

Detection and Assessment of Date-Like Anomalies in Clinical Laboratory Reference Ranges

Ahmed Ayad Saleh

Department of Computer Science, College of Computer Science and Mathematics, Tikrit University, Tikrit, Salah al-Din, Iraq
ahmed.ayad@tu.edu.iq

Ali Mahmood Khalaf

Department of Computer Science, College of Computer Science and Mathematics, Tikrit University, Tikrit, Salah al-Din, Iraq
ali.mahmood@tu.edu.iq

Alhakam Ayad Salih

Department of Computer Science, College of Computer Science and Mathematics, Tikrit University, Tikrit, Salah al-Din, Iraq
Alhakam.a.allawie@tu.edu.iq

Abstract

The interpretation of clinical laboratory data is greatly influenced by reference ranges. But these ranges may result in date-like expressions which can block direct numeric processing and obscure the original range limits. In this study, this is explored with the Multi-Center Clinical Laboratory Dataset (MCLD) by integration of Rule Based Detection (RBD), Numeric Parsing Validation (NPV) and Rule Based Candidate Reconstruction (RBCR). The expression for dates is discovered by RBD, the range for interpretation is checked by the NPV, and potential numeric intervals are created by the RBCR. The candidates generated are evaluated for consistency and ambiguity within the candidates. In 145,288 laboratory records examined, 5,054 (3.48%) had date-like reference-range expressions, representing 18 different combinations of test name, expression, and measurement units. 18 cases were coded as numeric candidates. These candidates matched the Normal and Abnormal classifications on 5054 affected records. An additional analysis found compatible alternative intervals in 14 of the 18 cases within a limited search grid of integer boundaries. The results demonstrate that there is not always a single reference interval associated with a particular classification. The framework enables identification of anomalies that fall outside of date ranges, some possible numeric corrections, and the assessment of internal consistency and ambiguity of those corrections.

Keywords: Clinical Laboratory Data, Reference Range, Data Quality, Date-Like Anomalies, Candidate Reconstruction.

1. Introduction

Reference ranges are very commonly used to interpret clinical laboratory results. These are used as a framework to differentiate values that fall within a range of acceptable values from values that need more clinical attention. Guidance for establishing and verifying laboratory reference intervals is provided by the Clinical and Laboratory Standards Institute (CLSI) [1]. Scientific data may be corrupt and difficult to interpret due to data formatting issues [2, 3]. Reference ranges represented in date formats in clinical laboratory data can cause difficulty with direct numeric processing and may make it hard to determine the true limits of the range. The Multi-Center Clinical Laboratory Dataset (MCLD) is a dataset that allows analysis of this issue in real laboratory records [4]. This study explores the possible numeric interpretation for the suspected date-like expressions in the reference-range field, based on the available test results and the recorded classification. The framework is based on a combination of Rule Based Detection (RBD), Numeric Parsing Validation (NPV) and Rule Based Candidate Reconstruction (RBCR). This study addresses a particular issue of clinical laboratory reference ranges, as opposed to general purpose data repair systems with the power to repair numerous types of errors. It makes an unprecedented effort to identify expressions suspected of being dates, makes a clear distinction between generating candidates for dates and verifying them, and analyzes whether the recorded classification can differentiate between the proposed time period and other time periods. All of this is not a new general purpose data cleaning algorithm introduced in the framework.

2. Related Work

Scientific data may be flawed because of data quality issues, and the conclusions drawn from scientific data may be incorrect. Van den Broeck et al. [5, 10, 11] defined data cleaning as the process of identifying, analyzing and fixing potential errors. Broman and Woo [2, 12] described some of the issues that arise with spreadsheet data and gave some rules for formatting the data in a consistent way. The examples cited by Ziemann et al. [3] demonstrated that data format changes can lead to errors in the date fields of gene names. There are some studies that propose ways to detect and repair errors in structured data. Chu et al. [6] surveyed the techniques of data cleaning using rules, constraints, patterns and stats. Rekatsinas et al. [7] also proposed a new method called HoloClean, which utilizes data constraints and statistical information to facilitate data repair. Dallachiesa et al. [13] proposed a data cleaning system (NADEEF) for detecting and repairing data errors according to the various types of data quality rules. Abedjan et al. [14] reviewed the current data error detection tools and pointed out their shortcomings in detecting different types of errors in real-world data.

