Targeted Data Quality Analysis for a Clinical Decision Support System for SIRS Detection in Critically Ill Pediatric Patients

Erik Tute[1] Marcel Mast[1] Antje Wulff[1,2]

[1] Peter L. Reichertz Institute for Medical Informatics of TU Braunschweig and Hannover Medical School, Hannover Medical School, Hannover, Niedersachsen, Germany
[2] Big Data in Medicine, Department of Health Services Research, School of Medicine and Health Sciences, Carl von Ossietzky University Oldenburg, Oldenburg, Niedersachsen, Germany

Address for correspondence Erik Tute, MSc, Peter L. Reichertz Institute for Medical Informatics of TU Braunschweig and Hannover Medical School, Hannover Medical School, Carl-Neuberg-Str. 1, 30625 Hannover, Germany (e-mail: Erik.Tute@plri.de).

Methods Inf Med 2023;62:e1–e9.

Abstract

Background Data quality issues can cause false decisions of clinical decision support systems (CDSSs). Analyzing local data quality has the potential to prevent data quality-related failure of CDSS adoption.

Objectives To define a shareable set of applicable measurement methods (MMs) for a targeted data quality assessment determining the suitability of local data for our CDSS.

Methods We derived task-specific MMs using four approaches: (1) a GUI-based data quality analysis using the open source tool *openCQA*. (2) Analyzing cases of known false CDSS decisions. (3) Data-driven learning on MM-results. (4) A systematic check to find blind spots in our set of MMs based on the *HIDQF* data quality framework. We expressed the derived data quality-related knowledge about the CDSS using the 5-tuple-formalization for MMs.

Results We identified some task-specific dataset characteristics that a targeted data quality assessment for our use case should inspect. Altogether, we defined 394 MMs organized in 13 data quality knowledge bases.

Conclusions We have created a set of shareable, applicable MMs that can support targeted data quality assessment for CDSS-based systemic inflammatory response syndrome (SIRS) detection in critically ill, pediatric patients. With the demonstrated approaches for deriving and expressing task-specific MMs, we intend to help promoting targeted data quality assessment as a commonly recognized usual part of research on data-consuming application systems in health care.

Keywords
- information
- data quality
- aggregation
- knowledge bases
- clinical decision support systems

Introduction

Data quality is recognized as an important topic for health care. ISO 8000–2:2020 defines data quality as "degree to which a set of inherent characteristics of data fulfils requirements."[1] Much research on health data quality originates from health research, e.g., focusing on clinical trials, health service research, epidemiology, or the quality of input data for machine learning approaches.[2–9] In the context of health research, good data quality requires that researchers can

received
June 30, 2022
accepted after revision
October 21, 2022
article published online
January 11, 2023

DOI https://doi.org/10.1055/s-0042-1760238.
ISSN 0026-1270.

Georg Thieme Verlag KG, Rüdigerstraße 14, 70469 Stuttgart, Germany

extract reliable knowledge from a dataset. In health care, the requirement for good data quality often is that data can be the basis for correct individual decisions.

Emerging digitalization in health care promotes use of data in medical decision-making. If individual decisions are made with the help of data, this ultimately necessitates a good quality of data. This is especially critical when data are not used directly but indirectly through tools, which advise decisions in health care. Clinical decision support systems (CDSSs) advise medical personnel by analyzing data captured throughout medical practice and presenting an alert, a suggestion, or new insight. The design of such systems—with respect to reusability in cross-institutional settings by integrating international and open interoperability standards for data representation—is an important research focus of our group. In the CADDIE project, we designed an interoperable CDSS for the detection of systemic inflammatory response syndrome (SIRS) and sepsis in pediatric intensive care settings.[10] An interdisciplinary team of clinicians, scientists, and administrators designed a knowledge-based approach able to detect critical phases throughout the clinical pathway of the critically ill child only by using routine data. The diagnostic accuracy of the approach was evaluated in a prospective diagnostic study at the Pediatric Cardiology and Intensive Care Medicine of the Hannover Medical School.[11] The study resulted in promising accuracies; however, also limitations potentially due to low data quality were revealed.

Since 2020, improvements are researched in the project ELISE (a learning, interoperable, and smart expert system for pediatric intensive care).[12] In ELISE, data-driven algorithms for prediction of SIRS/sepsis and associated organ dysfunctions are developed to explore the question of the extent to which CDSS can be used to optimize diagnostic and therapeutic workflows in pediatric intensive care. Amongst others, the routine dataset used in CADDIE is enhanced by more intensive care parameters resulting in a broad training dataset on which data-driven prediction algorithms are applied.[13,14] In the end, an open demonstrator of an interoperable, data-driven CDSS for detection and prediction of SIRS/sepsis and associated organ dysfunctions in critically ill children will arise which also will be evaluated in a clinical-driven study.

The ultimate performance measure for such a CDSS would be an improved patient outcome, e.g., a reduction of deaths. For this work, we assume that false decisions of the CDSS are not beneficial for the patient and thus use the rate of correct decisions as a measure for CDSS performance that can deteriorate due to low quality of the input data. For example, ELISE can only detect a SIRS episode if at least a current body temperature or laboratory value is present. Documentation processes may influence some of the data's characteristics, e.g., automatic vital sign measurements are typically present with higher frequency than manually documented values. Such different characteristics may not be a general problem in the data, but can still be an issue for a certain data usage like a CDSS. Thus, the task-dependent data quality can differ between clinical sites, even if all these sites have sound documentation processes. Besides spatial data quality differences, data quality can change over time, e.g., due to changed processes (cf. Sáez et al[15]).

