ELSEVIER

Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.

Computers in Biology and Medicine 73 (2016) 71–93

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/cbm

Data-mining to build a knowledge representation store for clinical decision support. Studies on curation and validation based on machine performance in multiple choice medical licensing examinations

Barry Robson [a,b,c,*,1], Srinidhi Boray [a,b]

[a] Ingine Inc., DE, USA
[b] The Dirac Foundation clg, Oxfordshire, UK
[c] St. Matthew's University School of Medicine, Cayman Islands

ARTICLE INFO

Article history:
Received 18 November 2015
Received in revised form
5 February 2016
Accepted 17 February 2016

Keywords:
Clinical decision support
Data mining
Knowledge representation
Natural language processing
Web surfing
Probability
Relevance
Medical licensing examinations
USMLE

ABSTRACT

Extracting medical knowledge by structured data mining of many medical records and from unstructured data mining of natural language source text on the Internet will become increasingly important for clinical decision support. Output from these sources can be transformed into large numbers of elements of knowledge in a Knowledge Representation Store (KRS), here using the notation and to some extent the algebraic principles of the Q-UEL Web-based universal exchange and inference language described previously, rooted in Dirac notation from quantum mechanics and linguistic theory. In a KRS, semantic structures or statements about the world of interest to medicine are analogous to natural language sentences seen as formed from noun phrases separated by verbs, prepositions and other descriptions of relationships. A convenient method of testing and better curating these elements of knowledge is by having the computer use them to take the test of a multiple choice medical licensing examination. It is a venture which perhaps tells us almost as much about the reasoning of students and examiners as it does about the requirements for Artificial Intelligence as employed in clinical decision making. It emphasizes the role of context and of contextual probabilities as opposed to the more familiar intrinsic probabilities, and of a preliminary form of logic that we call presyllogistic reasoning.

© 2016 Elsevier Ltd. All rights reserved.

1. Introduction

1.1. Medical knowledge in computer systems

The growth of the ability of computers to capture and use clinical and biomedical knowledge may represent an important transition in human history [1]. In particular, the wealth of data and knowledge on the Internet and its World Wide Web should lead to improved clinical decision support (CDS) by computer systems, i.e. to improved clinical decision support systems CDSS [1]. Prior to the growth of the Internet, software with similar goals, such as that of the pioneering Stanford MYCIN project [2] did, of course, exist, and it is notable that right from the outset, most such systems developed for medicine were seen as needing to consider probabilistic measures, such as degrees of certainty, to be associated with statements of clinical or biomedical knowledge [1,2]. However, these were *Expert Systems* that obtained their knowledge offline by useful statements about the world inputted with associated probabilities estimated by *human experts*, often seen as requiring a specialist human *knowledge engineer* to act as mediator, and overall representing a very time-consuming process [1,2]. We recently introduced a CDS application called MARPLE [3]. MARPLE stands for *Medical Automated Reasoning Programming Language Environment*. A common theme of work of this kind is that it involves a repository of knowledge in a form that computers can more readily use. Such a repository for prediction and decision making is said to be a *knowledge representation store* (KRS). Any kind of KRS is a set of syntactic and semantic conventions that describe things and relationships. Any specific example from such a store is a knowledge element or *KRS element*, that early Expert System designers might describe as a kind of *frame* [1,2]. MARPLE rests on a considerable body of previous work by ourselves, collaborators, and other workers, and these efforts close to its essential features are first reviewed here (this Section 1). We also introduce the new version, MARPLE 2, which has significant advantages in helping ensure the quality of the above knowledge, an important consideration as follows.

[*] Corresponding author.
E-mail address: robsonb@aol.com (B. Robson).
[1] Tel.: +1 345 928 7242; fax: +1 345 945 3130.
www.ingine.com, www.diractfoundation.org.

http://dx.doi.org/10.1016/j.compbiomed.2016.02.010
0010-4825/© 2016 Elsevier Ltd. All rights reserved.

1.2. The impact of data mining

Many matters discussed in this paper are not yet widely seen as significant pressing problems for the current CDSS industry, because the types of decision support that are currently most widely used are still largely limited to alerts, reminders, and tools designed to ease workflow or enhance cognition [4]. However, studies like MARPLE are timely. There has been an escalating interest in "Big Data" and rapid progress in *data mining* of it [1], including of electronic health records [1,5]. The input data being "mined" is usually conveniently classified as of two types (a) *structured data* as in spreadsheets but also including *relatively structured data* as in electronic health records (EHRs) or similar public health sources, and (b) so-called *unstructured data*, in the present case simply meaning that it is mainly represented by natural language text (NLT) on web pages and other medical text accessible in digital form. MARPLE gets some of its knowledge to use, test, or further curate if necessary from offline digital repositories of data and other knowledge collections [1,4,5], and from the automatic "surfing" of the Internet, particularly to extract knowledge from NLT as described in this paper. Because of the escalating quantity of information obtained by data mining, and with future applications to CDS in a real clinical setting in mind, the quality of it as usable and authentic knowledge is of concern. MARPLE draws on both structured and unstructured data not least because each has well known strengths and weaknesses. Notably, while structured data mining can efficiently provide probabilities to certain important kinds of knowledge, mining natural language text on the Internet usually faces the problem that prevalence may reflect matters of interest and newsworthiness in inverse relation to actual frequency of occurrences in the "real world". Combining elements of knowledge from various sources and exploring means of overcoming the above kind of probability problem are by no means unique to our efforts (e.g. Ref. [6]). Nonetheless, the escalation of collected knowledge makes it difficult to keep up with ensuring its quality.

1.3. Curation

Considerable focus in this present paper is placed on methods of ensuring good *provenance* of KRS elements, which is essentially a matter of demonstrating and ensuring the above quality. Especially with applications to CDS in real clinical settings in mind, knowledge elements should come as much as possible from good sources and adequately represent the originally intended information, but in a form usable by computers. In our definition, all aspects of this including tidying, correction, and even rejection if necessary, represent *curation*, briefly reviewed in Section 1.8. The MARPLE project is primarily a study to develop better methods for curation of KRS elements for CDS. In addition to structured and unstructured data mining, MARPLE also gets some of its knowledge from human experts just as was the case in Expert Systems like MYCIN [2] and as persists for many CDSS today [1,5]. However, with MARPLE, the role of the human expert has primarily become one of *auditor*, and of *curation not creation* of KRS elements already obtained automatically by data mining. Compared with early Expert Systems, the human role is now more responsive rather than proactive. This is important because having human experts provide knowledge in good form is long known to be time-consuming [1,2], and it is easier to automate a task of this well-defined nature.

1.4. Using medical licensing examinations

The prominent, unusual, and perhaps controversial feature of MARPLE [3] is that, as one of its tests of quality of knowledge, it attempts the kind of multiple choice examinations given to medical students as a major part of the process of satisfying medical licensing boards. These tests are undertaken by medical students to obtain a license to practice medicine, and as practice and self-assessment tests in preparation for taking these (see Section 1.9). Only secondarily is our project an investigation of how the same algorithms might be applicable to CDS, though this is potentially an important spin-off. As with real students, along with formally receiving knowledge, practicing these exams is part of the learning process. Similarly, MARPLE is told the official "correct answer", but only after it has made an attempt to answer the question, which contributes the official final exam score when we present results as if it were an exam taken by a medical student. Curation of the KRS to satisfy the criterion of good exam performance and learning from the examinations (as well as "learning" by receiving knowledge from data mining) become essentially the same thing. The level of automation of curation is already fairly high. There is only human intervention into the KRS when MARPLE persistently fails to answer an exam questions correctly. Before that happens, MARPLE takes considerable effort to seek out the knowledge required to answer the exam question, without human intervention.[2] By inspecting the question and most importantly the candidate answers, MARPLE queries the Internet and, having extracted knowledge from one web page, it explores more deeply by searching in turn on links found, including any found in a list of scientific references.