Several error detection methods and minimal user input are used to determine whether errors are present in tabular data using Raha, a system developed by Mahdavi et al. [8]. Reference ranges are directly involved in the interpretation of laboratory test results in clinical situations. Reference intervals are highlighted in the CLSI guideline EP28-A3c [1] which sets procedures for establishing

and verifying reference intervals. This is important when considering the potential for corrections, as the meaning of a reference range is dependent on the test and its clinical context. Heidari et al. [15] introduced the machine learning-based framework HoloDetect, which is designed to detect errors using few-shot learning and data augmentation.

These studies could be used to identify and fix errors in structured data. There are a number of data-quality related tasks that HoloClean and Raha solve, but the present study focuses on a more limited problem: expressions related to dates in clinical laboratory reference ranges. In this work, a generated interval is not considered a validated correction but rather is assessed against the recorded classifications and the ability of other intervals to agree with the classifications as well. This case-level ambiguity assessment is the main analytical contribution.

3. Materials and Methods

The Important section in the study where the dataset is described, the analysis of reference range formats is explained, the detection and numeric range validation are described and the reconstruction and assessment of candidate intervals are explained.

3.1 Dataset Description:

The Multi-Center Clinical Laboratory Dataset (MCLD) was used for this study, which included 145,288 laboratory records and 168 different test names. Analysis included the reference range expression, test name, the measurement unit, the numeric result, and the classification (Normal or Abnormal) as recorded. The file released contained two different values in its location field.

3.2 Reference Range Analysis:

Reference-range expressions were classified as numeric intervals, one-sided numeric limits, qualitative expressions, date-like expressions, or contextual/other expressions. These categories help to differentiate expressions that can be used for numeric interpretation from others.

3.3 Rule-Based Detection:

Rule Based Detection (RBD) was used to detect expressions that are a one- or two-digit number followed by a hyphen and a three-letter abbreviation of a month (in any order). These are such as 25-Apr and Aug-61. The expressions were flagged as suspected anomalies without changing the source records.

3.4 Numeric Parsing Validation:

Numeric Parsing Validation (NPV) determined if each reference-range expression could be interpreted as a numeric interval or as one-sided numeric limit. The rest of non-parseable expressions were captured separately and checked against the date-like patterns that were found by RBD.

3.5 Rule-Based Candidate Reconstruction:

Two explicit rules were used to produce candidate numeric intervals using Rule Based Candidate Reconstruction (RBCR): the first rule converts the month abbreviators to month numbers, the second rule sorts the resulting numeric components in ascending order. For instance, 4–25 was created for 25-Apr, but this interval was not confirmed as original reference values.

3.6 Candidate Assessment:

The numeric laboratory results were coded as 'with' range if the result was the lower bound and the upper bound for each candidate interval. The levels of these classifications were compared with the recorded normal and abnormal outcomes to evaluate internal consistency. Ambiguity of the candidates was explored by adding integer lower and upper bounds within five units of each bound that is generated. Lower limits were specified to be non-negative integers and upper limits were specified to be greater than lower limits. An alternative that reproduced all available classifications of the recorded case of test–expression–unit was compatible. Figure 1 below shows the scope analysis framework reference.

Reference-range analysis framework



Figure 1. Scope Analysis Framework Reference.

4. Results

The distribution of reference-range formats, the date-like anomalies detected, and the candidate reconstruction and assessment results are presented in this section.

4.1 Reference Range Formats and Detection Results:

Among the 145,288 records, numeric intervals accounted for 81,095 (55.82%), one-sided numeric limits for 55,621 (38.28%), qualitative expressions for 3,513 (2.42%), date-like expressions for 5,054 (3.48%), and contextual/other expressions for five records (less than 0.01%). 8,572 records were found by Numeric Parsing Validation (NPV) that could not be parsed numerically. We used Rule Based Detection (RBD) and it identified 5,054 suspected date records, which were all part of the NPV group. The remaining 3,518 non-parseable records were qualitative or contextual/other expressions, and were not counted as suspected date-like corruption. The 5,054 date-like were

58.96% of all non-numeric parsing failures, while the other 41.04% were other formats. These percentages refer to the parsing failures, not the number of actual errors.