It is important to prevent bad CDSS performance to avoid unsatisfied users or even harm to patients. Detection of unsuitable local data before rolling out a CDSS or during its operation allows us to react accordingly, e.g., by adapting documentation processes or at least by communicating realistic expectations about CDSS performance. As a growing number of health care organizations operate clinical data warehouses or some suitable alternative (e.g., in Germany data integration centers[16]), there is a growing chance that the needed data are available to conduct a targeted data quality assessment as preparation before CDSS rollout or even regularly to notice problematic changes in data quality. To conduct such a targeted data quality assessment, health care organizations need to know which inherent characteristics of the data are relevant for the use case and how to assess the degree to which these characteristics fulfil the requirements of the use case (cf. ISO definition of data quality). A well-known means to "describe actual and potential deviations from defined requirements"[7] is data quality indicators. Thus, a well-selected set of relevant data quality indicators and thresholds defining how to assess the deviations from requirements would be an ideal conception of how to express the needed knowledge for targeted data quality assessment. However, typically not all data quality-related information can be expressed as an indicator or at least the knowledge needed to define an indicator and its assessment is not available for all relevant aspects. Detection of data quality issues based on descriptors,[7] e.g., graphs or other outputs that need interpretation for the assessment, is common practice. Furthermore, data quality indicators and descriptors are often defined in a textual form, expressing the intention but leaving room for interpretation in their operationalization.

In the following, we present our work to identify and define an initial, shareable set of applicable measurement methods (MMs) with the purpose of supporting sites in analyzing their data's suitability for our CDSS for SIRS detection in pediatric patients. We refer to an MM as a specification of an applicable method that quantifies or describes an inherent characteristic of a dataset (cf. Johnson et al[3]), e.g., the number of values in a variable per patient and day, a plot showing the value distribution or the percentage of values outside of a given range. It is possible to combine MMs in multiple layers, e.g., an MM could use other MMs' results as input data to create an aggregated view or to add an assessment based on use case-specific thresholds for MM-results. Thus, MMs can express data quality indicators and descriptors (or similar concepts, e.g., operationalized assessment methods[5] or quality checks[9]) as well as information about how to assess the results. Beyond that, an MM specification is detailed enough to generate executable code from it. We refer to a compilation of MMs as a knowledge base as it represents shareable, applicable knowledge for data quality assessment for a certain use case. As the knowledge about data quality

requirements of the use case evolves, a knowledge base is subject to ongoing collaborative refinement based on new insights. Our initial knowledge base will be refined whenever the multidisciplinary team engaged in ELISE's development gains new insights on data quality requirements during further CDSS rollout and application.

Objectives

To define a shareable set of applicable MMs for a targeted data quality assessment determining the suitability of local data for the *ELISE CDSS*.

Methods

Dataset

The presented work used retrospective data from the CADDIE project and its successor ELISE. The used dataset consisted of approximately 12 million data points resulting from 2,029 days of stay of 168 patients at the Pediatric Intensive Care Unit and Pediatric Cardiology of Hannover Medical School. Clinical variables covered birth date, performed procedures, diagnoses, medications, heart rate, respiration rate, body temperature, pacing, temperature regulation, data on assisted ventilation, laboratory values for immature granulocytes, and white blood cell count (cf. clinical information models of ELISE[17]). The clinical data for the first iteration of data quality analysis (cf. ►Fig. 1) were available from a local *Better platform*,[18] an *openEHR* data repository. We retrieved the data using the *openEHR* REST-API[19] and the *Archetype Query Language* (AQL).[20] The CDSS for SIRS detection is in continuous development. Thus, it evolved from the first iteration of data quality analysis to the second. For simplicity, we refer to the CDSS version from the first iteration as *CADDIE CDSS* and to the CDSS version from the second iteration as *ELISE CDSS*. The data for the second iteration of data quality analysis covered the same patients and days, but *ELISE CDSS* considered more variables in its decisions (performed procedures, diagnosis data, and medications). The clinical data for the second data quality analysis were available in CSV files (the technical data access and the data format changed from the first to the second iteration due to different implementation phases of the technical infrastructures in the projects CADDIE and ELISE).

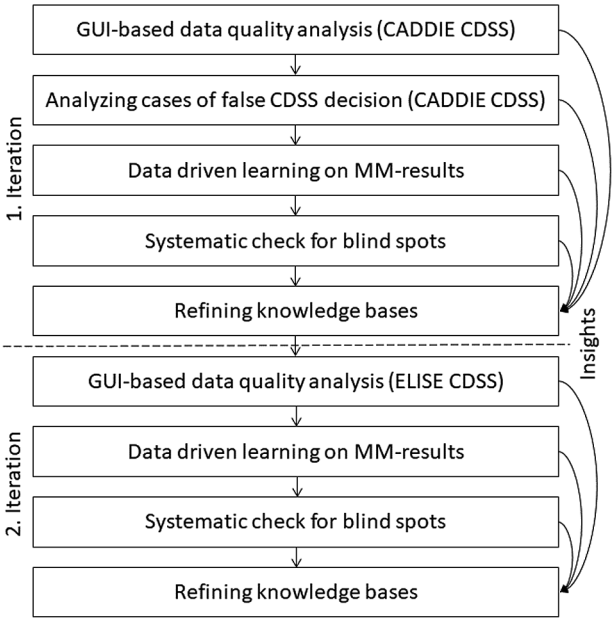

Fig. 1 Performed steps in chronological order.

