CED-KQN - Data quality assurance in digital clinical registries

J. de Laffolie[1], K. Sohrabi[2], H. Schneider[2], N.Schneider[2], K Zimmer and CEDATA-GPGE®

[1] General pediatrics & neonatology, Justus-Liebig-University, Gießen, Germany
[2] Institute of medical Computer Science, Justus-Liebig-University, Gießen, Germany

Members of the CEDATA-GPGE® study group: S. Buderus (Bonn, Deutschland), M. Claßen (Bremen, Deutschland), S. Dammann (Stuttgart, Deutschland), J. de Laffolie (Gießen, Deutschland), A. Hauer (Graz, Österreich), KM. Keller (Wiesbaden, Deutschland), S. Koletzko (München, Deutschland), A. Krahl (Darmstadt, Deutschland), M. Laaß (Dresden, Deutschland), T. Lang (Regensburg, Deutschland), C. Posovszky (Ulm, Deutschland), B. Rodeck (Osnabrück, Deutschland), S. Trenkel (Potsdam, Deutschland), KP Zimmer (Gießen, Deutschland)

Objectives and Study:

Clinical registries have been proven as a vital tool to further research concerning rare diseases like pediatric onset inflammatory bowel disease (PIBD) as they aggregate a sufficient amount of data for a better understanding of disease phenotypes and timeline and the support of clinical trials.

CEDATA-GPGE® is a large clinical patient registry for children and adolescent with PIBD by the Association for Pediatric Gastroenterology and Nutrition (GPGE e.V.) in Germany and Austria focusing on the improvement of the care. It contains data over 5,000 patients and over 50,000 contacts. As high data quality is an essential aspect for developing and maintaining clinical registries, the current CED-KQN project aims to analyze, assure and improve data quality.

Methods:

Verification of data quality in the CEDATA registry began by analyzing existing data for pre-existing errors. Features were examined regarding the expected value-type or range and their connection to other features. The results were transferred to a feature catalog, naming feature and its validity rules. An algorithm was implemented to find erroneous fields based on the catalog. These were corrected either by collecting data from the original medical records of the corresponding clinic or by deleting unrecoverable values. The corrected data was then migrated back into the registry. Afterwards, the previously identified errors were categorized by cause. Frequently occurring mistakes, were incorporated into a live validation system inside the registry.

Results:

When correcting the CEDATA registry we defined 443 plausibility rules for the registry features. In a total of 1174 fields across all forms, 114 were erroneous at least once compared to 1060 errorless columns.

The most common mistake was a missing decimal point in features such as weight which therefore is recorded in the wrong unit. Other common errors are values out of expected range. After a corrective migration no further errors were identified.

Data validation patterns are realized using a component-based form system. This system employs questionnaire elements specifically engineered to the feature type.

