International Journal of Artificial Intelligence in Education (2019) 29: 145–188
https://doi.org/10.1007/s40593-018-00172-w

ARTICLE

Ontology-Based Generation of Medical, Multi-term MCQs

J. Leo[1] · G. Kurdi[1] · N. Matentzoglu[1] · B. Parsia[1] · U. Sattler[1] · S. Forge[2] · G. Donato[2] · W. Dowling[2]

Published online: 25 January 2019

Abstract

Designing good multiple choice questions (MCQs) for education and assessment is time consuming and error-prone. An abundance of structured and semi-structured data has led to the development of automatic MCQ generation methods. Recently, ontologies have emerged as powerful tools to enable the automatic generation of MCQs. However, current question generation approaches focus on knowledge recall questions. In addition, questions that have so far been generated are, compared to manually created ones, simple and cover only a small subset of the required question complexity space in the education and assessment domain. In this paper, we focus on addressing the limitations of previous approaches by generating questions with complex stems that are suitable for scenarios beyond mere knowledge recall. We present a novel ontology-based approach that exploits classes and existential restrictions to generate case-based questions. Our contribution lies in: (1) the specification of procedure for generating case-based questions which involve (a) assembling complex stems, (b) selecting suitable options, and (c) providing explanations for option correctness/incorrectness, (2) an implementation of the procedure using a medical ontology and (3) and evaluation of our generation technique to test question quality and their suitability in practise. We implement our approach as an application for a medical education scenario on top of a large knowledge base in the medical domain. We generate more than 3 million questions for four physician specialities and evaluate our approach in a user study with 15 medical experts. We find that using a stratified random sample of 435 questions out of which 316 were rated by two experts, 129 (30%) are considered appropriate to be used in exams by both experts and a further 216 (50%) by at least one expert.

Keywords OWL · Ontology · Medical education · Student assessment · e-learning · Question generation

✉ J. Leo
jared.leo@manchester.ac.uk

[1] Computer Science, The University of Manchester, Oxford Road, Manchester, M139PL, UK

[2] Elsevier, 1600 John F. Kennedy Boulevard, Philadelphia, PA 19103, USA

Introduction

Multiple choice questions (MCQs) are widely used to measure achievement, intelligence, knowledge or skills of interests in tests that vary in purpose, size, and delivery format. Results obtained through these questions aid decision making, such as college admissions, graduation, and job placements. They also play an important role in evaluating how efficient the instructional activities are and how to revise these activities. In addition to their role as an assessment and evaluation tool, MCQs are used as learning and revision tools (e.g., drill and practice exercises).

There are, however, challenges involved in developing and using MCQs. One challenge is the continuous need to develop a large number of distinct MCQs in order to maintain their efficacy and test security. Reusing questions poses a threat to the validity of exams, since answers may become learned or memorised without representing real understanding or skills. The Computerised Adaptive Test (CAT), in which questions are tailored for test takers, is another context in which a large number of questions is needed. It is estimated that a CAT consisting of 40 MCQs administered twice a year requires 2000 questions minimally (Breithaupt et al. 2010 cited in Gierl et al. 2012).

Constructing high-quality MCQs is an error-prone process. An evaluation of 2770 MCQs collected from formal nursing examinations administered over a five-year period showed that about 46% of the questions contain one or more item-writing flaws[1] (Tarrant et al. 2006). This is explained by the fact, pointed out by Tarrant et al. (2006), that "*few faculty have adequate education and training in developing high-quality MCQs*". Item-writing flaws can destroy the validity of the questions (i.e. the extent to which they measure the construct of interest). For example, the similarity in wording between the question and the correct answer can cue test takers to the correct answer without them having the required knowledge.

To provide support for the construction of MCQs, automatic question generation (AQG) techniques were introduced. AQG has the potential to satisfy demand by producing large numbers of MCQs efficiently and therefore facilitating the preparation of different tests and decreasing the re-use of questions from previous years. AQG techniques can help educators in employing effective teaching and assessment strategies that are otherwise hindered by the formidable task of creating large numbers of items. It can facilitate providing students with MCQs as a form of drill and practice exercises. Utilisation of the benefits of repetition can be achieved using AQG methods that vary the question by using different scenarios, choices, and/or a new format. AQG can also fulfil the vision of adaptive (personalised) learning by providing personalised questions while taking into account learner ability and preferences. Furthermore, it can ease self directed learning by allowing learners to self validate their knowledge.

Ontologies, which are being increasingly used for representing domain knowledge, especially in the biomedical domain (Guardia et al. 2012), have emerged as a source for the construction of MCQs due to their precise syntax and semantics.

[1] Violations of best practices for authoring MCQs such as avoiding the option 'all of the above'.

We introduce a modular system called the EMMeT Multiple Choice Question Generator (EMCQG), for automatic generation of multi-term MCQs, specifically targetting the medical domain by making use of a medical ontology.

EMCQG is based on *The Elsevier Merged Medical Taxonomy* (EMMeT) knowledge base, and is capable of generating medical case-based questions which are standard in medical education because of their ability to invoke higher order thinking and problem solving skills. These questions mimic a real-life scenario and require integration of medical signs and symptoms in order to arrive at a diagnosis or a management decision. EMCQG is not open source, and thus not available for public review.

We also present an update on the contents of the current version of EMMeT-SKOS (v4.4) and its translation into an OWL Ontology named EMMeT-OWL, extending work carried out in Parsia et al. (2015).

Finally, we generate more than 3 million questions for four physician specialities and evaluate our approach in a user study with 15 medical experts.

The contributions of this work include the design, implementation, and evaluation for an ontology-based approach for generating case-based questions, which are a complex class of questions. We show that our approach for assembling a stem and selecting options generates questions that are appropriate to be used for assessment.

Background

MCQs

An MCQ consists of a short textual sentence or paragraph that introduces the question, called the stem, along with a set of plausible but incorrect options, the distractors, and a set of correct or best options, the keys. The conventional form of MCQs is what is called a single response question, having only a single key. Another popular form of MCQs is the multiple-response question which differs from single response questions by allowing for multiple keys. The structure of single response MCQs is illustrated in the following example *Q1*:

What is the capital of X?

Q1: What is the capital of France? ◄ ***Stem***

A. Cairo
B. Rome
C. Canberra

} **Distractors**

D. Paris ◄ **Key**

A high-quality MCQ is of an appropriate cognitive level and difficulty, discriminating, not guessable and error free. The cognitive level of questions is classified using existing taxonomies such as Bloom's taxonomy (Bloom et al. 1956), SOLO taxonomy (Biggs and Collis 2014), or Webb's depth of knowledge (Webb 1997).

Question difficulty, discrimination, and guessability are identified through a statistical analysis of responses to a particular question (i.e. item analysis) (Crocker and Algina 1986). The standard methods for item analysis are the item response theory and the classical test theory.

Distractors are a major determinant of MCQ quality. Distractors should be functional (i.e. selected by some examinees),[2] otherwise, the guessability of the questions will increase. A guessable question is invalid since it is not possible to differentiate between, based on its result, examinees who have the required knowledge from examinees who do not. Several MCQ writing guidelines emphasise the importance of avoiding errors that make distractors non-functional, such as grammatical inconsistency within the stem (Haladyna et al. 2002). As can be seen from *Q2* (below), which has the same stem and key as *Q1* but has a different set of distractors, the choice of distractors makes the question guessable.

What is the capital of X?

Q2: What is the capital of France? ◄ ***Stem***

A. Sky
B. Tree
C. Elephant

} **Distractors**

D. Paris ◄ **Key**

When considering MCQ generation techniques, we divide the stem into *stem components*. Stem components specify the characteristics of the relevant entities that appear in the stem (*stem entities*).[3] Analogous to a database, stem components can be seen as table schemas while stem entities can be seen as the actual data stored in the tables. Each stem component is defined by:

- an entity type,
- a relation that connects the question key to entities of the entity type,
- a relation annotation that can either indicate the empirical strength of the relation between the stem entities and the key, or the empirical strength and some restrictions on the relation.

In *Q1*, *'What is the capital of'* is a textual element that is fixed for all questions of this type. This type of question requires one stem component, whose entity is $France$. This stem component is defined as follows:

- it has an entity type *Country* (France is a country),
- it is connected to the key via a *has capital* relation (i.e., France *has capital* Paris),
- it does not have any strength considerations since there is no degree of strength on the relation *has capital* (however, many relations in the medical domain have a degree of strength, as will be seen later).

[2]Distractors selected by less than 5% of examinees are usually replaced or refined.

[3]Through this paper, we use '*multi-term questions*' to refer to questions with multiple stem entities.

We also adopt a similar definition of what we refer to as *option components* and *option entities*. Option components correspond to either a question's key or a question's distractors. The definitions of option components guide the selection of a valid key and plausible distractors. Referring back to *Q1*, each option component is defined as follows:

- it has an entity type *City* (Rome is a city),
- it must be connected to a *Country* via *has capital* relation (e.g. Italy *has capital* Rome).
- if the corresponding entity is the question's key, then it must be connected to the *stem entity* via the *has capital* relation.
- it does not have any strength consideration. However, it is possible to impose some restrictions, such as limiting the option entities to capital cities located in the same continent that the stem entity is located to increase their plausibility, as is the case in *Q3*:

What is the capital of X?

Q3: What is the capital of France? ◄ ***Stem***

A. London
B. Rome
C. Berlin
} **Distractors**

D. Paris ◄ **Key**

Case-Based MCQs

Case-based questions (also known as vignettes) are a popular type of MCQs. For example, they constitute a major part of questions used in medical education and medical licensing examinations which are used to judge readiness to practice. A study of types of questions used in German National medical licensing exam between October 2006 and October 2012 shows that among 1,750 questions, 51.1% were case-based questions (Freiwald et al. 2014). A real case-based question provided by the National Board of Medical Examiners (NBME 2017) is presented in *Q4* below:

What is the most likely dignosis?

Q4: A 50-year-old man has had gradually progressive hand weakness. He has atrophy of the forearm muscles, fasciculations of the muscles of the chest and arms, hyperreflexia of the lower extremities, and extensor plantar reflexes. Sensation is not impaired. Which of the following is the most likely diagnosis?

A. Amyotrophic lateral sclerosis ◄ **Key**
B. Dementia, Alzheimer type
C. Guillain-Barré syndrome
D. Multiple cerebral infarcts
E. Multiple sclerosis

The adequacy of case-based questions in assessing the skills required of medical graduates such as clinical reasoning and judgment have been a subject of ongoing research. While the suitability of case-based questions is not the subject of this paper, there is a good body of evidence which advocates their usage in assessment. Case-based questions are classified as testing higher order thinking and invoking problem-solving (Cunnington et al. 1997; Abdalla et al. 2011; Schuwirth et al. 2001). Furthermore, when compared to other question formats, these questions were able to discriminate better between low- and high information students (Carroll 1993; Lu and Lynch 2017). These questions can also be used to teach and train students on pattern recognition skill used by experts to solve clinical problems (Coderre et al. 2003; Elstein and Schwarz 2002).

Apart from assessment, case-based questions have also been used as an instrument to measure health professionals' adherence to clinical practice guidelines (Peabody et al. 2000; Veloski et al. 2005; Rutten et al. 2006; Converse et al. 2015). They are found to approximate costly approaches for measuring clinical decisions such as standardized patients[4] (Peabody et al. 2000). Additionally, there is evidence suggesting the consistency between responses to vintage cases and actual behaviour in real-life situations (Converse et al. 2015).

Several reasons make case-based questions a good, yet challenging candidate to computerised generation. In addition to their popularity and educational value, the structured format of these questions makes them suitable for automatic generation. Additionally, their stems consist of multiple terms and combining arbitrary terms randomly is expected to result in semantically incoherent questions (e.g. a child with a history of abortion, or a patient with a history of cancer and lung cancer). Hence, there is a challenge of coordination between these terms to get coherent questions.

Related Approaches

Automatic question generation from a variety of structured and unstructured sources is an active research area. Based on a recent systematic review (Alsubait 2015), text and ontologies are the most popular sources of auto-generated questions. Despite the fact that generating questions from textual sources has a longer history, studies utilizing texts are centred around either generating free response questions or multiple choice questions for the language learning domain.[5] Text-based approaches are suitable for generating free response questions because they do not require generating distractors which are difficult to find in the input text. They are also suitable for language questions because distractors can be generated by applying simple strategies such as changing the verb form or changing the part of speech of the key. Note that one of the limitations of text-based approaches is the high lexical and syntactic

[4] Standardized patients are actors trained to observe professional performance.

[5] 34 out of 39 studies as calculated from results provided in the systematic review on automatic question generation (Alsubait 2015).

similarity between generated questions and the input text. Paraphrasing questions requires text understanding and disambiguation (e.g. coreference resolution). On the other hand, ontology-based approaches are suitable for generating knowledge-related, free response or multiple choice, questions. Based on results provided in Alsubait (2015), 7 out of 11 studies that use ontology-based approaches generate domain-independent MCQs. In addition, ontology-based approaches allow the generation of questions that are varied in lexical and syntactic structures with lower effort. For example, using a synonym or abbreviation of a term in questions does not require disambiguation or using additional sources such as WordNet (Miller et al. 1990).

