

RESEARCH

Open Access



lab2clean: a novel algorithm for automated cleaning of retrospective clinical laboratory results data for secondary uses

Ahmed Medhat Zayed^{1,2*}, Arne Janssens¹, Pavlos Mamouris¹ and Nicolas Delvaux¹

Abstract

Background The integrity of clinical research and machine learning models in healthcare heavily relies on the quality of underlying clinical laboratory data. However, the preprocessing of this data to ensure its reliability and accuracy remains a significant challenge due to variations in data recording and reporting standards.

Methods We developed *lab2clean*, a novel algorithm aimed at automating and standardizing the cleaning of retrospective clinical laboratory results data. *lab2clean* was implemented as two R functions specifically designed to enhance data conformance and plausibility by standardizing result formats and validating result values. The functionality and performance of the algorithm were evaluated using two extensive electronic medical record (EMR) databases, encompassing various clinical settings.

Results *lab2clean* effectively reduced the variability of laboratory results and identified potentially erroneous records. Upon deployment, it demonstrated effective and fast standardization and validation of substantial laboratory data records. The evaluation highlighted significant improvements in the conformance and plausibility of lab results, confirming the algorithm's efficacy in handling large-scale data sets.

Conclusions *lab2clean* addresses the challenge of preprocessing and cleaning clinical laboratory data, a critical step in ensuring high-quality data for research outcomes. It offers a straightforward, efficient tool for researchers, improving the quality of clinical laboratory data, a major portion of healthcare data. Thereby, enhancing the reliability and reproducibility of clinical research outcomes and clinical machine learning models. Future developments aim to broaden its functionality and accessibility, solidifying its vital role in healthcare data management.

Keywords Electronic medical records, Clinical laboratories, Data integrity, Algorithms, Data preprocessing

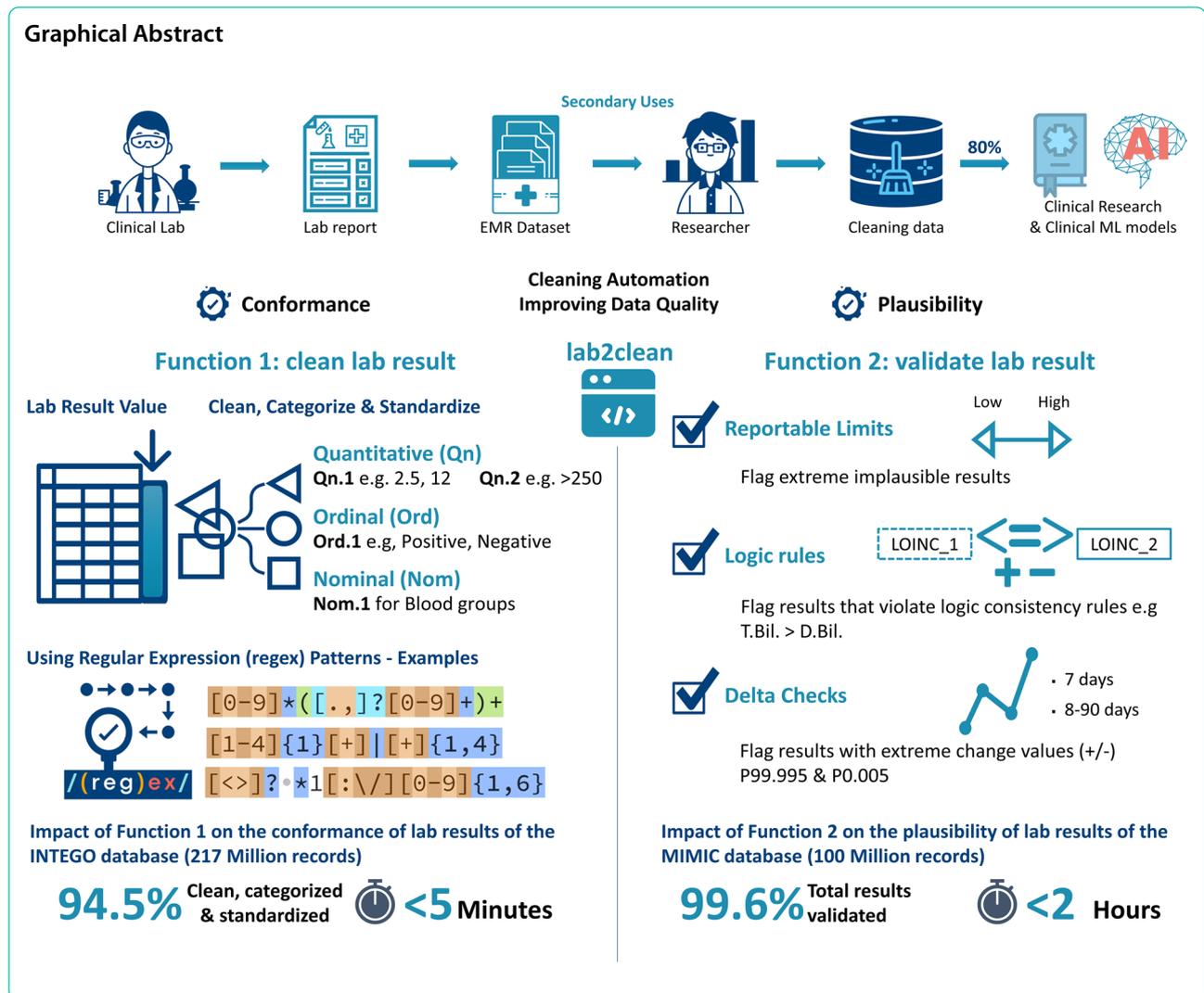
*Correspondence:

