Quality metrics for detailed clinical models

Sunju Ahn\textsuperscript{a,b}, Stanley M. Huff\textsuperscript{c,d}, Yoon Kim\textsuperscript{a,e,*}, Dipak Kalra\textsuperscript{f,g}

\textsuperscript{a} Department of Health Policy and Management, College of Medicine in Seoul National University, 28 Yeongeon-dong, Jongno-gu, Seoul, Republic of Korea
\textsuperscript{b} Graduate School, Interdisciplinary Program in Biomedical Engineering, Yonsei University, 50 Yonsei-ro, Seodaemun-gu, Seoul, Republic of Korea
\textsuperscript{c} Intermountain Healthcare, Salt Lake City, UT, USA
\textsuperscript{d} Department of Biomedical Informatics, University of Utah, Salt Lake City, UT, USA
\textsuperscript{e} Institute of Health Policy and Management, Seoul National University Medical Research Center (SNUMRC), 28 Yeongeon-dong, Jongno-gu, Seoul, Republic of Korea
\textsuperscript{f} Department of Health Informatics, University College London, London, United Kingdom
\textsuperscript{g} Centre for Health Informatics and Multiprofessional Education, University College London, United Kingdom

\section{Introduction}

To exchange and interpret clinical information consistently between electronic health record (EHR) systems, the data structures used to represent distinct items of information need to be standardized [1–5]. These data structure definitions, known as detailed clinical models (DCMs), specify how a particular kind of entry within an EHR has to be organized, for example, which clinical terms apply, what value range is valid, which measurement units may be used, which data items within the structure require a value [6,7]. The clinical

\* Corresponding author at: Department of Health Policy and Management, College of Medicine in Seoul National University, 28 Yeongeon-dong, Jongno-gu, Seoul, Republic of Korea. Tel.: +82 10 5024 6146; fax: +82 2 743 2009.
E-mail address: yoonkim@snu.ac.kr (Y. Kim).
1386-5056/$ – see front matter © 2012 Elsevier Ireland Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.ijmijournal.2012.09.006
element model by Intermountain Healthcare [8,9], templates defined by HL7 International [10,11], archetypes defined by the openEHR Foundation [12,13] and ISO 13606, the Clinical Information Model in the Netherlands [14], and the Clinical Content Model by the Center for Interoperable EHR in South Korea [15] are examples of published DCM formalisms that can be used to organize the clinical content of an EHR interoperability message and an EHR repository, share decision logic, and build a data capture form of a clinical application. Each instance of a DCM dictates how a corresponding generic EHR representation is to be used to represent particular types of clinical information. Examples of DCM instances might include representations for documenting a pain symptom, heart sounds, liver function tests, a prescribed drug, or a chest X-ray report.

Whereas DCMs have the potential to improve clinical decision support and clinical documentation in EHR system, the critical challenge is to identify the qualitative and quantitative requirements of DCMs. A few studies have suggested some quality requirements for DCMs (Table 1). For the flexibility and scalability of DCMs, general requirements for the system in which clinical data models are implemented demand the following [16,17]: (1) the addition of elements and attributes to the clinical model without the necessity of changing the underlying software or database schema; (2) use an existing formalism/syntax for the representation of the model; (3) tight binding of model attributes to standard terminology systems; and (4) the existence of a mechanism for stating ‘negation’. General principles of good modeling include: (1) adoption of standard terminologies for use in the models; (2) representing the models in standard modeling languages; (3) sharing and approving the DCMs with a community of clinical experts; (4) define decision modules that reference the models. DCM quality criteria [18] have also been proposed where the following qualities were identified as being important requirements of a good DCM: usefulness, desirability, the degree of use/acceptance in clinical services, reusability, the quality of clinical content, the degree of clinician introduction/validation, the use of vocabulary, mapping to information models, applicability, application to other technologies, and maintenance.

Principles for the development of DCMs [19] can be classified as principles pertaining to the structure of the DCM, principles for creating the DCM content, and principles for the DCM development process. The principles that pertain to the structure of DCMs contains information about the language formalism, description of binding of attributes to standard terminologies, a strategy for supporting semantic links among DCM instances, the definition of standard data types, and the description of standard units of measure. The principles for DCM content creation emphasize the granularity, reusability, correctness, and comprehensiveness of the models. Principles for the DCM development process emphasize evidence based model development, the need for proper use cases, use of meta data to track changes, and compliance to the syntax of the modeling language.