Table 1. Distribution of reference domain formats.

Format	Records	Percentage
Numeric interval	81,095	55.82%
One-sided numeric limit	55,621	38.28%
Qualitative	3,513	2.42%
Date-like (suspected)	5,054	3.48%
Contextual/other	5	<0.01%
Total	145,288	100%

4.2 Candidate Reconstruction and Assessment:

All 5,054 date-like records had 16 different expressions and 18 different test expression unit cases table 2. For each of the 18 cases, candidate numeric intervals were generated. The candidate intervals for all 5,054 records reproduced the recorded normal and abnormal classification. This comes from internal agreement and not a separately calculated reconstruction accuracy. All 81094 numeric interval records for which a similar classification could be done also matched their recorded limits, suggesting that the labels are not standalone evidence of the original limits. 1,201 records are affected by at least one alternative interval found as a result of the local integer-boundary search, which identified at least one alternative interval in 14 of the 18 cases. The other four cases within this limited grid were not made up of alternatives, so this does not mean that the candidates for those cases are uniquely correct.

Table 2. Summary candidate assessment.

Measure	Result
Date-like records	5,054
Distinct test-expression-unit cases	18
Cases with generated numeric candidates	18
Records agreeing with candidate classifications	5,054
Cases with local compatible alternatives	14
Records in cases with local alternatives	1,201

5. Discussion

There were 5,054 records (3.48%) in the MCLD reference-range field that were affected by date-like expressions. These expressions formed a small percentage of the data set, but were not expressible as numerical ranges directly. It was important to differentiate them from the non-numeric expressions so as not to mistake the qualitative reference ranges as formatting errors. The comparison between the Rule Based Detection (RBD) and Numeric Parsing Validation (NPV) provides a good illustration of why it is important to differentiate the type of errors. NPV found 8,572 non-parseable records with 5,054 (58.96%) records containing date-like patterns identified by RBD. The other 3,518 records were qualitative or contextual expressions. Therefore, counting all

non-parseable expressions as errors would be an overestimate of the number of possible anomalies of the date type. Rule Based Candidate Reconstruction (RBCR) resulted in numeric candidates for all 18 cases, which shared the classification with the recorded ones. The same data, had no records missing from the eligible eligibility intervals. The recorded labels might thus only represent reference ranges established in the source data and not independently confirm the reconstructed candidates. The reconstruction results are also difficult to interpret because of the ambiguity analysis. The same classifications were obtained in 14 of the tested local grid locations, using nearby alternative integer intervals. This finding illustrates that a candidate could be consistent with the data, but not uniquely identified by the data. In the other four cases, there is no alternative other than the search grid used here.

An external contextual review was conducted for six of the 18 cases. Published laboratory reference values can be used as a guide but cannot be used to set the original reference values of the contributing laboratories. For instance, the National Board of Medical Examiners (NBME) specifies that the level of free T4 (Free T4) is 12.0–21.9 pmol/L, while the candidate's level was in the range of 10–23 pmol/L. It also has sex-dependent limits for its prolactin [9]. No candidate was confirmed as the original reference interval without the assistance of other candidates. The external comparisons were therefore discussed in the context of clinical context and not in terms of accuracy of reconstruction. The benefit of this study is the separation of suspected-anomaly detection and structural candidate generation and evidence assessment for a defined laboratory data-quality problem. A practical limitation of label-based evaluation is found in the results: classified intervals can overlap with a number of candidate intervals. The study here gives a descriptive analysis of the anomalies detected in clinical laboratory data sets (MCLD) that are also beginning to outline the possibilities for and limitations in data reconstruction verification. Does not contain a direct comparison with other data-repair systems, e.g. HoloClean or Raha.. Raha..

6. Conclusion

In this study, we investigated date-like reference range anomalies in the Multi-Center Clinical Laboratory Dataset (MCLD) by applying a Rule-Based Detection (RBD), Numeric Parsing Validation (NPV) and Rule-Based Candidate Reconstruction (RBCR) interpretable framework. There were 5,054 date-like expressions in 145,288 records, in 18 different test expression unit cases (3.48%). The 18 cases were internally consistent in terms of their recorded laboratory classifications and numeric candidates were generated for each case. The local ambiguity analysis led to the identification of 14 alternative intervals that were compatible. This finding is consistent with the separation of detecting suspected anomalies and generating and independently verifying possible corrections. A further 58.96% of numeric parsing errors occurred for date-like expressions, a warning to those who are tempted to assume that anything that looks like a date is actually a date.