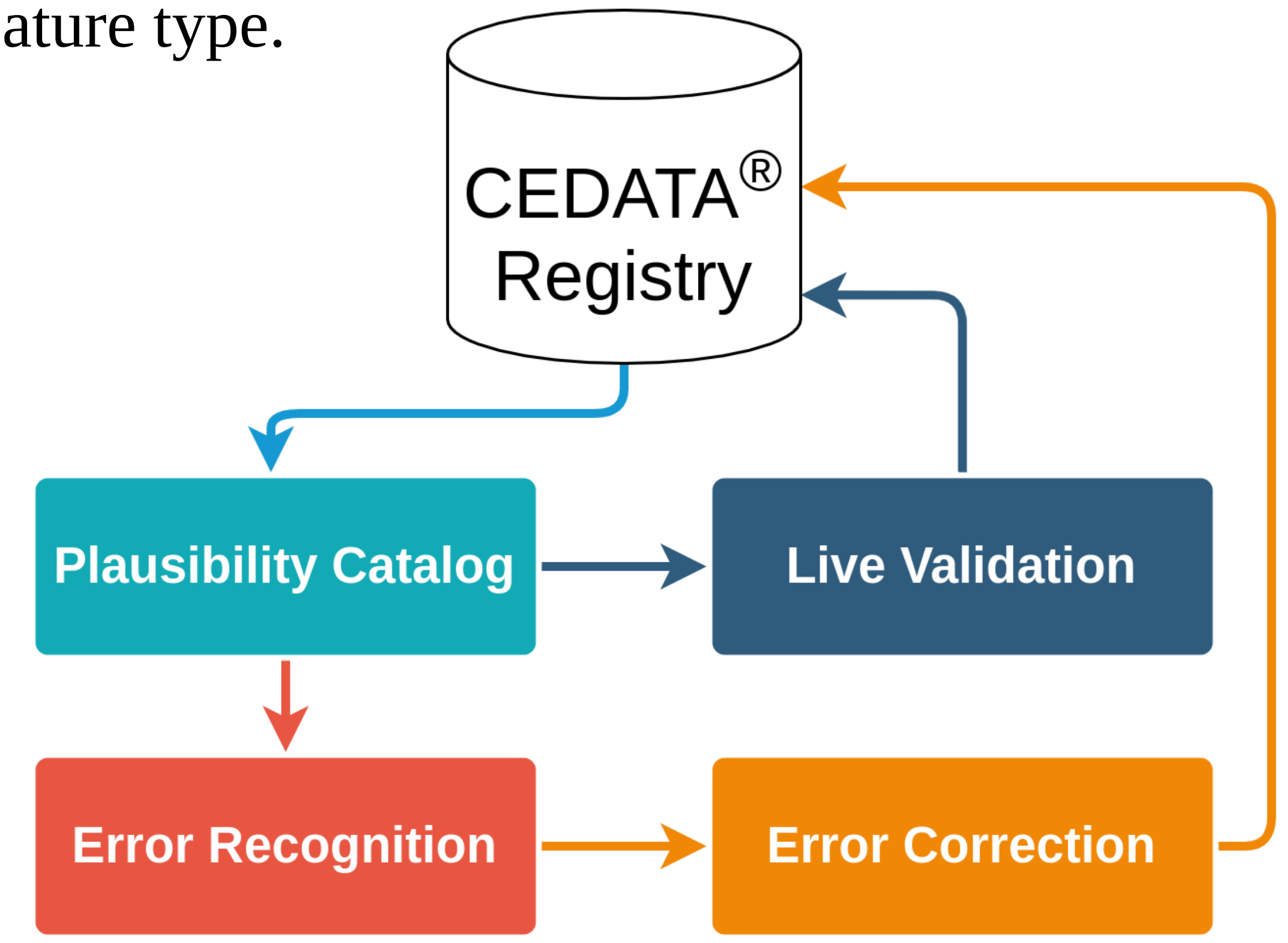


Figure 1: Workflow of the applied methods, beginning with the plausibility catalog as basis for error recognition and validation.

Discussion:

Deploying a property catalog of features while utilizing live data validation and user feedback provides an efficient system for error recognition and assurance of high- quality data in the digital CEDATA-GPGE® registry. This approach can be added incrementally to a registry without redesigning the entire application. An evolutionary approach can also include insights gained during an error recognition and correction process.

Based on an extensible catalog, other features like completion assistance can be added. The source of most errors on the clinical registry CEDATA is most likely human factor, comprised of omitted decimal points, misreading or typewriting mistakes. While these faults can severely corrupt a database, they are often easily fixable, if the user is provided with timely input feedback.

The underlying feature catalog however requires extensive interdisciplinary exchange to capture all the necessary features for each of the registry's parameters.

The implemented concept is a universal approach and an expandable architecture, that can be applied to other clinical registries.