Archetype representation requirements [20] published in ISO 13606-2 are focused more on the technological aspects of models. They are divided into requirements for archetype definition, archetype node constraints, and data value constraints. EuroRec, an organization that certifies EHRs in the European Union, led an EC funded research project to develop criteria for the quality classification of EHR systems [21]. The research resulted in a set of archetype quality criteria, which covered administrative, technical, and clinical components, and repository operation requirements. Furthermore, they emphasized the use of standard terms and modeling language, the construction of repositories for DCM sharing, and the importance of metadata. Among the clinical requirements, the requirement for clinical use suggests the listing of accurate use patterns of clinical concepts, specification of whether the corresponding archetype is used in a specific workflow, description of subject population groups, as well as expert groups using an archetype, etc.

These published requirements and criteria are valuable sources as a starting place to define the good quality characteristics of DCMs (Table 1). However, these existing quality criteria have different levels of detail. Furthermore, these criteria have not been specified to the level of precision that is needed to undertake formal and objective evaluation of the quality of DCMs. Current DCM criteria cannot compare the quality of DCMs due to the absence of validated and reliable measurable characteristics. There is, therefore, still a need to develop quality metrics that can be used to objectively quantify quality criteria of DCMs.

2. Objective

This study was conducted to develop objective, reliable, and reproducible quality metrics for DCMs drawing on the published quality criteria referenced above, and to test their validity.

3. Materials and methods

The study was conducted in four phases. Phase 1 developed quality metrics; phase 2 tested the validity of the quality metrics; phase 3 tested the reliability of the quality metrics; and phase 4 finalized the quality metrics. Fig. 1 shows the applied methods of the study (Fig. 1).

3.1. Phase 1: developing quality metrics

3.1.1. Structure definition

DCM quality domains were developed based on the domains of the Appraisal of Guidelines and Research and Evaluation (AGREE) [22] to systematically define the domains of DCM quality metrics. Among the 6 domains used in AGREE, we included 4 domains and excluded 2 domains. We added 4 new quality domains so that a total of 8 quality domains were defined. See Section 4.1.1 for an explanation of how the 8 domains were chosen.

3.1.2. Classification and specification of the quality criteria into quality domains

Existing quality criteria were categorized under the 8 defined DCM quality domains to develop DCM quality metrics (Fig. 2). From amongst these published requirements and criteria, the EuroRec archetype quality criteria were found to be the most suitable starting point for defining metrics since they define
<table>
<thead>
<tr>
<th>Requirement and criteria</th>
<th>General requirement for clinical data model</th>
<th>Requirement for DCM</th>
<th>DCM quality criteria</th>
<th>EuroRec archetype quality criteria</th>
<th>Archetype representation requirements</th>
<th>Requirement for clinical information model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scope and purpose</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>• An archetype shall specify any particular clinical scenarios or workflows for which it is particularly intended</td>
<td>• A formal statement defining the scope and clinical purpose of this archetype, expressed as a coded term or as free text in a given natural language</td>
<td>• Scope and usage of the model</td>
</tr>
<tr>
<td>Accuracy</td>
<td>–</td>
<td>• Accurate</td>
<td>• Clinical content quality</td>
<td>• An archetype shall specify the precise nature of the clinical entity (or set of entities) for which it defines a use pattern</td>
<td>–</td>
<td>• Accuracy</td>
</tr>
<tr>
<td>Comprehensive-ness</td>
<td>• The model must be comprehensive—it must accommodate representation of anything that can be stated about a patient</td>
<td>–</td>
<td>–</td>
<td>• An archetype's use pattern should be inclusive of all of the minor variations in clinical entity representation across its use cases, users and scenarios</td>
<td>–</td>
<td>• Comprehensive-ness</td>
</tr>
<tr>
<td>Vocabulary/terminologies</td>
<td>• There must be a tight linkage to standard terminologies</td>
<td>• The modeling language must have a mechanism of linking to standard coded terminologies</td>
<td>• Vocabulary use/mapping to coded terminology</td>
<td>• The clinical label for each node shall be drawn from a published controlled vocabulary</td>
<td>• Any node of an archetype may be mapped to any number of additional concepts, terms and synonyms from terminology systems</td>
<td>• Mapping to standard terminologies (SNOMED CT, LOINC, etc.)</td>
</tr>
<tr>
<td>Compliance with standards</td>
<td>• Standard terminologies</td>
<td>• Standard coded terminologies</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>• Compliance with international standards (ISO Data type, UCUM, etc.)</td>
</tr>
<tr>
<td>Formalism</td>
<td>• It must use an existing formalism (XML schema, ASN.1, Conceptual Graphs, etc.) without modification</td>
<td>• Standard modeling language</td>
<td>• XML fragment</td>
<td>• An archetype shall define a formal representation</td>
<td>• The formalism (including version) in which this constraint specification is represented</td>
<td>• Formal representation (XML, UML, OWL, RDF, ADL, etc.)</td>
</tr>
<tr>
<td>Declaration of reference</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>• An archetype shall be able to include references to one or more kinds of published knowledge that have informed its overall design, and/or to which it conforms</td>
<td>• A description, reference or link to the published medical knowledge that has underpinned the definition of this archetype</td>
<td>• Declare the reference in the meta data</td>
</tr>
<tr>
<td>Decision support</td>
<td>–</td>
<td>• Decision modules that the reference the model</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td></td>
</tr>
</tbody>
</table>