References

- [1] Clinical and Laboratory Standards Institute. Defining, Establishing, and Verifying Reference Intervals in the Clinical Laboratory. CLSI document EP28-A3c. 2010. <https://clsi.org/shop/standards/ep28/>
- [2] Broman KW, Woo KH. Data Organization in Spreadsheets. The American Statistician. 2018;72(1):2–10. <https://doi.org/10.1080/00031305.2017.1375989>
- [3] Ziemann M, Eren Y, El-Osta A. Gene name errors are widespread in the scientific literature. Genome Biology. 2016;17:177. <https://doi.org/10.1186/s13059-016-1044-7>
- [4] Mahmood ML, Rawf KM, Mohammed HM, Salih WM, Abdalqadir AH. Multi-Center Clinical Laboratory Dataset (MCLD). Mendeley Data. 2026;Version 1. <https://doi.org/10.17632/xxkd3kjfd.1>
- [5] Van den Broeck J, Argeseanu Cunningham S, Eeckels R, Herbst K. Data Cleaning: Detecting, Diagnosing, and Editing Data Abnormalities. PLOS Medicine. 2005;2(10):e267. <https://doi.org/10.1371/journal.pmed.0020267>
- [6] Chu X, Ilyas IF, Krishnan S, Wang J. Data Cleaning: Overview and Emerging Challenges. Proceedings of the 2016 International Conference on Management of Data. 2016:2201–2206. <https://doi.org/10.1145/2882903.2912574>
- [7] Rekatsinas T, Chu X, Ilyas IF, Ré C. HoloClean: Holistic Data Repairs with Probabilistic Inference. Proceedings of the VLDB Endowment. 2017;10(11):1190–1201. <https://doi.org/10.14778/3137628.3137631>
- [8] Mahdavi M, Abedjan Z, Castro Fernandez R, Madden S, Ouzzani M, Stonebraker M, Tang N. Raha: A Configuration-Free Error Detection System. Proceedings of the 2019 International Conference on Management of Data. 2019:865–882. <https://doi.org/10.1145/3299869.3324956>
- [9] National Board of Medical Examiners (NBME). Laboratory Values [Internet]. Available from: <https://www.nbme.org/laboratory-values/> (accessed 15 October 2026).
- [10] Rahm E, Do HH. Data Cleaning: Problems and Current Approaches. IEEE Data Engineering Bulletin. 2000;23(4):3–13. <https://dbis.uni-leipzig.de/research/publications/data-cleaning-problems-and-current-approaches>
- [11] Karr AF, Sanil AP, Banks DL. Data quality: A statistical perspective. Statistical Methodology. 2006;3(2):137–173. <https://doi.org/10.1016/j.stamet.2005.08.005>

-
- [12] Kandel S, Heer J, Plaisant C, Kennedy J, van Ham F, Henry Riche N, et al. Research directions in data wrangling: Visualizations and transformations for usable and credible data. *Information Visualization*. 2011;10(4):271–288. <https://doi.org/10.1177/1473871611415994>
- [13] Dallachiesa M, Ebaid A, Eldawy A, Elmagarmid AK, Ilyas IF, Ouzzani M, Tang N. NADEEF: A commodity data cleaning system. *Proceedings of the ACM SIGMOD International Conference on Management of Data*. 2013:541–552. <https://doi.org/10.1145/2463676.2465327>
- [14] Abedjan Z, Chu X, Deng D, Castro Fernandez R, Ilyas IF, Ouzzani M, et al. Detecting data errors: Where are we and what needs to be done? *Proceedings of the VLDB Endowment*. 2016;9(12):993–1004. <https://doi.org/10.14778/2994509.2994518>
- [15] Heidari A, McGrath J, Ilyas IF, Rekatsinas T. HoloDetect: Few-Shot Learning for Error Detection. *Proceedings of the 2019 International Conference on Management of Data*. 2019:829–846. <https://doi.org/10.1145/3299869.3319888>