1.5. Further purposes of the present paper

Having a computer tackle medical school exams might seem likely to require fairly advanced techniques in Artificial Intelligence (AI). However, our preliminary studies [3] suggested that while such exams obviously test some important qualitative aspects of captured knowledge, they represent a rather restricted and "artificially crisp" world, to allow the student every chance to verify his or her knowledge. Our study is allowing us to comment of features that are of educational interest, as well a dissecting out some issues that do suggest some useful tools for CDS that are at least of the flavor of AI. For example, such exams still clearly test knowledge, but usually the reasoning with that knowledge almost always only requires *pre-syllogistic logic*, a term that we introduce in more detail in Section 5.1 but provide the theoretical basis in Sections 2.3 and 2.4. Apart from calculation questions that are excluded in the present study, very few exceptions to that have been found. Whereas more complete logic will be important for applications like CDS, it is a powerful pre-filter. It relates to the somewhat surprising finding that so-called *contextual probabilities* were sufficient [3]. These were originally intended just to be rough empirical estimates of *prior probabilities* for identifying the most likely answer, but they worked well with or without the support of the more precise kinds of *intrinsic probabilities* that are much more familiar, and which we would ideally like to test for CDS as well as flat statements of knowledge. These are probabilities as degrees of truth or scope intrinsically associated with each such statement. MARPLE 2 has a much improved ability to separate parts of the calculations and the parts of the KRS so that we can perform "computer experiments" giving insight into the above issues, as described in this paper. We also describe technological progress since MARPLE 1. It is by processes of Internet searching, preliminary

[2] The deep relationship between MARPLE learning from examinations (by ultimately noting the official correct answer) and simply acquiring knowledge from the Internet (irrespective of being tested in any examinations) is revealed by a thought experiment. We imagine the best case that a web page contains text that is essentially the question with the correct answer provided in the course of discussion. Indeed, published clinical case studies are essentially of that nature (Section 1.9). In practice, MARPLE does not often hit upon anything like that which is directly relevant. However, given long enough, and noting that hundreds of thousands of extracts of knowledge can easily be generated in this way in a day, the required information is likely to be found, even if element by element. Unfortunately, accumulating too much knowledge in the rougher XTRACT form contributes noise (Section 1.6).

curation, later automatic sweeps of curation, and human intervention if necessary, that MARPLE 1 learned from examination questions as well as from the more general curated medical knowledge given [3]. However, the essential features of MARPLE 1 are not specific to medicine. For example, wherever first directed to search on the Internet, it could still ultimately start to gather more general knowledge of any kind. Although not without theoretical interest and indirect clinical value, the effort required is a heavy price to pay for success in medical school exams. MARPLE 2 is much more efficient by knowing how to focus on scholarly medical texts. Last but not least, the present paper explores curation as reduction of a kind of "noise" that MARPLE 2 helps reduce, as follows.

1.6. *The problem of noise in curation using medical school examination*

An educational analogy may be helpful. Generating and using well curated KRS elements as more formal teaching resembles *supervised learning* in schools and universities in formal classes, while knowledge as rougher extracts gathered from searching the Internet outside of formal class seems less organized, broader, diffuse, and vaguer, essentially analogous to that in *unsupervised learning* by a medical student reading and researching outside of formal class. When knowledge of a more general nature is collected over a long period, its use in an exam may seem like intuition and making hunches. It seems "noisy", in practice meaning that exam scores can deteriorate if there is too much reliance upon it. By "noise" in this present paper we essentially simply mean imperfections in the KRS evidenced by deteriorated examination performance, and repairing this is a major part of the curation and learning process. This arises from the fact that MARPLE can use information extracted from the Internet while still in a premature, rougher, and less well curated state. For this reason, MARPLE 2 benefits from being more focused. In the process of extraction there is some natural language processing as curation: extracts of source text that often but not always correspond to sentences are parsed and put into a canonical form so that MARPLE can at least sufficiently "understand" them. Some extracted elements will already essentially be in good form for that and even for future CDS applications. What the many remaining elements lack in sophistication, they can in principle make up in considerable abundance, as described throughout this paper. Exam performance indeed improves with Internet searching, since more time searching provides more knowledge likely to help answer the question, but as thousands or hundreds of thousands of extracts accumulate exam performance starts to deteriorate to a lower plateau of performance. It seems natural to use the above unsupervised learning analogy or, in a little more detail, assume that it represents fluctuations in information that do not in practice cancel out, as if there is an increasing probability of confusing or contradicting previous statements of knowledge. Theoretically, however, it seems somewhat of a paradox that is investigated in this paper.

Well curated KRS elements are demonstrably relatively "noise free", as discussed below, always present alongside the rougher Internet extracts except in certain "computer experiments", and playing a prominent role in obtaining good exam scores [3]. It may therefore be asked why the rougher knowledge the Internet is not better curated as it arrives. In practice, automatic curation takes some time for hundreds of thousands of knowledge elements if there is to be any degree of "smartness" or sophistication, e.g. taking account of semantic and knowledge considerations beyond simple basic correction grammatical processing. At the same time, the exam helps here by pinpointing what elements of knowledge need to be present and curated. The greater focus by MARPLE 2 on what is more immediately needed is therefore important. Since hundreds of thousands of extracts of knowledge can be generated in a day, the required information is likely to be found, even if found element by element, but it has to be recognized as "promising". That this brute force approach works at all is explicable as follows. The underlying algorithm attempts to see relevant knowledge by piecing a small chain of several KRS elements, i.e. a "chain of associations". Single elements of knowledge can answer a question, but that is usually insufficient except with questions that test simple matters of definition. More precisely, a small network of relevant knowledge elements tends to form between question and each answer, each contributing a different estimated probability or weight, and the overall weight of evidence ideally identifies the most probable candidate answer. The rougher elements freshly acquired from searching the Internet can be tolerated as contributing evidence in the overall argumentation process, somewhat analogous to use of circumstantial evidence in a court of law. Evidently, however, in the present case that can lead to noise in information of predictive or explanatory value that can cloud final judgment.

1.7. *Previous and related work: comparison with IBM's Watson*

IBM's Watson system [6–9] that beat humans in the TV quiz show "Jeopardy!" seems the most interesting other effort to compare because a general knowledge quiz like "Jeopardy!" has obvious similarities to taking a university examination. Also, a Google query *IBM Watson health* at time of writing this paper claims over 3 million results, showing that Watson is being preened for exploitation in healthcare. Technical comparison with medical versions of Watson is difficult because while the earlier reports (e.g. Refs. [7,8] are of a scientific nature, accounts of applications of Watson in medicine [9], and hints regarding medical licensing examinations [10,11], are essentially news items or anecdotal. It seems clear, however, that the issue of whether or not computers really can pass medical licensing examinations is in particular by no means closed by the performance in "Jeopardy!" That was a competition against humans, and scoring as incorrect the questions that Watson did not answer (as is appropriate in an examination), Watson fared less well [12]. Technical comparison has proven difficult in any case, because the aims and methods of Watson and MARPLE are somewhat different. Watson is a "Grand Challenge" selected to demonstrate the power of high performance computers as much as the skills of their programmers. MARPLE runs on a standard laptop and is part of a larger experimental prototype system for a future probabilistic "Thinking Web" starting with healthcare and intended to be distributed freely over more standard servers. Because the primary aim is the curation of knowledge elements for CDS, and the second is that some techniques developed to take part in exams may be applicable to CDS, MARPLE has no recriminations in searching the Internet for answers even during an exam (Section 1.2), a luxury forbidden to any contestant in "Jeopardy!" This is in any case important for MARPLE because it allows it to use the Internet as a "memory extension" by querying on the question and particularly each answer rather than rely only on limited personal computer resources. Knowledge in Watson is used in probabilistic way, but in the Watson as developed for "Jeopardy!", probabilities were assigned in order to "lay bets" on answers to quiz questions being correct using an empirical probability estimate based on different degree of confidence in the answers proposed by various subsystems, in the context of each question type. With future CDS applications in mind, the probabilities in our systems are so-called intrinsic probabilities mentioned in Section 1.2 that ideally come closer to those familiar in biostatistics, epidemiology and evidence based medicine such as risk factors [1], although it is contextual probabilities, also mentioned in Section 1.2, that emerge as better tested in the exam context [3]. The latter probabilities seem to come closer to Watson's. Nonetheless, in MARPLE this is currently governed by a single algorithm and essentially a single equation (Section 2.4).

