Formalization and computation of quality measures based on electronic medical records

Kathrin Dentler, Mattijs E Numans, Annette ten Teije, et al.

J Am Med Inform Assoc published online November 5, 2013
doi: 10.1136/amiajnl-2013-001921

Formalization and computation of quality measures based on electronic medical records

Kathrin Dentler,[1,2] Mattijs E Numans,[3,4] Annette ten Teije,[1] Ronald Cornet,[2,5] Nicolette F de Keizer[2]

▸ Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/amiajnl-2013-001921)

[1]Department of Computer Science, VU University Amsterdam, Amsterdam, The Netherlands
[2]Department of Medical Informatics, Academic Medical Centre, University of Amsterdam, Amsterdam, The Netherlands
[3]Julius Center for Health Sciences and Primary Care, University Medical Centre Utrecht, Amsterdam, The Netherlands
[4]Department of General Practice & Elderly Care Medicine, Academic Medical Center, VU University Amsterdam, Amsterdam, The Netherlands
[5]Department of Biomedical Engineering, Linköping University, Linköping, Sweden

Correspondence to
Kathrin Dentler,
AI Department, VU University Amsterdam, Amsterdam 1081 HV, The Netherlands;
k.dentler@vu.nl

Received 15 April 2013
Revised 18 October 2013
Accepted 22 October 2013

ABSTRACT

Objective Ambiguous definitions of quality measures in natural language impede their automated computability and also the reproducibility, validity, timeliness, traceability, comparability, and interpretability of computed results. Therefore, quality measures should be formalized before their release. We have previously developed and successfully applied a method for clinical indicator formalization (CLIF). The objective of our present study is to test whether CLIF is generalizable—that is, applicable to a large set of heterogeneous measures of different types and from various domains.

Materials and methods We formalized the entire set of 159 Dutch quality measures for general practice, which contains structure, process, and outcome measures and covers seven domains. We relied on a web-based tool to facilitate the application of our method. Subsequently, we computed the measures on the basis of a large database of real patient data.

Results Our CLIF method enabled us to fully formalize 100% of the measures. Owing to missing functionality, the accompanying tool could support full formalization of only 86% of the quality measures into Structured Query Language (SQL) queries. The remaining 14% of the measures required manual application of our CLIF method by directly translating the respective criteria into SQL. The results obtained by computing the measures show a strong correlation with results computed independently by two other parties.

Conclusions The CLIF method covers all quality measures after having been extended by an additional step. Our web tool requires further refinement for CLIF to be applied completely automatically. We therefore conclude that CLIF is sufficiently generalizable to be able to formalize the entire set of Dutch quality measures for general practice.

OBJECTIVE

We have previously developed[1–3] a method for clinical indicator (better known as quality measure) formalization (CLIF). CLIF supports its users in transforming quality measures—which are typically described in unstructured text—into precise queries that can be computed on the basis of patient data. The main envisioned users of CLIF are quality measure developers, but also those responsible for reporting measure results, as well as general practitioners and hospital physicians who are interested in the quality of care they deliver.

CLIF was originally inspired by the Logical Elements Rule Method (LERM),[4] a method used to assess and formalize clinical rules for decision support, as well as a method proposed by Stegers *et al*[5] to transform natural language into formal proof goals. We have successfully applied CLIF in the limited domain of colorectal cancer surgery to formalize a relatively small set of quality measures. In one of our previous studies,[2] we tested whether our method leads to reproducible results. We did this by having eight test subjects—who were previously unacquainted with the problem—formalize a sample measure, and by comparing their results with a reference standard that we developed together with domain experts. The study showed that CLIF can lead to reproducible results, but that unambiguous measures and the cooperation of trained experts with clinical as well as medical informatics expertise are required. The objective of the present study was to test whether CLIF is generalizable—that is, whether it is applicable to a variety of different types of quality measures in various domains.

SIGNIFICANCE AND BACKGROUND

In recent years, automated reporting of quality measures based on data collected during routine care has become a necessity. The sheer amount of quality measures demanded by governments, patient associations, accreditation organizations, and insurance companies to measure, compare, and improve the quality of delivered care has increased dramatically at a rate that makes their manual calculation unfeasible. Besides being time-intensive, manual calculation is also error-prone and can jeopardize the reproducibility, validity, interpretability, traceability, timeliness, and comparability of quality measure results.