In what follows, we briefly review relevant MCQ-generation approaches. Based on our observations about text-based approaches, we mainly focus on ontology-based approaches (Papasalouros et al. 2008; Žitko et al. 2009; Cubric and Tosic 2011; Jelenković and Tošić 2013; Alsubait et al. 2014; Al-Yahya 2014; Ellampallil and Kumar 2017) in addition to approaches that tackle question generation in the medical domain (Karamanis et al. 2006; Wang et al. 2007; Gierl et al. 2012; Khodeir et al. 2014).

One of earliest ontology-based approaches has been developed by Papasalouros et al. (2008). Questions generated by this approach follow three templates based on the knowledge that they intend to test, although the generated questions share the same stem *'Choose the correct sentence'*. The question below (*Q5*), taken from Papasalouros et al. (2008), is an example of questions that test examinees' knowledge about relationships between individuals. For this template, question keys are generated based on ABox axioms of the shape $R(a, b)$, where R is a relation, and both a and b are individuals. The authors propose multiple strategies for generating distractors. For example, the distractors in *Q5* were generated by selecting individuals who are members of a class equal to, or a subclass of, the range of R.

Choose the correct sentence:

Q5: Choose the correct sentence:

A. Kirillo Monina was sponsored by Kostaki Adosidi. ◄ **Key**

B. Kirillo Monina was sponsored by Herodotus.

C. Kirillo Monina was sponsored by Eupalinos.

D. Kirillo Monina was sponsored by Theofanis Arelis.

A recent approach has been presented in Alsubait et al. (2014), Alsubait (2015). The authors have developed five basic templates and rely on concept similarity to select question distractors. An example of questions generated by the approach is the question *Q6* (below) taken from Alsubait (2015). The key is a subclass of *hierarchy generation technique*. Each distractor is selected such that: 1) it is a non-subclass of hierarchy generation technique and 2) its similarity to the key is greater than a threshold.

Which is X?

Q6: Which of the following is a hierarchy generation technique?

A. Process laddering technique ◄ **Key**

B. State transition technique

C. Unstructured interview

D. Process map technique

Karamanis et al. (2006) have tackled the generation of medical questions using both text and the Unified Medical Language System (UMLS) thesaurus (Bodenreider 2004) as inputs. Question *Q7* below is generated from the sentence "Chronic hepatitis may progress to cirrhosis if it is left untreated". Questions are assembled from sentences of the 'SV(O)' structure that contain at least one term from the UMLS, after using the term frequency-inverse document frequency method to exclude terms such as 'patient' and 'therapy'. The sentence is transformed into a stem by replacing the UMLS term with a 'wh-phrase' selected based on the UMLS semantic type of the term. The UMLS semantic type and distributional similarity are used to select similar distractors.

Which disease or syndrome?

Q7: Which disease or syndrome may progress to cirrhosis if it is left untreated?

A. chronic hepatitis

B. hepatic failure

C. hepatic encephalopathy

D. hypersplenism

Wang et al. (2007) also investigates the generation of open-response questions about diseases, symptoms, causes, therapies, medicines and devices. Their generator takes a sentence as input, annotates the sentences with named entities using the UMLS thesaurus, and matches the annotated sentence with manually developed templates. The matching is done based on the presence of specific named entities and keywords in the annotated sentence. For example, the template "what is the symptom for DISEASE?" will be matched if the sentence contains named entities of the type disease and symptom, and one of the words feel, experience, or accompany. Finally, place-holders in the template are replaced by named entities from the annotated sentence.

Similar to our purpose, Gierl et al. (2012) focuses on generating medical case-based questions. Their method relies heavily on domain experts who start with a sign or a symptom and identify possible diagnoses (to be used as options) and conditions related to these diagnoses (to be used as stem entities). They use the information identified by experts to build templates (item models as named by the authors) and generate various questions per template. Taking the example of postoperative fever presented by the authors, experts identified six possible diagnoses: Urinary Tract

Infection (UTI), Atelectasis (A), Wound Infection (WI), Pneumonia (P), Deep Vein Thrombosis (DVT), and Deep Space Infection (DSI). Following this, experts identified information required to distinguish between these diagnoses such as the timing of fever and then set the possible values (1–2 days for A, 2–3 days for UTI, 2–3 days for WI, P, and DVT, and 4–6 days for DSI). Finally, the generator assembles questions by selecting a subset of the conditions provided by experts and values that match the selected conditions (setting the key to UTI and timing of fever to 3). Note that each template is specific to a sign or symptom and there is a slight variation between questions generated from the same template. From an exam perspective, questions generated from the same template substitute for one or possibly two questions in an exam because they cover the same topic. Also, most of the work is done manually, and the generator is used only to assemble all possible combinations of the model developed by experts.

Khodeir et al. (2014) also generate diagnostic questions using Bayesian network knowledge representation, such as question *Q8* below. As can be seen from the example, the stem consists of one stem component (the presenting symptoms). Patient demographics and histories which are standard in these questions are not included. In addition, its not clear how the most probable diagnosis is determined if, for example, two diseases $D1$ and $D2$ are related to two symptoms $S1$ and $S2$ where $D1$ is related to $S1$ with high probability and to $S2$ with low probability while $D2$ is related to $S1$ with low probability and to $S2$ with high probability.

Which is X?

Q8: If you have a case with maculopapular rash, sore throat, and rash fades choose and rank from the following diseases beginning by 1 to the highest likely diagnosis?

A. Measles

B. Rubella

C. Scarlet fever

D. Rosola infantum

E. Chickenpox

F. Infectios monoucleosis

Certain limitations were observed in current question generation approaches. Most notable is the simplicity of the structure of auto-generated questions when compared to hand-crafted questions. The questions generated in Papasalouros et al. (2008), Al-Yahya (2014), Jelenković and Tošić (2013), Cubric and Tosic (2011), Alsubait et al. (2014), Žitko et al. (2009), Wang et al. (2007), Karamanis et al. (2006) are restricted regarding their basic form, where the question stem contains at most two stem entities. They are also restricted regarding their cognitive level, where the majority of the generated questions in Alsubait et al. (2014), Al-Yahya (2014), Papasalouros et al. (2008), Ellampallil and Kumar (2017), Žitko et al. (2009) test only students' ability to recall learned

information (e.g., memorising definitions). This has also been highlighted by Khodeir et al. (2014) who stated that "*factual and definitional questions are the common types of questions in these [current] approaches*". There is a lack of questions that test higher forms of thinking such as applying learned knowledge to new situations, analysing learned knowledge and applying one's judgement which are valuable in many curricula (Tractenberg et al. 2013). While simple recall questions are still valuable, moving forward toward assembling complete exams, whether manually or automatically, requires questions that are varied in structure and cognitive levels.

Note that there is no simple relation between the number of stem entities and the cognitive complexity of questions. Having a stem with multiple stem entities does not necessarily raise the cognitive level of the question. Questions with a small number of stems entities, such as analogy questions[6] that have only two terms are higher in cognitive level than multi-term definition questions. Other factors also play a role in determining the cognitive level of questions. For example, prior exposure to questions at a high cognitive level in practice or sample exams may reduce the cognitive level to recall. However, from a computational perspective, generating multi-term questions is harder than generating a stem with one or two stem entities.

In this study, we focus on addressing the limitations observed in previous approaches, namely: 1) the simplicity of the structure of the generated questions and 2) the limited cognitive level of the generated questions.

EMMeT

The Elsevier Merged Medical Taxonomy (EMMeT) is a large clinical data set intended to act as a tool for search-based applications in a clinical setting. In its initial release, EMMeT was encoded entirely as a SKOS (Miles and Bechhofer 2009) knowledge base under the rationale of publishing the vocabulary in a standard format for publication on the Semantic Web.

We briefly outline the contents and structure of EMMeT v4.4.

EMMeT 4.4 Structure and Contents

Concepts EMMeT v4.4 contains over 900 K concepts covering clinical areas such as *anatomy, clinical findings, drugs, organisms, procedures*, and *symptoms*.

These concepts are defined in EMMeT by making use of the standard `skos` and `skosxl` terms.

Amongst these terms are elements to classify the concepts, e.g., `skos:Concept` and `skos:ConceptScheme`, as well as elements to provide human-readable representations of the concepts such as the `skosxl:prefLabel`. EMMeT also uses elements such as `skos:narrow`, `skos:broad` and `skos:exactMatch` to express relationships to concepts in external concept schemes or vocabularies, such as SNOMED-CT (Spackman et al. 1997) or ICD (World Health Organization 1992).

[6]Questions with stems of the form "A is to B as" and options of the for "C is to D". These questions require test takers to select the option with terms that share the same underlying relation as the terms in the stem.

Relations EMMeT contains over 1.4M `skos:broader`, `skos:narrower` and `skos:related` relations, that describe both hierarchical and associative relations between concepts. Whenever a custom property is needed, such as explicit semantic relationships (those that are more precise than the `skos:related` relation), EMMeT defines custom relations as sub-properties of standard W3C properties. EMMeT contains over 350 K custom clinical semantic relations, such as *hasClinicalFinding, hasDrug, hasDifferentialDiagnosis* and *hasRiskFactor*. The custom semantic relations come equipped (through reification) with a specified *ranking of importance* that the relation has in the general knowledge base. In its current application, the ranks are used in several ways, including to filter or order search results. Ranks are defined in the range of 0–100, where a higher number indicates in some way a stronger relation, however, the actual usage of the ranks are less granular and range only from 6–10.

Outside the SKOS terminology is a new experimental set of semantic relations between concepts called *Point Of Care (POC)* semantic relations. As with the custom semantic relations, POC relations are reified with a ranking, but are also reified with five additional attributes, namely: age, sex, conditions, genetics, and ethnicity. The additional attributes act as a set of constraints on the relation for which the relation itself applies to a specific population group.

POC relations are a separate terminology but are linked to EMMeT using IDs of related terms and are currently stored in CSV files. There are approximately 8 K POC relations which are set to be included in the next version of EMMeT, as reified custom semantic relations.

EMMeT Content Example To illustrate the content of EMMeT, consider Fig. 1. This extract displays the usage of the elements described above. The extract represents a graph between the following six concepts:

- `urethra disorders`
- `urethritis`
- `follicular urethritis`
- `obstetrics and gynecology`
- `hemorrhage of urethra`
- `african american`

4 narrower/broader relations:

- `<urethra disorders>skos:broader<urethritis>`
- `<urethritis>skos:narrower<urethra disorders>`
- `<urethritis>`
  `skos:broader<follicular urethritis>`
- `<follicular urethritis>`
  `skos:narrower<urethritis>`

along with 1 custom semantic relation and 2 POC relations:

- `(<urethritis> semrel:hasPhysicianSpecialty`
  `<obstetrics and gynecology>) rank:10.0`

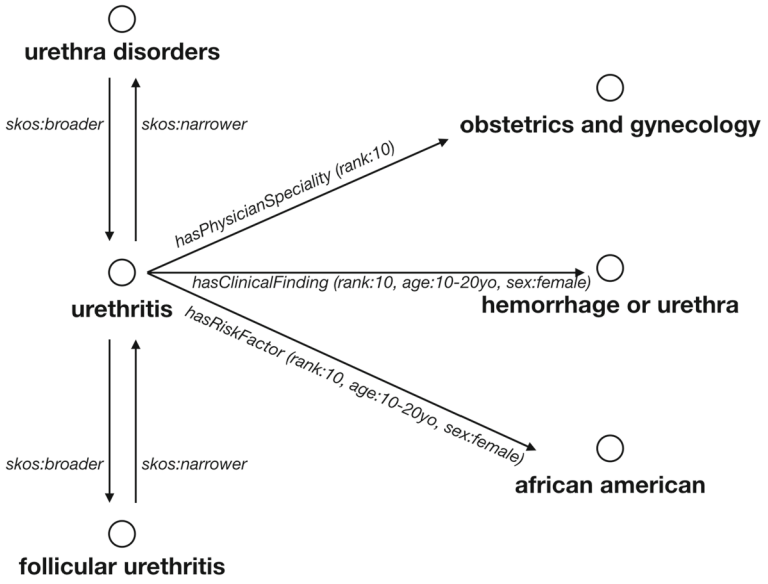

Fig. 1 A small extraction from EMMeT, illustrating the use of concepts and their relations which include their rankings and other associated data, such as sex and age

- `(<urethritis> semrel:hasClinicalFinding<hemorrhage of urethra>)`
  `rank:10.0, age:10-20yo, sex:female`
- `(<urethritis semrel:hasRiskFactor<african american>)`
  `rank:10.0, age:10-20yo, sex:female`

To demonstrate the meaning of the rankings, a rank of 10.0 for the relation *hasClinicalFinding* refers to a *most common* clinical finding (a rank of 9 would refer to only a *common* clinical finding), i.e., one of *urethritis*'s most common clinical findings is *hemorrhage of urethra*. A rank of 10.0 for the relation *hasRiskFactor* refers to a *strongly associated* risk factor (a rank of 9 would refer to a *commonly associated* risk factor).