A central task in this work was to investigate the relation between measurable dataset characteristics and CDSS performance, i.e., the number of false CDSS decisions. Thus, additionally to the clinical data, we used data on correctness of the CDSS SIRS detection for each patient and day. The *CADDIE CDSS* (in the first iteration) and *ELISE CDSS* (second iteration) were applied retrospectively on the data to detect SIRS episodes. For each patient and each day of stay on the intensive care unit, we derived a label specifying whether the CDSS's SIRS detection compared with the ground truth was true positive, false positive, true negative, or false negative. We refer to these data as CDSS performance data. In both iterations, CDSS performance data were available as a CSV file.

Formalization of MMs

We utilized previous work to express the MMs that represent our applicable knowledge on data quality assessment for *ELISE CDSS*.[21] This previous work proposed a formalization for MMs as 5-tuples with the objective to foster collaborative, interoperable, knowledge-based data quality assessment. The approach of the 5-tuples is to allow a flexible representation of all computable inherent characteristics about a dataset for data quality assessment but to stay maintainable by encapsulating the input data definition (*domain paths*) and by simplifying the representation of the most common operations (*checks* and *groupings*). ►Fig. 2 shows an MM formalized as 5-tuple (extended examples with explanations in ►Supplementary Appendix A [online only]). To compute an MM formalized as 5-tuple, the data specified in the *domain paths* are provided as R-vectors and the *check* (optional), *grouping* (optional), and *characterization* elements are inserted into a generic R-script template, a process which is simple to automate and applicable in different technical contexts (we tested the application of MMs without using *openCQA* for example in Kindermann et al[22]).

We needed to define MMs specifically targeting the CDSS's data quality requirements. The *check*, *grouping*, and *characterization* parts of the 5-tuple (cf. ►Fig. 2) allow for this with flexible definitions of the measurement process in R programming language while keeping most of the MMs simple, which eases understanding and adaptions. *ELISE CDSS* bases on *openEHR* clinical information models. Thus, for MMs using patient data as input, we used *openEHR* archetype paths in the *domain paths* to define the input data. This enabled direct application of our MMs on the data retrieved via AQL in our first iteration of data quality analysis (cf. ►Fig. 1). For the second iteration, where our data were available as CSV files, we adapted the column names in the CSV files to match the MMs' *domain paths*.

```
Tags:
standard_deviation,body_temp,per_subject_and_day

Domain paths:
# item0 (numeric) = dataset-row/[openehr-ehr-observation.body_temperature.v1]/.../magnitude
# item1 (string) = dataset-row/ehr/ehr_status/subject/external_ref/id/value
# item2 (string) = dataset-row/[openehr-ehr-observation.body_temperature.v1]/.../value.asDay

Check:
null

Grouping:
sprintf("%s_%s",item1,item2)

Characterization:
function(v) {sd(v,na.rm = TRUE)}
```

Fig. 2 MM formalized as 5-tuple. The MM's *tags* indicate that this MM quantifies the standard deviation of the body temperature for each patient and day. The *domain paths* specify the input data. *Check*, *grouping*, and *characterization* define the computation in R programming language. MM, measurement method.

We needed to organize all formalized MMs for targeted data quality assessment for ELISE CDSS into shareable sets. The concept of a knowledge base (in MM context) has the purpose to provide an organizing structure for MMs that eases their management and supports sharing. Thus, we organized our MMs in knowledge bases. Knowledge bases additionally allow adding some describing elements and the definition of the dataset for which the MMs are intended (cf. data need[23]). For example, this definition of the dataset could be an AQL query unambiguously defining which data to retrieve from a repository. We used these features to define the data need and to add explanations for each knowledge base.

GUI-Based Data Quality Analysis

We used *openCQA* the open source reference implementation for working with 5-tuple MMs. As depicted in ►Fig. 1, the GUI-based data quality analysis was our first step to analyze the data quality. As a starting point, we generated default MMs based on the variables' datatypes, e.g., count, minimum, maximum, mean, median, mode, standard deviation, and lower and upper quartiles for numeric values (cf. knowledge base "Add MMs based on datatype"[24]). We added a *grouping* (cf. ►Fig. 2) to the MMs to calculate results aggregated in the dimension levels *per_subject* and *per_subject_per_day* because our CDSS performance data were available per patient and day. This means that for example the MM calculating the standard deviation of the body temperature value did not return one value for the whole dataset but one value for each patient and day, e.g., six result values if we had three patients with 2 days of stay each. The MMs created and applied are openly available in our git repository (knowledge bases starting with "ELISE_"[24]). We inspected the calculated MM-results in *openCQA*. In case of suspicious dataset characteristics, e.g., negative minimum body temperature value, we checked if these were relevant for the CDSS, i.e., if these could cause a false CDSS decision, to determine if a targeted MM checking for these issues is sensible for the knowledge base.

