Caring is Sharing – Exploiting the Value in Data for Health and Innovation
M. Hägglund et al. (Eds.)

doi:10.3233/SHTI230070

Automatic Outlier Detection in Laboratory Result Distributions Within a Real World Data Network

Aída MUÑOZ MONJAS[a1], David RUBIO RUIZ[a,b], David PÉREZ-REY[a] and Matvey PALCHUK[b]
[a] *Biomedical Informatics Group, Universidad Politécnica de Madrid, Spain*
[b] *TriNetX , LLC, Cambridge, MA, USA*
ORCiD ID: Aída Muñoz Monjas https://orcid.org/0000-0003-4504-9775 , David Rubio Ruiz https://orcid.org/0000-0003-1880-8019, David Perez-Rey https://orcid.org/0000-0002-9021-2597, Matvey Palchuk https://orcid.org/0000-0002-7737-8752

Abstract. Laboratory data must be interoperable to be able to accurately compare the results of a lab test between healthcare organizations. To achieve this, terminologies like LOINC (Logical Observation Identifiers, Names and Codes) provide unique identification codes for laboratory tests. Once standardized, the numeric results of laboratory tests can be aggregated and represented in histograms. Due to the characteristics of Real World Data (RWD), outliers and abnormal values are common, but these cases should be treated as exceptions, excluding them from possible analysis. The proposed work analyses two methods capable of automating the selection of histogram limits to sanitize the generated lab test result distributions, Tukey's box-plot method and a "Distance to Density" approach, within the TriNetX Real World Data Network. The generated limits using clinical RWD are generally wider for Tukey's method and narrower for the second method, both greatly dependent on the values used for the algorithm's parameters.

Keywords. Outlier detection, laboratory test, real world data, LOINC, interoperability

1. Introduction

Over the last decades, healthcare systems have been undergoing a digitalization with significant implications for primary and secondary uses of clinical data. During this process, Health Care Organizations (HCOs) have mainly started storing their using Electronic Health Records (EHRs). The correct handling and processing of this data is essential, not only to guarantee patient safety, but also if these information sources are to be used for secondary purposes such as research [1].

In these institutions, laboratory data is commonly stored using local terminologies, due to the adaptability they enable. Associating local codes to standardized codes in terminologies like LOINC (Logical Observation Identifiers, Names and Codes) improves the potential uses of this data. To facilitate the study and understanding of laboratory tests, their results can be graphically represented. In quantitative laboratory tests where

[1] Corresponding Author: Aída Muñoz Monjas, e-mail: aida.munozm@alumnos.upm,es

the result is a number and a unit, the data can be plotted on histograms that represent the volume of results (on the y-axis) over the numeric value (on the x-axis). These histograms can have different shapes according to the nature of the test and specific characteristics of the population it is performed on.

In general, laboratory test results follow a Weibull distribution, which is defined by its two parameters: shape (k) and scale (λ) [2]. If k=5 and λ =1, the generated distribution approximates to a normal distribution, while for k=0.5 and λ =1 it approximates to an exponential distribution.

In a normal distribution, the expected result of the test is the number the distribution is centred on. Characteristics like mean, standard deviation, skewness and kurtosis are maintained throughout HCOs for each laboratory test, with positive skewness as the most common attribute of this type of distributions. In an exponential distribution, the expected result of the test is usually 0, an absence of the tested substance, having exponentially fewer positive results on the right tail of the distribution.

One of the characteristics of Real World Data (RWD) is the existence of outliers, observations with a great deviation from the rest of the observations registered [3] that add no value to the dataset. Outliers in a clinical context can be due to errors in the data registration, where the physician inputs the value with a unit that is not expected by the system, for example 500 g, but the data is represented with the same numeric value but different unit, 500 kg, in the distribution. Outliers can also be due to errors in the device that performs the lab test, resulting in negative amounts of substance measured, -100 mg/dL of glucose in blood, or physically impossible results, 10^{12} mg/dL of glucose in blood.