EMMeT-SKOS ⟶ EMMeT-OWL

A description of a bespoke translation process of the then current version of EMMeT (v3.8) into an OWL 2 (Motik et al. 2009) representation was described in Parsia et al. (2015). Since then, EMMeT has evolved to version v4.4, which now contains more validated content, pulling in additional data sources beyond the internal SKOS representation. Table 1 summarises the current translation mechanism from EMMeT-SKOS to EMMeT-OWL.

The translation from SKOS to OWL was entirely automated. The translation relied heavily on the strong relationship between both SKOS and OWL. For example, `skos:Concepts` were mapped directly to `owl:Classs`, since the definition states that the former is an instance of the latter. Similarly, the SKOS relations were

Table 1 A description of the automated translation process from EMMeT-SKOS to EMMeT-OWL

EMMeT-SKOS	EMMeT-OWL
s:Concept	$OWLCl$
s:broader, s:narrower, ***s***:related	$OWLObjectProperty$
s:Concept - s:broader - ***s***:Concept	$OWLCl \sqsubseteq \exists broader.OWLCl$
s:Concept - ***s***:narrower - ***s***:Concept	$OWLCl \sqsubseteq \exists narrower.OWLCl$
s:Concept - ***s***:related - ***s***:Concept	$OWLCl \sqsubseteq \exists related.OWLCl$
semrel, POC Relation	$OWLObjectProperty$
⟨ ***s***:Concept - *semrel* - ***s***:Concept⟩:Rank	$(OWLCl \sqsubseteq \exists semrel.OWLCl) : Rank$
s:broad***M***, ***s***:narrow***M***, ***s***:exact***M***	$OWLAnnotationProperty$
s:Concept - ***s***:broad***M*** - Data	$(OWLCl : (broadM : Data))$
s:Concept - ***s***:narrow***M*** - Data	$(OWLCl : (narrowM : Data))$
s:Concept - ***s***:exact***M*** - Data	$(OWLCl : (exactM : Data))$
s:prefLabel, ***s***:altLabel	$OWLAnnotationProperty$
s:Concept - ***s***:prefLabel - Data	$(OWLCl : (prefLabel : Data))$
s:Concept - ***s***:altLabel - Data	$(OWLCl : (altLabel : Data))$

semrel = Semantic Relation, $(\alpha) : Rank$ = a logical OWL axiom α annotated with a *Rank* (achievable in OWL 2), ***Cl*** = Class, ***s:*** = skos: and ***M*** = Match

mapped to OWL object properties and so on. Several design choices were made when considering what style of OWL axioms would be best suited for the corresponding SKOS assertions. One example includes using OWL axioms of the form $A \sqsubseteq \exists R.B$ for SKOS concept to concept relations, where A and B are OWL classes (converted from SKOS concepts), and R is an OWL object property (converted from a SKOS semantic relation).

An important design choice was made when considering how to enrich the class hierarchy of EMMeT. Although some form of a class hierarchy was described in EMMeT-SKOS (e.g., through hierarchical relations such as `skos:broader` or `skos:narrower`), it could not be transferred into an OWL class hierarchy as SKOS's hierarchical assertions are not the same as OWL subclass relations. For example, consider the EMMeT concepts *Abortion* and *Abortion Recovery*. It is clear that Abortion is a broader term than Abortion Recovery, hence the use of a `skos:broader` relation in EMMeT. However, to enforce that one is a subclass of the other is false: Abortion Recovery is not a kind of Abortion. The generation of a reliable EMMeT-OWL class hierarchy was automated by aligning the concepts with classes from an external source, namely SNOMED-CT (Spackman et al. 1997). SNOMED-CT is backed by a richly axiomatised OWL ontology and a long held focus on modelling domain relations correctly. Over 100 K EMMeT concepts contained mappings to equivalent SNOMED-CT classes (through `skos:exactMatch` elements). The alignment was achieved by adding subclass relations to existing classes in EMMeT-OWL wherever a subclass relation occurred between the equivalent classes in SNOMED-CT. This resulted in over 1M subclass relations being added to EMMeT-OWL.

For a complete description of the translation process of converting EMMeT-SKOS to EMMeT-OWL, which is used as the knowledge source for EMCQG, we refer the reader to Parsia et al. (2015).

EMMeT Quality & Control

To ensure both quality and correctness, EMMeT regularly undergoes development. Concepts, as well as their semantic type, are based on terms from reliable external vocabularies, such as SNOMED-CT (Spackman et al. 1997) and UMLS (Bodenreider 2004). Whenever changes occur in the external vocabularies, they are subsequently updated in EMMeT. Additional concepts and semantic types are also added based on Elsevier content, all of which are verified by experts in the related fields.

The custom semantic relationships in EMMeT are updated quarterly, which includes adding and removing relationship instances as well as adjusting rankings on the strength of the relationship instance. A group of medical experts in the EMMeT team, including physicians and nurses, create and maintain the relationships. Each relationship is manually curated and based on evidence in Elsevier content, which includes books, journals, and First Consult/Clinical Overviews. Potential relationships identified by each editor then pass through a second clinical EMMeT editor for medical-based quality assurance (QA) review. They are then passed to an EMMeT QA editor for technical and consistency checks. All phases of the quality control involve a combination of domain expertise and use of Elsevier sources.

EMCQG's Template System

EMCQG is an MCQ generation (MCQG) system built upon EMMeT-OWL that uses built-in *templates* to generate unique questions with varying difficulty, based on the classes, relations and annotations in EMMeT-OWL. Our presented work on MCQG is the first attempt to reuse EMMeT for a new application. In this section, we briefly describe EMCQG's template system and how it relates to EMMeT-OWL.

Question Templates

A question template acts as a generic skeleton of a question with place-holders that can be filled in with relevant question content to make various questions of a similar type. For example, given the following ontology (DL syntax):

- $England \sqsubseteq Country$
- $France \sqsubseteq Country$
- $Germany \sqsubseteq Country$
- $London \sqsubseteq City$
- $Paris \sqsubseteq City$
- $Berlin \sqsubseteq City$
- $Yellow \sqsubseteq Colour$
- $Sheep \sqsubseteq Animal$

- $London \sqsubseteq \exists capitalOf.England$
- $Paris \sqsubseteq \exists capitalOf.France$
- $Berlin \sqsubseteq \exists capitalOf.Germany$

where all appropriate classes are disjoint, the question:

What is the capital of X?

Q9: What is the capital of England?

A. London

B. Paris

C. Yellow

D. Sheep

would map to the following question template:

What is the capital of X?

What is the capital of ***Country****?*

A. $X : X \sqsubseteq City \sqcap \exists capital.$***Country***

B. $X : X \sqsubseteq City \sqcap \neg \exists capital.$***Country***

C. $X : X \sqsubseteq Colour$

D. $X : X \sqsubseteq Animal$

Similar questions can be made by substituting terms from the ontology:

What is the capital of X?

Q10: What is the capital of France?

A. Paris

B. Berlin

C. Yellow

D. Sheep

The more information in the ontology, the more questions can be mapped to the template.

With regards to medical question templates, experts from Elsevier identified four question templates that were representative of the type of questions used within their publications designed to help medical residents prepare for their board examinations. These publications, and therefore the questions used as a basis for the templates, were created by Elsevier authors who are practising medical doctors and/or professors of medicine and leading experts in their speciality area. All authors are acutely aware of the types of questions used on Board examinations.

EMCQG builds questions by filling in template skeletons with appropriate content from EMMeT-OWL, and calculates and varies the difficulty of the questions depending on the content that has been chosen from EMMeT-OWL.

As an example, consider the following question template associated with testing students' knowledge on a likely diagnosis given a patient scenario. An overview of the template is as follows:

Template 1: What is the most likely diagnosis?

A PATIENT-DEMOGRAPHIC *patient with* {HISTORY}* *presents with* {SYMPTOM}*. *What is the most likely diagnosis?*

A. Correct DISEASE

B. Incorrect DISEASE

C. Incorrect DISEASE

The template's stem entities include: PATIENT DEMOGRAPHIC, {HISTORY} and {SYMPTOM}. The PATIENT DEMOGRAPHIC refers to specific patient information such as the patient's age, sex or ethnicity. The {HISTORY} is usually a set of risk factors, observations or conditions that the patient has been diagnosed with previously, and {SYMPTOM} usually represents a set of presenting symptoms or clinical findings.

The option entities are diseases which are related (via clinical semantic relations) to the stem entities. The key would have the strongest relations to the stem entities (while satisfying the patient demographics), signifying that it would be the logical choice of satisfying the *most likely diagnosis* constraint. The distractors would either have no relations or weaker relations to the stem entities than those of the key's.

EMMeT-OWL can be used to fill in the template. Regarding the {HISTORY} and {SYMPTOM} sets, there exist two object properties *hasRiskFactor* (*hRF*) and *hasClinicalFinding* (*hCF*) that can help to identify entities in EMMeT-OWL that can be used as stem entities. *hRF* is a relation that relates Diseases or Symptoms to RiskFactors, which can, in turn, be Diseases, Symptoms, ClinicalFindings, Events, Procedures, Environments, SocialContexts, Substances or Drugs, each of which can be validated as a patient's history information. With regards to the {SYMPTOMS}, *hCF* is a relation that relates Diseases or Symptoms to Diseases, Symptoms or ClinicalFindings, each of which can be used as a patient's presenting symptoms. Both relations are used in both the standard ranked *semrels* and the POC *semrels* relation space. Although EMMeT-OWL does not have any specific classes containing sets of patient demographic information (specifically, groupings of ages, sexes and ethnicities), such information can be found as annotations on POC relations (restricting the POC attributes to only age, sex and ethnicity and excluding conditions and genetics). Therefore, the patient demographic information can be gathered from a POC relation's annotation content.

Using only this information, EMCQG can fill in a skeleton of the template with appropriate terms by simply querying. In this example, no reasoning is necessary as all of the required axioms are explicit in EMMeT-OWL. EMCQG does use OWL reasoning when validating possible terms to select for a question template; this is discussed in more detail in the next Section. The template is modelled according to the

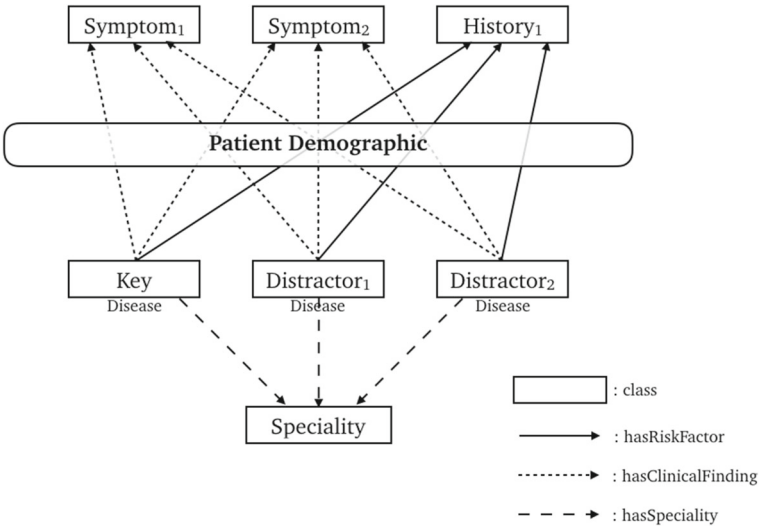

Fig. 2 The structure of the *What is the most likely diagnosis* template, using two symptoms and one history as stem entities

illustration in Fig. 2. Any terms that EMCQG chooses to fill in the roles for the option entities, the patient-demographic and the stem entities, must meet the following rules:

1. Each *hCF* and *hRF* relation from each option entity to each stem entity must be valid w.r.t the patient demographics, i.e., if the relation is a POC relation, then the attributes of the POC relation cannot conflict with the attributes of the chosen patient demographic.
2. The rank of a relation from any distractor to a stem entity must be less than or equal to the rank of the relation between the key and the same stem entity.
3. For any given distractor, the sum of its relations' ranks to all stem entities must be strictly less than the sum of the ranks of the key's relations to the stem entities.
4. Each symptom must be related to the key via a *hCF* relation and each history must be related to the key via a *hRF* relation.
5. Each option entity must have a *hasSpeciality* relation to a shared Speciality.

In a simplistic view, EMCQG searches for terms and axioms that match these rules and builds questions based on those terms. For example, the following question *Q11*:

What is the most likely diagnosis?

Q11: A 13-year-old African American female patient presents with Hemorrhage of urethra and Hematuria. What is the most likely diagnosis?

A. Dysmenorrhea

B. HIV infection

C. Urethritis ◀ **Key**

adopts the rules of template 1.

Figure 3 illustrates *Q11*. The three stem entities include the two clinical findings: *Hemorrhage of Urethra* and *Hematuria*, along with the risk factor *African American*, which are related to the option entities *Urethritis*, *Dysmenorrhea* and *HIV Infection* through both POC and ranked *hCF* and *hRF* relations. The patient demographic has the attributes *10–20, null, female* for age, ethnicity and sex respectively. The key for the question is the option entity *Urethritis*. It is easy to see that the rules are met according to the example. With over 920k concepts to choose from and over 350k relations, many varying questions can be generated.