Table 1 – Comparison of some published quality requirements and criteria for DCMs.
quality requirements more specifically and are the most comprehensive. To specify the quality metrics in a measurable way, the wording of each quality criterion was re-expressed as a formal test statement using the principles of ISO/IEC 9126 [23–25]. The ISO/IEC 9126 software quality standard provides a process for the specification and evaluation of the quality of software.

3.2. Phase 2: validity

3.2.1. Expert panel organization
We organized an international expert panel to test the validity and reliability of the quality metrics. First, a keyword search was performed from 1990 to 2010 in Medline using the keywords ‘DCM, archetype, Clinical Model, Quality Criteria,’ and the authors of these papers were selected for the expert panel. Second, a list of experts was obtained from DCM-related working groups in ISO, CEN, and HL7 International. The lists from the two sources were condensed into one list. Among the 13 experts identified, four were unable to participate in the panel for personal reasons; hence, the panel was finalized with nine experts. The panel was divided into two expert groups to encourage high participation since it was considered too demanding for one panel to test both the validity and reliability of all the quality metrics. Therefore, Expert Group 1 included those who had indicated that they could contribute more time than those in Group 2. Expert Group 1 consisted of 2 members and tested the face validity of the criteria. Expert Group 2, which included 7 members, tested content validity by applying the Delphi technique.

3.2.2. Face validity
The objective of the first face validity test was to formalize the first version of quality metrics. The draft metrics were tested by Expert Group 1 (two members). The face validity test includes three questions: “appropriateness of a selected metric as a quality evaluation metric,” “appropriateness as a quality improvement metric,” and “applicability as a quality evaluation metric for DCMs other than archetypes”. This third question was necessary since the original source quality criteria were developed specifically for archetypes, and it was important to confirm that the metrics derived from them are more widely applicable.

3.2.3. Content validity
We conducted two rounds of email-based Delphi surveys. If a structured questionnaire can be prepared based on an extensive literature review, a revised Delphi first round is possible and permissible [26–28]. Such an approach was possible here, using the metrics resulting from the face validity tests described above. The Delphi panel was provided a questionnaire with questions that were the same as those that were used for the first face validity test. In the first round, each panel member was asked to score each question on a 4-point scale (4: ‘Strongly agree,’ 3: ‘Agree,’ 2: ‘Disagree,’ 1: ‘Strongly disagree’). In the second round, feedback was provided to the panel participants from the initial round of scorings (Table 2). In this study, the base level for agreement among experts was defined as 70%. That is, if those who either replied ‘1’ or ‘2,’ or alternatively, ‘3’ or ‘4’ for each metric, comprised over 70% of the panel, it was taken to mean that an agreement had been reached [29].

