What can millions of laboratory test results tell us about the temporal aspect of data quality? Study of data spanning 17 years in a clinical data warehouse

Vincent Looten, Liliane Kong Win Chang, Antoine Neuraz, Marie-Anne Landau-Loriot, Benoit Vedie, Jean-Louis Paul, Laetitia Mauge, Nadia Rivet, Angela Bonifati, Gilles Chatellier, et al.

What can Millions of Laboratory Test Results Tell Us about the Temporal Aspect of Data Quality? Study of Data Spanning 17 Years in a Clinical Data Warehouse

Vincent LOOTEN[1,2], Liliane KONG WIN CHANG[3], Antoine NEURAZ[1,4], Marie-Anne LANDAU-LORIOT[5], Benoit VEDIE[5], Jean-Louis PAUL[5], Laëtitia MAUGE[6], Nadia RIVET[6], Angela BONIFATI[3], Gilles CHATELLIER[2], Anita BURGUN[1,2,4], Bastien RANCE[1,2]

[1] INSERM, Centre de Recherche des Cordeliers, UMRS 1138, Université Paris Descartes, Sorbonne Paris Cité, Paris, France;
[2] Hôpital Européen Georges Pompidou, Department of Medical Informatics, Assistance Publique - Hôpitaux de Paris (AP-HP), Université Paris Descartes, France
[3] LIRIS UMR CNRS 5205, Université Claude Bernard Lyon 1
[4] Hôpital Necker – Enfants Malades, Department of Medical Informatics, Assistance Publique - Hôpitaux de Paris (AP-HP), Université Paris Descartes, France
[5] Hôpital Européen Georges Pompidou, Department of Biochimistry, Assistance Publique - Hôpitaux de Paris (AP-HP), Université Paris Descartes, France
[6] Hôpital Européen Georges Pompidou, Department of Hematology, Assistance Publique - Hôpitaux de Paris (AP-HP), Université Paris Descartes, France

Corresponding author:

Bastien RANCE

Address: 20 rue Leblanc, 75015 Paris, France

E-mail: bastien.rance@aphp.fr

Phone number: +33 1 56 09 59 85

Keywords: Quality Control, Computational Biology/methods*, Information Storage and Retrieval, Humans, Clinical Laboratory Information Systems

ABSTRACT

Objective: To identify common temporal evolution profiles in biological data and propose a semi-automated method to these patterns in a clinical data warehouse (CDW).

Materials and Methods: We leveraged the CDW of the European Hospital Georges Pompidou and tracked the evolution of 192 biological parameters over a period of 17 years (for 445,000+ patients, and 131 million laboratory test results).

Results: We identified three common profiles of evolution: discretization, breakpoints, and trends. We developed computational and statistical methods to identify these profiles in the CDW. Overall, of the 192 observed biological parameters (87,814,136 values), 135 presented at least one evolution. We identified breakpoints in 30 distinct parameters, discretizations in 32, and trends in 79.

Discussion and conclusion: our method allowed the identification of several temporal events in the data. Considering the distribution over time of these events, we identified probable causes for the observed profiles: instruments or software upgrades and changes in computation formulas. We evaluated the potential impact for data reuse. Finally, we formulated recommendations to enable safe use and sharing of biological data collection to limit the impact of data evolution in retrospective and federated studies (e.g. the annotation of laboratory parameters presenting breakpoints or trends).

1 INTRODUCTION

1.1 Background

In the era of secondary use of healthcare data, big data and machine learning, data quality is a crucial element to build trust in the results of the analyses based on this type of data[1,2]. Among the vast quantity of biological and clinical data used, laboratory test results are probably the most ubiquitous. Laboratory test results are used in a wide variety of studies from classical epidemiology to modern data mining. Moreover, with the development of clinical data warehouses (CDWs) in the late 90's and 2000's, laboratory data represented a straightforward dataset to integrate.

Laboratory data as stored in CDWs are often leveraged in longitudinal retrospective studies. However, over long periods of time, different types of events may happen and impact the data: new equipment can be added to the pipeline of analysis, an automaton might be replaced, new scientific knowledge might lead to evolution of formulas, new legacy terminologies may be adopted and so forth. Such events could have consequences on individual patient results but more generally on the distribution of the laboratory test values (e.g., with sudden changes of all the values measured by the new instrument). In most data warehouses, the only elements of context provided for biological data are related to normal ranges. For example, in i2b2[3,4] the observation table has one optional attribute (the *valueflag_cd* column) designed to provide an annotation associated to the value (e.g. normal, high or low). In the OHDSI[5,6] system, the OMOP Common Data Model (OMOP CDM) proposes two optional attributes in the

measurement table: *range_low* and *range_high*. Both of these approaches strongly rely on the notion of normal values to provide context about the data itself. The definition of normal values was discussed early on[7], and continues to be explored in particular with the rise of stratified medicine[8,9]. The metadata currently provided in data warehouses contain little information regarding the context of generation and could prove to be insufficient to enable proper interpretation of the results in longitudinal studies (especially over long periods of time).

1.2 Related Works

Several aspects of data quality have been explored in medicine and biology. In their editorial of 2000, Brennan and Stead[10] discussed the notion of data quality of clinical records in terms of concordance, correctness and completeness. Weiskopf and Weng[11] identified two additional dimensions in their review of methods and dimension for data quality: plausibility and currency. In their recent survey of the vocabulary used in data quality, Kahn *et al.*[12] unified the data quality terminology in three main dimensions: conformance, completeness and plausibility. Other approaches exist like the data quality assessment organization proposed by Sáez *et al.* [13]. The notion of data quality has been largely explored in biology and medicine[14,15] and beyond in almost every realm of science. An international norm for data quality was proposed in 2002, and had its first component approved in 2008[16]. Hauser *et al.*[17] proposed a LOINC-based approach to standardize the results of laboratory tests in multicenter CDW. Nevertheless, they observed that “a process that standardizes the laboratory results in the CDW will necessitate frequent updates to stay current.” The notion of temporal quality was explored by Sáez *et al.*[18] on a Spanish public health mortality registry using methods based on information theory and geometry.