Ahmed Medhat Zayed
ahmed.zayed@kuleuven.be

Full list of author information is available at the end of the article



© The Author(s) 2024. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.



Background

Garbage in – garbage out, this phrase is widely recognized in computing and information systems and underscores the necessity of ensuring the quality of the data you provide a system to prevent flawed or nonsensical outputs [1]. This concept also applies to clinical research where the absence of high-quality input clinical data risks that research outcomes are unreliable or misleading. Such discrepancies can obstruct the effective translation of research findings into actionable evidence in clinical practice. Specifically, in clinical ML model development, high-quality data is non-negotiable. Notably, data preparation can command up to 80% of a project’s time, leaving only 20% for data analysis [2].

While clinical data primarily originates from daily healthcare interactions tailored to address immediate patient needs, the rise in the utilization of EMR systems has made this data increasingly accessible in digital

formats. However, repurposing this data from its primary clinical role to secondary research applications brings forth a lot of challenges [3]. Clinical laboratory data, a significant and mostly structured subset of EMRs, is indicative of these challenges. Despite the quality mandates that the clinical laboratories have in producing accurate and precise diagnostic reports, the laboratory data migrated for secondary uses remains messy and not entirely clean [4]. A primary concern is the lack of standardization, especially when integrating data from diverse clinical labs, each with its unique naming or coding conventions [3]. Furthermore, a majority of clinical laboratories do not adhere to consistent standards for result formatting and unit reporting [5]. Consequently, researchers find themselves investing a lot of time in pre-processing and cleaning laboratory data. Despite these efforts, optimal data quality remains elusive because most researchers, while skilled at data handling, are not

experts in laboratory medicine. To the best of our knowledge, there is no comprehensive framework that could act as a guiding reference for preprocessing any laboratory dataset intended for secondary use.

Clinical laboratory data are typically structured in a tabular format. In addition to the unique patient identifier and the timestamp, several variables characterize each lab test record. Among the most crucial are the test identifier, the test result value, and the test result unit for quantitative results. Collectively, we refer to these five elements as the fundamental variables integral to any laboratory dataset. Various standards have been established to address the test identifiers such as the Logical Observation Identifiers Names and Codes (LOINC) standard [6]. Similarly, the Unified Code for Units of Measure (UCUM) standard code system [7], has been created to standardize the units of measure. Given these established standards, the scope of this work has been narrowed down to primarily focus on the laboratory result values.

While some studies have explored automated methods to clean retrospective laboratory result values, they often focus on specific aspects, such as format [5] or plausibility [8, 9]. Moreover, these studies predominantly detail methods rather than offering ready-to-use tools for researchers. Consequently, there's a noticeable absence of a comprehensive automated tool tailored to address various data quality challenges of laboratory result values. Our work aims to bridge this gap by proposing a novel algorithm to automate and standardize the cleaning of clinical laboratory results. The presented lab2clean algorithm is available as an open-source R-package on CRAN [<https://CRAN.R-project.org/package=lab2clean>]. To evaluate its capacity, we tested the developed functions on the laboratory data tables from two EMR datasets, comparing the relevant data quality metrics before and after applying the tool's functions.

Methods

For improving the data quality of the laboratory result values, our focus was on improving the conformance and the plausibility Data Quality (DQ) categories as outlined by Kahn et al. in the harmonized data quality assessment framework tailored for secondary use of electronic health record data [10]. To achieve this, we programmed algorithms targeting these specific data quality dimensions and implemented them as two R functions; one to clean and standardize result formats (to improve conformance), and the other is to validate quantitative result values (to improve plausibility). Supplementary file 1 is added to guide users through the use of the package, providing clear examples of the 2 functions. Prior to applying the functions to the data, checking the structure of the laboratory data table is essential.