3.3. Phase 3: reliability

The reliability of quality metrics was tested based on the agreement of assessments between the two evaluators in Expert Group 1. In this study, the threshold kappa coefficient was set to 0.60, a value that falls within the “Good
agreement level” according to the guidelines for interpreting kappa coefficients [30]. We selected archetypes as a type of DCM for our reliability testing because archetypes are in wide use and they are based on an international standard. Among archetypes, we selected the Apgar score and blood pressure models for reliability testing because they have all the components and properties necessary for quality assessment [31].

3.4. Phase 4: finalizing DCM quality metrics

The purpose of phase 4 was to refine any quality metrics for which the results of the reliability test did not agree, in order to finalize the quality metrics. For this purpose, we first analyzed the causes of disagreement in cases where the results of reliability testing did not agree. The metrics were then refined according to the type of disagreement. The refined metrics were then tested again for their validity by Expert Group 1. After the validity test was completed, the DCM quality metrics were finalized.

4. Results

4.1. Quality metrics

4.1.1. Structure definition
Among the 6 domains used in AGREE, we included the following 4: the scope and purpose, stakeholder involvement, rigor of development, and clarity and presentation. We excluded the following 2 domains: applicability and editorial independence since these domains are not relevant in terms of guideline specificity. An attempt was made to map all of the archetype requirements in the published documents listed earlier to the four chosen AGREE domains, but some of the archetype requirements did not fit so four new domains were added: compliance to standards, general methodology, metadata, and management and maintenance. In total, 8 quality domains were defined.

4.1.2. Classification and specification of the quality criteria into quality domains
Existing quality criteria were classified into the 8 DCM quality domains (Fig. 3). Each allocated quality criterion was then refined to a concrete level. The concrete level quality criteria were then decomposed into multiple measurable quality metrics. Each metric was defined together with its attributes such as objects, evaluation method, and scoring method. Many of the existing quality criterion statements were found to cover more than one testable feature.

4.2. Validity

4.2.1. Face validity
Based on the results of the first face validity test from Expert Group 1, it was proposed that 2 metrics be deleted, 2 new ones added (Table 3). The newly proposed quality metrics were “the majority of clinicians participating in the development stage” and “the majority of clinicians participating in the review stage”. It was proposed to remove the metric “semantic link” because this is an unfamiliar concept even to experts.
and caused confusion, and to also remove the metric “patient’s medical condition” because this will usually be redundant if the DCM is used for the relevant populations. For some quality metrics, it was agreed to rewrite the statement using plainer terms.

4.2.2. Content validity
From the results of the content validity test conducted through the Delphi survey, the following quality metrics were considered inappropriate and were rejected: the “majority of clinicians participating in development or review” and the “qualification of clinicians participating in certification”. The reason for this rejection was that “Clinicians’ participation in DCM development, review, and certification processes” was already included as a metric for evaluating “stakeholder involvement”. These changes therefore resulted in improved wording and the removal of duplication but did not materially modify the coverage of the set of metrics.

4.3. Inter-evaluator reliability
Two archetypes, Apgar score and blood pressure, were evaluated by two evaluators. The resulting kappa coefficient of the evaluation of the two archetypes was 0.73, which fell within the result range “Good agreement level” (Table 4). When the reliability of the quality metrics was tested using the two archetypes, the number of metrics used for this test was 30 because the metric “Clinician-participants and clinician’s role involved in the development of DCM” was divided into two separate metrics (Table 3). The results agreed with 27 metrics but differed for the following 3: the “Appropriate support for negotiation,” “Appropriate provision of narrative text,” and “Professional discipline for use”. A characteristic of the quality metrics for which the assessments did not agree upon, “Appropriate support for negotiation,” might not be applicable to the archetypes, or its objects of evaluation may need to be expanded to the level of the reference model—a potential reason for the disagreement. In the current quality metrics, the reference model domain is not included in the objects of evaluation. Accordingly, it may not be appropriate to evaluate these at the archetype level but with the reference model. Another possible reason for disagreement is differences in personal opinion: this may be why the evaluation results for the “purpose of DCM” did not agree between the two evaluators. The non-agreement metrics were modified and validated by the two-member Expert Group 1 in terms of evaluation method. After the reliability test, the “Clinician-participants role” was merged with “Clinician-participants involved in the development of DCM”.