EMMeT's Suitability for Medical MCQ Templates

When considering ontology-based medical MCQG, the nature of the underlying ontology not only needs large coverage over clinical terms, but also clinical relations. As we have seen, the templates require both clinical terms and relations between those terms to not only fill the template skeleton, but to also ensure that the chosen terms meet the rules of the template. EMMeT-OWL is the perfect candidate for such a task. It is not only sufficient in its coverage of both clinical terms and their relations, but also in the high quality and level of detail of its relations (e.g., by providing strengths of its relations through its ranking system).

As far as we are aware, there exists no alternative medical ontology with the same level of detail as EMMeT-OWL. Candidates, such as SNOMED-CT, although rich in clinical terms, lack the desired relations.

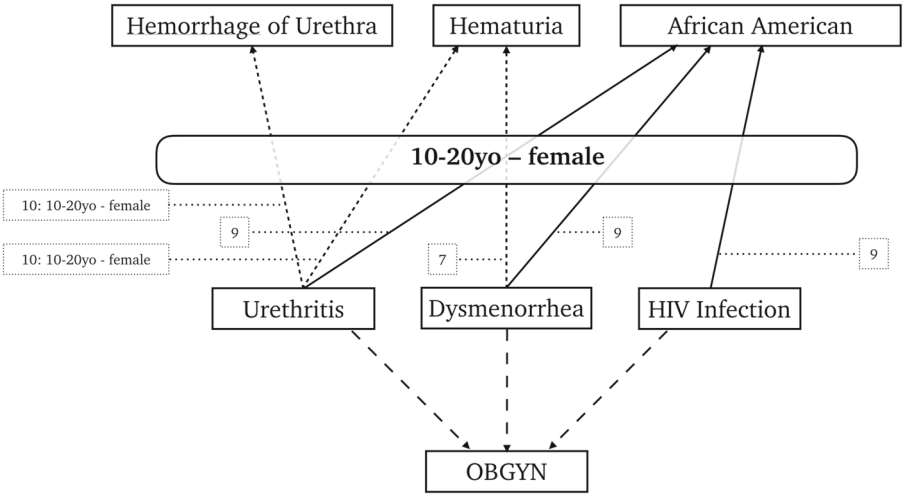

Fig. 3 A model based on axioms from EMMeT-OWL showing the *What is the most likely diagnosis* question. Note that the greater the rank, the stronger the relation is. A *hCF* of rank 10 indicates a most common clinical finding while a *hCF* of rank 7 indicates a rare clinical finding

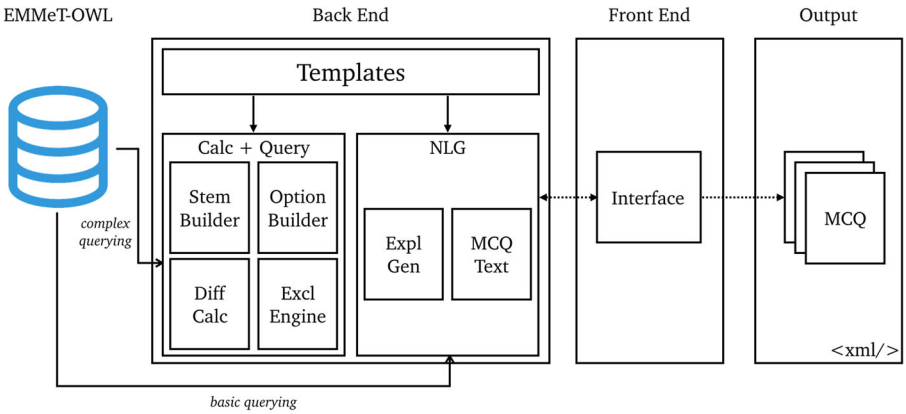

Fig. 4 A modular system diagram showing each major module in EMCQG and their position in the entire system

EMCQG—a System for Generating MCQs

EMCQG is built up of several modules that aid in the generation of case-based MCQs. One of the main modules consists of a templating system as introduced previously, and the remaining main modules act as *engines* to fill in and structure the template skeletons with content from EMMeT-OWL. The remaining six main modules consist of: (2) A *stem builder*; (3) an *option builder*; (4) an *exclusion engine*; (5) and *explanation generator*; (6) A *difficulty calculator*; and (7) a *question text generator*, each of which is described in the next Section.

A system diagram of EMCQG is depicted in Fig. 4.

EMCQG's Modules

Templates

Four medical question templates have been implemented in the current implementation of EMCQG. The naming of the templates is based on the core question the corresponding MCQ asks. The first template, introduced in the section titled "EMCQG's Template System" , is called *What is the most likely diagnosis?*.

The second template is called *What is the drug of choice?* and is presented as:

Template 2: What is the drug of choice?

A PATIENT-DEMOGRAPHIC *patient presents with* {SYMPTOM/DISEASE}. *What is the drug of choice?*

A. Correct DRUG

B. Incorrect DRUG

C. Incorrect DRUG

As with template 1, the question uses a patient demographic along with a single symptom or disease. However, no history (risk factor) information is used in this template. Also, instead of asking for the most likely diagnosis, the question asks for the *drug of choice*. Therefore, both keys and distractors are types of the EMMeT-OWL class *Drug* and are connected to the stem entities via the *hasDrug* (*hD*) relation. The key is the drug with the strongest relation to the stem entity, while each distractor either has no relation to the stem entity or one that is weaker than that of the key's.

The next template is called *What is the most likely clinical finding?*, and is presented as:

Template 3: What is the most likely clinical finding?

A PATIENT-DEMOGRAPHIC *patient presents with* {SYMPTOM/DISEASE}. *What is the most likely clinical finding?*

A. Correct CLINICAL FINDING

B. Incorrect CLINICAL FINDING

C. Incorrect CLINICAL FINDING

Again, the question uses a patient demographic and a single symptom or disease with no history information. The keys and distractors are types of *Clinical Finding* and rely on the *hCF* relation to relate them to the stem entity. Once again, the key would have the strongest relation to the stem entity, while each distractor would have either a weaker relation to the stem entity, or no relation at all.

The final template is called *What is the differential diagnosis?*, and is presented as:

Template 4: What is the differential diagnosis?

A PATIENT-DEMOGRAPHIC *patient with a history of* {HISTORY}* *presents with* {SYMPTOM}*. *What is the differential diagnosis?*

A. Correct DISEASE

B. Correct DISEASE

C. Incorrect DISEASE

D. Incorrect DISEASE

Unlike the previous questions, several keys may now appear in the option entities. As with the first template, the relations *hRF* and *hCF* are used to relate the option entities to the stem entities, as well as the relation *hasDifferentialDiagnosis* (hDDx) to interrelate the option entities.

As before, each template's keys and distractors will have to meet a set of rules to be valid w.r.t both the patient demographic and the question as a whole.

Stem Builder

Each template has its own stem builder responsible for providing a set of stem entities that are appropriate for the question stem. When given a speciality, key and patient

demographics, the stem builders retrieve a set of valid stem entities from EMMeT-OWL, which can be used in the stem of a question (collaborating with the Exclusion Engine by excluding stem entities that may invalidate a question). The stem builder implements the required rules of each question template. For example, considering Template 1, when the stem builder is tasked with finding a suitable set of symptoms, it queries the ontology for all subclasses of *Disease* and *Symptom* that are related to the key via the *hasClinicalFinding* relation. It will then exclude any classes from this list where the relation to the key violates the patient demographic. For example, if the patient demographic included an age restriction of 5–10 years old, and a class from the list was related to the key with a POC relation that contained the restriction 15–20 years old, said class would be excluded. It will then remove from the list any incompatible classes provided by the exclusion engine (see Section "Exclusion Engine"). The remaining classes are then classed as valid and can be placed in the question's stem.

Option Builder

Each template is assigned an option builder, responsible for providing a set of entities that can be used as possible answers in a template (whether they are keys or distractors). Given a speciality, key, patient demographic and a list of stem entities, each option builder will search EMMeT-OWL (again, in collaboration with an Exclusion Engine) to find entities that are valid w.r.t the rules of the template. Also, working with a difficulty calculator, the option builder will assign a difficulty to each option entity, dependant on the current question content.

Exclusion Engine

The purpose of an exclusion engine is to remove entities from potential stem or option entities that could *break* or *invalidate* the question if they were to be included. Suppose for example a patient demographic for a template included the age range of 5–10 years of age. Given a certain key, the stem builder may wish to choose the entity *Old Age* as a risk factor, but such an entity would invalidate the question. Depending on the task, the exclusion engine will provide a list of entities to exclude from potential results. As well as the age example, the exclusion engine also excludes entities w.r.t sex and also those entities derived by subclass relations in certain templates. For example, there should be no sub/superclass relation between distractors as this could make a distractor easy to eliminate.

Explanation Generator

The explanation generator acts as a simple natural language generator to provide explanations for the option entities as to why they are either correct or incorrect options. The explanation generator uses their relations to the stem entities, ranks and POC attributes to do so. Each template has its explanation generator. As an example, consider $Q11$, presented in the section titled "EMCQG's Template System". A simple explanation for the key *Urethritis* could involve a textual reading of its relations

to the stem entities as follows: "*Hemorrhage of urethra is a most common clinical finding for urethritis in 10–20 year old teenaged female patients and hematuria is a common clinical finding for urethritis in 10–20 year old teenaged female patients. African American is a commonly associated risk factor for urethritis.*". An explanation for the distractor HIV Infection could involve a comparison to the key as follows: "*Hemorrhage of urethra is not a clinical finding for HIV infection whilst it is a most common clinical finding for urethritis in 10–20 year old teenaged female patients and hematuria is not a clinical finding for HIV infection whilst it is a common clinical finding for urethritis in 10–20 year old teenaged female patients.*", where the textual representations of the relations' ranks are embedded in the generator (such as a rank nine clinical finding mapping to the description "common" while a rank ten clinical finding maps to the description "most common").

Difficulty Calculator

The difficulty calculator estimates the overall difficulty of a question using several calculations that measure different aspects of parts of the question which includes measures for the set of stem entities, individual option entities, and the set of option entities. This allows questions to be compared and placed into various categories (e.g., easy, medium, hard), and allows for users of EMCQG to understand how several terms can affect the difficulty of a question. Each template has its difficulty calculator which vary since each template has structurally different content.

Unlike previous approaches where difficulty is based on axiomatic concept similarity (Alsubait et al. 2014), the difficulty of EMCQG questions rely heavily on the ranking of relations over axioms. We introduced this adaptation to the difficulty model to account for the role of the stem in difficulty which was neglected in Alsubait et al. (2014). The stem entities' role in the difficulty calculation is to measure how indicative the stem entities are in identifying the key. The stronger their relations are to the key, the easier it will be to identify the key. The weaker their relations are to the key, the harder it will be to identify the key.

The role of the option entities in the difficulty calculation is to measure the difference between option entities' relations to the stem entities and the key's relations to the stem entities. The smaller the difference, the more indicative the stem entities are to the option entities, making them harder to differentiate from the key, and thus harder to eliminate. The larger the difference, the less indicative the stem entities are to the option entities, making them easier to differentiate from the key, and thus easier to eliminate.

The question difficulty is based on an average of the stem entities' difficulty and the option entities' difficulty.

As an example, the difficulty measure for the template *What is the most likely diagnosis?* are as follows:

Stem indicativeness ($stemInd$) is defined over two measures: the indicativeness of the symptoms ($sympInd$) and the indicativeness of the risk factors ($histInd$)

Definition 1 ($sympInd$) Let $\mathcal{S}$ be the set of symptoms and **k** be the key. Let *rank* be a function that returns the rank of any annotated axiom and let *min* and *max* be

functions that return the minimum and maximum ranks that a given relation can have (usually 7 and 10 respectively). $sympInd$ is defined as follows:

$$sympInd(\mathcal{S}, \mathbf{k}) = 1 - \left(\frac{\sum_{s}^{\mathcal{S}} (rank(\mathbf{k} \sqsubseteq \exists hCF.s) - min(hCF))}{|\mathcal{S}| \times (\max(hCF) - \min(hCF))} \right)$$

$histInd$ is calculated similarly:

Definition 2 ($histInd$) Let $\mathcal{H}$ be the set of histories and $\mathbf{k}$ be the key.

$$histInd(\mathcal{H}, \mathbf{k}) = 1 - \left(\frac{\sum_{h}^{\mathcal{H}} (rank(\mathbf{k} \sqsubseteq \exists hRF.h) - min(hRF))}{|\mathcal{H}| \times (\max(hRF) - \min(hRF))} \right)$$

Using these two measures allows $stemInd$ to be defined:

Definition 3 ($stemInd$) Let $\mathcal{H}$ be the set of histories, $\mathcal{S}$ be the set of symptoms and $\mathbf{k}$ be the key. $stemInd$ is defined as follows:

$$stemInd(\mathcal{S}, \mathcal{H}, \mathbf{k}) = \frac{sympInd(\mathcal{S}, \mathbf{k}) + histInd(\mathcal{H}, \mathbf{k})}{2}$$

The options entities' difference measure ($optDiff$) is defined in terms of each individual distractor difference ($disDiff$).