For these reasons, the automated computation and reporting of quality measures is included in the meaningful use of electronic medical records (EMRs), which is currently being put forward by the USA as a national goal.[6] The non-profit National Quality Forum (NQF) developed the Quality Data Model (QDM; http://www.qualityforum.org/QualityDataModel.aspx), an information model that defines concepts used in quality measures to automate their computation. The Centers for Medicare & Medicaid Services provide a web-based and QDM-driven measure authoring tool (MAT; https://www.emeasuretool.cms.gov/) for quality measure developers to create so-called 'eMeasures'. The MAT is a powerful tool that supports its users by offering a broad variety of functions and features. However, it is not based on a structured method that divides the highly complex task into clear, ordered subtasks.

A formalization method can help to guide users who were previously unacquainted with the problem of measure formalization through the

To cite: Dentler K, Numans ME, ten Teije A, *et al*. *J Am Med Inform Assoc* Published Online First: [*please include* Day Month Year] doi:10.1136/amiajnl-2013-001921

formalization process, and thereby help to ensure that the formalizations obtained faithfully represent the measure's intended meaning. Therefore, we propose our method, CLIF, as a complementary contribution.

MATERIALS AND METHODS

Set of quality measures

To answer our research question, we formalized the entire national set of 159 quality measures for general practice. This set is published in Dutch free text (http://www.nhg.org/themas/artikelen/download-indicatoren, last accessed October 2, 2013). The quality measures are defined on a national level, so that software providers can support the registration of required data and the reporting of measure results. The set of quality measures is heterogeneous, as it contains measures of various types, and addresses seven domains, such as 'asthma in adults' or 'diabetes mellitus'. Each domain contains a number of subdomains, such as 'HbA_{1c}' or 'smoking'.

Table 1 provides an overview of the quality measures, categorized according to Donabedian's trilogy: structure, process, and outcome.[7] Some measures have complementary measures, which we include. For example, the measure 'Diabetes patients for whom HbA_{1c} has been measured' has the complementary measure 'Diabetes patients for whom HbA_{1c} has NOT been measured'.

The quality measures are released in a narrative-based pseudo-formal format, and contain definitions such as 'age >40 and <80' and 'registration date <(reporting date−1 year)'. The reporting date is defined as the end of the reporting period, which is typically one reporting year. All quality measures are accompanied by relatively short lists of codes from the classification systems used (between one and 24 concepts per measure; approximately five on average). Two sample quality measures are presented in example 1. Online supplementary appendix 1 contains additional sample measures.

Example 1: Two sample quality measures (one process and one outcome measure)

Process measure 'Percentage of diabetes patients whose HbA_{1c} value has been measured within the previous 12 months'. Definitions:

- Patients younger than 80 years
- International Classification of Primary Care (ICPC) codes for diabetes mellitus: T90, T90.01 or T90.02
- ICPC codes for diabetes mellitus recorded before the end of the reporting period
- Patients registered with general practitioner for 12 months or longer (≥12 months)
- Code 2206 (main caregiver for diabetes mellitus); latest value for this code must be 48 (for general practitioner); ≥12 months
- HbA_{1c} measurement (code 2816) within the previous 12 months

This process measure is the basis for the outcome measure 'Percentage of diabetes patients whose latest measured HbA_{1c} value was below 53 mmol/mol', which only differs by one additional definition:

- HbA_{1c} value of last measurement below 53 mmol/mol (<53)

Patient data

We used an extract of anonymized routine healthcare data from the Julius General Practitioners' Network Database, which consists of administrative routine healthcare data extracted from the information systems of more than 60 primary healthcare centers (one to eight general practitioners per center) in the region of Utrecht, the Netherlands.[8] The administrative routine healthcare data were extracted locally from the general practitioner's EMRs by making use of the Mondriaan Client (http://www.projectmondriaan.nl/) and anonymized locally through a trusted third party (Custodix). This way, medical information cannot be used outside the practice location to identify individual patients by researchers or anyone else not directly involved in the treatment of the patients. Consultations, episodes, and diagnoses are encoded with International Classification of Primary Care (ICPC) codes, prescribed medications in the Anatomical Therapeutic Chemical (ATC) Classification System, and (laboratory) test results in a national coding system. We used data extracted from the 22 practices that use Promedico, a software system for general practices in the Netherlands. The other practices sharing data in the Julius General Practitioners' Network Database use other EMR software systems. Our database contains data related to 156 176 patients in the years between 2006 and 2011.