4.4. Quality metrics for DCMs
The DCM quality metrics finalized through the validity and reliability tests covered 8 quality domains and included 29 quality metrics (Table 5). The 8 DCM quality domains were “purpose and scope,” “stakeholder involvement,” “rigor of development,” “clarity and presentation,” “compliance to standard,” “general methodology,” “metadata,” and “management and maintenance”.

The meanings and functions of each domain, as well as the meanings and functions of the quality metrics within each domain were as follows. The “purpose and scope” domain evaluated whether the purpose and scope of a DCM was presented clearly and in a way that was understandable to users. The domain included quality metrics such as “whether the purposes, subjects, and patient groups of a DCM are described”; “whether a DCM user’s school system, major, and its applicable clinical environment are described”; “whether there are contents on applicable patient groups; and whether applicable ages and genders are specified”.  

### Table 3 – Summary of the metric change by validity and reliability.

<table>
<thead>
<tr>
<th>Initial metrics</th>
<th>Face validity</th>
<th>Content validity</th>
<th>Reliability</th>
<th>Face validity</th>
<th>Final metrics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>38</td>
<td>38</td>
<td>38</td>
<td>38</td>
<td>38</td>
</tr>
<tr>
<td>No. of accepted metrics</td>
<td>–</td>
<td>36</td>
<td>29</td>
<td>29</td>
<td>–</td>
</tr>
<tr>
<td>No. of rejected metrics</td>
<td>–</td>
<td>2</td>
<td>9</td>
<td>9</td>
<td>–</td>
</tr>
<tr>
<td>No. of proposed metrics</td>
<td>–</td>
<td>2</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

### Table 4 – Reliability test results.

<table>
<thead>
<tr>
<th>Agreement</th>
<th>Non agreement</th>
<th>Kappa</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apgar score</td>
<td>27</td>
<td>3</td>
</tr>
<tr>
<td>Blood pressure</td>
<td>27</td>
<td>3</td>
</tr>
<tr>
<td>Total</td>
<td>54</td>
<td>6</td>
</tr>
</tbody>
</table>
The “stakeholder involvement” domain evaluates whether clinicians participated in the development process and sufficiently reflected the needs of field users. This domain comprised metrics of whether clinicians had participated in DCM design (development), review, and approval processes.

The “rigor of development” domain evaluated whether all the data elements (or contents) of a DCM had been developed using evidence-based clinical knowledge. The data elements of DCMs are usually derived from clinical service guidelines, medical textbooks, care path guidelines, etc.; in addition, they may reflect current clinical practice patterns. Thus, this domain consists of metrics for evaluating whether users can see what reference data are utilized. The objects of the domain evaluated whether a DCM described clinical data elements (name, value, object and attribute) accurately and specifically for field users; its metrics were also related to these items.

The “clarity and presentation” domain evaluated whether the language and format of a DCM were described clearly. Its quality metrics included “whether to use a typified DCM modeling language,” “data coding for computer processing,” “consistent representation of data elements,” and “appropriate use of cardinality”.

The “compliance to standards” domain is a key requirement for guaranteeing DCM-based exchange and reusability of medical information. Therefore, it evaluated whether the DCM complied with current standards. This domain consisted of metrics evaluating the use of standard terminology, standard data types, and standard measurement units.

The “general methodology” domain classified the major attributes of DCMs and evaluated whether they were satisfied. Its metrics evaluated whether the DCM supports attributes such as modifiers, the source of information, and appropriate support for negation. While other domains evaluated individual DCMs, the general methodology domain evaluated the entire modeling methodology.

The “metadata” domain evaluated the faithfulness of the metadata. Metadata is not directly related to the quality of a DCM but provides critical information on its appropriate use. All of the panel members confirmed that metadata are essential. The quality metrics of this domain included the DCM developer, version, and the accessibility and scope of the metadata.

The “management and maintenance” domain evaluated activities for the process management and quality improvement of a DCM. Its quality metrics included “responsible maintenance institution” and “feedback mechanism for collecting users’ opinions on improvement”.