Tidiness check: laboratory data table structure

Checking for tidiness in laboratory data table structure is crucial. Our algorithm focuses on rectifying quality issues related to data content, and as such, is best applied to already tidy laboratory data tables. Therefore, input laboratory tables should be structured appropriately according to the criteria of tidy [11]. Firstly, every type of observational unit should have its own table. This means that all data associated with clinical laboratory tests should reside within one table, which exclusively contains clinical laboratory data, excluding other measurements like body weight or vital signs as the scope of the lab2clean algorithm is to clean only clinical laboratory test data. Secondly, every variable should be represented by a single column. This necessitates one column for at least each of the five fundamental variables; patient identifier, result timestamp, test identifier, result value, and the result unit. Finally, each observation should have its own row. This implies that every result value for a specific lab test on a particular date/date-time for an individual patient should be represented by a single record.

Function 1: clean_lab_result - improving conformance

Conformance refers to the extent to which data aligns with established internal or external formatting standards [10]. While LOINC and NPU provide standard result or scale types for laboratory test results, they don't prescribe specific standards for the format of these results. This grants laboratories the freedom to choose their preferred format for reporting result values, leading to a diverse range of formats of the same result such as '+', 'P', or 'positive', and sometimes descriptors like 'canceled' in place of a true result value [12].

To address this, we developed a three-step approach utilizing R's base regular expression functions, like 'grep' and 'gsub' [13]. Regular expressions, often abbreviated as regex, are powerful patterns used to match character combinations in strings. Our function employs these patterns to perform three key tasks: cleaning, categorizing, and standardizing result formats (Table 1). The initial step involves cleaning the data by identifying and removing extraneous variables accompanying the result value. These variables, such as measurement units (e.g., "%") or interpretative flags (e.g., "high" or "low"), violate the first rule of data tidiness mentioned earlier by representing multiple variables in a single column.

Subsequently, the function categorizes distinct result format into specific result types, adhering to LOINC's standard scale types such as Quantitative (Qn), Ordinal (Ord), Nominal (Nom). Within these scale types, our function does further subcategorization, such as differentiating simple numeric results (Qn.1) from inequalities (Qn.2), range results (Qn.3), or titer results (Qn.4) within the Quantitative scale.

Table 1 The three-step approach of the first function to clean, categorize, and standardize explained by some example lab result values

Function Input		Function Outputs			
Raw result		Result type		Clean Result	Cleaning Comments
=?	NA			NA	No result
Sample Hemolyzed	NA			NA	Test not performed
20%	20	Qn.1		20	Percent
H 25,6	25,6	Qn.1	25.6	25.6	flag
1,05		Qn.1	1.05	1.05	
> 12		Qn.2	>12	>12	
01-02		Qn.3	1-2	1-2	
1/80		Qn.4	1:80	1:80	
Negatief		Ord.1	Neg	Neg	
++		Ord.2	2+	2+	
A positief		Nom.1	A+	A+	

Clean
Categorize
Standardize

The final step involves standardizing the various formats detected for each result type (on the subcategory level) into a single, standard format. This standardization is based on the formatting guidelines developed by Hauser et al. [14], which describe preferred, acceptable, and discouraged reporting formats for each subcategory. Our function strictly converts acceptable or discouraged formats into their preferred counterparts as per these guidelines.

Moreover, this function is designed with a global perspective, supporting 19 distinct languages in representing frequently used terms such as “high,” “low,” “positive,” and “negative.” Furthermore, it addresses the variations in number formats with different decimal and thousand separators that arise due to locale-specific settings used internationally [15]. We chose to standardize these varying languages and locale-specific settings to English (EN).

Function 2: validate_lab_result - improving plausibility

Plausibility examines the credibility or truthfulness of data values [10]. In this context, atemporal plausibility is assessed by the value of a variable, especially when juxtaposed with another related variable. On the other hand, temporal plausibility is evaluated based on the sequence of values or transition of states. To reinforce the plausibility of laboratory data, this function executes three distinct checks (Fig. 1). The first two

predominantly address atemporal plausibility, while the third delves into temporal considerations by leveraging a delta analysis to pinpoint notable variations in result values over time. Those result values that fail to adhere to the criteria set by these three checks are promptly flagged by the function. It’s important to note that this function’s checks validate the result value variable, within the context of the four other fundamental laboratory data variables: test identifier, result unit, patient identifier, and result timestamp.

An initial cleaning step was undertaken to identify and flag duplicate result values for the same lab test at a particular timestamp for the same patient. These duplicate records breach the previously mentioned third data tidiness rule.

- 1) First Check: Setting boundaries with reportable limits.

We aimed to identify and flag any extremely low or high values for specific laboratory tests. To achieve this, we employed the reportable intervals of lab tests, which delineate the boundaries beyond which lab results of each test are considered incompatible with life and should ideally not be reported [16].

- 2) Second check: Logic consistency rules.