Outside of medicine, the computer science community has developed a large body of research on the topic of quality: for example, for data streaming[19–22], data outliers[23–25] or data cleaning process[26–28]. In addition to the intrinsic dimensions of data quality, data sharing and analysis through large federated networks of data warehouses emphasize characteristics required to make data reusable, like the presence of provenance information and shared vocabularies[29]. The FAIR principles[30] can be viewed as guidelines to help address this issue. For example, the same data object should be persistent and sufficiently well described to be findable and re-usable, thus with emphasis on their metadata.

1.3 Scope and Objectives

Despite this large body of work, there is still a need for reproducible methods to assess the quality of biological data in CDWs. To the best of our knowledge, there is no study describing, at a large scale, the potential evolution of laboratory test results in CDWs.

This study takes place in the context of data profiling[31]. We are interested in the longitudinal dimension of the data and in the impact of events external to the measurement process itself. We adopt Kahn *et al.*[12] definition of plausibility as "features that describe the believability or truthfulness of data values". The temporal plausibility considers the quality with regard to a reference. In this study, we are not interested in the quality of individual values at a given moment (correctness), or their presence when expected (completeness), but in the plausibility of a value over time with respect to the history of the distribution (serving as a reference). In this study, we are interested in bulk analysis (*i.e.* the evaluation of the quality at the level of the entire set of laboratory test results) and not at the level of individual patients. In the remaining of this article, we describe how we manually reviewed the data and identified common profiles of temporal evolution. We propose a semi-automated method to systematically track these profiles, and apply it to the CDW of the Georges Pompidou European Hospital (HEGP).

2 MATERIALS

2.1 The HEGP Clinical Data Warehouse and the LIMS

The European Hospital Georges Pompidou (HEGP) is a 700 beds public hospital located in Paris, France. The HEGP is specialized in oncology, cardiovascular diseases and emergency medicine. The Biology Department installed a laboratory information management system (LIMS) at the opening of the hospital in the year 2000[32]. The data are collected in majority directly at the analysis automaton or are entered manually by biologists and laboratory technicians. The hospital developed in 2008 a CDW integrating all the data produced in the hospital, including laboratory tests[33]. As of July 2017, the CDW contained more than 131 million laboratory test results for more than 445,000 patients and more than 11,000 distinct biological parameters. In this study, the data integrated ranged from the opening of the hospital in the year 2000 to July 2017. The laboratory data analyzed in this study are anonymized and timestamped results.

2.2 Inclusion criteria

We focused on laboratory data from the biochemistry and hematology departments from our hospital (located in a single site). All biological data were considered regardless of the type of encounter (*e.g.* inpatient, outpatient). We included biological parameters having at least 10,000 data points and formatted as numerical results. Periods of time with fewer than 100 values over a two-month period were excluded from the datasets.

2.3 Dataset of interest

We considered every laboratory result recorded in the CDW of the HEGP meeting the inclusion criteria. In the end, we obtained 192 "time series" comprised of pairs of timestamps and biological values. Each dataset represents a time series of the daily

distribution of results for a given biological parameter. In Figure 1, a single point represents the value for one patient at a given time. The overall scatter plot represents the evolution of the distribution of a single biological parameter over time.

2.4 Ethical Statement

This study uses only aggregated anonymous data from the HEGP CDW. In accordance with the French regulation on data privacy, the HEGP Institutional Review Board (HEGP IRB registration #00001072) has waived the requirement to obtain informed consent or specific approval for such studies[34].

3 METHODS

We developed a three-step approach to study elements of temporal data quality on HEGP biological data. In a nutshell: (step 1) we performed an expert review of the data set, and selected profiles of evolution of interest, (Step 2) we designed computational and statistical methods to detect and describe precisely the profiles defined during the step 1, (Step 3) we applied the methods developed in step 2 to detect the profiles of interest in the entire dataset of 192 biological parameters.