Each metric included the following attributes: definition, objects of evaluation, evaluation method, and scoring method. The evaluation method was direct evaluation. The objects of evaluation were usually documents describing a DCM, DCM

### Table 5 – The structure of the DCM Quality Metrics.

<table>
<thead>
<tr>
<th>Quality domains</th>
<th>Quality metrics</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Scope and purpose</td>
<td>1.1. Purpose of DCM</td>
</tr>
<tr>
<td></td>
<td>1.2. Appropriate description of application target of DCM</td>
</tr>
<tr>
<td>2. Stakeholder involvement</td>
<td>2.1. Clinician-participants involved in the development of DCM</td>
</tr>
<tr>
<td></td>
<td>2.2. Clinician-participants involved in the review/approval of DCM</td>
</tr>
<tr>
<td>3. Rigor of development</td>
<td>3.1. Mention of reference(s) used in DCM development</td>
</tr>
<tr>
<td></td>
<td>3.2. Correctness (embodiment of usage patterns)</td>
</tr>
<tr>
<td>4. Clarity and presentation</td>
<td>4.1. Formal syntax (codes of data elements written in ADL, XML, UML, OWL, etc.)</td>
</tr>
<tr>
<td></td>
<td>4.2. Codification of DCM text data elements</td>
</tr>
<tr>
<td></td>
<td>4.3. Appropriate use of cardinality</td>
</tr>
<tr>
<td></td>
<td>4.4. Precise mention of DCM domain</td>
</tr>
<tr>
<td>5. Compliance to standard</td>
<td>5.1. Use of international standard terminology</td>
</tr>
<tr>
<td></td>
<td>5.2. Use of international standard data types</td>
</tr>
<tr>
<td></td>
<td>5.3. Use of international standard units of measures</td>
</tr>
<tr>
<td></td>
<td>5.4. Examination of use of data types</td>
</tr>
<tr>
<td></td>
<td>5.5. Handling exception(s) to the use of international standard terminologies</td>
</tr>
<tr>
<td></td>
<td>5.6. Examination of the quality of mapping to international standard terminologies</td>
</tr>
<tr>
<td>6. General methodology</td>
<td>6.1. Use of modifiers</td>
</tr>
<tr>
<td></td>
<td>6.2. Information provenance</td>
</tr>
<tr>
<td></td>
<td>6.3. Appropriate support for negation</td>
</tr>
<tr>
<td></td>
<td>6.4. Appropriate support for past history</td>
</tr>
<tr>
<td></td>
<td>6.5. Appropriate provision of narrative text</td>
</tr>
<tr>
<td>7. Metadata</td>
<td>7.1. Developer of DCM</td>
</tr>
<tr>
<td></td>
<td>7.2. Appropriate description of discipline of DCM user</td>
</tr>
<tr>
<td></td>
<td>7.3. DCM version</td>
</tr>
<tr>
<td></td>
<td>7.4. Accessibility of metadata</td>
</tr>
<tr>
<td></td>
<td>7.5. Extent of metadata</td>
</tr>
<tr>
<td></td>
<td>7.6. Open access to the review and/or approval results of DCM</td>
</tr>
<tr>
<td></td>
<td>8.2. Existence of user feedback mechanism for DCM</td>
</tr>
</tbody>
</table>
contents (or instances), DCM metadata, DCM modeling language, as well as development policies and guidelines. In terms of scoring, ‘0’ means ‘Not satisfied’ and ‘1’ means ‘Satisfied’. A detailed description of the complete set of quality measures is available as supplementary material.

5. Discussion

This study has developed DCM quality metrics with high validity and reliability for the first time with the collaboration of an international expert panel. This study started with criteria found in the published literature, but the criteria were consolidated and converted into a coherent set of metrics that could be objectively evaluated, which distinguishes this study from previous work. This set of criteria can now serve as a basis for practical use and implementation and for further research and testing. Several interesting questions follow from this work.

First of all, will the use of the metrics have the intended result? That is, are models that meet these metrics more consistent, easier to use, and a better basis for interoperability than models that do not meet these criteria? A logical next step would be to apply the metrics to a large corpus of DCMs, some of which meet all criteria and some that do not, and then monitor the ease of use, frequency of use, rate of error correction, etc., for all of the models and see whether the models that meet the metrics are actually “better” in clinical use.