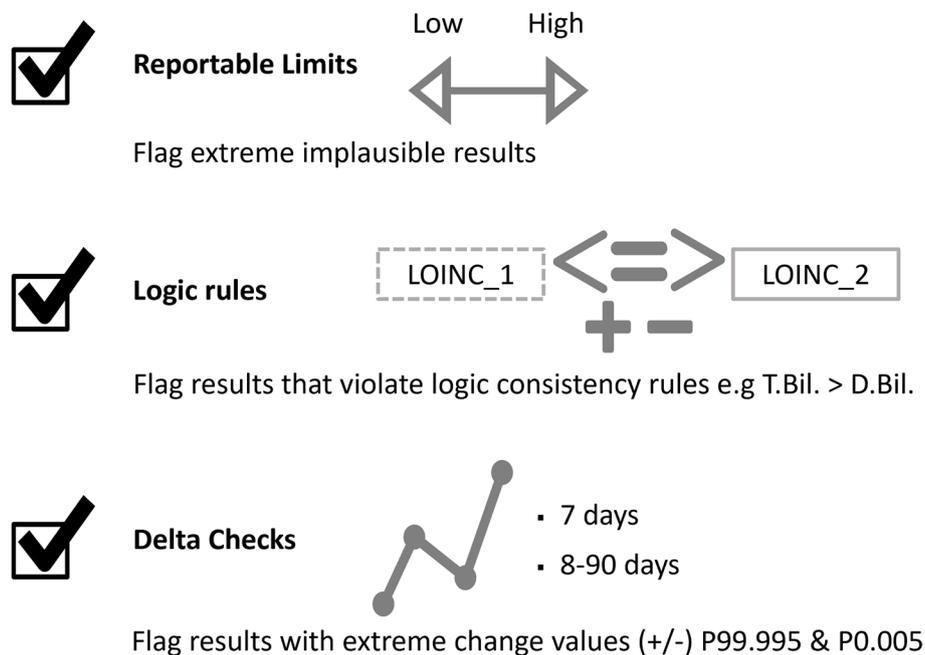


Fig. 1 Outline of the three validation checks of function 2

The second check focuses on the logical consistency of test records. When multiple tests are conducted for a specific patient at the same timestamp, there's an inherent logical relationship between the results of these tests. For instance, when both total bilirubin and direct bilirubin tests are ordered simultaneously for a patient, it's expected that the total bilirubin value isn't less than the direct bilirubin. To ensure such logical consistencies are maintained, our function adopts four established logic rules as introduced by Zhu et al. in the context of their application within auto-verification systems used in clinical laboratories [17]. Furthermore, while the function currently operates on these specific rules, we have designed it with scalability in mind. This ensures that as new insights or logic rules emerge, the function can be updated to accommodate them.

3) Third check: Delta change limits.

For over 50 years, laboratories have utilized delta checks by comparing changes in lab test result values from two sequential separate samples of the same patient against predefined thresholds to detect potential errors [18]. Our function utilizes delta checks to emphasize the temporal plausibility of laboratory test results data. It calculates both the absolute value difference and the time interval between consecutive result values of the same tests for specific patients. These time intervals are then categorized into short intervals suitable for hospitalized

patients (7 days or less) and longer intervals fitting for outpatient follow-ups (8–90 days). For tests with more than 1000 consecutive results in a time category, delta change limits are defined based on extreme percentiles for that time interval, flagging both the topmost and lowermost 0.25%, thus flagging an overall 0.5% (5/1000) of sequential results of the test that is eligible for a delta check. While this approach is effective in identifying potential data quality issues, we acknowledge that it may also flag clinically relevant changes. Therefore, it is crucial that users interpret these flagged results in conjunction with the patient's clinical context.

Evaluation datasets

We evaluated the functions of our algorithm on laboratory data tables from two contrasting retrospective clinical datasets spanning both sides of the Atlantic. The first, the INTEGO database, is a Flemish general practice-based morbidity registration network involving over 400 general practitioners (GPs) who utilize CareConnect® EMR software. Coordinated by the Academic Centre of General Practice (ACHG) at KU Leuven, Belgium, this network delivers a comprehensive view of primary care across Flanders [19, 20]. On the other hand, the Medical Information Mart for Intensive Care (MIMIC-IV v 2.0) offers a perspective from critical care. This dataset provides deidentified EMR data for more than 300,000 patients who were admitted to the intensive care units at the Beth Israel Deaconess Medical Center (BIDMC) in

the US [21–23]. The juxtaposition of these datasets not only bridges transatlantic healthcare systems but also provides insights spanning the continuum from primary to critical care.

The INTEGO database is multicentric, sourcing lab data from numerous participating laboratories within its network. This characteristic makes it particularly suited for evaluating our first function aimed at improving conformance to standardize formats of lab results from diverse sources. In contrast, the MIMIC database is single-centric, with results stemming from a single laboratory. This structure makes MIMIC an optimal choice for evaluating the secondary function, which seeks to validate result values. This validation largely hinges on other fundamental variables like the test identifier and unit of measure. Notably, these variables are more consistently standardized in the MIMIC database due to its single-centric nature.