3.1 Identifying profiles of interest

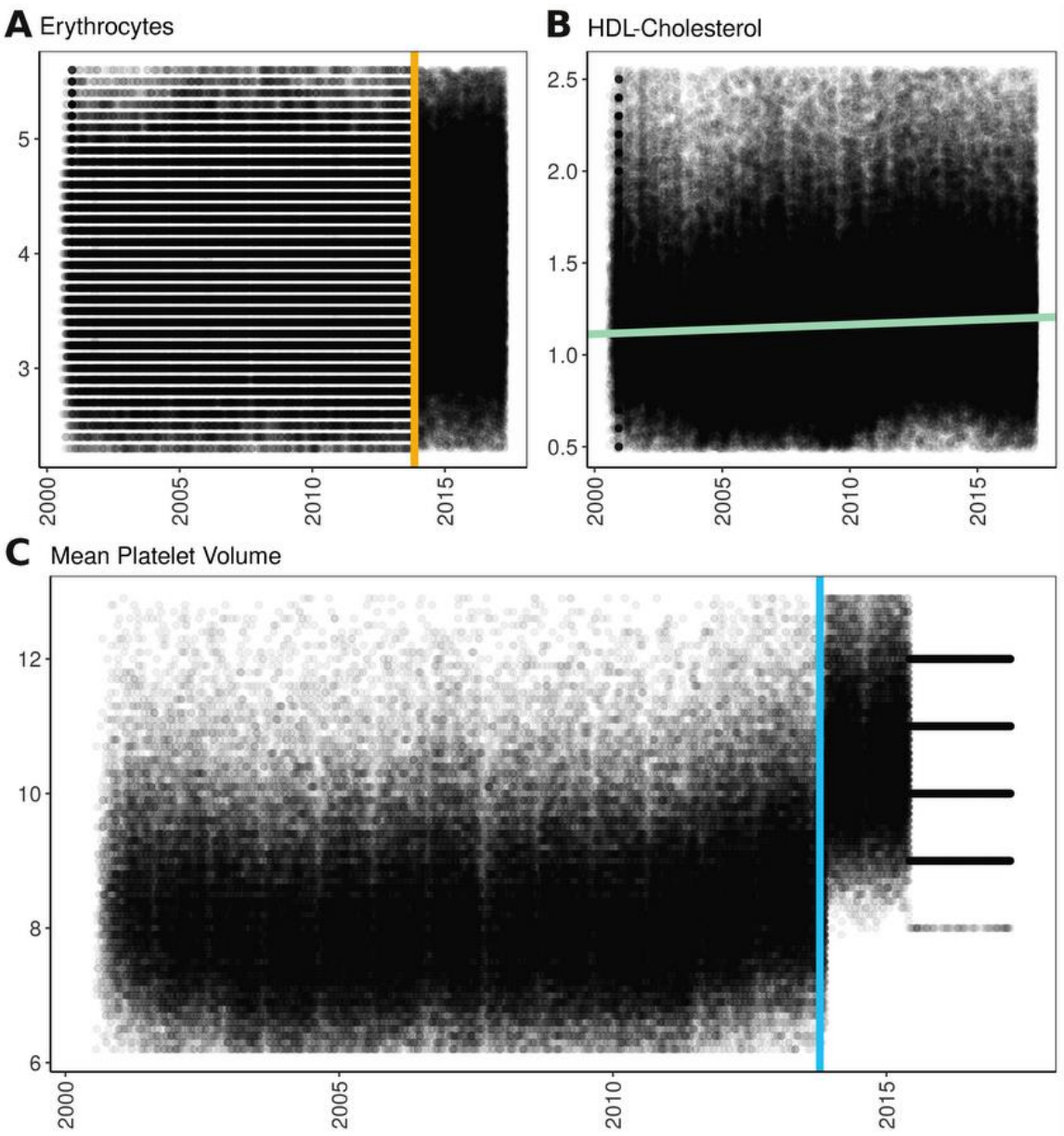

Figure 1: Examples of laboratory parameters impacted by temporal evolution (Discretization in orange plain lines, trends in green line, Breakpoint in plain blue line).

Medical informatics specialists (ABu, BR and VL), computer scientists (ABo, LK) and biologists (BV, JLP, and MAL) reviewed the graphical visualizations of the distribution of the 192 biological parameters. We identified and selected three profiles of particular interest that (a) could potentially have consequences in case of secondary use of the data and (b) are frequent in the dataset, and optionally (c) could probably be explained by prior knowledge of the hospital information system. We detailed the profiles as follows:

1. Discretization. Sudden change in the distribution of the data (including discretization): at a given point in time, the distribution is discretized; inversely, the distribution goes from discrete to continuous data (Figure 1A).
2. Trends. There is a continuous progressive evolution of the distribution of the data over time (i.e. identical standard deviation but evolving average) (Figure 1B).
3. Breakpoints. One or more breakpoints can be detected in the data. At a specific time, a clear modification of the mean of the distribution can be observed (Figure 1C).

Note that a distribution can exhibit several profiles, either simultaneously or sequentially (as can be seen in the additional discretization after 2015 in Figure 1C).

3.2 Detecting automatically evolution profiles

In this section, we detailed our methodology to detect quantitatively and extensively the three profiles identified above.

Detecting discretization. Our approach consisted in comparing the observed distribution of the data over time. We leveraged the generalization of the Benford's law to identify bias in the distribution. For a random distribution, we expect a similar ratio of each digit at the last position (namely 10%). If a bias occurs, for example caused by the rounding of a number, one number will be overrepresented. For example, if a uniform distribution of numbers between 0 and 100 is rounded to the nearest decade, we will observe a higher ratio of 0 than expected under the uniform hypothesis.

We defined classes based either on the number of decimals (for floating numbers), or on the number of digits (for integers). We then computed the number of occurrence of the last decimal (resp. digit) within the class, and the ratio of each digit by class weighted by the frequency of the class itself. A vector is constructed for each every window of 60 days

We computed the cosine similarity of consecutive vectors (*i.e.* at time *t* and *t+1*) of the ratio of last digits within classes weighted by their frequency. If no evolution is detected, a distance close to zero is observed. By opposite, for large change in distribution (including discretization), we observe a plunge in the values of the similarity. Figure 2 shows an example, with a change in distribution occurring on the 1000^{th} day.

The output of the algorithm provides the dates of detected distribution alterations.

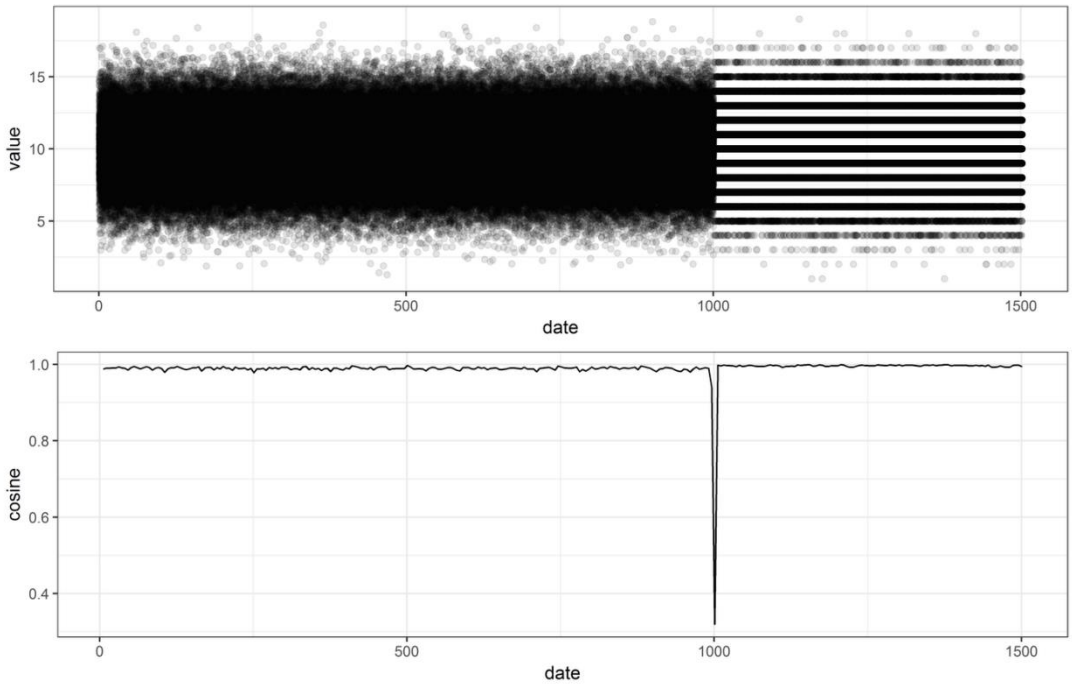

Figure 2: Representation of the evolution of the cosine similarity leveraging the Benford's law to capture changes in distribution (discretizations). The top graph represents the distribution of the data, the bottom one shows the value of the cosine similarity.