Secondly, it would be good to know which criteria contribute the most to “goodness.” One would suspect that the metrics are not of equal value. For example, one could speculate that tracking of meta data and model versions is essential to model use while mention of references might be nice but less important. It would be interesting to design an experiment that would allow some evaluation of the value of each metric in producing the “best” and most usable models.

A third question would be how much it costs in time and resources to fulfill the criteria for each metric, and whether the cost is equal or greater than the value obtained by meeting the criterion. For example, one could speculate that it may be costly for modelers to involve practicing clinicians at every step of model development, or that it may be time consuming to collect and record literature references for the models. The cost of meeting these criteria could slow or stop the development of models and could outweigh the benefits. More research is needed into the cost and benefits of each metric to ensure that each metric is not only “good”, but that it provides value that is commensurate with its development cost.

Finally, one could question whether all quality criteria and quality domains have been discovered. Are the 29 metrics and 8 domains sufficient? Greater experience with a large corpus of models could well lead to the discovery of additional metrics and quality domains.

There are a couple of limitations to the study. Although the experts agreed that these quality metrics are applicable to the quality evaluation of any DCM, the objects of evaluation in this study were limited to two archetypes. However, we think that these quality metrics can be used for evaluating quality of other DCMs because the structure and content of archetype is more complex and more comprehensive than other DCMs. In addition, two archetypes are one of the typical ones in terms of their structure and attributes. Although the quality metrics were developed from a broad set of published criteria, most of the metrics ultimately came from the archetype literature. The metrics might be best suited to evaluating archetypes. Considering those limitations, external validity of the quality metrics can be improved when its validity is tested for a larger set of archetypes and other DCMs. These quality metrics can be also improved when the metrics showed low agreement rates in the reliability test were retested to check their completeness. It would be good to do a follow on study that would evaluate whether the metrics would work as well and be as applicable when reviewing HL7 templates, CDA templates, clinical element models, etc.

The Clinical Information Modeling Initiative (CIMI) is an international collaboration that is dedicated to providing a common format for detailed clinical models so that semantically interoperable information may be created and shared in health records, messages and documents [32]. CIMI is committed to making these specifications openly available in a number of formats, beginning with the Archetype Definition Language (ADL) from the openEHR Foundation (ISO 13606.2) and the Unified Modeling Language (UML) from the Object Management Group (OMG), with the intent that the users of these specifications can convert the models into their local formats. With increasing research and implementation of DCMs in international standards organizations and local health information systems as well, it is important to ask how the models’ quality can be objectively measured. If DCMs are to adequately support the EHR documentation needs of clinical practice, be endorsed by clinical professional bodies and health services, and be adopted by vendors, these models have to be of good quality, trusted, and in the future, certified [33].

6. Conclusion

A set of quality metrics for DCMs were developed and validated using existing published quality criteria and an international panel of experts. Given that the metrics were validated by a panel of DCM experts, we expect them to be used to support rational decision-making by DCM developers who will now have the essential qualitative and quantitative quality requirements to use as guidelines as they create new content. Clinical users can then use the quantitative assessments as they select models for specific use cases and implement them in their clinical systems.

Author’s contributions

Sun-Ju Ahn contributed in the design and conduct of this research and the writing of this manuscript. Stan Huff contributed as a panel member and provided valuable comments to the study. Yoon Kim is an adviser to this paper. He provided valuable comments to the study. Dipak Kalra contributed as a panel member and provided valuable comments to the study.
Summary points
What was already known on the topic before the study?

- Detailed clinical model supports health information exchange and reusability.
- In order for DCMs to adequately support the EHR documentation needs of clinical practice, to be endorsed by clinical professional bodies and health services, and to be adopted by vendors, these models have to be of good quality, trusted and, in the future, certified.
- To implement DCM in clinical practice, one of the most critical conditions is that user should evaluate the quality of the DCM systematically.

What this study has added to our knowledge?

- A set of quality metrics for DCMs have been developed and validated using existing published quality criteria and an international panel of experts.
- Using these DCM quality metrics, any DCMs may be compared and evaluated in a systematic and objective manner.
- The quality metrics for DCM also provides a qualitative and quantitative quality requirement to DCM developers, users and evaluators.

Competing interests
None of the authors have any competing interests to declare.

Appendix A. Supplementary data
Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ijmedinf.2012.09.006.

References