CLIF

CLIF is a method for formalizing natural-language quality measures as computable queries based on formally defined concepts, information model, and selection criteria. The original version of CLIF[1] consists of eight steps, which are presented together with the formalization of the sample outcome measure in table 2.

Web tool

We have built a web tool (http://clif.mash-it.net; login and password are both 'test') that implements CLIF to guide its users through all eight steps and stores the formalized criteria in a

Table 1 Overview of the set of quality measures used

Domain	Measures	Subdomains	Type			
			Structure	Process	Outcome	Not specified
Asthma in adults	13 (8%)	3	3	9	1	0
Chronic obstructive pulmonary disease	14 (9%)	3	3	9	2	0
Cardiovascular risk	23 (14%)	6	3	16	4	0
Diabetes mellitus	50 (31%)	10	0	27	18	5
Depression and anxiety	12 (8%)	0	10	2	0	0
Prevention	15 (9%)	2	4	0	11	0
Prescription	32 (20%)	9	0	27	0	5
All	159 (100%)	33 (26 distinct)	23 (14%)	90 (57%)	36 (23%)	10 (6%)

Table 2 Steps of the original version of CLIF

Step	Description	Example
1. Concepts	Extraction of clinical concepts (eg, diagnoses, procedures) from the quality measure text. Depending on the measures and the patient data, standard terminologies such as SNOMED CT,[9] ICD or ICPC, or local/national coding systems can be used	The ICPC codes for diabetes mellitus (T90, T90.01, or T90.02), and the national codes for the main caregiver (2206) and HbA_{1c} (2816) are elaborated in the quality measure definition
2. Information model	Binding of concepts from the previous step to the concepts of the information model. Depending on the measures and the patient data, standard information models such as the QDM or openEHR archetypes, or local database schemas can be used	Here, we define query variables (aliases), such as diabetes, for the local database table that stores ICPC entries, and we bind this variable to the three diabetes mellitus concepts identified in the previous step
3. Temporal criteria	Formalization of temporal criteria	The sample measure contains various temporal criteria. Patients must be below 80 years to be included. They must be registered for 12 months or longer, and the general practitioner must have been the main caregiver for 12 months or longer, the diagnosis must be present at the reporting date or before, and the HbA_{1c} value must have been measured within the previous 12 months. Finally, the values of both the code for the main caregiver and the HbA_{1c} measurement must be the latest within the specified time frames
4. Numeric criteria	Formalization of numeric criteria	The sample measure contains two numeric criteria: the value of the code for the main caregiver must be 48 for general practitioner, and the HbA_{1c} value must be below 53 mmol/mol
5. Boolean criteria	Formalization of Boolean criteria	Our sample measure does not contain any Boolean criteria
6. Boolean connectors	Grouping of criteria by Boolean connectors	The three different codes for diabetes are connected by OR. Other criteria are connected by AND
7. Exclusion criteria/ negations	Definition of exclusion criteria/negations	Our sample measure does not contain any exclusion criteria/negations
8. Numerator only	Identification of criteria that only aim at the numerator	The difference between the numerator and the denominator is not explicitly defined for this measure. We define it as the HbA_{1c} value being below 53 mmol/mol

CLIF, clinical indicator formalization; ICPC, International Classification of Primary Care; QDM, Quality Data Model.

dedicated database. To test CLIF's generalizability, we use this web tool as a starting point to formalize the set of quality measures. Importantly, the user can record comments for each step, which is indispensable in cases when a measure is ambiguous and the user needs to decide on how to operationalize it. The user can create so-called query variables (aliases) for database tables, and then attach one or more codes (eg, those specified for diabetes) to these variables. In subsequent steps, the user defines which criteria need to be valid for a patient to be included in the quality measure result. To increase usability, the underlying database schema or information model is used to populate options for each step. For example, for temporal criteria, only database fields that have temporal data types are preselected. Also, criteria are colored (eg, red for exclusion criteria/negation), and can be deactivated so that they are not included in the automatically constructed Structured Query Language (SQL) query. During or after the formalization process, users can run the query (if the tool is connected to a database). SQL was chosen because of the format of our underlying patient database, but other query languages, such as the SPARQL Protocol and RDF Query Language, or standards-based output formats, such as the Health Quality Measures Format (HQMF), which is used for eMeasures, could also be an option. The screenshots contained in online supplementary appendix 2 show how the sample measure is formalized step by step.