Detecting trends. We performed median regressions using the method developed by Koenker and d'Orey[35]. We select a random sample of 100,000 data points per type of laboratory parameter. We performed a sensibility analysis on 10 distributions to ensure that the sampling size did not alter the results. The unit of time considered for regression was the day. If prior breakpoints were detected, we performed the median regressions by intervals (to avoid trends explained by the breakpoint). We detected trends if the p-value of the Wald-test was lower than 0.05/192 using the Bonferroni correction for multiple testing.

We computed the *total estimated relative change* as the total estimated difference between the minimum and the maximum values over the inclusionperiod divided by the median.

Detecting breakpoints. For each of the 192 datasets we standardized the data by computing a z-score. The z-score is the distance between the raw score (raw lab values) and the population mean (for a given biological parameter) divided by the standard deviation of the population. We computed a moving daily median with a window of 60 days to limit the impact of outliers and random noise. In summary, for each day *d*, a median is computed using all the biological values collected between 30 days prior *d*, and 30 days after *d.* Eventually, we obtained time series with a single value of moving median for each day.

A breakpoint is defined as a sudden increase or decrease of the entire distribution for a given parameter (for example in Figure 1C, the increase occurs in 2014). We applied the Pruned Exact Linear Time (PELT) segmentation to detect breakpoints on the standardized moving median time series[36]. The outcomes of this algorithm are the dates for which breakpoints were detected.

3.3 Defining a threshold for the discretization detection method

The discretization detection method is based on the cosine similarity. This approach requires a threshold to build a decision rule. If the cosine similarity is lower than the threshold, a discretization is detected. Simulations were performed to help define a threshold. We simulated time series of values following a uniform law U(0,100), and identified the optimal threshold of 0.7 (in terms of optimization of precision and recall). The simulation program and associated codes are provided as supplementary material.

3.4 Implementation.

All analyses were performed using the R statistical software. We leveraged the R packages changepoint[37] and quantreg[35] for the detection. We provide an algorithm to simulate realistic distributions presenting one or a combination of evolution profiles. Data integration, analytical scripts, and the simulation algorithms of datasets are available in supplementary materials and at the URL: http://github.com/equipe22/BioQuality.

4 RESULTS

4.1 Global results

Overall, we assessed the longitudinal quality of 192 biological parameters. The number of biological test values covered by these 192 parameters was 87,814,136. The global profiling is presented in Table 1.

Table 1 - Summary of the profiles observed in the data. * Note that a laboratory distribution can fit into multiple categories

Category*	# of biological parameters	Example of laboratory parameter impacted	Example
Discretization	32 (16.7%)	Erythrocytes	Figure 1A
Breakpoints	30 (15.6%)	Mean platelet volume	Figure 1B
Trends	79 (41.1%)	HDL Cholesterol	Figure 1C

Evolution over time. Of the 192 observed biological parameters, 135 (with a total of 57,750,902 values) presented at least one profile among breakpoints, trends and discretization.

We plotted the number of events detected over time for each category (breakpoints, discretization) (see Figure 3). We did not represent events associated with trends, because no specific date can be assigned to the evolution.

4.2 Detailed results

We detected a total of 39 discretizations in 32 unique biological parameters, and 49 breakpoint events in 30 unique parameters. Figure 1 represents some typical profiles detected.

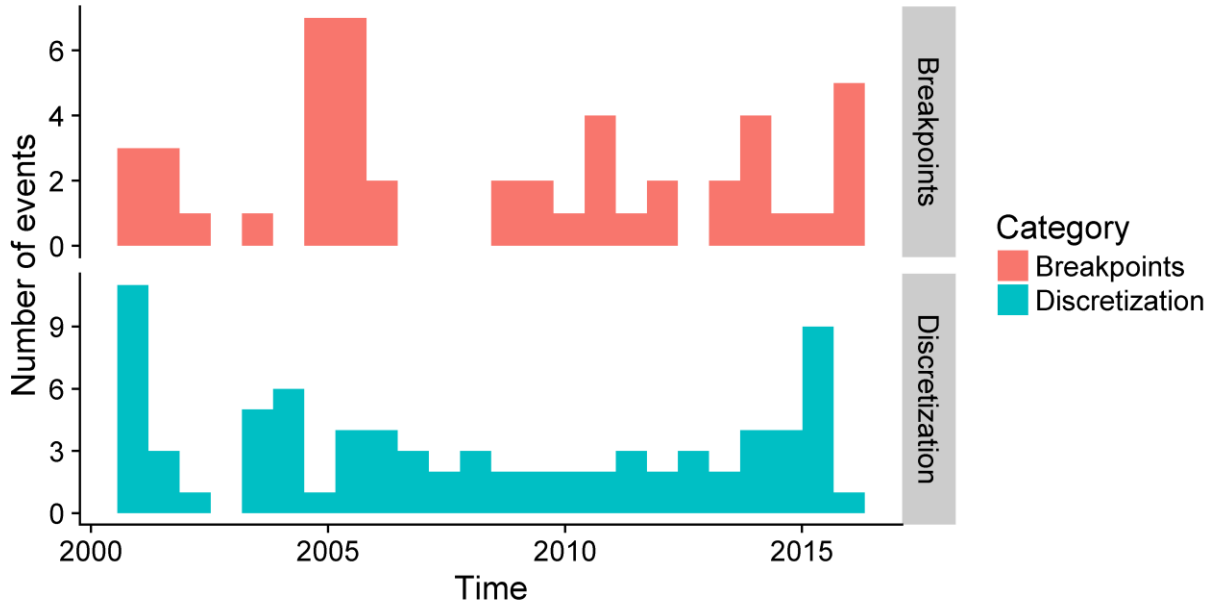

Figure 3 Aggregated count of evolution profiles detected per period of time

Trends. We detected 126 trends for 79 distinct biological parameters presenting a significant trend over the entire period of study or on parts of it (sub periods are split on breakpoints, see Figure 4). The repartition of the total estimated relative changes is described in Table 2 and Figure 4.

Table 2 - Distribution of the total estimated relative change of over the 79 biological parameters. Intervals correspond to quantiles of the variation distribution (min, max, median, and quartile).