1.8. Previous and related work. theoretical and technical basis

For discussing the theoretical basis, we continue to use the notation standard and associated probabilistic algebra due to theoretical physicist Paul Dirac. It was developed in the 1930s to 1940s [13] for quantum mechanics (QM) [14] and author Robson and colleagues have adapted it for classical inference over several years [15–23]. MARPLE owes a debt to a considerable body of work going back to efforts in early data mining in bioinformatics in the 1970s, as reviewed in Refs. [20,23,27]. Dirac notation may be an accepted standard in physics and so its use here may please physicists, but to most readers it is still largely an unfamiliar approach and somewhat less essential for understanding here because QM is about intrinsic probabilities, and it is extrinsic probabilities that played the important role in exams (as discussed above, and in Ref. [3]). However, QM represents a huge of amount of done work, integrated in a way that provides a unified formal picture, to which we can refer. Notably, an important consideration for CDS is that Dirac's work leads to mathematics that allows an inference network to be a general graph, including cyclic paths, rather than artificially restricted to a directed acyclic graph as is traditionally (and actually by definition) characteristic of a Bayes Net approach [20]. MARPLE was developed in that context and it is a *Q-UEL application*, testing knowledge in a Q-UEL system. The "Q" in Q-UEL is a reference to QM and Dirac's notation and algebra on which Q-UEL is based. The "UEL" is a reference to the "XML-like" *Universal Exchange Language* requested for healthcare by the (US) President's Council of Advisors on Science and Technology (PCAST) in 2010 [24]. This resulted in the development of Q-UEL interoperability language and the prototype Q-UEL system based on it [25–28]. MARPLE KRS elements are, or are readily interconvertible with, Q-UEL communication artifacts called *tags* (by analogy with XML). All Q-UEL tags are not only based on Dirac notation, but can also have analogous algebraic roles. They are usually in the common Dirac "bra-operator-ket" form $< subject| relationship|object >$, associated with a complex number that encodes two probabilities, the first for the statement as read, and the second as read with subject and object switched. It is also called a *Semantic Triple* (ST) because it has three parts, analogous to subject-verb-object (SVO) clause. The knowledge represented in these can interchange with preexisting Q-UEL applications such as POPPER [22], and DiracBuilder [23], and its sources are human experts inputted via POPPER HELPER [23] and structured data mining such as DiracMiner [23], as well as other Q-UEL applications and clinical data repositories [27,28].The knowledge from the Internet which is said to be less well curated form (Section 1.3) is captured as attribute values called *XTRACTs* in Q-UEL XTRACT tags generated by the automatic browser and surfer called XTRACTOR [27]. It is these XTRACTs that are the rougher knowledge elements that are extracted from source text. When an XTRACT is finally well curated, it may become one or more STs, but usually it becomes one or more Linear *Semantic Multiples* (LSMs). Seen as the parsed form of a sentence, each LSM is a linear graph, a single path, extracted from it (the parsed form could be a linear path already). In format the LSMs looks like a break with Dirac notation and Q-UEL's so-called *general specification* [27], but they can merely be interpreted as convenient abbreviated forms of certain Dirac expressions. Nonetheless, there are significant practical advantages (see Section 2.3). Notably, the LSM helps avoid or reduce a *combinatorial explosion* of possibilities that similarly arises in the Feynman Path Integral [29]. Roughly speaking, this is a kind of inference net in QM, with analogies to the exam problem.

1.9. Previous work: comparison with XTRACTOR and MARPLE 1

XTRACTOR is an invisible automatic browser that "autosurfs" the Internet to obtain the QUEL tags called XTRACT tags [27]. An XTRACT typically, but by no means always, corresponds to an original sentence in natural language text on a web page. It is reparsed into a more canonical form that is as close to an LSM as possible, grammatically annotated. However, this retains source links, and subsequently XTRACTOR can surf automatically on links including those in scientific citations. Hence the impression is that XTRACTs can spawn other XTRACTS from the links. If it later encountered text that had no links, an additional small application enabled it to build new queries from the content and so "keep on surfing". In MARPLE the exam itself generates the queries. MARPLE can use content of the exam question to do that but initially it works through the essential content of candidate answers as queries, one at a time. Since it can, if permitted, keep on re-adressing. exams, with or without new questions, it can keep on accumulating, and helping curate, knowledge from the Internet. An exam will of course give searches a more specific focus, the topics covered by questions. MARPLE 2 is more efficient because it checks that web pages are of scholarly medical character before any XTRACTs are generated. While general knowledge is important, it comes at a high price of distracting from the main curation activity and by adding to the above combinatorial explosion (Section 1.8). It draws from two new two dictionaries of words and phrases that are, and that are not, characteristic of "serious medical writing", and using a *common topic* algorithm it assesses the degree of appropriateness of every web page before XTRACTs are obtained. XTRACTOR tended to draw on Wikipedia [27] because its analytic techniques were well set up to understand Wikipedia web pages, but it was by no means confined to it, and might not use it at all. Sources like PubMed were a common option [27]. MARPLE 2 *always* starts by submitting queries to Wikipedia and by using its own essentially simple query system, not riding on that of Google as MARPLE 1 could do, and as was done extensively in our earlier XTRACTOR efforts [27]. MARPLE 2 also keeps track of links to web pages it has examined before and avoids revisiting them in a single run, i.e. in the course of an exam. It has a list pre-specified sources to search if it ever gets stuck, e.g. if there are no links, or the sources do not look like scholarly medical text. Nonetheless, in routine ongoing operation it is encouraged to revisit web pages that may be rewritten by experts over the years, in order to capture many different ways of expressing same knowledge. When reading each item on a KRS to find those relevant to an exam question, MARPLE 2 repairs more obvious format damage and checks that noun and relationship phrases are in the right slots between delimiters and in correct sequence (but it does not delete the element if there are difficulties). Unlike its predecessor, MARPLE 2 can convert quantitative clinical data in questions to low, normal, or high ranges. It considers that low, normal or high, not the original numeric value, is the important description in a knowledge element. Last but not least, MARPLE 2 facilitates switching on and off of various algorithmic contributions and blocks of knowledge so that some surprising findings in MARPLE 1 can be explored.

1.10. Previous and related work: curation

Review of curation is hampered because definitions vary. *Digital curation* in general is a better defined discipline [30–31], but it emphasizes interoperability and extensibility that Q-UEL has already sought to address, and we have already reviewed these (e.g. Refs. [27,28]). The word "curation" does not appear as often as one might expect in literature about the Semantic Web (SW) [32], probably simply because curation of knowledge into a widely usable canonical form is the whole point of the SW. It is common to read that an author is "using the tools of the SW" to curate other digital data (e.g. Ref. [33]). MYCIN had facilities for capturing and curating knowledge from human experts [2]. The larger INTERNIST Expert System effort [34] provided facilities for capturing and curating tens of thousands of statements of knowledge, perhaps now 100,000 or more, for diagnosis in internal medicine, over many years [34]. Both these systems [2,34] could use something like an exam or test to help curate knowledge and made comparison with human experts under same conditions, and since these were genuine experts in specific fields these could be regarded as much harder tests. However, an overall comprehensive exam as taken by MARPLE is not necessarily easier, because medical experts in one field often forget their medical lessons in other fields. Within our own Q-UEL effort, there has significant amount of work related to facilities for manual curation (e.g. by POPPER HELPER [22]) and automatic curation (e.g. using aids like THESAURUS [27]) in the Q-UEL project, Ref. [6] provided a good example by other workers of the fairly general sense in which we also interpret curation.