The main objective of this work is to automatically eliminate these outliers from laboratory result distributions to facilitate the understanding, visualization, and analysis of the obtained results.

2. Methods

Outlier detection in a laboratory result distribution brings the focus to the relevant information by defining a set of limits that exclude the majority of these inconsistent values. Several methods have been used in the past to define these limits [3]. In this work, two different methods are compared: Tukey's box-plot method, one of the most used approaches, and a "Distance to Density" method, proposed by Last et al. [4].

Tukey's method for outlier detection flags a value located between the inner and outer limits as a possible outlier, while a value outside the outer limit is a clear outlier of the distribution [5]. The inner and outer limits are calculated according to Eq. (1),

$$inner\ limits = \begin{cases} Q3 + 1.5 \cdot IQR \\ Q1 - 1.5 \cdot IQR \end{cases} \qquad outer\ limits = \begin{cases} Q3 + 3 \cdot IQR \\ Q1 - 3 \cdot IQR \end{cases} \tag{1}$$

where *Q1* and *Q3* are the first and third quartiles respectively, and *IQR* is the interquartile range calculated as *IQR* = *Q3* - *Q1*. Tukey's method is useful for skewed distributions, as it does not depend on the mean or standard deviation of the distribution [3].

The Distance to Density method [4] defines the reliability of each data element based on its distance to the values nearby and the weight of its frequency on the distribution. Reliabilities close to 0 characterize outliers, that will be filtered according to a threshold α [4]. Reliability is calculated from above (μ_{RL}) and below (μ_{RH}), according to Eq. (2).

$$\mu_{RLj} = \frac{2}{1+e^{\frac{\left(\beta \cdot M \cdot D \cdot \left(V_{j+1} - V_j\right)\right)}{\left(N_j \cdot \left(V_{j+M+1} - V_{j+1}\right)\right)}}} \qquad \mu_{RHj} = \frac{2}{1+e^{\frac{\left(\beta \cdot M \cdot D \cdot \left(V_j - V_{j-1}\right)\right)}{\left(N_j \cdot \left(V_{j-1} - V_{j-M-1}\right)\right)}}} \quad (2)$$

where D is the total number of observations, V_j is the value of index j with frequency N_j, β is the shape factor representing the attitude towards the distance between succeeding values and M is the lookahead. This method relies heavily on the values of parameters β, M and α, which should be adjusted to fit the results to the desired usecase.

3. Results

In order to assess the quality of the distribution limits generated by these two methods in RWD, limits were generated for the result distributions of sixteen laboratory LOINC codes. The limits generated by both methods were compared against a list of manually curated limits by an expert based on recent literature. Each histogram depicts the numeric results obtained in a laboratory test versus their frequency, grouping the results into bins. The histograms represent the studied data without modifications, which makes distributions difficult to analyse in certain cases, such as code 2744-1 in Figure 1, when outliers are present.

The processed data was obtained from TriNetX's global research network [6], by aggregating the data of EHRs from over 110 million patients across 125 HCOs. The histograms studied in this work are not raw data from a single organization, but the aggregation of multiple sources, ensuring the protection of clinical data and meeting the legal requirements for its handling, i.e., HIPAA, GDPR, etc. [7].

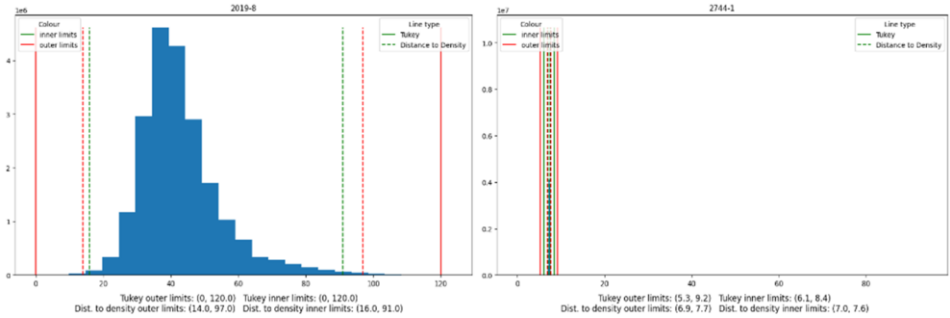

Figure 1. Visualization of generated limits for LOINC codes 2019-8 (left) and 2744-1 (right).