Total estimated relative change	# of detected trends
Lower than -2%	42 (21.9%)
Between -2% and 0%	28 (14.6%)
Between 0 and 5.2%	11 (5.7%)
Larger than 5.2%	45 (23.4%)

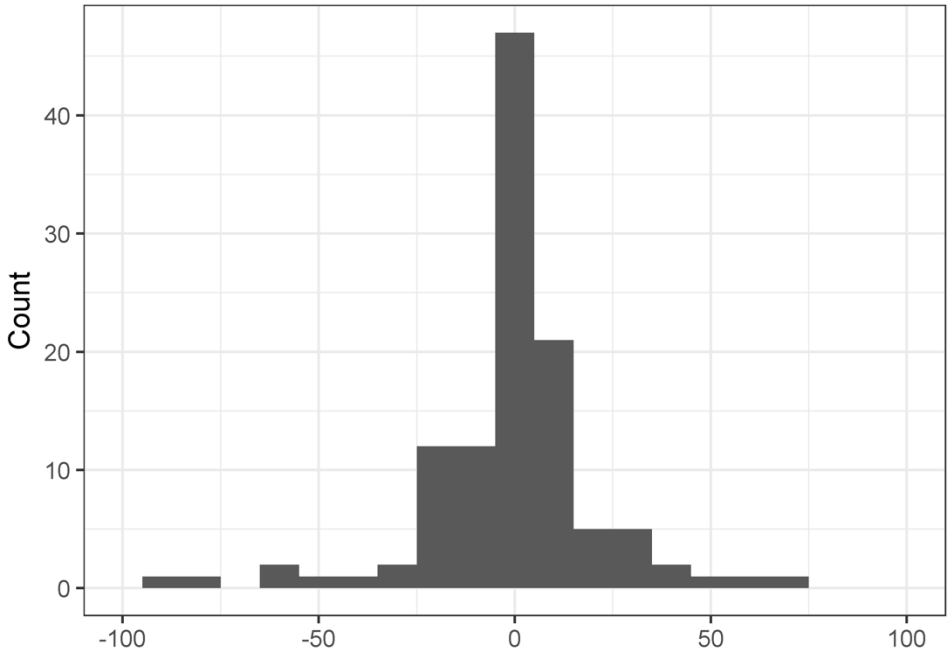

Figure 4 Distribution of the total relative change of the trends detected over time

5 DISCUSSION

5.1 Discussion of the results

Overall, 70% of the parameters presented at least one detectable evolution. The observation of the number of transformations over time reveals several noticeable peaks of breakpoints over the study period.

We did not attempt to identify the exact cause of each of the events detected, but looked for causes affecting a large number of laboratory measurement. Through discussions with the biology teams and considering the history of the information system, five major events emerged as probable causes of the observed profiles: (1) the early upload of historic values in 2000-2001. The HEGP hospital replaced three former hospitals, and a portion of the patients was transferred from these locations to the HEGP. A large volume of retrospective results was uploaded on a single day and share the same timestamp. The import may be the cause of several discretizations detected: the data imported were originated from different laboratories with different sets of normal ranges. In consequence, the distribution of results is noticeably different from the surrounding period, similarly breakpoints are detected for the same reason. (2) The replacement of an automaton in 2005 has been connected to several breakpoints

detected. (3) An accreditation visit of the laboratory in 2013 led to evolution of the formula used to compute the anion gap, and caused a breakpoint. (4) The replacement of the LIMS in 2015 caused a surge of discretizations. Another peak of breakpoints can be observed in 2010-2011, and is currently under investigation.

Trends. We observed a large number of trends. However, the effect of the evolution is often limited (with 50% of the trends between -2 and 5%). Regarding explanations, trends are more difficult to explain. In our running example, we hypothesize that the trend observed for HDL-Cholesterol could be explained by the quality improvement in healthcare.

5.2 A dashboard to explore single biological parameters over time

The output of our algorithms provided us with a classification for each dataset (impacted or not impacted). The breakpoint, distribution alteration detection algorithms also provides the dates of the occurrence of the phenomenon. We designed a dashboard with visualizations (Figure 5) of the evolution profile of each of the parameters to help the interpretation and contextualize the events detected.

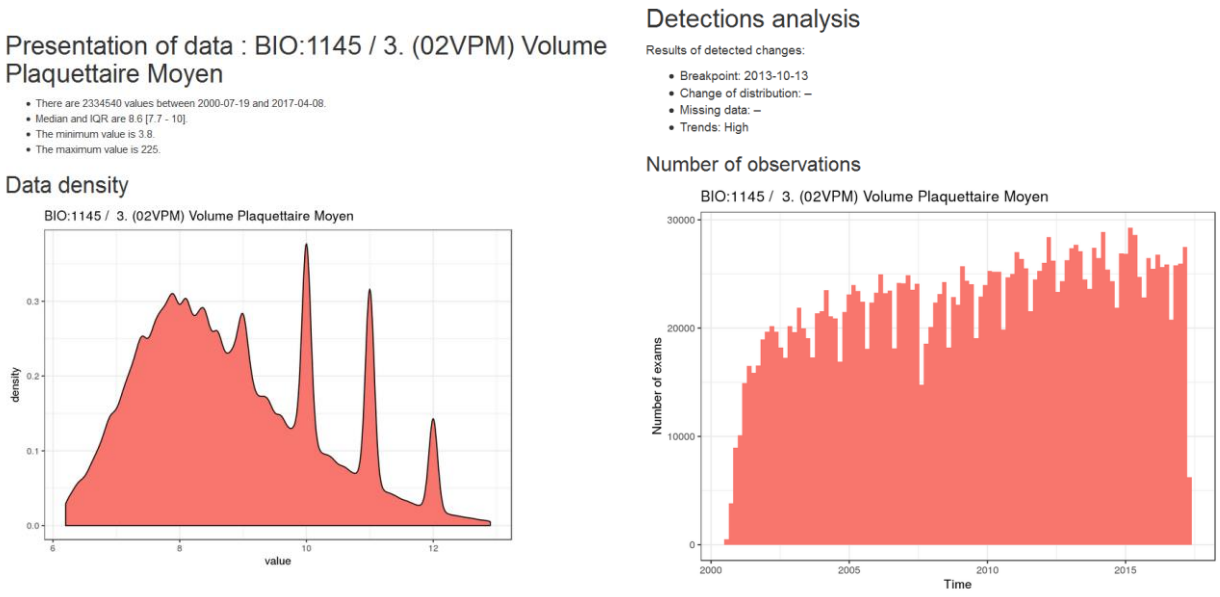

Figure 5. Visualization of the evolution profile of the mean platelet volume over time. This visualization provides a context to the distribution shown in Figure 1C

5.3 Consequences of the observed profiles

Statistical consequences. In this section we evaluated the potential statistical impact of the different profiles of evolution observed in the data.