1.11. Other Work and issues related to medical licensing exams

In education science, there are always ongoing efforts improve the fairness and testing power of multiple choice medical examinations (e.g. Ref. [35]). The questions presented by such sources are excellent for testing and training MARPLE, because they provide a more realistic clinical scenario and because of the demands they place on the discerning power of the KRS elements. For example, there is increasing interest in "Extended-Matching (R-Type) Items", basically a single large sent of answers to which many different questions are directed [35]. However, they are not at time of writing typical of qualifying exams, and the more typical and recent questions actually used in real final medical licensing exams are not so easy to obtain and share in any significant quantity. Licensing boards like that for the USMLE hold copyright on actual exam and practice questions [36]. While they make these available for practice via the Internet, they usually prohibit and digitally block copying and distribution. This is somewhat impeding research and sharing in the field of automated tackling of medical school exams, but since the field is barely emergent it is hardly seen as a problem. Fortunately many professors construct their own questions and make them available in text books and lecture presentations [37], though excessive copying and distribution still risks raising copyright issues. The examples used in our papers are chosen cautiously to respect original authors and publishers and do not represent the full set used in testing. Published clinical case studies [38] can be readily adapted to form exam questions by presenting the symptoms and tests and removing the clinician's interpretations, and these are especially valuable for MARPLE by testing the decisions made in real clinical practice. However, the text forming the question part can represent up to a third or half of the original published paper, raising even stronger copyright issues.

Certain style points desirable in setting the more standard exam questions are well known, and some relate to further aspects of the present study. Grammarists commonly exemplify how the injudicious use of determiners and pronouns can lead to misinterpretations (e.g. Ref. [39]), and inspection of medical documents such as radiologists' reports often shows that their use naturally tends to be minimized, giving the stilted but unambiguous style [40]. Nonetheless, determiners like "a", "the", "some", "many" can be used to help estimate an intrinsic probability that the remaining part of the statement is considered true or generally applicable [41]. The use of negatives like "no" is simply a more obvious case. This should all be important for CDS. However, the importance in the exam scenario does not appear to have been explored in detail by any computational means, so this is also touched upon below.

2. Theory

2.1. General overview

The general principles needed to answer medical school exam questions should not be expected to be too hard to understand. For fairness to the student, an exam normally requires only the kind of mental manipulations that humans do naturally and fairly well. Notable here are the simpler uses of syllogistic logic. For example, given the two statements that rubeola virus causes measles and that the typical manifestation of measles is a rash, we can deduce that a rash may indicate rubeola virus but only sometimes, because, for example, very similar rashes may be due to other causes such as rubella. Each statement in a syllogism can be of explicit or implicit universal ("all"), or existential ("some"), including universal negative ("no") or existential negative ("not all", "some not") character. Q-UEL can express the extent of any of these so-called quantifications in several truly quantitative equivalent ways.

$$\begin{aligned} <\mathrm{A}|\mathbf{R}|\mathrm{B}> &= \{\mathrm{P}(\text{“A R B”}),\ \mathrm{P}(\text{“B R A”})\} \quad \text{(probability dual)} \\ &= <\mathrm{B}|\mathbf{R}|\mathrm{A}>^{*} = \{\mathrm{P}(\text{“B R A”}),\ \mathrm{P}(\text{“A R B”})\}^{*} \quad \text{(relation to the complex conjugate)} \\ &= \boldsymbol{\imath}\mathrm{P}(\text{“A R B”}) + \boldsymbol{\imath}^{*}\mathrm{P}(\text{“B R A”}),\ \ \boldsymbol{\imath} = ½(1+\boldsymbol{h}),\ \ \boldsymbol{\imath}^{*} = ½(1-\boldsymbol{h}), \quad \text{(physicists’ spinor projectors)} \\ &= ½[\mathrm{P}(\text{“A R B”}) + \mathrm{P}(\text{“B R A”})] + ½\boldsymbol{h}[\mathrm{P}(\text{“A R B”}) - \mathrm{P}(\text{“B R A”})] \quad \text{(Hermitian commutator form)} \\ &= \text{Existential}\ (<\mathrm{A}|\mathbf{R}|\mathrm{B}>) + \boldsymbol{h}\ \text{Universal}(<\mathrm{A}|\mathbf{R}|\mathrm{B}>) \quad \text{(probabilistic semantic form)} \end{aligned} \qquad (1)$$

Here ***h*** is the hyperbolic imaginary number ***h*** such that ***hh*** = +1 rediscovered by Dirac under several guises. <A| and |B> are Dirac's bra and ket vectors and **R** an operator. For brevity see Refs. [20–23,27]. By "A R B" in Eq. (1) we mean, for example, "obese patients *are* type 2 diabetics" which is a matter of probability P("A R B"), and so does <A|**R**|B> as similarly read like a sentence in English. We correspondingly mean that "type 2 diabetics *are* obese patients" is a matter of probability P("B R A"), switching subject and object expressions. When **R** is a verb of conditional, categorical, causal character, or implies a propagation of effect, we can usually say things like, for example, <A | **are** | B> = <B | **if** | A> = < B | **is caused by** | A> = <A | **causes** |B> = <B | **causes*** | A> = <A | B>* = <B | A>. We can then also compute probabilities involved in syllogisms, e.g. <A | **R** | C> = <A | **R** | B> <B | **R** | C>. However, if the **R** are other kinds of verbs, such as of action, the product is not really the value of the "conclusion" but merely reflects the degree of collective truth of the propositions. Examples of Q-UEL compatible MARPLE KRS elements that reflect the above ideas are

$$\begin{aligned}
&< \text{overeating Pfwd:=0.8} \mid \textbf{causes} \mid \text{obesity Pbwd:=0.7} > \\
&< \text{obesity Pfwd:=0.2} \mid \textbf{causes} \mid \text{type 2 diabetes Pbwd:=0.85} > \\
&< \text{‘acute anterior uveitis’ Pfwd:=0.3} \mid \textbf{if} \mid \text{‘ankylosing spondylitis’ Pbwd:=0.2} > \\
&< \text{osteoporosis Pfwd:=0.06} \mid \textbf{if} \mid \text{‘ankylosing spondylitis’ Pbwd:=0.04} > \\
&< \text{HLA} - \text{B27 Pfwd:=0.8} \mid \textbf{if} \mid \text{‘ankylosing spondylitis’ Pbwd:=0.2} > \\
&< \text{HLA} - \text{B27 Pfwd:=0.6} \mid \textbf{if} \mid \text{‘acute anterior uveitis’ Pbwd:=0.1} >
\end{aligned} \tag{2}$$

Such forms using tag values Pfwd and Pbwd are canonicalized Q-UEL forms of information found on the Web. When similarly used in programming mode in POPPER [22] we typically see an assigned probability dual as e.g. < obesity | **causes** | type 2 diabetes > = 0.2, 0.85. This might cautiously be given the interpretation that P(type 2 diabetes | obesity) = 0.2 and P(obesity | type 2 diabetes) = 0.85. In many examples below, Pfwd and/or Pbwd are absent. The implied default is probability 1, to indicate ignorance or an assertion yet to be refuted, consistent with information theory, the philosophy of Popper, and other considerations [20,22,27].