Results

After developing the functions of lab2clean, the capacity of the developed functions was evaluated applying them to the laboratory results data of the evaluation datasets.

Tidiness check results

We assessed the structure of the laboratory data tables from both the MIMIC and INTEGO datasets using the predefined tidiness rules (Table 2). Both datasets include columns capturing the five essential variables of laboratory data. In the MIMIC database, result values are split into two columns: one presenting raw results and another reflecting these results in numeric formats. Surprisingly, records exhibiting NULL values in these two columns often contained text result values (e.g. positive, negative) within the ‘comments’ column. This distribution means that the result values span across three columns. Such an arrangement goes against the second tidiness rule, which necessitates that all result values (as one variable) -whether quantitative, ordinal, or any other type

- should be consolidated into a single column. To address this tidiness issue, we merged the numeric values with the text-based results from the ‘comments’ column, only for records with missing numeric values. This approach ensured a unified column for result values, ensuring no missingness.

Impact of Function 1 on the conformance of lab results

From the INTEGO database, we obtained around 218 million lab result records (Table 3). After de-duplication, 927,838 distinct result values remained. Function 1 processed these records in less than 5 min, successfully cleaning, classifying, and standardizing 394,337 distinct values (42.5% of the distinct result values). This standardization affected 94.5% of the total result records (frequency-based; 206 million records out of the initial 218 million).

The cleaning process involved the removal of extraneous variables such as interpretative flags or units, affecting 27,300 distinct result values. A significant majority, 90.8%, of the standardized result records were classified as Quantitative results, with the most common subtype being simple numeric results (Qn.1). Ordinal results comprised 3.6% of the total standardized records. For a comprehensive evaluation of the data quality improvement, we evaluated the rates of conformance (percentage of raw result values confirmed by the function to match the standard format) and standardization (percentage of non-standard result values standardized by the function) of both the distinct result values (presented as conformance and standardization rates) and the total result records (presented as frequency-based conformance and standardization rates) as shown in Fig. 2. It is important to note that result values which could not be classified or standardized by the function are not represented in Fig. 2; these unclassified values are further analysed in Table 3.

Table 2 Tidiness check of the evaluation datasets

	Intego	MIMIC-IV
Tidiness rule #1		
One table for all laboratory results	Yes	Yes
Tidiness rule #2		
One column for each variable	Yes	Three columns for the result value
Columns for the five fundamental variables	Yes	Yes
Tidiness rule #3		
Records (rows)	~218M records	~123M records

Table 3 Standardization results of function 2

	Result records		Distinct values	
Total results	217,917,631	(%)	927,838	(%)
Total standardized	206,033,234	94.5%	394,337	42.5%
Total not standardized	11,884,397	5.5%	533,501	57.5%
No result	451,927	0.2%	398	0.0%
Multiple results	53,259	0.0%	4,182	0.5%
Test not performed	449,870	0.2%	24,960	2.7%
Not standardized result values with <100 records	1,452,961	0.7%	497,929	53.7%
Not standardized result values with >=100 records	4,962,801	2.3%	6,030	0.6%
"volgt"	4,513,579	2.1%	2	0.0%

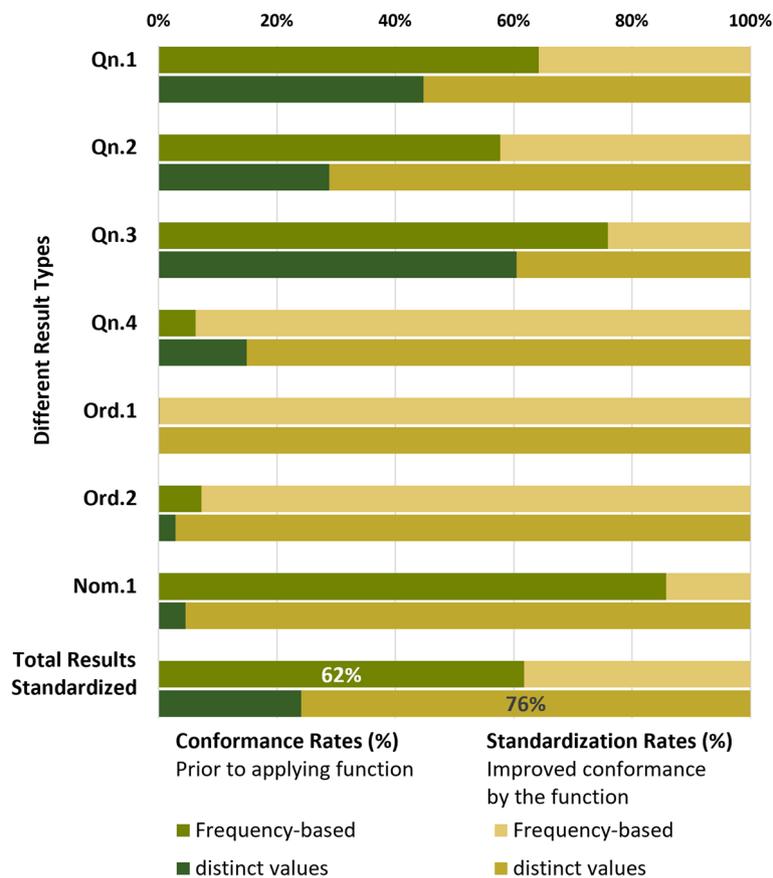


Fig. 2 Improved conformance (standardization rates) of result values by applying function 1 on the INTEGO database. Color map: ‘bamako’ by Fabio Cramer [24]