Table 1 contains the obtained results for a group of LOINC codes and their reference values. These reference values depict highest or lower recorded measures or values that are not physiologically possible, or represent manually curated limits generated by an expert. A very small percentage of patients are expected to have lab values outside these limits, and they can be considered good reference values for the exclusion of outliers in each of these tests.

In order to compute these limits, Tukey's method was applied as explained, and two different values for α were used in the Distance to Density method to obtain the inner and outer limits. The parameter M (lookahead) took a value of 1% of the data values, while the inner and outer α (thresholds) were $0.75 \cdot 10^{-4}$ and $0.75 \cdot 10^{-8}$ respectively,

with a β of 0.01. The generated limits were plotted against the distribution to visually assess their quality, as seen in Figure 1.

These inner and outer limits are used in TriNetX to define the plotting area of the distribution (inner limits) and to select the data that will be included when calculating measures such as average mean and standard deviation (outer limits).

Table 1. Obtained inner and outer limits with each method for seven LOINC codes. Comparison with the reference values and manually curated limits

LOINC code	Method	Inner limits		Outer limits	
		generated	reference	generated	reference
14749-6: Glucose in Ser/Pl. (mmol/L)	tukey	(0, 27)	(2.3 [8], 30)	(0, 39)	(0.7 [8], 40)
	distDen	(3.3, 8.3)		(2.8, 9.9)	
2019-8: CO2 in Art.Bld (mm[Hg])	tukey	(0, 120)	(8, 115[9][10])	(14, 97)	(0, 240)
	distDen	(16, 91)		(2.8, 9.9)	
33959-8: Procalcitonin in Ser/Pl. (ng/mL)	tukey	(0, 150)	(0, 9.7 [11])	(0, 150)	(0, 20)
	distDen	(0, 5.1)		(0, 7.2)	
8302-2: Body Height ([in us])	tukey	(0, 110)	(15, 90) ((38.1, 228.6) cm)	(0, 150)	(0, 107.09 [12]) ((0, 272) cm)
	distDen	(47, 76)		(33, 78)	
26881-3: Interleukin-6 in Ser/Pl. (pg/mL)	tukey	(0, 740)	(0, 800)	(0, 1200)	(0, 2221 [13])
	distDen	(0, 1200)		(0, 7600)	
2744-1: PH of Arterial Blood	tukey	(6.1, 8.4)	(4, 9)	(5.3, 9.2)	(0, 9)
	distDen	(7, 7.6)		(6.9, 7.7)	
12841-3: Prostate Specific Ag free/total in Ser/Pl. (%)	tukey	(0, 88)	(0, 80)	(0, 88)	(0, 100)
	distDen	(2, 67)		(0.87, 67)	

The limits for these sixteen LOINC codes were generated using both methods. Tukey's outer limits included over 95% of observations in every code, while the outer limits generated by the Distance to Density method included the same percentage in 13 out of the 16 codes. Both methods included over 80% of the observations in every code between the outer limits, and over 75% between the inner limits.

4. Discussion

Both methods studied in this work managed to shorten the interval of relevant information for the tested LOINC codes, generating limits that can separate outliers from the rest of the distribution in current and future data.

The limits generated by Tukey's method are generally wider than those manually generated by experts in the field, but provide an acceptable approximation to considerably improve the analysis and visualization of these distributions.

The Distance to Density method performs correctly with normal distributions, while being more unreliable with exponential distributions. This is likely due to the values of the algorithm's parameters, which were mainly selected to fit normal distributions, as exponential distributions limits are not clearly defined concerning outliers. When comparing the limits generated by this method with the limits generated by Tukey's method, it can be observed that these limits are closer to the distribution peaks, while Tukey's limits are generally wider.