2.2. Determiners and intrinsic probability in MARPLE

There seems at least little doubt that at least the categorical qualifiers, “all”, etc, along with negations, are important determiners, and no less so in Q-UEL context, as Eq. (1) shows. Adhering to Dirac’s notation, we could see the determiners as operators, e.g. to be used as follows (note however the discussion in regard to Eq. (7) below).

$$\begin{aligned}
&< \textbf{all } \$\text{B} \mid \$\mathbf{R}^* \mid \textbf{all } \$A > \; = \; < \textbf{all } \$\text{A} \mid \$\mathbf{R} \mid \textbf{all } \$\text{B} > \\
&< \textbf{all } \$\text{B} \mid \$\mathbf{R}^* \mid \textbf{some } \$\text{A} > \; = \; < \textbf{some } \$\text{A} \mid \$\mathbf{R} \mid \textbf{all } \$\text{B} > \\
&< \textbf{some } \$\text{B} \mid \$\mathbf{R}^* \mid \textbf{all } \$\text{A} > \; = \; < \textbf{all } \$\text{A} \mid \$\mathbf{R} \mid \textbf{some } \$\text{B} > \\
&< \textbf{some } \$\text{B} \mid \$\mathbf{R}^* \mid \textbf{some } \$\text{A} > \; = \; < \textbf{some } \$\text{A} \mid \$\mathbf{R} \mid \textbf{some } \$\text{B} >
\end{aligned} \tag{3}$$

However, MARPLE takes the default as **all** in the bra apart < …| and as **some** in the ket part | .. > if they are not specified. There are Dirac rules about manipulating such operators, but we can be consistent by suitable choice of type of operator. There is also a relation between these determiners and the Pfwd and Pbwd attributes (Section 5.4). Interpreting the role of any determiner can be thought of as assigning a point in the relevant two-dimensional region enclosed by the path $0+0\boldsymbol{h}=0\rightarrow0.5+0.5\boldsymbol{h}\rightarrow1+0\boldsymbol{h}=1\rightarrow0-0.5\boldsymbol{h}$ and back again to 0. In probability dual notation, for example, we can write

$$\begin{aligned}
&\{\text{P(“no A are B”), P(“no B are A”)}\} = \{0,0\} = 0 \\
&\{\text{P(“All B are A”), P(“some B are A”)}\} = \{1,0\} \\
&\{\text{P(“A equals B”), P(“B equals A”)}\} = \{1,1\} = 1 \\
&\{\text{P(“All B are A”), P(“some B are A”}\} = \{0,1\}
\end{aligned} \tag{4}$$

However, for formal reasons we should replace 0 throughout by ~0 meaning “approximately zero” (See Eq. (7) and associated footnote). Links in chains of reasoning in Q-UEL are most commonly computed in the same way as syllogisms, which can be implemented by a metastatement [22,27].

$$< \$\text{A} | \$\mathbf{T} | \$\text{C} > \; = \; < \$\text{A} | \$\mathbf{R} | \$\text{B} > \; < \$\text{B} | \$\mathbf{S} | \$\text{C} > \tag{5}$$

With < A | **R**| B > = {w, x} and < B | **S** | C > = {y, z}, the use of Eq. (5) then implies

$$\{wy,\ xz\} = \{w,\ x\}\ \{y,\ z\} \tag{6}$$

and so < A | **T** | C > = {wy, xz}, if **R**, **S**, and **T** are of conditional, categorical, causal or related nature, otherwise we can only say that < $A | $**R**| $B > < $B | $**S** | $C > = {wy, xz}, as the collective truth expressed in probability dual notation. In the above account so far, there are several issues arising that are readily addressed[3], but seemingly more troublesome is that, in principle, there are two kinds of way conforming to the above. One could work with P: type (POPPER) statements or M: type (MARPLE) statements, in such a way that we can write

$$\begin{aligned}
&< \text{P: A} \mid \textbf{are} \mid \text{B} > \; = \{0,0\} = 0 = \; < \text{M: } \textbf{no} \text{ A} \mid \textbf{are} \mid \text{B} > \\
&\quad = \; < \{\text{P(“no A are B”), P(“no B are A”)}\} < \text{P: A} \mid \textbf{are} \mid \text{B} > \\
&\quad = \{1,\alpha\} = \; < \text{M: } \textbf{all} \text{ A} \mid \textbf{are} \mid \text{B} > \; = \{\text{P(“All B are A”),} \\
&\qquad \text{P(“All B are A”)}\} \\
&< \text{P: A} \mid \textbf{are} \mid \text{B} > \; = \{1,1\} = 1 = \; < \text{M: A} \mid \textbf{equals} \mid \text{B} > \\
&\quad = \{\text{P(“A equals B”), P(“B equals A”)}\} \\
&< \text{P: A} \mid \textbf{are} \mid \text{B} > \; = \{\alpha,1\} = \; < \text{M: } \textbf{all} \text{ B} \mid \textbf{are} \mid \text{A} > \\
&\quad = \{\text{P(“All B are A”), P(“All B are A”)}\}
\end{aligned} \tag{7}$$

MARPLE currently simply sees statements as only initially as of M: type, with the aim of helping a human expert users curate the knowledge representation tags derived from natural language text, and to assign probabilities that are more quantitative. On occasion there are process of curation in which we need to reconcile two or more statements into one when detected as having essentially same meaning but, for various reasons, do not necessarily come with the same probability values:

$$< \text{A}|\mathbf{R}|\text{B} >_{\text{n}} \leftarrow \; < \text{A}|\mathbf{R}|\text{B} >_{\text{n}-1} + < \text{A}|\mathbf{R}|\text{B} >_{\text{n}} - < \text{A}|\mathbf{R}|\text{B} >_{\text{n}-1} < \text{A}|\mathbf{R}|\text{B} >_{\text{n}} \tag{8}$$

There is a formal relationship here with determiners in a corresponding non-recursive form relating to the binomial function via the dual {P(A|B), P(B|A)} = $\{1\text{- }(1\text{- n[B]}^{-1})^{\text{n[A, B]}}, 1 - (1\text{- n[A]}^{-1})^{\text{n[A, B]}}\} = 1\text{- }\{(1\text{- n[B]}^{-1}), (1\text{- n[B]}^{-1})\}^{\text{n[A, B]}}$. Similar terms can sometimes be seen in our diagrams interpreting the meaning of accessible regions of ***h***-complex space in terms of many kinds of determiners [22,41]. However, further theoretical discussion is deferred to elsewhere because of the dominance, in taking exams, of the following.

2.3. Contextual probability, linear semantic multiples, and automatic curation

The considerations in the above title are not linked together in any essential way, but in practice they are used together. Computation of contextual probability would often run into formidable combinatorial problems (Section 1.8) if it were not for linear semantic multiples (LSMs), and the conversion of

[3] Examples are as follows. There are cases where < A | **R** | B > = < B | **R*** | A > can seem to lead erroneous assumption (“Beavers build dams”, and “Dams are built by beavers”, but the Hoover dam is not). However, note that in noun expressions A, B, C, etc, the determiners explicit or implicit are part of the expression: they travel with the noun (“Some dams are built by all beavers”). For truly accessible values in Eq. (4) we should again write {1, ~0} and {~0, 1} with ~0 meaning “approximately zero” because if we say that 0 truly means zero, then if P(A|B) = 0, we likely mean that P(A, B) in P(A|B) = P(A,B)/P(B) is zero, and hence we should have P(B|A) = P(A,B)/P(A) = 0, not 1. Dirac’s original rules require that, for example, < **all** A| **are** |B > = < A | $\textbf{all}^{\dagger}$ **are** |B > = < A | **are** | $\textbf{all}^{\dagger}$ B >, where † indicates a distinct adjoint form, but it is perfectly possible to define our determiner as an entity such that $\textbf{all}^{\dagger}$ = **all**.