- Different result types include simple numeric results (Qn.1), inequalities (Qn.2), ranges (Qn.3), titres (Qn.4), simple ordinal results (Ord.1), grades of positivity (Ord.2), and blood group types (Nom.1).
- Conformance rates represent the percentage of raw result values that were confirmed by the function to match the standard format.
- Standardization rates represent the percentage of non-standard result values that were standardized by the function (= 100% - Conformance rates).

Across all result types, the overall frequency-based conformance rate among the total result records was 62%, indicating that standard results were more frequently encountered than non-standard ones. In contrast, the standardization rate among the distinct result values reached 76%. Typically, the frequency-based conformance rates exceeded the conformance rates for distinct results because non-standard result values were less frequent than standard ones—a reflection of the effectiveness of the standardization process.

Notably, the highest standardization rates were observed within ordinal result types, which can be attributed to their inherently limited standard formats (e.g., “Pos,” “Neg,” or “Normal” for Ord.1, and “+1,” “+2,” “+3,” or “+4” for Ord.2). This restriction in variability inherently simplifies the standardization process for these types of results, as opposed to more complex quantitative measures. The function achieved a 72% reduction in overall variability among the standardized records. Prior to standardization, there was an average of 19 distinct result values per 10,000 records. Following the application of the function, this variability decreased, resulting in an average of only 5 distinct values per 10,000 records.

The result values that the function could not classify and standardize were further analyzed in Table 3. In the cleaning process, after removing typographical errors and extraneous characters, some values became empty of characters, so they were considered to have no result. Other non-standard values included multiple results (e.g. negative and positive words in the same value) indicating long text result reports. Words that suspect that the test

was not performed (e.g. describing sample or specimen) marked 2.7% of the result values. The majority of result values that were left unstandardized by the function had a frequency of less than 100 records, which indicates the rare formats representing 93% of the unstandardized distinct result values. Out of the result values that had a frequency of more than 100 records, we observed that the unstandardized text value “volgt” - a Dutch word recorded to denote that the result value will follow - represented 2% of the total result records. This observation highlights that in certain healthcare systems, specific words or abbreviations can be frequently utilized to express the status or value of lab results. If these specific words are identified and standardized, it can further improve the standardization rate.

Impact of function 2 on the plausibility of lab results

From the MIMIC database, we obtained 100,870,373 quantitative lab result records. The function processed these records in less than 2 h, flagging 409,921 records (0.406%) as potentially inaccurate. The specifics of these flagged records are comprehensively detailed in Table 4. This demonstrates the function's capacity to validate a vast majority of the dataset, with 100,460,452 records (99.594%).

The initial step of Function 2 involved systematically identifying duplicate records. It flagged 333,574 records (0.331% of the total dataset), accounting for 81% of all flagged records. This high number of duplicates indicates potential issues in the mapping of test identifiers that may lead to assigning the same LOINC code to different tests, a violation of data tidiness principles. Ensuring no duplicate values for the same test, timestamp, and patient is crucial for data integrity.

After addressing duplicates, Function 2 conducted three detailed validation checks across the dataset. The check that resulted in the most flags was related to delta change limits, which accounted for 72,004 records (18%

of flagged records). This finding is particularly relevant considering MIMIC's nature as a critical care database where close monitoring of patients is a standard practice. The other checks, including logic consistency rules and reportable limits, flagged a smaller percentage of records but were nonetheless essential for maintaining data accuracy and consistency.

Discussion

In our study, we describe the development and assess the capacity of the lab2clean algorithm to address the challenges of utilizing clinical laboratory data for secondary research purposes. lab2clean aims to automate and standardize the complex process of cleaning clinical laboratory results, improving data quality. This focus is critical in optimizing the potential of clinical laboratory data in both clinical research and clinical ML model development.

By automating the cleaning process, lab2clean saves time and effort, typically expended in data preprocessing. Testing the developed functions on substantial laboratory datasets, encompassing over 300 million records in total, demonstrated the algorithm's fast performance and efficiency, completing tasks in less than 2 h. This efficiency allows clinical and ML researchers to shift their focus to more complex analytical tasks, thereby accelerating the research process and fostering greater productivity across various studies. However, it should be noted that a direct comparison of lab2clean with other tools for cleaning clinical laboratory data was not conducted due to the absence of comparable methodologies available for analysis.