1. *Breakpoint*: Breakpoints shift the mean and could have an effect on linear regression and statistical tests based on the mean. Moreover, a shift causes heteroscedasticity that violates the hypothesis of numerous regression models.

In addition to erroneous models and statistical test interpretation, breakpoints render the use of threshold-selection erroneous (e.g. all patients with a mean platelet volume superior to 12 does not have the same meaning over time).

2. *Trends*: Trends reflect a more profound change, often linked to larger epidemiological reasons. Trends seem to reflect either population evolution in the hospital (e.g. occurring with the opening of a new department), or even in the population (e.g. the progressive but moderate increase of HDL Cholesterol level). Most of the trends observed remained, however, moderate and would have no impact on statistical tests in subpopulation studies.
3. *Discretization*: The discretization could have an impact for patients at the edge of a threshold (in a gray zone). We evaluated the impact of discretizations (by rounding values) on variance-based statistical tests. Rounding values at 10^p decreases asymptotically the variance of $\frac{(10^p)^2}{12}$ in the hypothesis of a uniform distribution of decimals and independence between integer part and decimal part. In this same hypothesis, rounding values has asymptotically a little effect on the mean. In consequence, the impact on tests or regressions is negligible if p is small. Note that there is a quadratic decrease of the variance for larger p (*i.e.* important discretization) that could disable the availability of variance-based statistical tests. In practice, the observed discretizations do not shift the distribution and keep the values within the range of error of the measures. In those conditions, the impact of the discretization is smaller than the impact of the random noise.

Clinical impact. Our study focuses on the impact of evolution of distribution for secondary use of care data. None of the alteration of the distributions could have had any impact on the patient care. Breakpoints were accompanied with new normal ranges, discretized measures are interpreted with regards to the overall state of the patient. Finally, trends may reflect epidemiological evolution, or the evolution of the population of the hospital. We were able to find explanations for every phenomenon observed, and ensured the absence of consequence during the everyday care.

5.4 Limitation

Size of time windows. In the preprocessing step we chose a window size of 60 days. We experimented with periods of 15 and 30 days, the balance between the volume of results and rate of events of interest was maximized with 60 days. However, the setting of this window size remains arbitrary.

Discretization detection. In our simulation scenario, the optimal threshold for cosine similarity was equal to 0.7. The combination of rounded and non-rounded distribution influence the optimum threshold (from 0.7 to 0.85 in our simulations).In this study, we decided to limit the detection to case with clear discretizations (i.e. 100% of the values

were rounded after the discretization event). In real data, we performed sensitivity analysis. A threshold of 0.5, returns 20 biological parameters presenting at least one discretization time and a threshold of 0.85 returns 59 biological parameters. Graphically, we observed a point of inflection at 0.7. The threshold remains a parameter of the algorithm; we did not find theoretical reasons to fix a unique threshold.

Trends detection. Our approach for trend detection is influenced by 3 factors: the p-value, the effect size and the quality of the linear model. Despite the Bonferroni correction, a large number of biological parameters still present trends. However, as stated by Tatonetti *et al.*[38], the p-value is not an optimal criteria for detection of phenomenon in the big data paradigm (large datasets increase the power and tend to detect small and meaningless effects with statistical significance), for this reason, we included in our results the total relative change. The quality of the models was not assessed and trends could detect evolution that would be better explained by non-linear models.

Other profiles. In this study, we used a pragmatic approach to define a limited number of evolution profiles. Numerous other profiles could be considered: evolution of variances, presence of multimodal distributions, missing data and so forth. Our method could be easily adapted to capture some of these profiles (e.g. using moving variance instead of median). However, a careful evaluation of the performance would be needed. Other evolution profiles could be derived from the data themselves using unsupervised clustering, for example.

Additionally, our study did not consider categorical results. Further methodological development will be needed to explore all the facets of biological data.

Other dimensions of data quality. We limited our analysis to the observation of temporal plausibility. Most of the other dimensions of data quality have been extensively explored (e.g. for completeness[39,40], for outlier detection[41]). The added context provided by the longitudinal aspect of the data could help reveal new issues.

Guideline-based quality evaluation. Similarly to what was proposed for DQ of EHRs in [42], a standardization of the process of control of biological data in CDW would be beneficial to address potential issues during secondary use.

5.5 Proposed actions

We propose a series of actions to limit the impact of the evolution of the distributions over time in studies re-using retrospective routine data:

- *Automatic annotation of altered distributions*. In the data warehouse, we suggest to mark the concepts presenting an alteration of the distribution over time using a data quality flag. We could store the nature of the transformation using the three

categories described before. In a separate table, we suggest storing the identified date of occurrence of breakpoints and discretization. Trends and discretizations could be recorded as separate alterations. This approach is simple and can be applied to other types of alterations.

- *Correction of altered distribution*. We propose to normalize values of distribution in which breakpoint were observed. One problem might be the normalization at the edges of transition periods if not precisely identified. This solution is less trivial to put in place and can be applied only when the precise date of occurrence can be identified and explained, and for which a meaningful standardization exists.

- *Education of data warehouse users and specialists*. In addition to the automated annotation of the distributions in the warehouse, dedicated training sessions should be proposed to the CDW users.

- *Creation of metadata required for biological parameter reuse*: metadata related to the exams should be stored: identification of the automaton, normal and outbound values, date of installation and retirement of automaton. We also recommend adding information regarding the protocols in which the biological parameters were included, and the history of their evolution.

6 CONCLUSION

In this study, we proposed a semi-automatic method for the data profiling of longitudinal laboratory data. We placed our analyses in the context of temporal plausibility and search for three types of events influencing the quality of the data over time: breakpoints, discretization and trends.