Definition 4 ($disDiff$) Let $\mathcal{S}$ be the set of symptoms, $\mathcal{H}$ be the set of histories, $\mathbf{d}$ be a distractor and $\mathbf{k}$ be the key. $disDiff$, is defined as follows:

$$disDiff(\mathcal{S}, \mathcal{H}, \mathbf{k}, \mathbf{d}) = \frac{2}{\left(\frac{\sum_{s}^{\mathcal{S}} (rank(\mathbf{k} \sqsubseteq \exists hCF.s) - \mathbf{d}_s) \times \mathbf{d}_s}{|\mathcal{S}|} + \frac{\sum_{h}^{\mathcal{H}} (rank(\mathbf{k} \sqsubseteq \exists hRF.h) - \mathbf{d}_h) \times \mathbf{d}_h}{|\mathcal{H}|} \right)}$$

where 2 is the number of stem components (specifically the histories and symptoms in this template), $\mathbf{d}_s = rank(\mathbf{d} \sqsubseteq \exists hCF.s)$ and $\mathbf{d}_h = rank(\mathbf{d} \sqsubseteq \exists hRF.h)$

Using this measure allows $optDiff$ to be defined:

Definition 5 ($optDiff$) Let $\mathcal{D}$ be the set of distractors. $optDiff$ is defined as follows:

$$optDiff(\mathcal{D}, \mathcal{S}, \mathcal{H}, \mathbf{k}) = \sum_{d}^{\mathcal{D}} \left(disDiff(\mathcal{S}, \mathcal{H}, \mathbf{k}, \mathbf{d})^2 \right)$$

Finally, question difficulty ($queDiff$) is defined as simply the average of the stem indicativeness and the option entities' difference:

Definition 6 ($queDiff$) Let $\mathcal{S}$ be the set of symptoms, $\mathcal{H}$ be the set of histories, $\mathcal{D}$ be the set of distractors and $\mathbf{k}$ be the key. $queDiff$ is defined as follows:

$$queDiff(\mathcal{S}, \mathcal{H}, \mathcal{D}, \mathbf{k}) = \frac{stemInd(\mathcal{S}, \mathcal{H}, \mathbf{k}) + optDiff(\mathcal{D}, \mathcal{S}, \mathcal{H}, \mathbf{k})}{2}$$

As an example, consider the question $Q11$ illustrated in Fig. 3. $stemInd$ is defined as the mean of the $sympInd$ and $histInd$. $sympInd$, intuitively representing the degree to which the symptoms are indicative of the key (the more indicative, the easier), can be computed as follows: $1 - \frac{(10-7)+(10-7)}{2*(10-7)} = 0$ (highly indicative). Similarly, the indicativeness of the risk factors is calculated as $1 - \frac{(9-7)}{1*(10-7)} = .33$. Hence, the stem indicativeness is $\frac{0+.33}{2} = .17$. Next, we calculate the difference of the option entities, which is, intuitively, the sum of the individual distractor differences. The individual distractor differences (Definition 4) capture how close, or similar, a distractor is to the key. This closeness is again defined regarding the empirical strength of the distractor's relations to the symptoms and risk factors, when compared to those of the key's. To capture the fact that higher degrees of closeness makes the task of excluding a distractor considerably harder, we chose to, for the lack of an empirically validated coefficient, square the individual distractor difference (thereby giving considerably more weight to a distractor which is very similar to the key). It is not useful to list the whole set of equations for the individual distractor difficulty at this point, so we restrict ourselves to an example. The difficulty of the distractor $disDiff$ 'Dysmenorrhea' is $\frac{2}{\frac{(10-6)*6+(10-7)*7}{2} + \frac{(9-9)*9}{1}} = .09$ (where 6 is the rank of a non-relation). The overall distractor set difference is: $optDiff = .09^2 + .08^2 = .015$. Lastly, the overall question difficulty (Definition 6) is defined simply as the mean of the stem indicativeness and the option entities' difference: $\frac{.17+.015}{2} = .092$.

The goal of introducing a difficulty measure is to allow users of EMCQG to generate questions for different levels of expertise. However, in this paper, we do not provide a formal evaluation of the effectiveness of our difficulty measure. We believe that a cursory understanding of our difficulty calculator helps to gain an intuitive sense how difficulty *can* be estimated. Whether or not our approach generates quality questions is independent of whether or not we can accurately predict their difficulty. A formal investigation of how well our models capture real difficulty is part of future work.

Question Text Generator

The question text generator is another natural language generator whose purpose is to generate the overall question text of the template, i.e., the suitable text that would be placed in an exam. Each template has its question text generator. Although the rules of the template are fixed, the way that stem entities appear in the question will differ based on their type (the general superclass they belong to). For example, in the *What*

is the most likely diagnosis template, if a Population Group is used as a history (risk factor), then the history will not appear in the history list, but rather as a demographic of the patient. To illustrate, instead of the question reading "*A patient with a history of African American presents with...*", the module will check if the risk factor (African American) is a subclass of any specified classes (in this case, the *PopulationGroup* class), and then proceed to reorder the question text to read "*An African American patient presents with...*". Similar rules exist in the question text generator for risk factors including age and sex. Reordered risk factors appear in the following order: 1) age, 2) population groups/ethnicities and 3) sex.

Together, these seven modules (along with various other minor modules) make up the internal structure of EMCQG.

Materials and Methods

We evaluate our approach across two dimensions. *Effectiveness* quantifies the number of distinct questions we can hope to generate from a knowledge base such as EMMeT. *Question quality* quantifies the degree to which our approach generates appropriate questions for assessment. We operationalise appropriateness as acceptances by medical instructors.

We have not considered evaluating EMCQG in comparison to existing approaches for the following reasons. The questions generated by EMCQG are more complex than questions generated by the approaches outlined in Wang et al. (2007), Karamanis et al. (2006), Khodeir et al. (2014), and thus, the performance is not comparable. We also did not compare our questions with the case-based questions produced by Gierl et al. (2012) because this approach is mainly dependent on domain experts as explained in the section titled "Related Approaches". Therefore, no quality issues will be found in their questions except errors made by domain experts. In addition, generated questions are not publicly available and the replication of the generation methodology is expensive since it requires a heavy engagement from domain experts.

We generated questions with EMCQG, underpinned by EMMeT-OWL, with the following parameters, broken down by each applicable template:

All templates:

- Questions were generated for four physician specialities: gastroenterology and hepatology, cardiology, internal medicine and orthopaedics.
- For questions involving symptoms, the symptoms were combined in such a way that at least one symptom did not belong to the class of "*commonly occurring symptoms*", with a commonality threshold of 100[7] to avoid questions such as "*A patient presents with fever and pain, what is the most likely diagnosis?*".[8]

[7]Symptoms that have at least 100 incoming hCF relations

[8]The symptoms *fever* and *pain* are so common that it would be extremely difficult to determine the key.

What is the most likely diagnosis template:

- Generated questions involved the following stem sizes (#History|#Symptom): 1|1, 2|1, 1|2, 2|2, 3|2, and 2|3, against the following number of distractors: 3 and 4.

What is the most likely clinical finding template* and *What is the drug of choice template:

- Generated questions involved the following number of distractors: 3 and 4.

What is the differential diagnosis template:

- Questions were generated with the following stem sizes (#History|#Symptom): 1|1, 2|1, 1|2, and 2|2, against the following number of keys: 1, 2, and 3, and the following number of distractors: 2 and 3.

Method Effectiveness: How Many Questions Can We Generate from a Knowledge Base?

We quantify the effectiveness of our method by comparing the density of available ontological relationships with the number of resulting questions. For example, for *What is the most likely diagnosis* questions, diseases and clinical findings are needed that are connected by the *hasClinicalFinding* relationship. The number of questions that we can generate is therefore bound by the total number of *hasClinicalFinding* relations.

Quantifying the effectiveness of the method serves two purposes. Firstly, it indicates how restrictive the constraints imposed on the generation are (e.g. all distractors must be related to the key via hDDx relation in differential diagnosis template). If the number of generated questions found to be very small compared to the number of ontological relations, then loosening the constraints to increase the number of generated questions would be one possible solution. Although this could produce flaws in questions (e.g. some non-plausible distractors), we expect that these questions can be revised, or used as seeds for other questions. We also expect that the time needed for revision is less than the time needed for writing questions from scratch. In addition, showing the relation between the properties of the knowledge base and the number of generated questions are important for ontology modellers who are interested in developing or using existing medical ontologies for case-based question generation. It serves as a guide on how an ontology should be structured along with the coverage of the ontologies clinical knowledge to get the desired number of questions. However, it is important to note that the density of stem entities is only one of the factors that affect the number of generated questions. Other factors such as the distribution of similarities between concepts, and the depth of the inferred class hierarchy are also expected to affect the number of generated questions.

Since quantifying effectiveness this way does not take into account the blacklisting and filtering of entities and ignores possible interactions of different relations, it will not serve as a precise grounds for interpolating to arbitrary ontologies and should, therefore, be viewed as an estimate.

Quality Assessment

To evaluate the quality of the questions, we conducted a study with 15 qualified medical experts who were paid for their participation. All experts have teaching experience and the majority of them have exam construction experience. See "Demographic Characteristics of Domain Experts" in the Appendix for their demographic characteristics.

Given one hour per expert, a sample size of 435 was selected for review. Our decision about the sample size was based on the following estimation. We estimated the time needed to review each question, including time needed to solve it, to be about two minutes. This estimation was made considering a similar study (Alsubait 2015) where it was reported that experts spend around one minute per question. We added another minute considering that more aspects of the questions need to be evaluated in our study.

We used a stratified sampling method in which questions were divided into groups based on the following strata: speciality, question template, the number of distractors (key-distractor combinations in the case of differential diagnoses questions), the number of stem entities, and predicted difficulty. Since the size of some groups in the population was relatively small, we selected the sample size for each group disproportionally to the size of that group in the population targeting an equal sample size from each group. This decision was made to ensure that we had enough questions from all groups in our sample. However, it is important to note that group population sizes were unequal and the size of some groups was smaller than the target sample size for these groups. To rectify this issue, we redistributed the extra slots among the other groups evenly. We then randomly selected the questions from the different groups.

Since our purpose of the sampling is to provide a proof-of-concept of the ability of the proposed method to generate high-quality questions of different types (that differ in templates, specialities, and stem size) and given the short supply of medical experts available, we decided to use disproportional stratified sampling. We were also interested in knowing whether some groups within the population are of high or low quality and how they compare with the other groups. For example, whether differential diagnosis questions are as useful as diagnostic questions or not. If a random sampling or proportional stratified sampling were used, most of the subgroups of interest would be less likely to appear in the sample[9] due to a large number of generated questions in some groups (see Table 2). Another reason behind our decision to use this sampling technique is that we are interested in features underlying question difficulty. Capturing as many possible combinations of features in our sample as possible will allow us later to investigate the feasibility of building a predictive model using machine learning techniques.

[9]For example, the number of generated diagnostic questions is 3,199,830 while the number of differential diagnosis questions 444. Considering random sampling, this means that the probability of picking a diagnostic question is approximately .94 (3,199,830/3,407,493) compared to 0.0001 (444/3,407,493) for a differential diagnosis question.

Each expert reviewed approximately 30 questions in their speciality, considering resident specialists or practising specialists as the target audience of the questions. Whenever possible, we collected two reviews per question except for the speciality of orthopaedics, due to the lack of a second reviewer.

A web-based system was developed to conduct the review. First, with no time constraint, the reviewer was asked to answer the displayed question and submit his/her answer. After the answer had been submitted, the reviewer was told whether he/she responded correctly or incorrectly and was provided with an explanation of the incorrectness of the selected option in the latter case. The reviewer was then asked to evaluate the appropriateness of the questions. If the reviewer rated the question as inappropriate due to one of the four options provided (see the Appendix "Survey Questions"), no additional survey questions appeared, and the reviewer could move to the next question. In cases where the reviewer rated the question as appropriate, he/she was asked to complete additional ratings about the difficulty of the question, the quality of the distractors, and the medical accuracy of the explanations provided. Reviewers were also asked to indicate whether the question contained *clustered distractors* or not. Clustered distractors are distractors that have a high degree of similarity to each other as a result of them exhibiting similar features so that once one of them is excluded, all the others can also be excluded (Kurdi et al. 2017). The survey questions are provided in the Appendix "Survey Questions". Each question was followed by an optional comment box in case the reviewer wanted to elaborate further. Comments provided by the reviewers were analysed by reading through them and extracting common and important themes.

Results and Discussion

We generated 3,407,493 questions using our approach as implemented by EMCQG. A breakdown by speciality and template can be found in Table 2.

Method Effectiveness

As an approximation, the number of base questions[10] with a single stem entity (questions belonging to template 2 and 3) is expected to be equal to the number of the ranked relations that are used for identifying the question key. This approximation is very rough because it assumes that the ontology contains concepts, other than the key, that satisfy distractor selection criteria (see Section "Question Templates" for an example). As the number of potential distractors increases, the number of variants of the base questions increases.