Computation of quality measures

To compute the formalized quality measures based on our patient data, we automatically constructed SQL queries based on the criteria that are stored in the database of CLIF's web tool. When our web tool did not support a construct, we applied CLIF manually by directly translating the respective construct into SQL. Subsequently, for every measure and reporting year (2007–2011), one query for the numerator and one for the denominator were constructed automatically. These queries were used to compute the measures, and to generate plots for all computed measures to visualize how the percentages develop over the course of the reporting years. The query in online supplementary appendix 3 represents the numerator of our sample outcome measure 'Percentage of diabetes patients whose latest measured HbA_{1c} value was below 53 mmol/mol' for the reporting year 2011.

Evaluation of results

Apart from assessing the face validity of our computed measure results based on the generated plots, we evaluated our result set computed for the reporting year 2011 for the quality measures in the domain diabetes mellitus, which is the largest domain contained in the measure set. We compared our result set with the result sets computed independently by two other parties for the same reporting year based on a large subgroup of general practices of the Julius General Practitioners Network that are working together in diabetes care. At the request of these practices, one of the measure result sets was provided by an academic institution specializing in the reuse of routine primary care data for research purposes (Integrated Primary Care Information (IPCI), Rotterdam, the Netherlands (http://www.erasmusmc.nl/med_informatica/research/555688/?lang=en)). The other measure result set was provided by a software company specializing in generating management reports based on extracted routine primary care data that are used to support reimbursement of diabetes care with the healthcare insurance companies that pay for it (Proigia, Ede, the Netherlands (http://www.proigia.nl/)). In both cases, as well as in our own procedure, all data are anonymized at source, in the practices, before it is shared.

The comparison gives a first indication of the comparability of our computed results. However, a strict evaluation of the computed results is not possible because of the absence of a gold standard. As ambiguous quality measure definitions allow different interpretations, it is hard to distinguish right and wrong formalizations. Ultimately, definitions should be based on a broad consensus, and formalization helps to identify open issues and make them explicit. Another hindrance is that we only computed results based on 22 general practices which use Promedico, whereas the other parties computed results based on twice as many practices which use various information systems for general practitioners.

RESULTS

We formalized the entire set of 159 quality measures, which took, on average, 10 min per measure and resulted in a total of 849 concepts and 1283 criteria, with the help of the web tool. This way, 86% of the quality measures could be formalized fully, while the remaining measures required the manual application of CLIF. We computed all measures except for two numerators that combined a number of other quality measures and had to be canceled because of excessive run times. In the following, we quantify our results according to CLIF's steps. Note that one new step 'Textual criteria' has been added to the original version of CLIF. Table 3 presents an overview of the results.

Step 1: Concepts. In this step, we entered 849 (148 distinct) concepts.

Step 2: Information model. In this step, we defined 465 query variables and connected them to the concepts entered in the previous step. Of the 106 distinct variables, 60 were related to measured (laboratory) values, 33 to ATC medications, and 13 to ICPC diagnoses. In the set of quality measures, entire variables can be negated. For example, one measure asks for all patients whose HbA_{1c} value has not been measured. We implemented this option in the web tool and made use of it 10 times.

Step 3: Temporal criteria. 1068 criteria were temporal. Many quality measures pertain to the latest value of a measurement before the reporting date. We implemented this temporal abstraction in the web tool. This functionality was required 145 times.

Step 4: Numeric criteria. We formalized 206 numeric criteria, all of them simple value comparisons. A number of quality measures included numeric quantification over (temporal) criteria. For example, the 'chronic' intake of a prescribed drug is defined as 'at least three prescriptions or one prescription with a duration of 6 months or longer during the previous 12 months'. Even harder to formalize is the construct 'multiple' chronic intake, which is defined as chronic intake of five or more different drugs. Another criterion that comprises numeric quantification is 'at least two resurgences during the previous 12 months'. As we did not implement this option in our web tool, we manually formalized the respective parts of the nine measures that required numeric quantification over (temporal) criteria.

