springer; 2007.

38. Tortajada S, Fuster-Garcia E, Vicente J, Wesseling P, Howe FA, Julià-Sapé M, et al. Incremental Gaussian Discriminant Analysis based on Graybill and Deal weighted combination of estimators for brain tumour diagnosis. J Biomed Inform. 2011 Aug;44(4):677–87.

39. Svolba G, Bauer P. Statistical quality control in clinical trials. Control Clin Trials. 1999;20(6):519–30.

40. Gassman JJ, Owen WW, Kuntz TE, Martin JP, Amoroso WP. Data quality assurance, monitoring, and reporting. Control Clin Trials. 1995;16(2):104–36.

41. Knatterud GL. Management and conduct of randomized controlled trials. Epidemiol Rev. 2002;24(1):12–25.

42. Sayer DC, Goodridge DM. Pilot study: assessment of interlaboratory variability of sequencing-based typing DNA sequence data quality. Tissue Antigens. 2007 Apr;69:66–8.

43. Pompe B, Blidh P, Hoyer D, Eiselt M. Using mutual information to measure coupling in the cardiorespiratory system. Eng Med Biol Mag IEEE. 1998;17(6):32–9.

44. Kopylova Y, Buell DA, Huang C-T, Janies J. Mutual information applied to anomaly detection. J Commun Netw. 2008 Mar;10(1):89–97.

Author version of manuscript published in Journal of the American Medical Informatics Association (http://dx.doi.org/10.1093/jamia/ocw010)