XTRACTs to one or more LSMs is the most prominent kind of of curation. We can understand the impact of contextual probability as might be seen from the perspective of mutual information, i.e. as the last term in

$$\begin{aligned} &\mathrm{I}(<\mathrm{A}|\mathbf{R}|\mathrm{B}>,\ \ <\mathrm{B}|\mathbf{S}|\mathrm{C}>)\\ &\quad=\mathrm{I}(<\mathrm{A}|\mathbf{R}|\mathrm{B}>)+\mathrm{I}(<\mathrm{B}|\mathbf{S}|\mathrm{C}>)+\mathrm{I}(<\mathrm{A}|\mathbf{R}|\mathrm{B}>;\ <\mathrm{B}|\mathbf{S}|\mathrm{C}>) \end{aligned} \tag{9}$$

In certain cases such as < A | **if** | B > which can be written as [ıP (A)+ ı*P(B)] K(A; B), there is a useful relationship between association constant K rewritten as K(A; B | C) = $e^{I(A;\ B\ |\ C)}$ conditional on a context C that we shall discuss elsewhere. Although we do not have such information or that in Eq. (9) directly, its effect might still be capable of estimation from some function f( ) such that, for example,

$$\left\{\mathrm{P}(\mathrm{A}\mathbf{R}\mathrm{B},\ \ \mathrm{B}\mathbf{S}\mathrm{C}),\ \ \mathrm{P}(\mathrm{A}\mathbf{S}\mathrm{B},\ \ \mathrm{B}\mathbf{R}\mathrm{A})\right\}=\mathrm{f}(<\mathrm{A}|\mathbf{R}|\mathrm{B}>,<\mathrm{B}|\mathbf{S}|\mathrm{C}>) \tag{10}$$

Another way of thinking about the matter is that we might define the LSM as follows.

$$<\mathrm{A}|\mathbf{R}|\mathrm{B}|\mathbf{S}|\mathrm{D}>\ =\ <\mathrm{A}|\mathbf{R}|\mathrm{B}>\ <\mathrm{B}|\mathbf{S}|\mathrm{C}> \tag{11}$$

We can say that the LSM really just a convenient abbreviation for Dirac notation by imagining the following equality, and note that the vertical bars take on the role of indicating a relationship operator, more traditionally done by using bold font of a cap '^'.

$$\begin{aligned} &<\mathrm{A}|\mathbf{R}|\mathrm{B}|\mathbf{S}|\mathrm{C}|\mathbf{T}|....\big|\mathrm{Z}>\ =\ <\mathrm{A}\big|\mathbf{R}|\mathrm{B}>\ <\mathrm{B}|\mathbf{S}|\mathrm{C}>\ <\mathrm{C}|\mathbf{T}|....|\mathrm{Z}>\\ &\quad=\ <\mathrm{Z}|\ldots\mathbf{T}|\mathrm{C}|\mathbf{S}|\mathrm{B}|\mathbf{R}|\mathrm{A}>^{*}\\ &\quad=\ <\mathrm{Z}|\ldots\mathbf{T}^{*}|\mathrm{C}|\mathbf{S}^{*}|\mathrm{B}|\mathbf{R}^{*}|\mathrm{A}> \end{aligned} \tag{12}$$

Advantages of LSMs are that that all knowledge in an LSM comes from the *same source context*, via an XTRACT. Also by < A | **R** | B | **S** | C| **T** | D |….. > rather than forms < A | **R** | B >, < B | **S** | C > etc. that can be rearranged in different combinations, we avoid the combinatorial explosion that can come from multiple possible pathways linking question to each answer, analogous to the problem encountered by the Feynman path integral in QM [29].

In practice, by far the greater part of curation at the present time is the curation of XTRACTs to become LSMs as the more complicated examples of "well curated KRS elements. Several kinds of correction can be applied automatically. The main one used in MARPLE is to ensure that the XTRACT and hence resulting LSM element have noun phrases A, B, C, etc. and relationship phrases **R**, **S**, **T** etc., i.e. verbal or prepositional phrases, are in the correct slots, and that as a point of style negation is associated with the verb. The main feature of the common content function below (Section 2.4) is that < A | **R** | B | **S** | C | **T** | D > not only implies e.g. < C | **S*** | B > where **S*** is the active-passive inverse of **S** and *vice versa*, but also that this "has something to do with" < C | **U** | B >, as well as many other forms. These will score less with the common content function, i.e. *have a lower contextual probability*, but MARPLE does not have to find them elsewhere if it if it has that larger form, and contextual probabilities are calculated at moment of use. With applications to CDS in mind rather than forcing solutions to exam questions, it is important that knowledge is not contrived, but almost always based on what is actually found, and that if any modification is required, that it reflects common sense. It includes taking an initial mis-parsed effort by XTRACTOR such as

< damage | **to** | 'right lenticulostriate arteries' | **causes** | left 'spastic hemiparesis' >

and changing it to the following

< left 'spastic hemiparesis' | **'is caused by'** | damage | **to** | 'right lenticulostriate arteries' >

to reflect the fact that it is more logically the damage that causes the disease. Note that XTRACTOR could not deduce that from grammar of the source text alone. While these are checked manually, automatic processing can take place using the easily applied *metastatements*, e.g.

< $C | **'is caused by'** | $A | **to** | $B > = < $A | **to** | $B | **causes** | $C >

In the Q-UEL general specification [27] there is the more general form in which **$R** replaces "causes" and **$R***, its complex conjugate (and in POPPER usually its adjoint) so that **causes*** means "is caused by". Usually there would have to be a prior definition < $B | **'is caused by'** | $A > = < $A | **causes** | $B >. This is because adjoint forms can be a matter of vocabulary, as for prepositions like "on" and "under". However, POPPER allows the following more general "standard English" solution: < $B | **'is $Red by'** | $A > = < $A | **$Rs** | $B >, where **$R** binds to a root such as "cause". Irregular forms still have to be explicitly dealt with. Some common and basic operations of curation are, however, "hardwired" into MARPLE 2. It is of theoretical interest that the essential features of the above can be justified in terms of the Dirac notation [22].

2.4. Contextual probability and the common content function

Contextual probabilities are computed by MARPLE empirically using the concept of *common content*, i.e. words or phrases that are the same that crop up between two knowledge elements (ST, LSM, or corresponding XTRACT), or between a knowledge element and question, or a knowledge question or answer. It is a *minimum requirement* for logic of syllogistic kind as discussed in Section 2.3 above. In addition, the role of determiners, "a", "the" "some" etc., is demoted (except for negatives). Details as implemented in MARPLE 2 are as follows.

(I) First, in this preliminary step, content is represented in a simpler and rather more canonical form. So-called "trivial words" that are removed are actually really ignored, because they correspond to the determiners such as "a", "an", "the", "some", "all", many etc. that we may (or may not) wish to use them later in curation. However, negative forms such as "no", are not ignored.

(II) All the *noun fields* F, i.e. A, B, C, etc. between relators in a LSM are each kept intact in word order, and held separately, but as above, each has so called "trivial words" ignored.

(III) This is repeated in terms of *individual words* W, again after all so-called "trivial words" have been removed. It may be helpful to think of a larger number of fields, each now a single word, but for clarity they are still called words rather than fields.

(IV) N_F is now computed as the number of fields that are common to any two specified tags, or between a specified tag and the question, or a specified tag and a specified answer.

(V) N_W is now computed as the number of words that are common to any two specified tags, or between a specified tag and the question, or a specified tag and a specified answer.

(VI) We now consider separately the number of fields, and then words, when at least one of the two tags (or question or an answer) contain *at least one positive verb form* **R**, **S**, **T**, etc., as N_{F+} and N_{W+} respectively.

(VII) We also consider separately the number of fields, and then words, when at least one of the two tags (or question or an answer) contain *at least one negative verb form* **R**, **S**, **T**, etc., as N_{F-} and N_{W-} respectively.