The results in Table 2 shows that for template 2, the number of questions exceeds the number of relations by a factor of 1.5 on average while for template 3, it exceeds the number of relations by a factor of 2.4 on average. It is important to note here that

[10] An intermediate representation of questions that composes of a stem, a key, and all possible distractors that satisfy distractor selection rules. Different questions can be assembled from base questions by combining different distractors.

Table 2 Number of questions per generated template

T	Specialty	#hCF	#hRF	#hDDx	#hD	#questions
1	Cardiology	12,767	347	NR	NR	11,264
	Gastroenterology	13,549	867	NR	NR	111,556
	Internal medicine	34,224	3,271	NR	NR	3,072,820
	Orthopedics	11,131	374	NR	NR	4,180
	All	71,671	4,859	NR	NR	3,199,830
2	Cardiology	NR	NR	NR	3,692	6,103
	Gastroenterology	NR	NR	NR	4,090	6,344
	Internal medicine	NR	NR	NR	9,092	11,137
	Orthopedics	NR	NR	NR	3,419	4,457
	All	NR	NR	NR	20,293	28,041
3	Cardiology	12,767	NR	NR	NR	35,615
	Gastroenterology	13,549	NR	NR	NR	29,496
	Internal medicine	34,224	NR	NR	NR	90,724
	Orthopedics	11,131	NR	NR	NR	23,343
	All	71,671	NR	NR	NR	179,178
4	Cardiology	12,767	347	95	NR	33
	Gastroenterology	13,549	867	431	NR	208
	Internal medicine	34,224	3,271	1,505	NR	203
	Orthopedics	11,131	374	211	NR	0
	All	71,671	4,859	2,242	NR	444

T: Template, hCF: hasClinicalFinding, hRF: has RiskFactor, hDDx: hasDifferentialDiagnosis, hD: hasDrug, NR: Not relevant for the template

the difference between the number of relations and the number of questions is similar across the four specialities which indicate that our method performed consistently.

With regards to questions belonging to template 1 (with multiple stem entities), it can be seen that as the number of relations increase, the number of questions increases significantly. This is expected since we can construct one question of size 1H|2S for a concept that is related to one risk factor and two symptoms compared to six distinct questions for a concept with two risk factors and three symptoms.

Finally, the number of questions belonging to template 4 is lower than the number of hDDx relations (the most important relation for this template). This is due to the low number of hDDx relations compared to other relations. The reason behind the low number of hDDx relations is that the relations themselves are experimental relations and they are still being developed. Additionally, unlike other templates, template 4 requires the keys to be connected to each distractor via hDDx relations. By inspecting the ontology, we found that the number of concepts that can be served as potential keys is much lower than the number of hDDx relations. For example, only 14 cardiology diseases have at least one risk factor, one symptom, and three

differential diagnoses which nominate them as keys for stems with one risk factor and one symptom. Considering stems with more risk factors and symptoms, the number of potential keys will decrease further.

Quality Assessment

A total of 435 questions were reviewed, of which 316 questions were reviewed by two reviewers, while 119 were reviewed by one reviewer (751 reviews in total).

Review Time

It is important to consider whether or not the time spent in reviewing automatically generated questions is less than the time usually spent on constructing these questions manually. Of the reviews, 58% were completed in less than two minutes, and 83% were completed in less than four minutes. With regards to the time spent in solving the questions, they were solved in less than one minute in 89% of the reviews and in less than two minutes in 97%. This indicates that reviewing questions takes much less time than is estimated for constructing MCQs manually (about 7 minutes to 1 hour per MCQ (Mitkov et al. 2006; Brady 2005; Bridge et al. 2007)).

Question Quality

To analyse question quality, we compared the number of questions found to be appropriate, by one or both domain experts, to the number of those that were not. We also compared the number of questions solved correctly to the number of questions solved incorrectly by experts. Numbers were broken down further by templates, stem size, and specialities for specific analyses as will be seen below. We used the number of questions that reviewers agreed/disagreed on and unweighted kappa statistics[11] to assess agreement between reviewers.

With regards to the appropriateness of the questions, 79% (345) of them were rated as appropriate by at least one reviewer. Figure 5 illustrates reviewers ratings of questions' appropriateness. Figure 5 further illustrates the agreement between reviewers. Although reviewers disagree on the appropriateness for 40% (127) of the questions, a high percentage of disagreement can be explained. Note that average Cohen's Kappa indicates more than chance agreement (details in the Appendix "Agreement Between Domain Experts").

By investigating the difficulty ratings of the questions causing disagreement, we found that 42.5% (54) of these questions were rated as easy, and 11% (14) as difficult by the reviewers who believed they were appropriate. We anticipated that these questions were found to be too easy or too difficult, and therefore inappropriate, by the other reviewers, which was suggested by some of the reviewers' comments.

[11] We used unweighted kappa since question appropriateness encompasses two categories (appropriate/inappropriate).

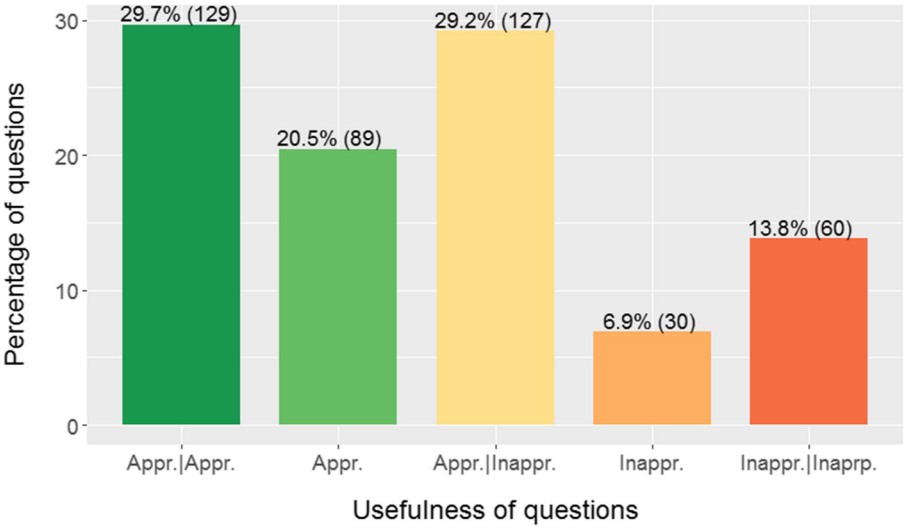

Fig. 5 Results of the evaluation of question appropriateness. Raw numbers are presented between parentheses. Appr.|Appr. = appropriate by two reviewers; Appr. and Appr.|Inappr. = appropriate by one reviewer; Inappr. = inappropriate by one reviewer; Inappr.|Inaprp. = inappropriate by two reviewers

To further understand the reasons behind reviewer disagreement, we inspected the reasons selected by reviewers who rated the questions as inappropriate. We found that 16 questions rated as inappropriate because they are guessable while 11 questions rated as inappropriate because they do not require medical knowledge. This explained around 22% (27 questions) of disagreement. Furthermore, 35% (45 questions) of the remaining disagreement came about when one reviewer thought the question was confusing, while the other thought it was appropriate. We attribute this to language issues in the questions. For example, the question $Q12$ presented below was rated as appropriate by one of the reviewer and as 'inappropriate/confusing' by the other. The reviewer who rated the question as confusing stated that "*patient cannot present with functional tricuspid regurgitation*" and suggested to add the string 'exam revealed' before the term *functional tricuspid regurgitation*. Questions like *Q12* are still useful since they require minor lexical changes to be considered appropriate.

What is the most likely diagnosis?

Q12: A patient with a history of alagille syndrome presents with fatigue and functional tricuspid regurgitation. What is the most likely diagnosis?

A. Hashimoto disease

B. hypertension

C. cardiac tamponade

D. pulmonary valve stenosis ◄ **Key**

Although the reviewers' comments suggest the need for more advanced language generation techniques, some of the linguistic issues can be traced back to the modelling of concepts in the ontology. For example, the syntactic issues in the stem: "... patient with a history of Family history: Sudden infant death (situation) presents with ..." can be fixed by introducing rules for rewriting the history component whenever a concept contains the string 'family history' is included. A better solution is introducing a class such as *genetic risk factor* and adding axioms such as: $suddenInfantDeath \sqsubseteq geneticRiskFactor$. Although the later solution still requires incorporating rules for writing the history components, the benefit of this approach is that more knowledge is added into the ontology and the rules are based on precise knowledge rather than on string matching. Other syntactic issues can simply be fixed by renaming concepts. For example, the risk factor *hospital patient* can be renamed to *being hospitalised* which will result in the stem "... patient with a history of being hospitalised ..." instead of the stem "... patient with a history of hospital patient ...". On the axioms level, the axiom: $beingHospitalised \sqsubseteq riskFactor$ (being hospitalised is a risk factor) reads better then the axiom: $hospitalPatient \sqsubseteq riskFactor$ (hospital patient is a risk factor).

Another main issue identified by reviewers was the need for more context about some of the stem entities, mostly the presenting symptoms, such as their location, duration, or description (i.e. colour, shape, or size). An illustrative example is the question *Q13*, where the reviewer recommended adding the location of one of the symptoms: "*the question stem needs to better identify where the heaviness sensation is - it is confusing as written*". This information is not contextualised in the current version of EMMeT since the current application of EMMeT does not require this level of specificity. Enriching EMMeT with this kind of specific information and investigating whether it leads to improvement in the quality of questions are areas for future work.

What is the most likely diagnosis?

Q13: A patient with a history of smoking presents with heaviness sensation and pruritus. What is the most likely diagnosis?

A. angioedema and/or urticaria

B. end stage renal disease

C. dermatomyositis

D. jaundice

E. stasis dermatitis ◄ **Key**

One of our main interests was the quality of questions with multiple stem entities. Questions with multiple stem entities performed as well as questions with a single stem entity. Figure 6 shows that the distribution of appropriate questions is similar across all templates. For example, 51.7% (163) of *What is the most likely diagnosis* questions were rated as appropriate, compared to 50% (23) of *What is the drug of choice* questions.

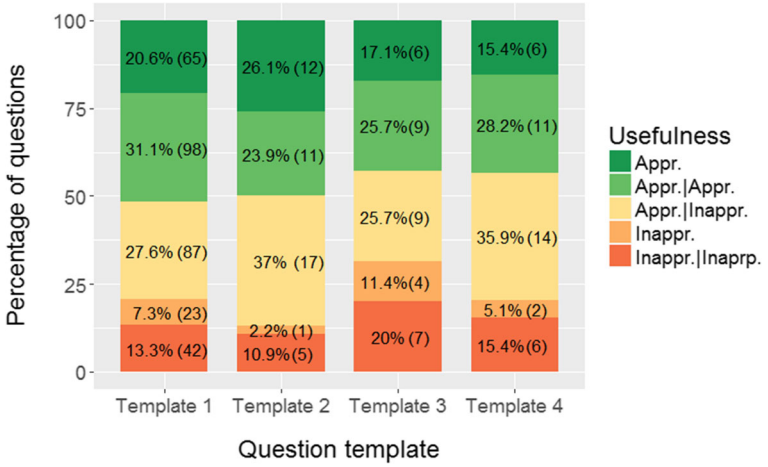

Fig. 6 Performance of different templates generated by the system. Template 1 = "What is the most likely diagnosis?", Template 2 = "What is the drug of choice?", Template 3 = "What is the most likely clinical finding?", Template 4 = "What is the differential diagnosis?"

We were also interested to know whether the quality of multi-term questions was related to the number of stem entities or not, for example, whether diagnostic questions with one history and one symptom are as useful as questions with more history and symptom elements. Although we suspected questions with one history and one symptom to be vague compared to questions with a higher number of histories and symptoms, this was not the case. We broke down the number of appropriate/inappropriate questions by stem size but could not find any association between the two variables.

With regards to the relation between appropriateness and speciality,[12] internal medicine and cardiology questions outperformed gastroenterology questions. We attribute this to intrinsic differences to the nature of the specialities. It is easier to view the signs and symptoms of diseases belonging to cardiology and internal medicine whereas symptoms of gastrointestinal diseases are vague since they are internal and the patient cannot pinpoint the exact cause and the diagnostic symptoms are normally defined by images/biopsies. Another possible cause of the difference in quality is the richness of the specialities in the knowledge base. We found that internal medicine and cardiology are richer than gastroenterology regarding the number of concepts. In addition, laboratory findings such as histology and image results which benefit gastroenterology are not fully covered in EMMeT compared to symptoms which benefit internal medicine and cardiology.

Another indicator of quality issues involves experts' performance on questions. Questions solved incorrectly by reviewers are of interest due to the fact that they could possibly point to flaws in the ontology or the generation process, under the expectation that reviewers should have the required knowledge to answer the

[12] We excluded orthopaedics questions due to the lack of a second reviewer, see "Quality Assessment"

questions correctly. An example of a question solved incorrectly due to incorrect knowledge in the ontology (*Q14*) states:

What is the most likely diagnosis?