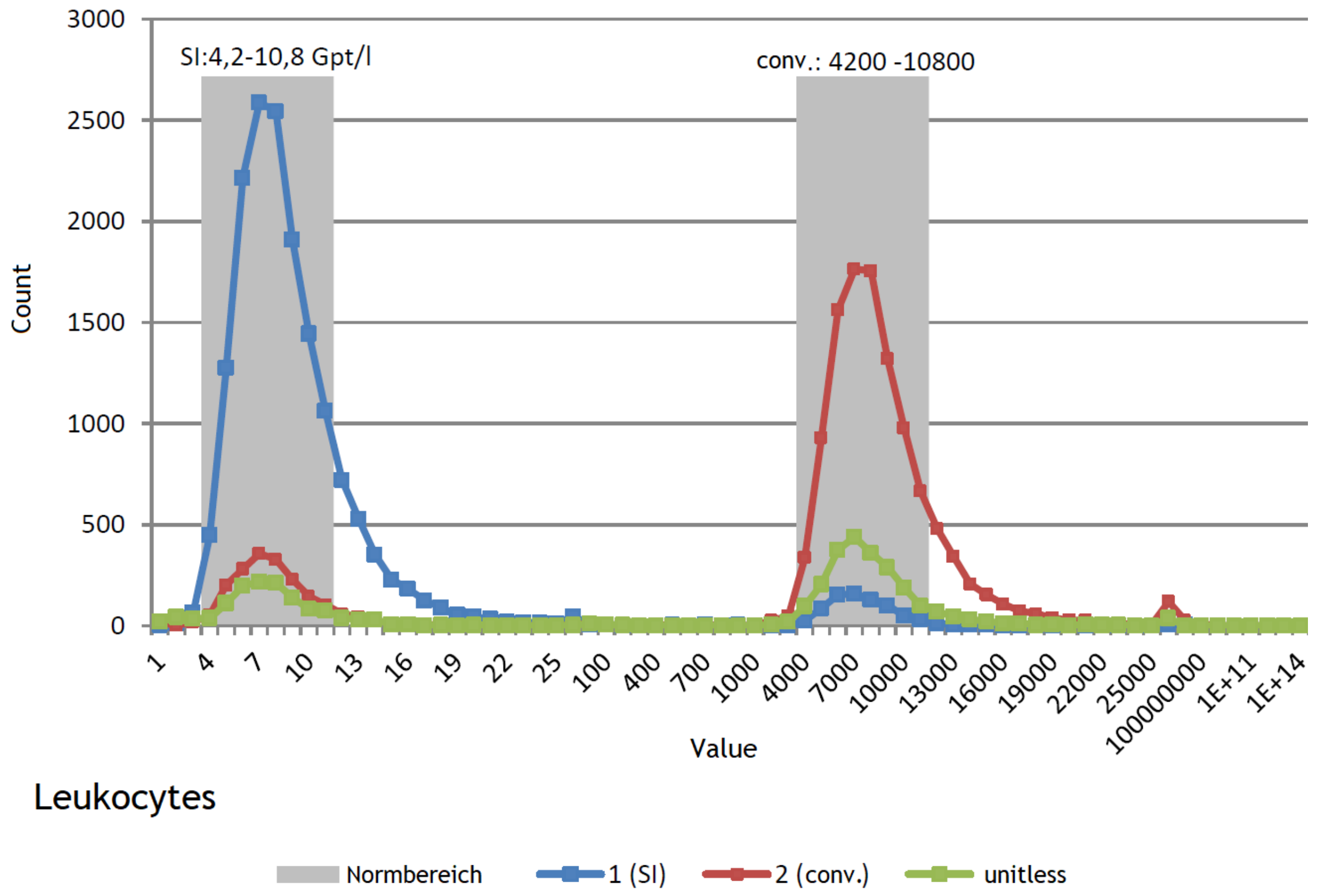