Step 5: Boolean criteria. Owing to the schema of our database, no Boolean criteria were required.

Step 6: Textual criteria. This step had to be added to CLIF because some data elements —for example, gender and smoking behavior—were stored in a text field in the patient database. Textual criteria were currently required for nine measures. Note that one needs to consider whether textual data elements can be transformed into coded form, in which case the step 'Concepts' would be adequate.

Step 7: Boolean connectors. The step for Boolean connectors is not implemented in CLIF's web tool in a way that users can manipulate them. In the query generation, the standard connector is 'AND', and we automatically detect groups of criteria that must be combined by 'OR'. This is the case whenever only one value at a time is possible. For example, one entry in the medication database can have only one ATC code. Therefore, when a query variable is assigned to two or more ATC codes, they are automatically combined by 'OR'. The same is the case for value comparisons that are based on mutually exclusive categories. For example, the smoking status cannot be 'yes' and 'never' at the same time. This simple mechanism covered most of the required Boolean connectors. However, exceptions occurred: for example, one of the asthma measures covers patients with persistent asthma OR patients who smoke. Two different query variables must be defined, as these entities must fulfill different criteria. Also, criteria for patients with a valid reason for an absent cervical screening—such as refusal or pregnancy—are to be connected by 'OR'. Finally, custom Boolean connectors can also be applicable to values of codes. For example, because the smoking status must be updated yearly only for (ex-) smokers, the quality measure 'smoking habits known' measures the percentage of patients whose last recorded value for smoking was 'never' regardless of the registration date, OR 'previously' OR 'yes' during the reporting year. Likewise, there is a need to nest previously defined criteria, as required for the construct 'at least three prescriptions OR one prescription with a duration of 6 months or longer during the previous 12 months'. Therefore, the step to combine criteria by Boolean connectors has been performed manually for 17 measures.

Step 8: Exclusion criteria/negations. 66 criteria were marked as exclusion criteria/negations.

Step 9: Numerator only: 714 previously defined criteria only aim at the numerator.

Table 3 Overview of results per step

Step	Used	Additionally implemented	Manual formalization
1. Concepts	849 (148 distinct)	–	–
2. Information model	465 (106 distinct) variables	Negated query variables (used 10 times)	–
3. Temporal criteria	1068 (83% of criteria)	Latest value (used 145 times)	–
4. Numeric criteria	206 (16% of criteria)	–	Numeric quantification: 9 (5% of measures)
5. Boolean criteria	–	–	–
6. Textual criteria	–	9 (1% of criteria)	–
7. Boolean connectors	1914 AND; 567 OR (only counted occurrences in numerators)	–	Custom connectors and nesting: 17 (11% of measures)
8. Exclusion criteria	66 (5% of criteria)	–	–
9. Numerator only	714 (56% of criteria)	–	–

To summarize, we added the step 'Textual criteria' to our method, and extended our web tool by this step as well as the possibility to negate entire query variables in the step 'Information model' and to specify the latest value of a measurement in the step 'Temporal criteria'. The functionalities of numeric quantification over (temporal) criteria and custom grouping of Boolean connectors have not been implemented, and we therefore applied them manually. This enabled us to formalize 100% of the quality measures.

Evaluation of results

Figure 1 shows the results for the 43 of the 50 diabetes measures that have been computed by all three parties (the other parties did not compute complementary measures). Our results are generally higher than the ones computed by the other two parties, with strong correlations according to Pearson's correlation coefficients.

The observed differences are explainable by differences in the approach to computing the quality measures: while we preserved the original measure definitions, the other parties adapted the measure definitions to match the data as much as possible. An example of this is the outlier on the top left, which is due to the fact that, in our dataset, the code specified in the measure narrative for diabetes mellitus type 1 only occurs twice. The problem is probably caused by versioning differences in the ICPC codes used to describe the data in the Promedico system. The other parties adapted the ICPC code from T90.01 to T90.1 to match the data, and thereby included many more patients with diabetes mellitus type 1, while we did not. This difference may also have influenced the results of other measures, as patients who have diabetes mellitus, or diabetes mellitus type 1 or 2, are the basis for the denominators of the subsequent measures. Also different interpretations and definitions, such as alcohol usage registered as 'ever' instead of 'within the past 5 years', can influence the results. Further differences may be explainable by different approaches to handling missing data, and by different decisions on defining the denominator.