Analyzing Cases of False CDSS Decision

The next step was to look at the cases where a CDSS decision was wrong, i.e., a label for a patient and day was false positive or false negative. This step used *CADDIE CDSS* performance data. An analysis of these cases had already been performed with the objective to detect potential enhancements in *CADDIE CDSS*, e.g., to be more robust against outliers. As the outlier example shows, the differentiation between data quality issues and reasoning issues in the CDSS is often ambiguous. As part of our targeted data quality analysis, we discussed the identified reasons for CDSS failure to determine if an enhancement of the CDSS or an MM targeting the dataset characteristic were suitable reactions. CDSS enhancements were preferred if the problem was easy to catch or not related to identifiable characteristics of the dataset. Features and development of the CDSS are not subject of this article. Wulff et al presented the design[10] and evaluation of the *CADDIE CDSS*[11] and will publish an updated article on the *ELISE CDSS* in a timely fashion. We derived targeted MMs whenever catching in *ELISE CDSS* was not suitable and an identifiable characteristic in the data directly caused CDSS failure or at least enhanced the probability for CDSS failure.

Data-Driven Learning on MM-Results

During analysis in *openCQA* and during analysis of false CDSS decisions, we experienced a problem: even if we suspected a correlation between an MM's results and a lower CDSS performance, deriving information like thresholds from two tables with >1,000 rows each is not practical without methodical support. This is why we applied a data-driven learning approach for data quality assessment.[25] This approach used the MMs' results as features to train a decision tree and the CDSS performance data as labels. Each row of the features in the training dataset consisted of all MMs' result values for the respective patient and day. This is why we only used MMs' results as input for the data-driven approach that were aggregated in dimension levels *per_subject_per_day* or

per_subject. We used *rpart*[26] as decision tree implementation in the language R. Analysis of variance (ANOVA) determined the best splits. Each split in such a decision tree indicates a correlation in the dataset/subset between a feature and the label. The highest label-value difference is at the splitting condition. The idea behind the data-driven approach is: if we trained the decision tree on MMs' results with CDSS performance data as labels. Then, each of the decision tree's splits indicates a correlation of a certain MM's result values with a difference in CDSS performance, where the splitting value is a promising threshold to separate between good and bad MM-results. Note that the only purpose of the resulting tree is to indicate MMs' results that possibly deserve attention and not to actually perform a prediction or classification.

We applied the machine learning approach separately on two divided sets from the MM-results: one set containing the false-positive and true-negative cases. The second set contains the true-positive and false-negative cases. This was necessary because the effects of data quality in the dataset were small compared with the effects of clinical differences between days with SIRS episodes and non-SIRS days. Thus, the decision trees would consist of splits that are relevant for the distinction between positive and negative cases instead of showing splits that are relevant to distinguish true and false decisions. For both training datasets, the CDSS made considerably more correct than false decisions. Thus, both training datasets were imbalanced (e.g., in the second iteration 85 false-positive/1,494 true-negative cases for the first set and 414 true-positive/36 false-negative cases in the second set). To attenuate the risk of misleading results, we tested using weights as well as using under-sampling techniques in our training procedures. Both had only minor effects on the resulting trees and did not change the information derived from the trees.

In the first iteration (cf. ►Fig. 1), we just trained one decision tree on all 214 features (197 after filtering feature vectors only consisting of NAs or with identical values to other feature vector), i.e., all MMs' results, for the false-positive/true-negative cases. In the second iteration, we trained two trees: the first decision tree with all 216 features (190 after filtering feature vectors) for the false-positive/true-negative cases; the second one on the false negative/true positive cases. The intention of training the second tree was to look in particular for thresholds for MMs quantifying *value coverage*. Since the number of false-negative cases was small, we selected only the MMs' quantifying *value coverage* as features (eight resp. seven features) for the second decision tree. The R-scripts for training the trees are openly available in our git repository (files with ". R" ending in the root folder[24]).

The most crucial point in applying the data-driven approach is to interpret the resulting tree, i.e., to decide for each split whether it indicates a sensible data quality-related information. For all trees, we inspected each split to decide if it indicated possibly sensible data quality-related information or just an overfit to the dataset. We considered in this decision if we could make any sense of the split, looked at the number of rows separated, and inspected the respective MM's results in *openCQA*. Inspecting the split in *openCQA* often indicated that the split happened, because it simply by chance separated a set of cases with a high number of false decisions. To confirm such an assumption, we excluded the respective MM from the dataset and trained the same tree again. If the tree split the same cases, but on another unrelated condition, we classified this as an overfit to our data and discarded information from the respective split.

Systematic Check for Blind Spots

To check if our set of MMs considers the most common data quality categories, we consulted the *HIDQF* data quality framework proposed by Kahn et al (original publication,[27] update[8]). We considered all of their categories/subcategories and selected *Conformance*, *Completeness*, *Uniqueness Plausibility*, *Atemporal Plausibility*, and *Temporal Plausibility* (cf. ►Table 1 in Kahn et al.[27] ►Fig. 6 in Liaw et al.[8]) as relevant for our context. Analogous to Diaz-Garelli's systematic data quality assessment process (cf. ►Table 1 in Diaz-Garelli[23]), we also considered the aggregation in dimension levels (called granularity levels in Diaz-Garelli's work), e.g., if MMs addressed the category with values aggregated per patient and day, per patient, or for the whole dataset. Although it is not necessary to cover all combinations of data quality categories and aggregations with MMs, it is reasonable to give thought to each combination while considering the purpose of the data quality assessment. This way, we could become aware of possible blind spots worth addressing.