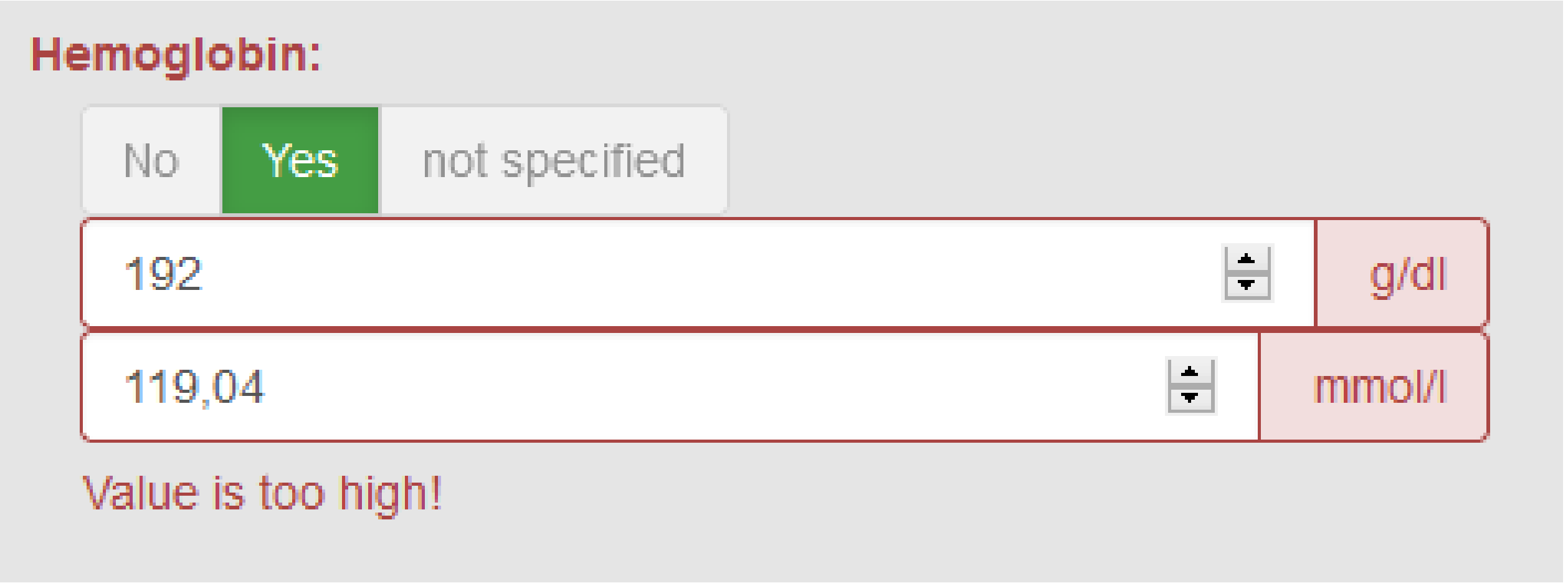