More practical detail with preferred options used in this study is given in Methods Section 3. Note that the above process (a)-(e) is repeated for each answer in turn and the scores *accumulate* for each answer. Note that MARPLE resets the accumulative counts of $N_{F+,}$ N_{W+} $N_{F-,}$ and N_{W-} each back to zero only when moving on to a new question. The Common Content Function is computed in terms of the Riemann zeta function ζ (s, n) partially summated to n rather than using ζ(s) = ζ (s, n = ∞).

$$\begin{aligned} I(F+:F-)=&\zeta(s=1,\ \ N_{F+}+N_{W+}+v)-\zeta((s=1,\ \ N_{F-}+N_{W-}+v)\\ &-\zeta(s=1,\ \ N_{W+}+v)+\zeta(s=1,\ \ N_{W-}+v) \end{aligned} \tag{13}$$

This can be applied several times for each question as described below, each viewing the contextual importance of the tag from a slightly different perspective, so note that MARPLE resets back to zero the accumulative counts of $N_{F+,}$ N_{W+} $N_{F-,}$ and N_{W-} only when moving on to a new question. See Section 2.5 next, regarding v. For s= 1 we have the Euler series that gives an expected information measure [20,23]

$$\zeta(s=1,\ n)=1+1/2+1/3+\ldots 1/(n) \tag{14}$$

2.5. Weighting by quality

When sources of knowledge are combined by Eq. (13), they are *not* otherwise weighted according to provenance, i.e. according to extent and quality of curation of each knowledge element as judged automatically or by humans. However, there are *automatic* techniques or equivalent with somewhat analogous effect considered below and in next Section 2.6. Another is that new things of potential interest usually have to be observed several times before they start to have significant impact. Eq. (13) indicates that a virtual frequency v is added to the counts immediately before use by the zeta function to represent a kind of absolute prior frequency (in actuality v is added inside the zeta subroutine function as called, but the above is the appropriate mathematical representation). The value of v is rather larger than has been the case in previous studies, as this quenches the considerable noise due to use of knowledge representation tags that may as yet be poorly curated. This should be seen in the context of an additional scoring and bonus systems described later below that modifies the value of counts $N_{F+,}$ N_{W+} . As a probability, MARPLE computes an associated *probit* probability, i.e. that implies the use of predictive odds distorted monotonically into a probability form by imposing normalization.

$$P(F+)=e^{I(F+:F-)}/Z \tag{15}$$

Here Z is the sum of all e $^{I(F+:\ F-)}$ computed, one for each answer, such that the sum of the probabilities of each answer sum to 1. That there are ample opportunities for finding many common fields is ensured (a) by collecting as many tags as possible with fields that express equivalent content in different scientifically acceptable ways, and (b) by in the first pass by removing the so called “trivial words” represented by determiners (except negative determiners). If no tags are found in the knowledge store that satisfy a path between question and answer, then an automatic consequence of the above is that all answers have equal probability, i.e. 1/n for n answers. In addition one sees that probability of 1/n for all answers if no common fields are found that ultimately link the question to any of the answers, or if for those that are found there is an equal number of positive and negative relationships.

The above measure is also weighted not as to source and extent of curation but as to the quality of the match found, in the following sense. Values of $N_{F+,}$ N_{W+} $N_{F-,}$ and N_{W-} can be modified by a weight W. This is a multiplicative scaling factor to reflect the fact that observations of occurrence some circumstances carry more weight than others, and in effect such scaling assigns “bonus scores”. Everything generally scores 1, e.g. if F+ is seen 8 times then the N_{F+} = 8 and 8 + v is used as the independent variable or argument of the zeta function, with the following exceptions that imply the use of the weightings or “bonus scores”. However, When there is common content with the question and with the set of answers the value of N_{F+} replaced by W × N_{F+} such as10 × 8=80, and similarly for N_{F-}. Note that this is prior to adding v. Recall that the incrementing of counts and weighting N_{F+} and $N_{F-,}$ is applied to fields, and then repeated for N_{W+} and N_{W-} for individual words as fields according to Eq. (13). Subsequently, this all repeated yet again using more challenging criteria for commonality of content (Section 3.3) especially the requirement that there common content with a *specific* answer. Again recall that the counts including weighted counts are accumulative, and that MARPLE resets the accumulative counts of $N_{F+,}$ N_{W+} $N_{F-,}$ and N_{W-} each back to zero only when moving on to a new question.

2.6. Contextual probability and the topic relevance function

To focus, MARPLE 2 starts searches on key elements of the *answer* string being addressed, and continues to check that web pages searched relate to keywords and phrases in the examination and its particular question and answers. MARPLE 2 has a large so-called “buzzword’ list of medical roots, words, and phrases to try and ensure that the context is medical or statistical, and a large “badword” list that reflects many common and average non-medical sites which the first version of MARPLE accessed in the absence of such guidance. Analysis of, and surfing from, a webpage only continues if the web page contains (a) key elements of the original search string *and* (b) yields an estimate of relevance of medical and statistical content that exceeds a required value. The *relevant topic function*, or *topic relevance function*, that yields that estimate is as follows.

$$\begin{aligned} I(OS+:OS-)=&\,\zeta(s=1,\ OS++v)-\zeta(s=1,\ OS-+v)\\ &-\zeta(s=1,\ ES++v)+\zeta(s=1,\ ES-+v) \end{aligned} \tag{16}$$

OS+ and OS− are the number of strings observed in the web page accessed and about to be processed that are characteristic of a medical web page (buzzwords) and characteristic of a non-medical web page (badwords) respectively. The expected strings ES+ are the expected number of occurrences, estimated from the number of strings on the buzzwords file multiplied by the number of characters on the web page divided by the number of characters on the web page and buzzword file combined. The expected strings ES− is similarly the number of strings on the badwords file multiplied by the number of characters on the web page divided by the number of characters on the web page and badwords file combined. In contrast to counts for the common content function, the above counts are initialized to zero prior to reading each web page. Virtual frequency as above is still applied, with the same value as for the common content function, but no weight factors are applied to give “bonus scores”. The notion of a negative relationship still stands, but is somewhat different, i.e. as a relationship with something that is “off topic”. I(OS+ : OS−) is an estimate of an amount of information currently used to reject web pages from detailed examination and knowledge tag extraction when less than a critical value, i.e. a *decision constant*. The value of this, set as 1.5 for the current report. was optimized empirically along with v and W over a large number of exam-taking sessions. Although the end user could examine the web pages displayed and this kind of score to make a partly manual selection of relevant web pages found, MARPLE executed wholly automatically in the present study. XTRACTOR technology can obviously be readily applied to detailed parsing of the question and querying by contented extracted from it, but canonicalizing the question seems to provide little advantage.

3. Methods

3.1. Main input

The main inputs are the examination and the knowledge required to answer it. See also Section 3.4 regarding dictionaries. The “exam paper” is on the exam file exam.txt, which is simply an exam paper in digital form, typical of those used as input when students take exams by computer. It is therefore an exception to the fact that any Q-UEL file can be written and read by any Q-UEL application. It consists of numbered questions followed by typically 5-25 numbered candidate answers. A variety of possible reasonable formats can be accepted and are used in reporting the exam in output. Each question and candidate answer set is usually followed by the correct answer according to the examiner, in order to assess performance, but it is not of course "seen" by the algorithm that takes the examination. The main sources of knowledge elements reside on, and "shuffle around" on, various files according to the modes of use described in Section 3.2 and the curation cycle in Section 3.3, but the normal and direct input is usually a mix of well curated KRS elements and the rougher XTRACTs on the working KRS archive knowledge.txt. There we may see boundaries like the snapshot below, which in this particular case also represents the source XTRACT and the LSM automatically curated from it [4].

< Q-UEL-Xtractor34 "Age [0https://en.wikipedia.org/wiki/Aging] |^is| `a major |^risk| factor |for| `most `common neurodegenerative [0https://en.wikipedia.org/wiki/Neurodegeneration] _diseases) (token subject)

|including|

Mild cognitive impairment [0https://en.wikipedia.org/wiki/Mild_cognitive_impairment]

Alzheimer's _disease [0https://en.wikipedia.org/wiki/Alzheimer%27s_disease] cerebrovascular _disease

[0https://en.wikipedia.org/wiki/Cerebrovascular_disease] Parkinson's _disease

[0https://en.wikipedia.org/wiki/Parkinson%27s_disease] Lou Gehrig's _disease [0https://en.wikipedia.org/wiki/Lou_Gehrig%27s_disease]"

(source:='https://en.wikipedia.org/wiki/Aging_brain' time:='Sat Oct 17 10:32:36 2015' extract:=0) Q-UEL-Xtractor34 >

< 'neurogenerative disease' | includes | 'mild cognitive impairment' | and | Altzheimer's disease' | and | 'Parkinson's disease' | and | 'Lou Gehrig's disease' >