Q14: A patient with a history of increased glomerular filtration rate presents with fatigue and blurred vision. What is the most likely diagnosis?

A. cardiac tamponade

B. diabetes mellitus

C. stroke

D. primary pulmonary hypertension

E. diabetic neuropathy ◄ **Key**

The reviewer pointed out that "*diabetics have decreased glomerular filtration rate, not increased*" which made the question confusing. Questions solved incorrectly by reviewers may also not be subject to such issues, but are instead very difficult, which raises a question about their appropriateness for assessing the knowledge of the intended cohort. Another example question *Q15*, which is above the level of the targeted exam audience is:

What is the most likely diagnosis?

Q15: A male patient with a history of [taking][a] azacitidine presents with hepatomegaly and malaise. What is the most likely diagnosis?

A. carcinoid tumor

B. amebiasis

C. hepatitis C

D. fatty liver ◄ **Key**

[a] A grammatical correction that is manually added to the question.

One of the reviewers stated that "*azacitidine is not a common drug that a medical resident would know with regard to side effects*".

Overall, reviewer(s) solved 78.8% (343) of the questions correctly while 19.3% (84) were solved incorrectly (see Table 3 for details). Among the questions solved incorrectly, 76.19% (63) were rated as inappropriate by at least one reviewer. Of the questions solved incorrectly and rated as inappropriate, 59% (38 questions) were confusing according to the reviewers which is mainly attributed to the linguistic issues discussed before.

Another category of interest here is questions solved incorrectly but rated as appropriate by the reviewers. We expect that the reviewers made mistakes in solving these questions but rated them as appropriate because they agreed with the answers. Questions in this category were 31% of the questions solved incorrectly. Among these, 42% (11 questions) were rated as difficult by at least one reviewer.

Table 3 Statistics about correctness of answers given by domain experts

	Correct	Partially correct	Incorrect	None	Total
Correct	41.1%	3.9%	14%	19.8%	78.8%
	(179)	(17)	(61)	(86)	(343)
Partially correct		1.4%	0	0.5%	1.9%
		(6)	(0)	(2)	(8)
Incorrect			12.2%	7.1%	19.3%
			(53)	(31)	(84)

Raw numbers are presented between parentheses. None indicates that the questions were reviewed by one reviewer

The questions with the highest percentage of correct responses[13] were the *What is the differential diagnosis* questions and the *What is the most likely diagnosis* questions (84.6% and 83.2% respectively). This again highlights that multi-term questions are sensible. Surprisingly, the percentage of correct responses to the *What is the most likely clinical finding* questions was relatively low (51.5%). A possible interpretation is that these questions consist of one stem entity and therefore the number of hints they provide is limited compared to questions with multiple stem entities.

Distractor Quality

The analysis of distractor quality is built around the number of distractors within each quality category (i.e. not plausible, plausible, difficult to eliminate, or cannot eliminate) while considering reviewer agreement (the number of cases showing agreement/disagreement and unweighted kappa statistics) and the relation between assessment of distractor quality and explanation correctness. We also analysed the number of questions with clustered distractors.

An important component of MCQs is their distractors. We define appropriate distractors as those rated as plausible (regardless of them being easy to eliminate or difficult to eliminate). This category accounted for 73% (859) of distractors, as rated by at least one reviewer, as can be seen in Fig. 7. Among these, 80.2% (689) were rated as easy, 13% (112) were rated as difficult, and 6.7% (58) were rated as easy by one reviewer and difficult by the other. Having more easy distractors has also been the case in Alsubait et al. (2014) who attributed this to the rarity of distractors with a very high similarity to the key in ontologies.

On the other hand, inappropriate distractors are those rated as not plausible or unable be eliminated. Implausible distractors accounted for 89.8% of all inappropriate distractors (among those, 6% (27) were selected by reviewers when answering the questions) while a low percentage of distractors (9.2%) were inappropriate due them being equally as correct as the key. To find whether or not distractor inappropriateness results from errors in the ontology, we looked at the correctness of their explanations.

[13] Cases where one of the reviewers solved the question correctly were considered as correct.

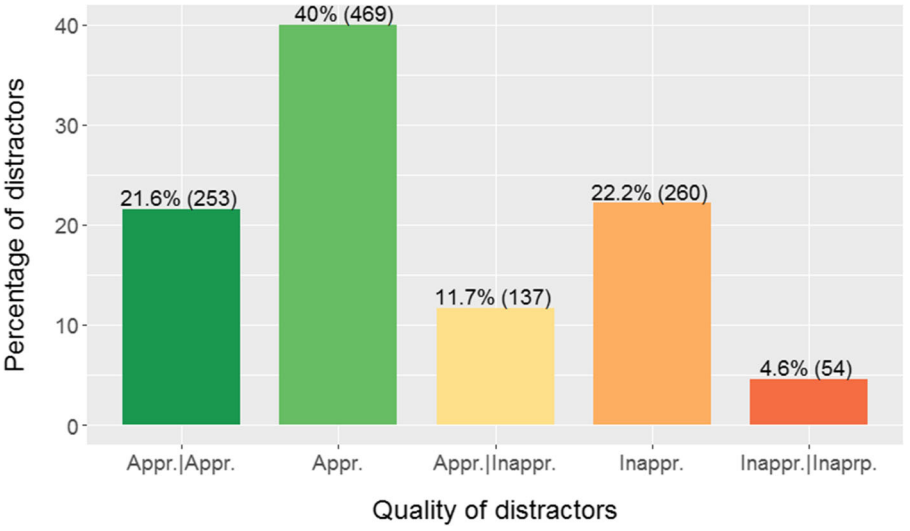

Fig. 7 Results of evaluating question distractors. Raw numbers are presented in parentheses

Of the distractors rated as inappropriate by at least one reviewer, 22.4% (101 distractors) had incorrect explanations according to at least one reviewer. Another reason for distractor inappropriateness was incompatibility with the patient demographics (31%). This is due to the unavailability of demographic restrictions for these distractors. Once additional POC relations are added to the ontology, such distractors will be eliminated.

Regarding agreement on distractor appropriateness, we distinguish between strong and weak disagreement as below:

- Strong: including two cases:

  NP|D One of the reviewers rated a distractor as not plausible (NP) while the other rated it as difficult (D);

  E|K One of the reviewers rated a distractor as easy (E) while the other rated it as a key (K).

- Weak: including two cases:

  NP|E One of the reviewers rated a distractor as not plausible while the other rated it as easy;

  D|K One of the reviewers rated a distractor as difficult while the other rated it as a key.

Overall, reviewers agreed on 69% (307) of cases (Fig. 7). The percentage for weak disagreement was 88.3% (121) and that for strong disagreement was 11.7% (16). Of the distractors causing disagreement between reviewers, 24.1% (33) have incorrect explanations according to one reviewer at least.

Finally, 4% of the questions suffer from clustering, as indicated by at least one reviewer. This is a low percentage considering that a previous evaluation we conducted (Kurdi et al. 2017) had identified clustering as a prevalent issue in automatically generated questions from ontologies. We speculate that this low percentage of clustering is a result of a restriction we imposed on distractor selection. The restriction avoids generating questions with distractors that have sub/superclass relations between them since these distractors are likely to exhibit clustering (due to the shared features between the subclass and the superclass).

Methodological Reflection

For practical reasons, we had to restrict our analysis to four specialities which were not randomly selected. Although other specialities might be less mature, which in turn will affect the number and the quality of generated questions, we believe that given specialities that have the same size (regarding the number of concept and relations) and shape as the selected specialities will yield similar results.

For the purpose of this paper, we adopt an expert-centred study to evaluate generated questions. While expert approval is the level of criteria for acceptance of hand-written questions, we are aware that it provides preliminary evidence for the exam-readiness of questions. To get further evidence, the questions need to be administered to a sample of students and their properties (empirical difficulty, discrimination, and reliability) need to be analysed. However, expert review is a necessary prior step to filter invalid questions (questions that are ambiguous, guessable, or do not require medical knowledge).

A source of possible bias in the expert review is paying experts to evaluate 30 questions each. Since inappropriate questions require less review time than appropriate questions,[14] reviewers could be biased to rate questions as inappropriate to minimize review time. But even with such a bias in mind, the results suggest that we were successful at generating case-based questions.

Another biasing factor is requiring experts to solve the question and showing them the key before they rate the appropriateness of the questions and the distractors. Solving a question incorrectly could bias expert judgement on quality. To find out whether this was the case or not, we ran two analyses of the correlation between: (1) expert performance (i.e. whether they got a question right or not) and their rating of question appropriateness (i.e. whether they rate a question as appropriate or not) and (2) distractor selection (i.e. whether they select a distractor or not) and their rating of distractor appropriateness (i.e. whether they rate a distractor as appropriate or not). The Spearman's coefficient is .30 (p-value = 0) for the correlation between performance and question appropriateness and −.02 (p-value = .40) for the correlation between distractor selection and appropriateness. This indicates that expert judgements were

[14] Reviewers are required to evaluate more aspects of questions they rate as appropriate (see "Quality Assessment").

not systematically biased. A possible adaptation of the experimental protocol requires experts to evaluate question quality before solving the questions, showing them the key, then allowing them to edit their rating of appropriateness while keeping track of the changes. This will allow the discovery of systematic biases if they existed or finding out which part of the question is believed to be problematic.

Although the number of questions reviewed by experts was larger than any sample used in other experiments (Papasalouros et al. 2008; Al-Yahya 2014; Alsubait et al. 2014; Wang et al. 2007; Khodeir et al. 2014; Karamanis et al. 2006; Gierl and Lai 2013), using stratified sampling results in having a small number of questions in multiple groups. Therefore, these results should be dealt with as preliminary rather than confirmatory. Further experiments are needed to strengthen our confidence in the results. Once this has been done, probability weights can be used to adjust the sample distribution to match the population distribution (i.e. distribution of all generated questions), which in turn will allow making claims about the whole population.

Conclusions and Future Work

We have presented the design, implementation, and evaluation of a new ontology-based approach for generating MCQs. What distinguishes our approach from previous work is its ability to generate case-based questions which require more than recall of information to be solved. These forms of questions are a valuable addition to the existing forms as their structure is a move toward a more sophisticated structure (i.e., multi-term) when compared to the simple structure (at most two terms) of questions generated by other current approaches. We also believe that this approach could be applied to other kinds of diagnostic questions outside of the medical domain provided that suitable knowledge bases are available.

Unlike other studies which use hand-crafted ontologies for question generation (Alsubait et al. 2014), we demonstrate the feasibility of our approach using a pre-existing ontology.[15] The results are promising and suggest that, given appropriate ontologies, our approach can generate four types of medical case-based questions successfully. Our approach is also less expensive than existing approaches for generating medical questions as it does not involve reliance on domain experts (apart from revision) or using both ontologies and text. Also, evaluating the quality of the generated questions highlighted different areas where the ontology can be enriched. This suggests that these questions can be used, in addition to their role as an assessment tool, as a modelling and validation-assistant tool.

As a next step, we plan to administer the generated questions to a student cohort and collect statistical characteristics of the questions such as difficulty and discrimination. These statistics will provide further evidence of question quality and allow us to validate our difficulty model.

[15] Although it could be argued the EMMeT-OWL is a *hand-crafted* ontology, it was a direct translation of an existing SKOS knowledge base into an OWL ontology with minimal intervention due to the close relation between SKOS and OWL.

Acknowledgements This work was funded by an EPSRC grant (ref: EP/ P511250/1) under an Institutional Sponsorship (2016) for The University of Manchester, along with a partial contribution from Elsevier. The funding acts as a secondment to an initial EPSRC grant (ref: EP/K503782/1) awarded as an Impact Acceleration Account (2016) for The University of Manchester.

Appendix

Survey Questions

How would you rate the usefulness of the question?

- Appropriate: The question is appropriate as a Board exam question; the level of knowledge required to answer the question is that of a resident specialist or practicing specialist.
- Inappropriate/no medical knowledge needed: Can be answered correctly by people having little to no medical knowledge, (far) below the level of targeted exam audience.
- Inappropriate/guessable: The correct answer is guessable based on syntactic clues. For example, similar words between the stem and the key can clue examinees to the correct answer.
- Inappropriate/confusing: The syntax or terminology is not intelligible and/or the key does not logically follow from the question stem.
- Inappropriate/other: The question is inappropriate for other reasons.

How would you rate the difficulty of the question?

- Easy: More than 70% of examinees would be expected to answer the question correctly
- Medium: 30% to 70% of examinees would be expected to answer the question correctly
- Difficult: Less than 30% of examinees would be expected to answer the question correctly

How would you rate the quality of the MCQ distractors? (reviewers answered this question for each distractor)

- Not plausible: will not selected by any examinees.
- Plausible, but easy to eliminate: examinees with minimum amount of knowledge will be able to eliminate this distractor.
- Difficult to eliminate: Only examinees with sufficient amount of knowledge will be able to eliminate this distractor.
- Cannot eliminate: The correctness of this distractor is equal to the correctness of the key.