Refining Knowledge Bases

Based on the insights from the previous steps (cf. ►Fig. 1), we adapted the existing MMs or added new targeted MMs to the

Table 1 Overview of covered data quality categories and dimension levels

HIDQF (sub)category / Dimension	Conformance	Completeness	Uniqueness plausibility	Atemporal plausibility	Temporal plausibility
Overall			Covered	Covered	
Per subject		Covered			
Per subject per day	Covered	Covered		Covered	Covered
Per subject per hour		Covered		Covered	
Per age group				Covered	

Abbreviation: HIDQF, harmonized intrinsic data quality framework.

knowledge bases. Finally, we sorted the set of MMs to have the most important MMs at the beginning and added a description for each knowledge base. The description explained the general rationale of the knowledge base and highlighted the most important MMs.

Results

GUI-Based Data Quality Analysis

Inspection of MM-results in *openCQA* revealed some suspicious data characteristics. Surprisingly, most of them were not relevant for *CADDIE CDSS* and none of them was relevant for *ELISE CDSS* (due to increased robustness of the decision algorithm). We still derived new MMs from this step, as we missed a few general visualizations and descriptive measures, e.g., an overview about the overall amount of data or the number of patients per age group.

Furthermore, we noticed that our initial dimension levels *per_subject* and *per_subject_per_day* were not sufficient. The daily granularity was too coarse for some characteristics. Thus, we added MMs aggregating their results *per_subject_per_hour*.

►Fig. 3 shows the median pacing per age group. The huge difference between the values for different age groups illustrates that sometimes it was necessary to consider the patients' age, e.g., to assess value distribution plausibility. Accordingly, we added MMs showing results *per_age_group*. Typically, the *grouping* function (cf. *Methods - Formalization of MMs*) should be used to group results into age groups. However, for plots this would create one plot per age group. To achieve the plot in ►Fig. 3, showing results for all age groups in one plot, we defined the stratification in the *characterization* function.

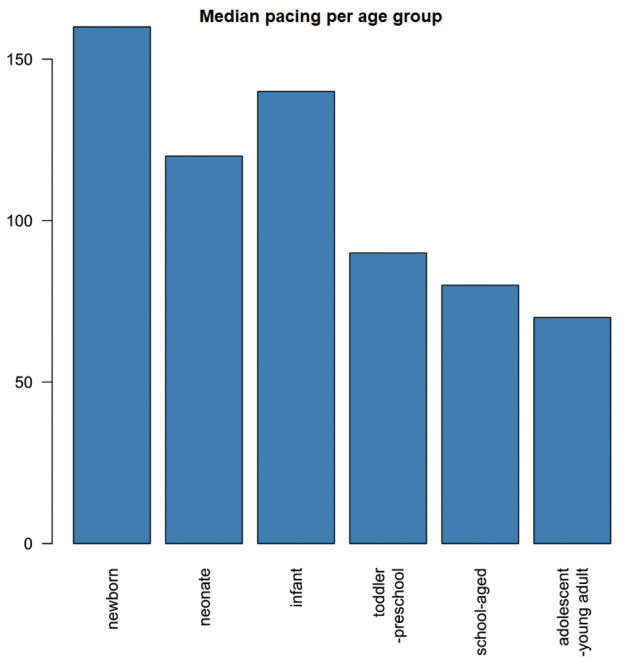

Fig. 3 Example MM-result. Plot showing the median pacing per age group. MM, measurement method.

Analyzing Cases of False CDSS Decision

The only identified reason for false-negative CDSS decisions was absence of current values—for the CDSS decision, the two laboratory values expire after 24 hours and vital signs after 1 hour. The most important variables for the CDSS SIRS decision are the laboratory and body temperature values. It may be perfectly correct that there are no data available, but if there is a timespan of the intensive care unit stay, not covered by these values, there is the possibility that the CDSS misses a SIRS episode resulting in lower sensitivity. As a reaction to this issue, we added MMs checking the *value coverage* for the laboratory values and the vital signs (cf. knowledge bases starting with "ELISE," MMs with tag "value_coverage"[24]). Those MMs specify for each variable the rate of time covered with values. Especially MM-results indicating low *value coverage* of laboratory values and body temperature may be critical for CDSS sensitivity, i.e., a higher rate of undetected SIRS episodes is possible.

Sometimes the CDSS detected a false-positive SIRS episode for patients due to values that were abnormal because of another disease or a recent medical procedure. Examples for this were low or high white blood cell counts or a low body temperature. This is only a data quality problem if the corresponding diagnosis or procedure data are missing. Thus, we added MMs checking diagnosis and procedure availability.

A related and indisputable local data quality problem was that documented procedure time points were often imprecise or incorrect. We were not able to add an MM targeting this issue, since we found no possibility to determine when this was the case (no other data to compare or to triangulate).

Finally, missing respiratory rate values sometimes caused false-positive SIRS detection resulting in lower specificity. The *value coverage* MMs already covered this issue. Results indicating low *value coverage* for respiratory rate could warn about an increased rate of false-positive SIRS detections.

Data-Driven Learning on MM-Results

In the first iteration (cf. ►Fig. 1), the MMs' results used as training data for the decision tree were unspecific for the use case. As a result, the decision tree did not indicate any sensible information regarding data quality.

From the first tree of the second iteration ("decision tree no feature selection"), we derived thresholds that deserve attention for low (<4.2) and high (≥ 17) white blood cell counts per patient and day. From the second tree ("decision tree value_-coverage"), we derived a threshold for low white blood cell *value coverage* (<77) and for low body temperature *value coverage* (<93). Where a low white blood cell count value coverage means that for this patient current values (<24 h) exist for less than 77% of days. Similarly, low value coverage for body temperature means that for this patient and day, current values (<1 h) exist for less than 93% of 1-hour intervals. Because these thresholds seem sensible to separate MM-results correlating with lower CDSS performance, we expect datasets with a high rate of such MM-results to be problematic. The added MMs calculate the rate of patients (and days) that are below/above these values (cf. knowledge bases "ELISE_body_temp" and "ELISE_WBCs," MMs with tags "below_soft_limit"/"above_soft_limit"[24]). The resulting trees from

the second iteration are available in our git repository (files with ".pdf" ending in the root folder[24]).