We observed potential data quality issues in 135 out of 192 biological parameters studied that need to be acknowledged and addressed for secondary use. We proposed actions to limit the impact of the evolution in retrospective and federated studies.

COMPETING INTERESTS

The authors have no competing interests to declare.

CONTRIBUTORS

Substantial contributions to the conception or design of the work, or the acquisition, analysis, or interpretation of data for the work: Conception of the work: ABo, ABu, AN, BR, GC, VL. Acquisition of data: BR, LK, VL. Interpretation of data: BR, BV, JP, LM, ML, NR, VL. Drafting the work: ABo, ABu, AN, BR, VL. Revising the work critically for

important intellectual content: all authors. Final approval of the version to be published: all authors. Agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved: all authors.

ACKNOWLEDGMENT

The authors would like to thank Eric Zapletal, PhD for his help and expertise. BR was supported in part by the SIRIC CARPEM research program.

The authors would like to sincerely thank anonymous reviewers for insightful comments and implication in improving this study.

REFERENCES

[1] S.R. Sukumar, R. Natarajan, R.K. Ferrell, Quality of Big Data in health care, Int. J. Health Care Qual. Assur. 28 (2015) 621–634. doi:10.1108/IJHCQA-07-2014-0080.

[2] A.W. Toga, I.D. Dinov, Sharing big biomedical data, J. Big Data. 2 (2015) 7. doi:10.1186/s40537-015-0016-1.

[3] S.N. Murphy, M. Mendis, K. Hackett, R. Kuttan, W. Pan, L.C. Phillips, V. Gainer, D. Berkowicz, J.P. Glaser, I. Kohane, H.C. Chueh, Architecture of the open-source clinical research chart from Informatics for Integrating Biology and the Bedside., Annu. Symp. Proceedings. AMIA Symp. (2007) 548–52.

[4] i2b2: Informatics for Integrating Biology & the Bedside, (n.d.). https://www.i2b2.org/.

[5] G. Hripcsak, J.D. Duke, N.H. Shah, C.G. Reich, V. Huser, M.J. Schuemie, M.A. Suchard, R.W. Park, I.C.K. Wong, P.R. Rijnbeek, J. van der Lei, N. Pratt, G.N. Norén, Y.-C. Li, P.E. Stang, D. Madigan, P.B. Ryan, Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers., Stud. Health Technol. Inform. 216 (2015) 574–8. http://www.ncbi.nlm.nih.gov/pubmed/26262116 (accessed September 7, 2017).

[6] OHDSI, (n.d.). http://ohdsi.org/.

[7] B.J. Bock, C.T. Dolan, G.C. Miller, W.F. Fitter, B.D. Hartsell, A.N. Crowson, W.W. Sheehan, J.D. Williams, The Data Warehouse as a Foundation for Population-Based Reference Intervals, Am. J. Clin. Pathol. 120 (2003) 662–670. doi:10.1309/W8J85AG4WDG6JGJ9.

[8] A.K. Manrai, C.J. Patel, J.P.A. Ioannidis, In the Era of Precision Medicine and Big Data, Who Is Normal?, JAMA. (2018). doi:10.1001/jama.2018.2009.

[9] N. Rappoport, H. Paik, B. Oskotsky, R. Tor, E. Ziv, N. Zaitlen, A.J. Butte, Creating ethnicity-specific reference intervals for lab tests from EHR data, 2017. doi:10.1101/213892.

[10] P.F. Brennan, W.W. Stead, Assessing Data Quality: From Concordance, through Correctness and Completeness, to Valid Manipulatable Representations, J. Am. Med. Informatics Assoc. 7 (2000) 106–107. doi:10.1136/jamia.2000.0070106.

[11] N.G. Weiskopf, C. Weng, Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research, J. Am. Med. Informatics Assoc. 20 (2013) 144–151. doi:10.1136/amiajnl-2011-000681.

[12] M.G. Kahn, T.J. Callahan, J. Barnard, A.E. Bauck, J. Brown, B.N. Davidson, H. Estiri, C. Goerg, E. Holve, S.G. Johnson, S.-T. Liaw, M. Hamilton-Lopez, D. Meeker, T.C. Ong, P. Ryan, N. Shang, N.G. Weiskopf, C. Weng, M.N. Zozus, L. Schilling, A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data, EGEMs (Generating Evid. Methods to Improv. Patient Outcomes). 4 (2016) 18. doi:10.13063/2327-9214.1244.

[13] C. Sáez, J. Martínez-Miranda, M. Robles, J.M. García-Gómez, Organizing data quality assessment of shifting biomedical data., Stud. Health Technol. Inform. 180 (2012) 721–5. http://www.ncbi.nlm.nih.gov/pubmed/22874286 (accessed April 3, 2018).

[14] R. Khare, L. Utidjian, B.J. Ruth, M.G. Kahn, E. Burrows, K. Marsolo, N. Patibandla, H. Razzaghi, R. Colvin, D. Ranade, M. Kitzmiller, D. Eckrich, L.C. Bailey, A longitudinal analysis of data quality in a large pediatric data research network, J. Am. Med. Informatics Assoc. 24 (2017) 1072–1079. doi:10.1093/jamia/ocx033.

[15] K. Lee, N. Weiskopf, J. Pathak, A Framework for Data Quality Assessment in Clinical Research Datasets., AMIA ... Annu. Symp. Proceedings. AMIA Symp. 2017 (2017) 1080–1089. http://www.ncbi.nlm.nih.gov/pubmed/29854176.

[16] ISO/TS 8000-110:2008. Data quality -- Part 110: Master data: Exchange of characteristic data: Syntax, semantic encoding, and conformance to data specification, n.d. https://www.iso.org/standard/50800.html.

[17] R.G. Hauser, D.B. Quine, A. Ryder, LabRS: A Rosetta stone for retrospective standardization of clinical laboratory test results, J. Am. Med. Informatics Assoc. 25 (2018) 121–126. doi:10.1093/jamia/ocx046.

[18] C. Sáez, O. Zurriaga, J. Pérez-Panadés, I. Melchor, M. Robles, J.M. García-Gómez, 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. Informatics Assoc. 23 (2016) 1085–1095. doi:10.1093/jamia/ocw010.