The standardization of data cleaning across different datasets is another cornerstone of lab2clean. This uniform approach to handling vast amounts of lab data improves consistency in data quality, regardless of its source, which is crucial for the reproducibility of research outcomes and clinical ML models. The importance of

Table 4 Validation results of function 2

	Result records	
Total results	100,870,373	(%)
Total validated	100,460,452	99.594%
Total flagged	409,921	0.406%
Duplicate records (Preprocessing check)	333,574	0.331%
Result values with extreme delta change values (Check 3)		
Short interval delta flag (<=7 days)	55,397	0.055%
Long interval delta flag (8-90 days)	16,607	0.016%
Records violating logic consistency checks (Check2)	4,479	0.004%
Unreportable values (Check 1)		
Extremely low	428	0.000%
Extremely high	56	0.000%

lab2clean is further emphasized by the high volume of clinical laboratory data, estimated to comprise 70% of EMR data [25], highlighting its vital role in improving the quality of a major portion of healthcare data.

The scope of this work has focused on the laboratory result values assuming that the test identifiers and the units of measure are compliant with the LOINC and UCUM standards, respectively. This assumption, however, poses challenges in heterogeneous, multi-centric databases like INTEGO, where such standardization is often lacking. Consequently, the evaluation of function 2 for validating lab result values was impeded by the diversity in unit formats and test identifiers. To address this current limitation in the lab2clean package, future developments include additional functions for cleaning and standardizing result units and mapping test identifiers to LOINC codes. These developments aim to transform lab2clean into a more comprehensive tool for clinical laboratory data cleaning, meeting a wider range of research needs and standards.

In parallel, expanding the Lab2clean package to other programming languages other than R, notably Python, addresses a current limitation in its accessibility to researchers preferring Python. Future work includes expanding lab2clean into a Python library. Moreover, the potential integration of lab2clean algorithms into different software applications may broaden its applicability beyond retrospective data cleaning. These applications may include real-time data monitoring and quality control in clinical decision support systems (CDSS) that rely on accurate and consistent data for effective analysis and decision-making to generate accurate recommendations and support. Additionally, its use in healthcare data integration projects could facilitate better interoperability between different healthcare information systems, improving overall healthcare data management. The publication of lab2clean as an open-source software package under a general public license [26] furthers its potential for wide adoption and adaptation in various healthcare data management scenarios.

Conclusions

The lab2clean algorithm represents an advancement in the field of medical informatics, offering an automated, standardized approach to cleaning clinical laboratory results. Its capacity to improve data quality, as demonstrated in the study, underscores its potential to facilitate the utilization of clinical laboratory data for more reliable and reproducible clinical research outcomes and clinical ML models. Future work will focus on expanding its functionality and accessibility, further improving its role as a tool in healthcare data management.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12911-024-02652-7>.

Supplementary Material 1.

Acknowledgements

Authors acknowledge INTEGO and MIMIC-IV for their invaluable contribution in providing access to their datasets, enabling the comprehensive analysis and insights presented in this study.

Authors' contributions

AZ contributed to the conceptualization of the study, drafted the initial manuscript, and programmed the functions for the study. AJ assisted in programming the second function, enhancing the software's functionality. PM supervised the programming efforts, contributed to the development of the R package, and created tutorials for the package to facilitate its use. ND contributed to the general supervision and conceptualization of the study. All authors critically reviewed and approved the final manuscript, ensuring the integrity and accuracy of the work presented.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Availability of data and materials

No datasets were generated or analysed during the current study.

Declarations

Ethics approval and consent to participate

As the study was a retrospective analysis using secondary patient data only without reporting on protected health information, the requirement for obtaining informed consent was waived. The INTEGO procedures were approved by the KU Leuven Ethics Committee (nr. ML1723) and by the Belgian National Privacy Commission's Sectoral Committee (decision nr. 13.026 of March 19, 2013, last modified on April 17, 2018). The Institutional Review Board at Beth Israel Deaconess Medical Center reviewed the gathering of patient information and the development of the MIMIC-IV dataset. They approved the data sharing initiative and granted a waiver of informed consent.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Author details

¹Department of Public Health and Primary Care, KU Leuven, Leuven, Belgium.

²Laboratory Medicine Department, Menoufia University National Liver Institute, Shebin El-Kom, Egypt.