Figures 2 and 3 show plots for the process measure 'Percentage of patients whose HbA_{1c} has been measured' and the outcome measure 'Percentage of diabetes patients whose latest measured HbA_{1c} value was below 53 mmol/mol'. The red line depicts the aggregated percentages for all included practices, and the other lines represent the individual practices. A high but decreasing variability can be observed. The lines in the plots do not suggest a trend but only connect the measurements per reporting year to increase readability.

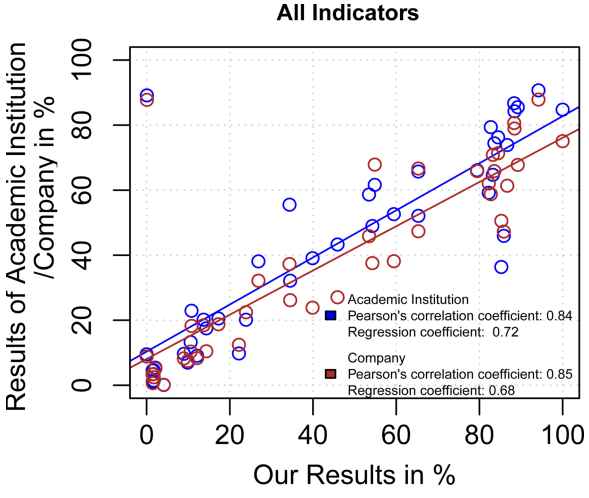

Figure 1 Comparison of our results (in percentages) with the results computed by an academic institution and a commercial company.

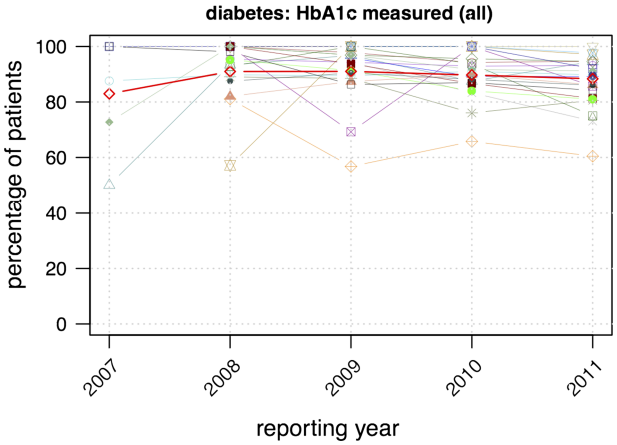

Figure 2 The target value desired by insurance companies is an HbA_{1c} measurement for 95% of the patients with diabetes.

DISCUSSION

After we extended CLIF by the additional step to formalize textual criteria, which was not available in its original version,[1] our method covered the formalization of all quality measures. Our web tool, however, required additional functionality. Even though the web tool was not complete, we could apply CLIF manually by directly translating the missing constructs into SQL, enabling us to fully formalize 100% of the measures. This leads us to conclude that CLIF is sufficiently generalizable to be able to formalize the entire set of Dutch quality measures for general practice.

Observations during our study

Quality measure definitions

Repetition and reusability

We observed considerable repetition. Many quality measures shared the same denominator, numerators were used as denominators for subsequent measures, and measures of a number of subdomains such as smoking and body mass index were

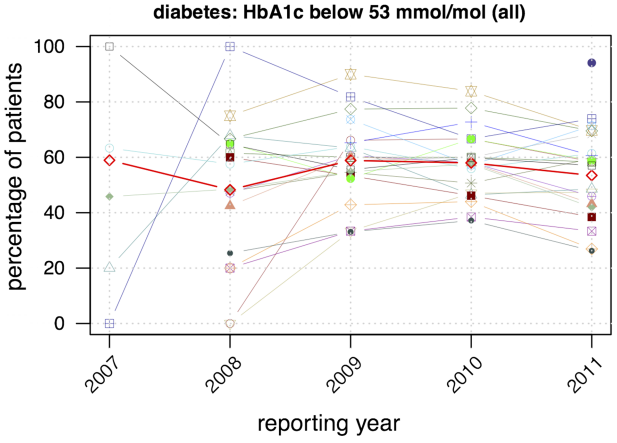

Figure 3 The measured value is best when it is below 53 mmol/mol. However, values up to 69 mmol/mol are acceptable.

applicable to a number of domains such as diabetes mellitus and cardiovascular risk. This is advantageous, as the respective criteria only have to be formalized once and can be reused thereafter. Also concepts and query variables can be reused—for example, our measure set made use of 849 concepts, but only 148 distinct ones, so that 701 (82%) have been reused.