Figure 2: Diagram of leukocyte values in the CEDATA registry by frequency of occurrence with additional accentuation of the normal range.

Laboratory Features	Conversion Factor
ALAT	1 µmol/sl → 60 U/I
ALB	1 g/l → 0,1 g/dl
Hemoglobin	1 mmol/l → 1,61 g/dl
Creatinine	1 µmol/l → 0,0113 mg/dl
Leukocytes	1 Gpt/l → 1000 1/µl
MCV	1 fl → 1 µm^3

Table 1: Extraction from the conversion table used for multi unit laboratory inputs.


Hemoglobin:

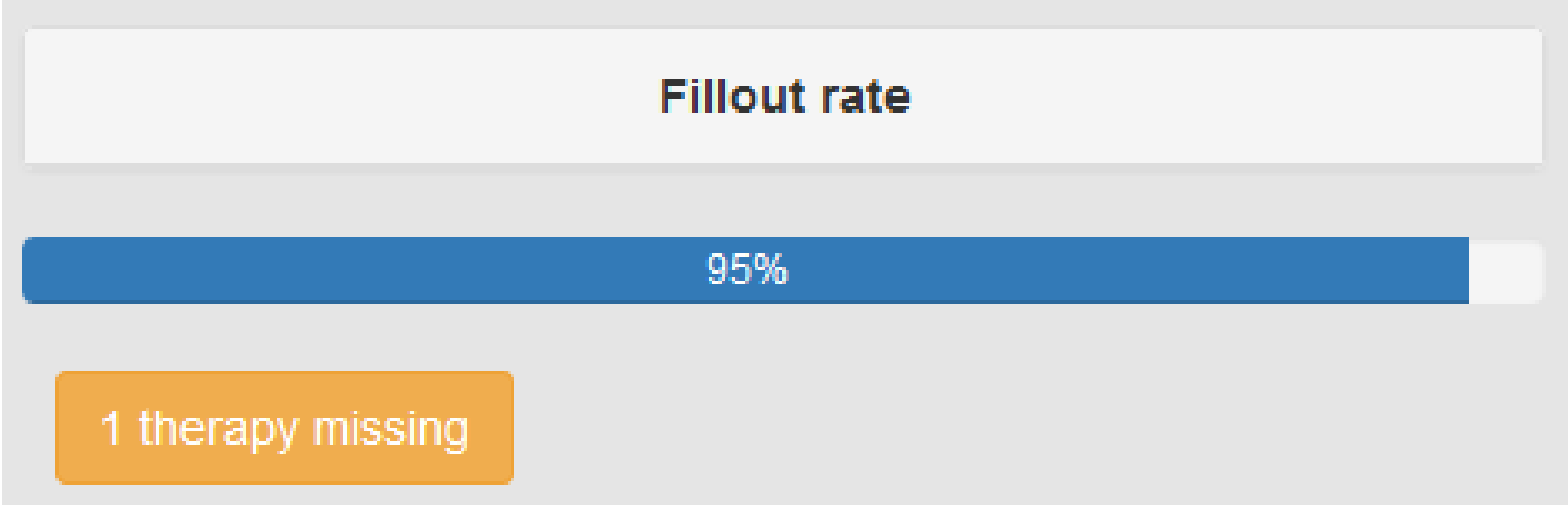
No | Yes | not specified

192 g/dl

119,04 mmol/l

Value is too high!

Figure 3: Multi-unit component with erroneous user feedback for entering laboratory values in the CEDATA registry.


Fillout rate

95%

1 therapy missing

Figure 4: Fill out rate of the opened documentation. Progress is shown in percent, missing values are listed and also linked.

Summary:

A performant and adaptable validation and error recognition system is implemented in the clinical digital registry using an extensible feature catalog. While clinical registries have been proven vital to further research concerning rare diseases like PIBD they are highly depended on a high data quality throughout their recorded observations. Errors in existing data have been identified and corrected, newly entered data is checked on input and feedback is displayed to the user including scores and laboratory values conversed to common metrics. Additional input assistance is provided, enabling a timelier entry and insights into the progress of documentation completion. The incorporation of the error recognition results in the validation system enabled the prioritization of the most common mistakes, most of which are caused by human error. The achieved improvement of data quality is essential for further research using the CEDATA registry, especially when employing statistical models or machine learning.

Acknowledgments: We thank all members of the CEDATA-GPGE® study group for their support in recruitment as well as our patients and their families. CEDATA-GPGE® was endorsed by the „Falk Foundation", „Vifor Pharma", Takeda, Abbvie and the „Deutsche Morbus Crohn/Colitis ulcerosa Vereinigung e.V". „DCCV". The CED-KQN project is funded by the „Gemeinsamen Bundesausschuss"(Federal Joint Committee Germany), CEDKQN, VSF17054

Literature:

1 Buderus, Stephan; Scholz, Dietmar; Behrens, Rolf; Classen, Martin; Laffolie, Jan de; Keller, Klaus-Michael et al. (2015b): Inflammatory bowel disease in pediatric patients: Characteristics of newly diagnosed patients from the CEDATA-GPGE® registry. In: Deutsches Ärzteblatt international 112 (8), S. 121–127. DOI: 10.3238/arztebl.2015.0121.

2. Stenzhorn H, Weiler G, Brochhausen M, Schera F, Kritsotakis V, Tsiknakis M, et al. The ObTiMA system - ontology-based managing of clinical trials. Stud Health Technol Inform. 2010;160(Pt 2):1090–4.

3. Kodra Y, Posada de la Paz M, Coi A, Santoro M, Bianchi F, Ahmed F, et al. Data Quality in Rare Diseases Registries. Adv Exp Med Biol. 2017;1031:149–64.

e-mail: jan.delaffolie@paediat.med.uni-giessen.de

The authors declare, that no conflict of interest exists.

Presented: @ WCPGHAN 2021