Systematic Check for Blind Spots

►Table 1 gives an overview of covered data quality categories and dimension levels in the final knowledge bases. We considered all combinations of HIDQF categories (*Conformance*, *Completeness*, etc.) with the dimensions (*overall* (whole dataset), *per_subject*, etc.). For each cell in ►Table 1, "covered" indicates whether our created knowledge bases contain MMs for the respective combination of category and dimension. As mentioned, not all combinations might be sensible to address with MMs. For example, we do not think it is worthwhile to implement any MMs checking some variable's values conformance to some eligible-value constraint and aggregate the number of violations depending on patient age. However, aggregating such constraint check results on the same aggregation level as the CDSS performance data seemed promising because it is reasonable to assume that problematic data instances could cause false CDSS decisions for the respective patient and day. Thus, MMs that fit into the *Conformance* category and aggregate their results per subject and day are part of our created knowledge bases.

In our setting, CDSS SIRS assessment was performed retrospectively. Thus, there was no potential for timeliness issues, i.e., value not available in time. Nevertheless, timeliness is obviously a critical requirement. Every value that is not available during CDSS decision has the same effect on CDSS performance as a missing value.

Final Knowledge Bases

►Table 2 lists the 13 final knowledge bases. We organized the 394 MMs into one knowledge base providing a quick overview over the whole dataset and one for each CDSS-relevant clinical concept.

Most of the MMs are only relevant for detailed inspection in case of a data quality issue. The knowledge bases are sorted to display the most informative MMs first. Additionally, a description in each knowledge base briefly explains the rationale and mentions important MMs to inspect. For clinical concepts likely to have many values, a tag "optional" gives the possibility to filter optional MMs to reduce the overall runtime of a knowledge base. Ten of the knowledge bases specify the data need using AQL. The final knowledge bases are openly available in our git repository (knowledge bases starting with "ELISE"[24]).

Discussion

Specific MMs can support a targeted data quality assessment for a certain data use, e.g., to check whether local data are suitable for a CDSS for SIRS detection in critically ill, pediatric patients. It is a well-known problem that there are no established standards on how to derive, express, or share such task-specific MMs[3] (or similar concepts, e.g., operationalized assessment methods,[5] data quality indicators/descriptors,[7] quality checks[9]). We derived specific MMs and expressed them as 5-tuples. That way, we demonstrated approaches for deriving and expressing task-specific MMs in a real-world use case and we provide shareable, applicable knowledge on data quality assessment for *ELISE CDSS*.

Table 2 Final knowledge bases for targeted data quality assessment for *ELISE CDSS*

Knowledge base name	Number of MMs	Data need specified
ELISE_overview	8	Yes
ELISE_body_temp	54	Yes
ELISE_date_of_birth	4	Yes
ELISE_diagnosis	2	Yes
ELISE_IG	27	Yes
ELISE_medication	3	No
ELISE_pacing	61	Yes
ELISE_procedure	1	No
ELISE_pulse	53	Yes
ELISE_respiratory_rate	53	Yes
ELISE_respiratory_rate_setting	48	Yes
ELISE_temperature_regulation	50	Yes
ELISE_WBCs	30	Yes

Abbreviations: IG, immature granulocytes; WBC, white blood cell count.

Data Quality Assessment Knowledge for ELISE CDSS

We created a set of MMs specific for data quality assessment for *ELISE CDSS*. These MMs can already help to decide whether local data are suitable for *ELISE CDSS* initially before CDSS rollout or regularly during CDSS operation, e.g., to notice changes in data quality due to changed documentation processes.[15] However, we are just at the start of a continuous improvement process. The initial knowledge bases will be refined based on new insights. We derived these MMs from insights that we gained from CDSS application on data from the Pediatric Cardiology and Intensive Care Medicine of Hannover Medical School. We expect that the planned rollout of *ELISE CDSS* to more clinical sites will provide more insights into data quality requirements. Furthermore, so far we have only integrated the perspective of medical informatics experts. The physicians performing their decisions based on the same data each day have an invaluable knowledge about the data's characteristics and their effect on SIRS detection. We plan to integrate their perspective as one of the next steps. Therefore, we are confident that we will be able to add MMs in future versions of the knowledge bases that allow for more direct decisions about the data's suitability for *ELISE CDSS*, for example, an MM that summarizes important MM-results as a table indicating with colors green, yellow, and red if certain characteristics of the dataset need attention.