Ambiguities

Ambiguities in quality measure narratives leave freedom for interpretation, which is especially problematic when values of locally computed measures are compared. Also, results computed for ambiguous measure definitions are hard to assess, as it is unclear what exactly has been computed. Here, a structured formalization method, ideally with tool support, can help users to resolve ambiguities, and to document all steps and decisions.

A major issue during the formalization was that in the set of quality measures, it is generally not explicitly stated how the denominators are defined. For example, it was unclear whether the denominator of the outcome measure 'Percentage of diabetes patients whose latest measured HbA_{1c} value was below 53 mmol/mol' is 'Diabetes patients for whom HbA_{1c} has been measured' or 'Diabetes patients'. Discussions with several experts showed that opinions vary about which denominator would be the correct one.

Another problem is the absent definition of qualitative terms such as 'high' dosage, as well as constructs such as 'indication for cervical screening' and 'gynecological intervention affecting the cervix'. Likewise, it is not clear whether medication dates in the measures refer to the prescription date, start date, or dispense date. We presented our findings to those responsible for the measure definitions.

Mismatches between quality measures and data

We detected several mismatches between quality measure definitions and our data. Some codes that were specified in the measures did not occur in the data. Examples include the ICPC code T90.01 for diabetes mellitus type 1 and the measurement code PAP XP BV for cervical screenings. Likewise, because of the absence of standardized codes, reasons for a cervical screening not being carried out were encoded depending on the underlying EMR, impeding an EMR-independent formalization a priori. Finally, some values should be encoded by 1 for 'yes', but they were stored as 'yes' in the database. The use of a standard information model for both measures and data might help to bridge mismatches between quality measures and data.

Quality measure results

Quality measure results should be treated with caution, especially when they are used to compare healthcare institutions. Percentages can be similar even if the numbers used to compute them are very different, affecting statistical significance. In our case, the numbers on which the percentages were based typically increased with the reporting years, but this is not evident in the plots. Similarly, although data quality improved over time in our dataset, data quality and missing values can influence the results.

Related work

The complexity of eligibility criteria in clinical trials has been analyzed in previous studies[10 11] and seems to be comparable to criteria for quality measures. For example, Conway *et al*[11] report a 'heavy reliance on nested Boolean logic, complex temporality and ubiquitous (…) codes'. Weng *et al*,[12] as well as Tu *et al*,[13] proposed semi-automated approaches to transforming free-text eligibility criteria into computable criteria. Milian *et al*[14] addressed the problem of formalizing eligibility criteria and derived a set of patterns that are the basis for a semi-formal representation. A pattern that Ross *et al*[10] also detected is the 'if-then' construct. Quality measures themselves can be rewritten into 'if *denominator* then *numerator*' constructs, and LERM[4] is applicable for such scenarios. With regard to phenotyping algorithms, Thompson *et al*[15] encountered 'non-Boolean logic' — for example, 'at least two of four criteria must be true', which did not occur in our measure set.

Limitations

One of the main limitations of our work is that the use of the tool presumably influenced our study. In retrospective, it is impossible to determine the actual influence of the tool on the formalization process and consequently on the obtained formalizations.

Another limitation of our work is that our results are limited to Dutch quality measures for general practice. However, in this and previous studies,[1–3] we have formalized a variety of heterogeneous quality measures (structure, process, and outcome measures) for both hospitals and general practitioners in various domains, using a variety of standard and non-standard coding systems and information models. This experience suggests that CLIF might also be sufficiently generalizable to be able to formalize other sets of measures, but the level of complexity of Dutch measures may differ from sets in other countries.

Future work

More research may provide further insights into the generalizability of our method—for example, by formalizing international sets of quality measures, such as the meaningful use measures put forward by the USA.