Received: 29 April 2024 Accepted: 26 August 2024

Published online: 03 September 2024

References

1. Garbage. in, garbage out. In: Wikipedia. 2023. Available from: https://en.wikipedia.org/wiki/Garbage_in,_garbage_out. Cited 2024 Feb 12.
2. Kandel S, Heer J, Plaisant C, Kennedy J, van Ham F, Riche NH, et al. Research directions in data wrangling: visualizations and transformations for usable and credible data. *Inform Visual*. 2011;10(4):271–88.
3. Abhyankar S, Demner-Fushman D, McDonald CJ. Standardizing clinical laboratory data for secondary use. *J Biomed Inform*. 2012;45(4):642–50.
4. Dixon BE, McGowan JJ, Grannis SJ. Electronic laboratory data quality and the value of a health information exchange to support public health reporting processes. *AMIA Annu Symp Proc*. 2011;2011:322–30.

5. Hauser RG, Quine DB, Ryder A. LabRS: a Rosetta stone for retrospective standardization of clinical laboratory test results. *J Am Med Inf Assoc.* 2017;25(2):121–6.
6. The Logical Observation Identifiers Names and Codes (LOINC) Committee. LOINC Users' Guide. Version 2.76. Regenstrief Institute, Inc. 2023. Available from: <https://loinc.org/kb/users-guide>.
7. Schadow G, McDonald CJ. The Unified Code for Units of Measure (UCUM) Specification Version: 2.1. Available from: <https://ucum.org/ucum>. Cited 2024 Feb 12.
8. Shi X, Prins C, Van Pottelbergh G, Mamouris P, Vaes B, De Moor B. An automated data cleaning method for Electronic Health Records by incorporating clinical knowledge. *BMC Med Inf Decis Mak.* 2021;21(1):267.
9. Estiri H, Klann JG, Murphy SN. A clustering approach for detecting implausible observation values in electronic health records data. *BMC Med Inf Decis Mak.* 2019;19(1):142.
10. Kahn MG, Callahan TJ, Barnard J, Bauck AE, Brown J, Davidson BN, et al. A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data. *eGEMs.* 2016;4(1):18.
11. Wickham H. Tidy data. *J Stat Softw.* 2014;59:1–23.
12. Hauser RG, Gisriel S, El-Khoury J. The surprising absence of a laboratory result Standard. *Am J Clin Pathol.* 2022;157(5):642–3.
13. regex function -. RDocumentation. Available from: <https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/regex>. Cited 2024 Feb 12.
14. Hauser RG, Quine DB, Iscoe M, Arvisais-Anhalt S. Development and implementation of a standard format for clinical laboratory test results. *Am J Clin Pathol.* 2022;158(3):409–15.
15. Wilcock J, Dempsey P, Sriramajeyam S, Dodiya K, Sharkey K, Gwyneth M et al. Format numbers - Globalization. 2024. Available from: <https://learn.microsoft.com/en-us/globalization/locale/number-formatting>. Cited 2023 July 28.
16. Zayed AM, Saegeman V, Delvaux N. Establishing the reportable interval for routine clinical laboratory tests: a data-driven strategy leveraging retrospective electronic medical record data. *J Appl Lab Med.* 2024;9(4):776–88.
17. Zhu J, Wang H, Wang B, Hao X, Cui W, Duan Y, et al. Combined strategy of knowledge-based rule selection and historical data percentile-based range determination to improve an autoverification system for clinical chemistry test results. *J Clin Lab Anal.* 2022;36(2):e24233.
18. Randell EW, Yenice S. Delta Checks in the clinical laboratory. *Crit Rev Clin Lab Sci.* 2019;56(2):75–97.
19. Truyers C, Goderis G, Dewitte H, Akker M, vanden, Buntinx F. The Intego database: background, methods and basic results of a flemish general practice-based continuous morbidity registration project. *BMC Med Inf Decis Mak.* 2014;14(1):48.
20. Intego. [intego.be](https://www.intego.be/). cited 2024 Feb 12. Intego - Vlaams huisartsenregistratienetwerk - INTEGO. Available from: <https://www.intego.be/>.
21. Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PCh, Mark RG, et al. PhysioBank, PhysioToolkit, and PhysioNet. *Circulation.* 2000;101(23):e215–20.
22. Johnson AEW, Bulgarelli L, Shen L, Gayles A, Shammout A, Horng S, et al. MIMIC-IV, a freely accessible electronic health record dataset. *Sci Data.* 2023;10(1):1.
23. Johnson A, Bulgarelli L, Pollard T, Horng S, Celi LA, Mark R. MIMIC-IV (version 2.0). *PhysioNet.* 2022. <https://doi.org/10.13026/7vcr-e114>.
24. Cramer F. Geodynamic diagnostics, scientific visualisation and StagLab 3.0. *Solid Earth*; 2018. Available from: <https://gmd.copernicus.org/preprints/gmd-2017-328/gmd-2017-328.pdf>. Cited 2024 Jan 10.
25. Hallworth MJ. The '70% claim': what is the evidence base? *Ann Clin Biochem.* 2011;48(6):487–8.
26. The GNU General Public License v3.0 - GNU Project. - Free Software Foundation. Available from: <https://www.gnu.org/licenses/gpl-3.0.en.html>. Cited 2024 Feb 12.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.