Approaches for Deriving and Expressing Task-Specific MMs

We used four approaches to derive our task-specific MMs: in the GUI-based data quality analysis, (1) we started the first

iteration (cf. ►Fig. 1) by exploring the data with generic MMs generated in *openCQA*. This simple possibility to create MMs and to inspect their results enabled us to identify some specific MMs (cf. *Results—GUI-based data quality analysis*). Thus, it was a valuable contribution for our work. Analyzing cases of false decisions (2) was worthwhile as well (cf. *Results—Analyzing cases of false CDSS decision*). The data-driven learning approach (3) contributed by complementing the insights from the first two steps. While it could not indicate any new characteristics to assess with MMs, it provided threshold values for noteworthy values in already identified characteristics, such as the *value coverage*. We were not able to derive these thresholds manually or by logic reasoning about the CDSS algorithm. This endorses an assumption from testing the data-driven approach with artificial data: the data-driven approach is valuable in particular for cases where the data quality issue is nothing obviously odd (like outlier values) or a perfect cause of failure (like a CDSS always failing with error).[25] Since decision trees basically just perform a lot of statistical tests (ANOVA) on all feature variables, they are a handy method to get a grip on hidden correlations between MM-results and CDSS performance. The approach's current limitation is the uncertain reliability of the derived information from the decision trees. Interpreting a tree is a subjective task since there is not enough experience with this method. Tangible rules or suitable measures, e.g., a measure like the area under the curve for classifier models, to help decide whether a tree is sensible enough for interpretation or for the decision if a certain split indicates sensible data quality related information, are missing. Thus, the derived thresholds still need to demonstrate their value in future targeted data quality analyses. Our theory-based check for blind spots (4) did not yield any new MMs but made us aware of the limitation that the retrospective research context of the CDSS application was not suitable to identify any requirements regarding timeliness (cf. *Results—Systematic check for blind spots*). As soon as the embedding of the CDSS into real clinical processes is planned, it will be necessary to analyze the resulting requirements regarding timely availability of values.

In summary, each of the four approaches contributed to our knowledge on data quality assessment for *ELISE CDSS* justifying their application. Approaches (1), (2), and (4) can be conducted with limited effort and the value of resulting MMs is evident to the knowledge base's curator from the way they are derived. Thus, we would recommend to apply these approaches in each applicable real world use case. Since the data-driven approach (3) requires more effort and can only provide insights if the MM-results used as training data contain relevant information, we would recommend a selective application if three conditions are met: first, performance data (label data for the trees) are available in a suitable granularity. Second, MMs that are promising to calculate results containing relevant information in a suitable granularity are already implemented. Third, a correlation between measurable data quality issues and the performance data of the use case is expected. Established measures to support evaluating the reliability of trained decision trees and derived MMs would be desirable to improve the value of the method.

One insight from our work on this use case was that some data quality assessments require specific MMs to be informative, an experience that is in line with Blacketer et al,[9] who mentioned limited possibilities for adaptions on and adding of data quality checks as important requirements for the development of the *Data Quality Dashboard* in the *OHDSI* network. Since the scope of 5-tuple MMs includes such data quality checks, these requirements apply to MMs as well. Other researchers strive to define preferably generic sets of data quality indicators and descriptors[7] because harmonized sets of generic indicators and descriptors could enhance comparability of results. Comparability is indisputably more limited for task-specific MMs, although the structure and unambiguous definitions of the 5-tuple-formalization aim to foster comparability.[21] However, 5-tuple MMs could represent implementations of generic indicators and descriptors as proposed by Schmidt et al and our concept supports the integration of existing generic MMs already defined as 5-tuple into use case-specific knowledge bases (cf. *Methods—GUI-based data quality analysis*). Additionally, MMs allow making use of already implemented generic R-functions, for example, from R-packages targeting data quality.[21] That way, we could attenuate the comparability issue of task-specific MMs, since we could make use of generic, more comparable MMs and R-functions wherever possible and only create task-specific MMs where existing generic solutions were insufficient. Besides improving comparability, supporting usage of existing implementations and concepts such as for example OHDSI's quality checks or Schmidt's indicators, descriptors, and R-functions gives the possibility to benefit from existing experiences and invested work, while supporting management and sharing of targeted MMs for use case-specific data quality assessment.

Conclusions

We have created a set of shareable, applicable MMs that can support targeted data quality assessment for CDSS-based SIRS detection in critically ill, pediatric patients. This initial knowledge base will be refined based on new insights during further CDSS rollout and application. Preventing data quality-related failure of CDSS could improve user satisfaction and avert harm from patients.

The demonstrated approaches for deriving and expressing task-specific MMs have the potential to foster targeted data quality assessment for a variety of use cases. Our ultimate goal would be to promote task-specific data quality assessment as a commonly recognized usual part of research on data-consuming application systems in health care.

Ethical Considerations

All methods were performed in accordance with relevant guidelines and regulations. All study participants, their parents, or legal guardians gave written informed consent. Both CADDIE and ELISE were approved by the Ethics Committee of Hannover Medical School (No.

7804_BO_S_2018 and No. 9891_BO_S_2021). All authors had a valid permission (Datenzugriffsvereinbarung) to work with the dataset.

Funding

Development of the used methods and tools was partly done within project "HiGHmed" (German MI-Initiative), funded by BMBF (Grant No. 01ZZ1802C). This work was funded by the Federal Ministry of Health (Grant No. 2520DAT66A).

Conflict of Interest

None declared.

Acknowledgments

We would like to thank the ELISE Study group for its input. Moreover, the assistance provided by the MHH Information Technology was greatly appreciated.