We have shown that openEHR archetypes can facilitate the semantic integration of routine patient data from several sources and patient data and quality measures to automatically compute measures.[3] In the future, it would be interesting to analyze whether new information model standards, such as the QDM as currently used for eMeasures, could be integrated into our approach, and how they would affect its generalizability.

CONCLUSION

The formalization of quality measures with the help of CLIF forces the user to disambiguate unclear parts of quality measures that are documented in inherently ambiguous natural language, and to precisely define the difference between denominator and numerator. Additionally, a formalized measure ensures that it can be computed automatically, and that the same query is used across several locations to compute a measure, making the computed results reproducible, comparable, traceable, and interpretable. Therefore, we propose that quality measures should be released in a formalized form, and ideally based on standard information models and terminologies. CLIF has been shown to be a useful method for achieving this goal.

Acknowledgements We thank the general practitioners participating in the Julius General Practitioners Network, Utrecht, the Netherlands, for sharing anonymized electronic medical records with the Julius General Practitioners Network Database. We thank Stephanie Medlock for language editing.

Contributors KD, NFdK, AtT and RC conceived and designed the study. KD conducted the experiments and wrote the first draft of the paper. MEN contributed the data and domain expertise. All authors approved the submitted revised version. KD is guarantor for the study.

Competing interests None.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1 Dentler K, ten Teije A, Cornet R, *et al*. Towards the automated calculation of clinical quality measures. *Knowledge Represent Health Care* 2012;LNCS 6924: 51–64.
2 Dentler K, Cornet R, ten Teije A, *et al*. The reproducibility of CLIF, a method for clinical quality measure formalisation. In: J Mantas *et al*, editor. *Stud Health Technol Inform* 2012;180:113–17.
3 Dentler K, ten Teije A, Cornet R, *et al*. Semantic integration of patient data and quality measures based on openEHR archetypes. *ProHealth 2012/KR4HC 2012* 2013;LNAI 7738:85–97.
4 Medlock S, Opondo D, Eslami S, *et al*. LERM (Logical Elements Rule Method): a method for assessing and formalizing clinical rules for decision support. *Int J Med Inform* 2011;80:286–95.
5 Stegers R, ten Teije A, van Harmelen F. Managing Knowledge in a World of Networks, Lecture Notes in Computer Science. EKAW'06 2006;4248:51–8.
6 Blumenthal D, Tavenner M. The "meaningful use" regulation for electronic health records. *N Engl J Med* 2010;363:501–4.
7 Donabedian A. Evaluating the quality of medical care. *Milbank Q* 2005;83: 691–729.
8 Grobbee DE, Hoes AW, Verheij TJM, *et al*. The Utrecht Health Project: optimization of routine healthcare data for research. *Eur J Epidemiol* 2005;20:285–7.
9 Cornet R, de Keizer N. Forty years of SNOMED: a literature review. *BMC Med Inform Decis Mak* 2008;8(Suppl 1):S2.
10 Ross J, Tu MSS, Carini MSS, *et al*. Analysis of eligibility criteria complexity in clinical trials. *AMIA Summits Transl Sci Proc* 2010;2010:46–50.
11 Conway M, Berg RL, Carrell D, *et al*. Analyzing the heterogeneity and complexity of Electronic Health Record oriented phenotyping algorithms. *AMIA Symp* 2011;2011:274–83.
12 Weng C, Wu X, Luo Z, *et al*. EliXR: an approach to eligibility criteria extraction and representation. *J Am Med Inform Assoc* 2011;18(Suppl 1):i116–24.
13 Tu SW, Peleg M, Carini S, *et al*. A practical method for transforming free-text eligibility criteria into computable criteria. *J Biomed Inform* 2011;44:239–50.
14 Milian K, Bucur A, Ten Teije A. Formalization of clinical trial eligibility criteria: evaluation of a pattern-based approach. *2012 IEEE International Conference on Bioinformatics and Biomedicine*; 2012:1–4.
15 Thompson WK, Rasmussen LV, Pacheco JA, *et al*. An evaluation of the NQF Quality Data Model for representing Electronic Health Record driven phenotyping algorithms. *AMIA Annu Symp Proc* 2012;2012:911–20.