The main limitation of both algorithms is the appropriate selection of the values of their parameters, as these greatly affect the obtained results. Generating limits that are too narrow can exclude relevant information, while generating overly wide limits produces a less readable graph and increases the complexity of the histogram representation.

5. Conclusions

Outliers are inherent to RWD, but still allow data analysis. The detection and exclusion of outliers cleans up the numeric results obtained in laboratory distributions and facilitates the analysis and interpretation of the result histograms. The main aim of this work is to provide a comparison between the results obtained by Tukey's approach to outlier detection and the "Distance to Density" method in clinical RWD. Tukey's method proved to generate limits wider than those annotated by an expert, while the Distance to Density method generated narrower limits. Both methods produced an acceptable approximation to improve the analysis and visualization of the distributions, but additional manual curation is recommended for optimal results.

References

[1] Miriovsky B, Shulman L, Abernethy A. Importance of Health Information Technology, Electronic Health Records, and Continuously Aggregating Data to Comparative Effectiveness Research and Learning Health Care. Journal of clinical oncology: official journal of the American Society of Clinical Oncology. 2012 10;30.

[2] Hallinan AJ Jr. A Review of the Weibull Distribution. Journal of Quality Technology. 1993;25(2):85-93. Available from: https://doi.org/10.1080/00224065.1993.11979431.

[3] Seo S. A Review and Comparison of Methods for Detecting Outliers in Univariate Data Sets. 2006 Aug. Available from: https://www.semanticscholar. org/paper/A-Review-and-Comparison-of-Methods-for-Detecting-in-Seo/ cb868f0b242b9623b7544a58b6a21647dfa138a5.

[4] Last M, Kandel A. Automated Detection of Outliers in Real-World Data. 2002 03.

[5] Dawson R. How Significant is a Boxplot Outlier? Journal of Statistics Education. 2011;19:2. Available from: https://doi.org/10.1080/10691898.2011.11889610.

[6] Topaloglu U, Palchuk M. Using a Federated Network of Real-World Data to Optimize Clinical Trials Operations. JCO Clinical Cancer Informatics. 2018 02;2:1-10.

[7] TriNetX. Home; 2022. Available from: http://trinetx.com.

[8] Cryer P. Hypoglycemia, functional brain failure, and brain death. The Journal of clinical investigation. 2007 04;117:868-70.

[9] Goila A, Pawar M. The diagnosis of brain death. Indian journal of critical care medicine : peer-reviewed, official publication of Indian Society of Critical Care Medicine. 2009 03;13:7-11.

[10] Wu Z, Luo Z, Luo Z, Ge J, Jin J, Cao Z, et al. Baseline Level and Reduction in PaCO2 are Associated with the Treatment Effect of Long-Term Home Noninvasive Positive Pressure Ventilation in Stable Hypercapnic Patients with COPD: A Systematic Review and Meta-Analysis of Randomized Controlled Trials. International Journal of Chronic Obstructive Pulmonary Disease. 2022 04;Volume 17:719-33.

[11] Rangeard O, Audibert G, Perrier JF, Loos-Ayav C, Lalot JM, Agavriloaie M, et al. Relationship Between Procalcitonin Values and Infection in Brain-Dead Organ Donors. Transplantation Proceedings. 2007;39(10):2970-4. Available from: https://www.sciencedirect.com/science/article/pii/S0041134507009670.

[12] Records GW. Tallest man ever;. https://www.guinnessworldrecords.com/world-records/tallest-man-ever.

[13] Amado J, López-Espadas F, Vázquez-Barquero A, Salas E, Riancho JA, López-Cordovilla JJ, et al. Blood levels of cytokines in brain-dead patients: Relationship with circulating hormones and acute phase reactants. Metabolism. 1995;44(6):812-6. Available from: https://www.sciencedirect.com/science/article/pii/0026049595901981.