How would you rate the medical accuracy of the explanations? (reviewers answered this question for each explanations)

- Correct: the explanation provided for the correctness or incorrectness of the option is accurate.
- Incorrect: the explanation provided for the correctness or incorrectness of the option is inaccurate.

Does the question contain clustered distractors?

- Yes: the question contains incorrect options that are very similar to each other and once one of them is excluded, all the other can be excluded. For example, the correct answer is a heart disease while all other options are lung disease. Once examinees exclude any disease related to the lung, they can exclude all the incorrect options at once.
- No
- Don't know

Demographic Characteristics of Domain Experts

Table 4 Demographic characteristics of domain experts

Demographic characteristics	Categories	Number
Speciality	Internal medicine	5
	Gastroenterology	4
	Cardiology	5
	Orthopedics	1
Level	Resident	1
	Generalist	7
	Specialist	7
Experience as a practitioner	None	2
	Less than 1 year	0
	1–3 years	4
	3–6 years	3
	More than 6 years	6
Teaching experience	None	0
	Less than 1 year	1
	1–3 years	6
	3–6 years	3
	More than 6 years	5
Exam construction experience	None	4
	Less than 1 year	6
	1–3 years	2
	3–6 years	1
	More than 6 years	2

Agreement Between Domain Experts

The following tables provides information about agreement between domain experts. Kappa values were interpreted according to Viera and Garrett's guideline (Viera et al. 2005).

Table 5 Agreement between pairs of reviewers on questions appropriateness

Experts	N	Kappa	Interpretation
Internal medicine			
i2 and i3	32	.28	Fair agreement
i2 and i4	20	.29	Fair agreement
i3 and i5	28	.08	Slight agreement
i4 and i5	27	-.30	Less than chance agreement
Gastroenterology			
g1 and g2	28	.13	Slight agreement
g1 and g3	44	.20	Slight agreement
g2 and g4	29	-.11	Less than chance agreement
Cardiology			
c1 and c2	41	.38	Fair agreement
c3 and c4	46	.28	Fair agreement
Average		.13	

Table 6 Agreement between pairs of reviewers on distractors appropriateness

Experts	N	Kappa	Interpretation
Internal medicine			
i2 and i3	49	.19	Slight agreement
i2 and i4	29	.20	Slight agreement
i3 and i5	32	-.06	Less than chance agreement
i4 and i5	50	.23	Fair agreement
Gastroenterology			
g1 and g2	12	.00	Chance agreement
g1 and g3	67	.53	Moderate agreement
g2 and g4	25	.20	Slight agreement
Cardiology			
c1 and c2	104	.09	Slight agreement
c3 and c4	67	.41	Moderate agreement
Average		.20	

References

Abdalla, M.E., Gaffar, A.M., Suliman, R.A. (2011). Constructing A-type multiple choice questions (MCQs): step by step manual. *Blueprints in Health Profession Education Series*.

Al-Yahya, M. (2014). Ontology-based multiple choice question generation. *The Scientific World Journal*. https://doi.org/10.1155/2014/274949.

Alsubait, T. (2015). Ontology-based question generation. PhD thesis, University of Manchester.

Alsubait, T., Parsia, B., Sattler, U. (2014). Generating multiple choice questions from ontologies: lessons learnt. In *OWLED* (pp. 73–84).

Biggs, J.B., & Collis, K.F. (2014). *Evaluating the quality of learning: the SOLO taxonomy (structure of the observed learning outcome)*. New York: Academic Press.

Bloom, B.S., Engelhart, M.D., Furst, E.J., Hill, W.H., Krathwohl, D.R. (1956). *Taxonomy of educational objectives, handbook I: the cognitive domain* (Vol. 19). New York: David McKay Co Inc.

Bodenreider, O. (2004). The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, *32*(Database-Issue), 267–270. https://doi.org/10.1093/nar/gkh061.

Brady, A.-M. (2005). Assessment of learning with multiple-choice questions. *Nurse Education in Practice*, *5*(4), 238–242. http://www.sciencedirect.com/science/article/pii/S1471595305000065.

Breithaupt, K., Ariel, A.A., Hare, D.R. (2010). *Assembling an inventory of multistage adaptive testing systems* (pp. 247–266). New York: Springer.

Bridge, P., Appleyard, R., Wilson, R. (2007). Automated multiple-choice testing for summative assessment: what do students think? In *The international educational technology (IETC) conference*.

Carroll, R.G. (1993). Evaluation of vignette-type examination items for testing medical physiology. *Advances in Physiology Education*, *264*(6), S11. PMID: 8328552. https://doi.org/10.1152/advances.1993.264.6.S11.

Coderre, S., Mandin, H., Harasym, P.H., Fick, G.H. (2003). Diagnostic reasoning strategies and diagnostic success. *Medical Education*, *37*(8), 695–703.

Converse, L., Barrett, K., Rich, E., Reschovsky, J. (2015). Methods of observing variations in physicians' decisions: The opportunities of clinical vignettes. *Journal of General Internal Medicine*, *30*(3), 586–594. https://doi.org/10.1007/s11606-015-3365-8.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Belmont: Wadsworth Publishing.

Cubric, M., & Tosic, M. (2011). Towards automatic generation of e-assessment using semantic web technologies. *International Journal of e-Assessment*, *1*(1).

Cunnington, J.P.W., Norman, G.R., Blake, J.M., Dauphinee, W.D., Blackmore, D.E. (1997). Applying learning taxonomies to test items: is a fact an artifact?. In A. J. J. A. Scherpbier, C. P. M. van der Vleuten, J. J. Rethans, A. F. W. van der Steeg (Eds.) *Advances in medical education* (pp. 139–142). Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-011-4886-3_40.

Ellampallil, V.V., & Kumar, P. (2017). Automated generation of assessment tests from domain ontologies. *Semantic Web*, *8*(6), 1023–1047. https://content.iospress.com/articles/semantic-web/sw252.

Elstein, A.S., & Schwarz, A. (2002). Clinical problem solving and diagnostic decision making: selective review of the cognitive literature. *BMJ: British Medical Journal*, *324*(7339), 729–732.

Freiwald, T., Salimi, M., Khaljani, E., Harendza, S. (2014). Pattern recognition as a concept for multiple-choice questions in a national licensing exam. *BMC Medical Education*, *14*(1), 232. https://doi.org/10.1186/1472-6920-14-232.

Gierl, M.J., & Lai, H. (2013). Evaluating the quality of medical multiple-choice items created with automated processes. *Medical Education*, *47*(7), 726–733. https://doi.org/10.1111/medu.12202.

Gierl, M.J., Lai, H., Turner, S.R. (2012). Using automatic item generation to create multiple-choice test items. *Medical Education*, *46*(8), 757–765. https://doi.org/10.1111/j.1365-2923.2012.04289.x.

Guardia, G.D., Vêncio, R.Z., de Farias, C.R. (2012). A uml profile for the obo relation ontology. *BMC Genomics*, *13*(5). https://doi.org/10.1186/1471-2164-13-S5-S3.

Haladyna, T.M., Downing, S.M., Rodriguez, M.C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education*, *15*(3), 309–333. https://doi.org/10.1207/S15324818AME1503_5.

Jelenković, F., & Tošić, M. (2013). Semantic multiple-choice question generation and concept-based assessment. In *The first international conference on teaching english for specific purposes*.

Karamanis, N., Ha, L.A., Mitkov, R. (2006). Generating multiple-choice test items from medical text: a pilot study. In *Proceedings of the fourth international natural language generation conference* (pp. 111–113). Association for Computational Linguistics.

Khodeir, N., Wanas, N., Darwish, N., Hegazy, N. (2014). Bayesian based adaptive question generation technique. *Journal of Electrical Systems and Information Technology*, *1*(1), 10–16. http://www.sciencedirect.com/science/article/pii/S2314717214000087.

Kurdi, G., Parsia, B., Sattler, U. (2017). *An experimental evaluation of automatically generated multiple choice questions from ontologies*, (pp. 24–39). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-54627-8_3.

Lu, Y., & Lynch, J. (2017). Are clinical vignette questions harder than traditional questions in gross anatomy course?. *Medical Science Educator*, *27*(4), 723–728. https://doi.org/10.1007/s40670-017-0473-6.

Miles, A., & Bechhofer, S. (2009). SKOS simple knowledge organization system reference, available at http://www.w3.org/TR/skos-reference/, retrieved August 4, 2015.

Miller, G.A., Beckwith, R., Fellbaum, C., Gross, D., Miller, K.J. (1990). Introduction to wordnet: an on-line lexical database. *International Journal of Lexicography*, *3*(4), 235–244.

Mitkov, R., Le An, H., Karamanis, N. (2006). A computer-aided environment for generating multiple-choice test items. *Natural Language Engineering*, *12*(2), 177–194.

Motik, B., Patel-Schneider, P.F., Parsia, B., Bock, C., Fokoue, A., Haase, P., Hoekstra, R., Horrocks, I., Ruttenberg, A., Sattler, U., Smith, M. (2009). Owl 2 web ontology language: structural specification and functional-style syntax, W3C.

NBME (2017). Subject examinations: content outlines and sample items. https://www.nbme.org/pdf/SubjectExams/SE_ContentOutlineandSampleItems.pdf.

World Health Organization. (1992). *The ICD-10 classification of mental and behavioural disorders: clinical descriptions and diagnostic guidelines* (Vol. 1). Geneva: World Health Organization.

Papasalouros, A., Kanaris, K., Kotis, K. (2008). Automatic generation of multiple choice questions from domain ontologies. In *IADIS international conference e-learning* (pp. 427–434).

Parsia, B., Alsubait, T., Leo, J., Malaisé, V., Forge, S., Gregory, M.L., Allen, A. (2015). Lifting emmet to OWL getting the most from SKOS. In V. A. M. Tamma, M. Dragoni, R. Gonçalves, A. Lawrynowicz (Eds.) *Ontology engineering - 12th international experiences and directions workshop on OWL, OWLED 2015, co-located with ISWC 2015, Bethlehem, PA, USA, October 9–10, 2015, Revised Selected Papers, vol. 9557 of Lecture Notes in Computer Science* (pp. 69–80). Springer. https://doi.org/10.1007/978-3-319-33245-1_7.

Peabody, J.W., Luck, J., Glassman, P., Dresselhaus, T.R., Lee, M. (2000). Comparison of vignettes, standardized patients, and chart abstraction: a prospective validation study of 3 methods for measuring quality. *JAMA*, *283*(13), 1715–1722. https://doi.org/10.1001/jama.283.13.1715.

Rutten, G.M., Harting, J., Rutten, S.T., Bekkering, G.E., Kremers, S.P. (2006). Measuring physiotherapists' guideline adherence by means of clinical vignettes: a validation study. *Journal of Evaluation in Clinical Practice*, *12*(5), 491–500. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2753.2006.00699.x.

Schuwirth, L.W.T., Verheggen, M.M., Van Der Vleuten, C.P.M., Boshuizen, H.P.A., Dinant, G.J. (2001). Do short cases elicit different thinking processes than factual knowledge questions do? *Medical Education*, *35*(4), 348–356. https://doi.org/10.1046/j.1365-2923.2001.00771.x.

Spackman, K.A., Campbell, K.E., Côté, R.A. (1997). Snomed rt: a reference terminology for health care. In *AMIA 1997, American medical informatics association annual symposium, Nashville, TN, USA, October 25–29, 1997/ AMIA*. http://knowledge.amia.org/amia-55142-a1997a-1.585351/t-001-1.587519/f-001-1.587520/a-127-1.587635/a-128-1.587632.

Tarrant, M., Knierim, A., Hayes, S.K., Ware, J. (2006). The frequency of item writing flaws in multiple-choice questions used in high stakes nursing assessments. *Nurse Education Today*, *26*(8), 662–671. Proceedings from the 1st nurse education international conference. http://www.sciencedirect.com/science/article/pii/S0260691706001067.

Tractenberg, R.E., Gushta, M.M., Mulroney, S.E., Weissinger, P.A. (2013). Multiple choice questions can be designed or revised to challenge learners' critical thinking. *Advances in Health Sciences Education*, *18*(5), 945–961.

Veloski, J., Tai, S., Evans, A.S., Nash, D.B. (2005). Clinical vignette-based surveys: a tool for assessing physician practice variation. *American Journal of Medical Quality*, *20*(3), 151–157. https://doi.org/10.1177/1062860605274520.

Viera, A.J., Garrett, J.M., et al. (2005). Understanding interobserver agreement: the kappa statistic. *Family Medicine*, *37*(5), 360–363.

Žitko, B., Stankov, S., Rosić, M., Grubišić, A. (2009). Dynamic test generation over ontology-based knowledge representation in authoring shell. *Expert Systems with Applications*, *36*(4), 8185–8196. http://www.sciencedirect.com/science/article/pii/S0957417408007392.

Wang, W., Hao, T., Liu, W. (2007). Automatic question generation for learning evaluation in medicine. In *International conference on web-based learning* (Vol. 4823, pp. 242–251). Berlin: Springer.

Webb, N.L. (1997). Criteria for alignment of expectations and assessments in mathematics and science education. Council of Chief State School Officers. Washington, DC.