References

1 International Organization for Standardization. ISO 8000–2:2020. Data quality—Part 2: Vocabulary. Geneva, Switzerland: ISO International Organization for Standardization; 2020
2 Nonnemacher M, Nasseh D, Stausberg J. Datenqualität in der medizinischen Forschung. 2., aktual. u. erw. Aufl. Berlin: Medizinisch Wissenschaftliche Verlagsgesellschaft; 2014
3 Johnson SG, Speedie S, Simon G, Kumar V, Westra BL. Application of an ontology for characterizing data quality for a secondary use of EHR data. Appl Clin Inform 2016;7(01):69–88
4 Khare R, Utidjian L, Ruth BJ, et al. A longitudinal analysis of data quality in a large pediatric data research network. J Am Med Inform Assoc 2017;24(06):1072–1079
5 Weiskopf NG, Bakken S, Hripcsak G, Weng C. A data quality assessment guideline for electronic health record data reuse. EGEMS (Wash DC) 2017;5(01):14
6 Meng XL. COVID-19: a massive stress test with many unexpected opportunities (for data science). Harvard Data Sci Rev. 2020. DOI: https://doi.org/10.1162/99608f92.1b77b932
7 Schmidt CO, Struckmann S, Enzenbach C, et al. Facilitating harmonized data quality assessments. A data quality framework for observational health research data collections with software implementations in R. BMC Med Res Methodol 2021;21(01):63
8 Liaw ST, Guo JGN, Ansari S, et al. Quality assessment of real-world data repositories across the data life cycle: a literature review. J Am Med Inform Assoc 2021;28(07):1591–1599
9 Blacketer C, Defalco FJ, Ryan PB, Rijnbeek PR. Increasing trust in real-world evidence through evaluation of observational data quality. J Am Med Inform Assoc 2021;28(10):2251–2257
10 Wulff A, Haarbrandt B, Tute E, Marschollek M, Beerbaum P, Jack T. An interoperable clinical decision-support system for early detection of SIRS in pediatric intensive care using openEHR. Artif Intell Med 2018;89:10–23
11 Wulff A, Montag S, Rübsamen N, et al. Clinical evaluation of an interoperable clinical decision-support system for the detection of systemic inflammatory response syndrome in critically ill children. BMC Med Inform Decis Mak 2021;21(01):62
12 Peter L Reichertz Institut für Medizinische Informatik: PLRI | ELISE [Internet]. Braunschweig: Peter L. Reichertz Institut für Medizinische Informatik der Technischen Universität Braunschweig und der Medizinischen Hochschule Hannover; c2022. Accessed June 24, 2022 at: https://plri.de/forschung/projekte/elise
13 Wulff A, Mast M, Bode L, Rathert H, Jack T; ELISE Study Group. Towards an evolutionary open pediatric intensive care dataset in the ELISE project. Stud Health Technol Inform 2022;295:100–103
14 Wulff A, Mast M, Bode L, et al. ELISE - An open pediatric intensive care data set. Accessed June 24, 2022 at: https://publikationsserver.tu-braunschweig.de/receive/dbbs_mods_00070468
15 Sáez C, Gutiérrez-Sacristán A, Kohane I, García-Gómez JM, Avillach P. EHRtemporalVariability: delineating temporal data-set shifts in electronic health records. Gigascience 2020;9(08):giaa079
16 Semler SC, Wissing F, Heyder R. German medical informatics initiative. Methods Inf Med 2018;57(S 01):e50–e56
17 Clinical Knowledge Manager [Internet]. Heidelberg: HiGHmed e. V. Accessed June 21, 2022 at; https://ckm.highmed.org/ckm/projects/1246.152.38
18 Platform | Better care [Internet]. Ljubljana: Better d.o.o.; c2022. Accessed June 21, 2022 at: https://www.better.care
19 API Overview [Internet]. London: openEHR Foundation; c2022. Accessed June 21, 2022 at: https://specifications.openehr.org/releases/ITS-REST/latest/overview.html
20 Language AQ. (AQL) [Internet]. London: openEHR Foundation; c2022. Accessed June 21, 2022 at: https://specifications.openehr.org/releases/QUERY/latest/AQL.html
21 Tute E, Scheffner I, Marschollek M. A method for interoperable knowledge-based data quality assessment. BMC Med Inform Decis Mak 2021;21(01):93
22 Kindermann A, Tute E, Benda S, Löpprich M, Richter-Pechanski P, Dieterich C. Preliminary analysis of structured reporting in the HiGHmed use case cardiology: challenges and measures. Stud Health Technol Inform 2021;278:187–194
23 Diaz-Garelli J-F, Bernstam EV, Lee M, Hwang KO, Rahbar MH, Johnson TR. DataGauge: a practical process for systematically designing and implementing quality assessments of repurposed clinical data. EGEMS (Wash DC) 2019;7(01):32
24 Server app/client/public/knowledge_base · SIRS_CDSS_KB · Erik Tute / openCQA · GitLab [Internet]. Braunschweig: Peter L. Reichertz Institut für Medizinische Informatik der Technischen Universität Braunschweig und der Medizinischen Hochschule Hannover; c2022. Accessed June 29, 2022 at: https://gitlab.plri.de/tute/openehr-dq/-/tree/SIRS_CDSS_KB/Server%20app/client/public/knowledge_base
25 Tute E, Ganapathy N, Wulff A. A data driven learning approach for the assessment of data quality. BMC Med Inform Decis Mak 2021;21(01):302
26 CRAN - Package rpart [Internet]. Wien: Institute for Statistics and Mathematics of WU (Wirtschaftsuniversität Wien). Accessed June 21, 2022 at: https://cran.r-project.org/web/packages/rpart/index.html
27 Kahn MG, Callahan TJ, Barnard J, et al. A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data. EGEMS (Wash DC) 2016;4(01):1244