[19] T. Dasu, G.T. Vesonder, J.R. Wright, Data quality through knowledge engineering, in: Proc. Ninth ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. - KDD '03, ACM Press, New York, New York, USA, 2003: p. 705. doi:10.1145/956750.956844.

[20] T. Dasu, T. Dasu, S. Krishnan, S. Venkatasubramanian, K. Yi, An information-theoretic approach to detecting changes in multi-dimensional data streams, PROC. SYMP. INTERFACE Stat. Comput. Sci. Appl. (2006).

[21] T. Dasu, S. Krishnan, D. Lin, S. Venkatasubramanian, K. Yi, Change (Detection) You Can Believe in: Finding Distributional Shifts in Data Streams, in: Springer, Berlin, Heidelberg, 2009: pp. 21–34. doi:10.1007/978-3-642-03915-7_3.

[22] L. Berti-Equille, T. Dasu, D. Srivastava, Discovery of complex glitch patterns: A novel approach to Quantitative Data Cleaning, in: 2011 IEEE 27th Int. Conf. Data Eng., IEEE, 2011: pp. 733–744. doi:10.1109/ICDE.2011.5767864.

[23] M.M. Breunig, H.-P. Kriegel, R.T. Ng, J. Sander, LOF, in: Proc. 2000 ACM SIGMOD Int. Conf. Manag. Data - SIGMOD '00, ACM Press, New York, New York, USA, 2000: pp. 93–104. doi:10.1145/342009.335388.

[24] E.M. Knorr, R.T. Ng, V. Tucakov, Distance-based outliers: algorithms and applications, VLDB J. Int. J. Very Large Data Bases. 8 (2000) 237–253. doi:10.1007/s007780050006.

[25] E.M. Knorr, R.T. Ng, Notion of Outliers: Properties and Computation, in: KDD Proc., 1997.

[26] M. Yakout, A.K. Elmagarmid, J. Neville, M. Ouzzani, I.F. Ilyas, Guided data repair, Proc. VLDB Endow. 4 (2011) 279–289. doi:10.14778/1952376.1952378.

[27] M. Stonebraker Mit, D. Bruckner, I.F. Ilyas Qcri, G.B. Qcri, M. Cherniack, S. Zdonik, A. Pagan, S. Xu, Data Curation at Scale: The Data Tamer System, in: Bienn. Conf. Innov. Data Syst. Res., 2013.

[28] Xu Chu, I.F. Ilyas, P. Papotti, Holistic data cleaning: Putting violations into context, in: 2013 IEEE 29th Int. Conf. Data Eng., IEEE, 2013: pp. 458–469. doi:10.1109/ICDE.2013.6544847.

[29] J.S. Brown, M. Kahn, D. Toh, Data Quality Assessment for Comparative Effectiveness Research in Distributed Data Networks, Med. Care. 51 (2013) S22–S29. doi:10.1097/MLR.0b013e31829b1e2c.

[30] M.D. Wilkinson, M. Dumontier, Ij.J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L.B. da Silva Santos, P.E. Bourne, J. Bouwman, A.J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C.T. Evelo, R. Finkers, A. Gonzalez-Beltran, A.J.G. Gray, P. Groth, C. Goble, J.S. Grethe, J. Heringa, P.A.. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S.J. Lusher, M.E. Martone, A. Mons, A.L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van

Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M.A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, B. Mons, The FAIR Guiding Principles for scientific data management and stewardship, Sci. Data. 3 (2016) 160018. doi:10.1038/sdata.2016.18.

[31] T. Dasu, T. Johnson, Exploratory Data Mining and Data Cleaning, 2003.

[32] P. Degoulet, The HEGP component-based clinical information system, Int. J. Med. Inform. 69 (2003) 115–126. doi:10.1016/S1386-5056(02)00101-6.

[33] E. Zapletal, N. Rodon, N. Grabar, P. Degoulet, Methodology of integration of a clinical data warehouse with a clinical information system: the HEGP case., Stud. Health Technol. Inform. 160 (2010) 193–7. http://www.ncbi.nlm.nih.gov/pubmed/20841676 (accessed November 3, 2016).

[34] A.-S. Jannot, E. Zapletal, P. Avillach, M.-F. Mamzer, A. Burgun, P. Degoulet, The Georges Pompidou University Hospital Clinical Data Warehouse: A 8-years follow-up experience, Int. J. Med. Inform. 102 (2017) 21–28. doi:10.1016/j.ijmedinf.2017.02.006.

[35] R. Koenker, quantreg: Quantile Regression, (2017). https://cran.r-project.org/package=quantreg.

[36] R. Killick, P. Fearnhead, I.A. Eckley, Optimal Detection of Changepoints With a Linear Computational Cost, J. Am. Stat. Assoc. 107 (2012) 1590–1598. doi:10.1080/01621459.2012.737745.

[37] R. Killick, I.A. Eckley, changepoint : An R Package for Changepoint Analysis, J. Stat. Softw. 58 (2014). doi:10.18637/jss.v058.i03.

[38] N.P. Tatonetti, Translational medicine in the Age of Big Data, Brief. Bioinform. (2017). doi:10.1093/bib/bbx116.

[39] N.G. Weiskopf, G. Hripcsak, S. Swaminathan, C. Weng, Defining and measuring completeness of electronic health records for secondary use, J. Biomed. Inform. 46 (2013) 830–836. doi:10.1016/j.jbi.2013.06.010.

[40] H. Estiri, K.A. Stephens, J.G. Klann, S.N. Murphy, Exploring completeness in clinical data research networks with DQe-c, J. Am. Med. Informatics Assoc. 25 (2018) 17–24. doi:10.1093/jamia/ocx109.

[41] C.C. Aggarwal, P.S. Yu, Outlier detection for high dimensional data, in: Proc. 2001 ACM SIGMOD Int. Conf. Manag. Data - SIGMOD ’01, ACM Press, New York, New York, USA, 2001: pp. 37–46. doi:10.1145/375663.375668.

[42] N.G. Weiskopf, S. Bakken, G. Hripcsak, C. Weng, A Data Quality Assessment Guideline for Electronic Health Record Data Reuse, EGEMs (Generating Evid. Methods to Improv. Patient Outcomes). 5 (2017) 14. doi:10.5334/egems.218.