Both the above are Q-UEL-compatible tags [22,27]. The former would ultimately be removed to avoid redundancy and to free up storage, although if both are reasonable quality representations of the same knowledge, leaving the XTRACT has been found not significantly impair exam performance. They can certainly be ordered and partitioned into categories for human convenience, as with the "relevancy sets" of POPPER [22], but MARPLE does not require it. Recall that Q-UEL tags relate to Dirac notation and so are formally algebraic objects with a dual probability value [20,22,27]. Again, if a probability value is not explicit, the value is 1. In this report, tags come from DiracMIner [23] in the structured case and XTRACTOR [27] in the unstructured case, and also from Q-UEL archives especially those that have been built up by POPPER over the past five years, which includes a lot of knowledge entered by human experts using the POPPER HELPER interface [22]. These are essentially already MARPLE KRS elements, differing only in the format for presenting probabilities, so that conversion is trivial. They comprise some 800,000 well curated clinical KRS elements from structured data mining sources including patient medical records such as Ref. [5] and public health studies [23], plus several million of a pharmaceutical nature. It also includes older direct entries by human experts using the POPPER HELPER interface [22]. However, since MARPLE 1 came into operation, the great majority of entries started out as XTRACTs gathered only over the last year and subsequently curated automatically in MARPLE or in POPPER using metastatements [22] (Section 2.3), or in troublesome cases manually by POPPER HELPER. When reporting the nature and numbers of both well curated KRS elements and XTRACTs below, only those obtained in the present study are used and counted unless stated otherwise, in which case the older preexisting elements are referred to as *legacy knowledge elements*.

3.2. MARPLE's modes of use and their control

Usually, MARPLE 2 works in three well distinguished modes.

(1) "Popper" mode – Offline during exam. Uses "medical encyclopedia" of well curated Q-UEL "Popper tags", originally from other sources but validated at POPPER HELPER [22] as knowledge elements, if needed.
(2) Hybrid Mode – Online during exam. Uses above curated tags *and* searches Internet to supplement the curated tags with precurated XTRACT tags [27]. It keeps XTRACT tags for curation, to add to future knowledge. This is the common mode in routine use.
(3) Crude Web Mode – An alternative mode for testing and comparison purposes, online during exam. This is not allowed to use the "medical encyclopedia" of well curated Q-UEL tags as knowledge elements. It keeps XTRACT tags for curation, to add to future knowledge.

Both 1 and 2 really represent learning modes. Knowledge content from curated tags is seen as *supervised learning*, as from lectures that adhere to rigid syllabus, while crude XTRACT tags obtained from surfing the Web can be considered as *unsupervised learning*, like student reading of text books and journals. In addition, these modes and associated files can be manipulated so that information gathered in "training" MARPLE to answer classes of question can be removed and results compared. Because this often involves taking out information related to test questions but putting it back into answer others, "experiments" of this kind are called *jackknifing*. The flow of information in such modes and studies is influenced by input.txt. Normally, this file contains just a link (URL), or an html page, or natural language text extracted from a web page or, in the case of the present study, a query to the Wikipedia system. If it commences with an HTML DOCTYPE specification it is taken as a page in HTML, and if it starts with http:// or https:// it gets the required page in HTML. If it is the former and it contains signs that it is page in response to a query to Wikipedia, and that it has successfully found a relevant Wikipedia entry, then subroutine Xtract extracts the first reference to that entry. Further control as to specific sites can be exerted by changing a text variables called $HitList and $StartList. Normally, a link is memorized in $HitList, so that the link, and hence the corresponding web page, is only used once. If a query via the examination answer and topic relevance functions persistently fails on relevance tests it can be reset is to restart surfing from a list of URLs to start from, and judged by the end-user as likely to be productive, e.g. 'https://en.wikipedia.org/wiki/ Pathology'.

3.3. The curation cycle

The overall curation process is one of progressive refinement most succinctly described by the reference to three files used cyclically as follows.

(1) *The Prior Knowledge File* (PrK) is the only KRS file used in Popper mode (i). It is held on a file usually called knowledge.txt and comprises knowledge obtained mainly from the Internet as above, tided by curation to cover progressively the medical licensing board syllabus taught and examined.
(2) *The Posterior Knowledge File* (PoK) file is input and is a KRS file usually called knowledgeDynamic.txt. It contains the tags generated by the XTRACTOR component. It is normally this file that is read is read for each answer to each question in the examination. In the hybrid mode (ii), the PrK file, usually

[4] Here and in results below, we will not usually follow our Q-UEL convention of writing operators, including relationship expressions, in bold font, but rather as they look on the basic ASCII flat-file. Some exceptions will be made for readability. Q-UEL applications can display the bold font method, however, and organize the layout generally [28], not unlike what happens when one opens an .xml file.

knowledge.txt, is copied to this file to initialize it, i.e. before it is extended by KRS element tags by the action of the XTRACTOR. In other words, it contain well curated PrK conent, and the rougher new XTRACT content extended progressively.

(3) Curation is usually applied to a *curated Knowledge File* (CuK) curated.txt which is output as far as initial main runs are concerned, but becomes the input PrK in future runs, which therefore grows in a cyclic way. It arises because both on writing and reading the above XTRACT tags in the course of an examination, there is correction, repair and curation so that tags have correct canonical form, by methods described in Section 3.2.

In a larger cycle over longer periods, knowledge is constantly usefully flowing between MARPLE, POPPER [22] and DiracBuilder [23], because they all create, curate and use and test knowledge in somewhat different ways.

3.4. Dictionaries

There are also a number of auxiliary files of the character of dictionaries that not only play more obvious important roles in natural language processing but also in assessing relevance of text sources. The *nominators file* supports determiners and other similar words that appear in noun phrases but are not themselves nouns, e.g. *The, A , An, This, That, These, Those, All, Both, Half, Either, Neither…* etc. The *relators file* is more structured in content and contains entries such as: < *irregular verbs* >, *awake, awoke**, *awoken** , *be**, *being**, *am**, *is**, *are**, *was**, *were**, *been**, *beat, beaten**, *become, became**, *begin, began**, *begun**, … etc. Here the asterisk indicates irregular forms that cannot be deduced from roots by word generation routines inherent in XTRACTOR [27]. The *compounds file is* essentially a collection of commonly associated words that imply a compound concept, primarily verbal, prepositional, some verbal and some noun phrases, and particularly those that are useful in describing relationships, such as "in the absence of". If it is empty it will be created in a run and contain at least computed legal or potentially legal combinations of words that can serve as relators according to English grammar rules, as computed by MARPLE, and more elaborate multiword entries and phrases drawn from the relator.txt file. The *buzzwords file* is so-called because it contains "buzzwords" characteristic of the general topic of interest, here medicine and related topics like statistics. This file includes a comprehensive dictionary of medical roots, words, and phrases. *aden/o gland, adenoid/o adenoids, adip/o fat, adren/o adrenal gland … etc,* The *badwords file* contains roots words and phrases common in web pages that are not of the above topic type, i.e. in the present case it contains roots, words and phrases prominently found in non-medical text on the web. Examples include: *restaurant, menu, beds, hotel, garden, holiday, vacation, guest, blog, log in, TV, drama, news, quotes, FaceBook, Twitter, vacation, music, spectacular,…*Note that it is a matter of balance of evidence so just a few occurrences of words from the latter will not prohibit access.

3.5. Q-UEL algorithms used

The main subroutines that are used in MARPLE 2 are as follows. We expect that the algorithms and service functions that they represent will be necessary in any comparable approach.

GetQuestion - Reads the exam paper (typically on exam.txt) and extract question, answers, and indication of the correct answer.

GetCrudeKnowledgePrior – Examines the KRS file which with Popper Mode set on is the PrK file, usually knowledge.txt. However, with Popper Mode set to "no", it is the PoK file which will progressively fill with tags extracted by XTRACTOR from the Internet, and which *GetCrudeKnowledgePrior* will take as input. The basic counting required to evaluate the common content function is directly computed in this routine and is set to score +1 for field and word hits score as described in the Theory Section 2.4, but up to two additions of an adjustable bonus score taken as 10 in the present study as follows. However, in all cases below, any added score including bonus is negative if the tag being examined contains at least one negative relator filed, e.g. "is not", "do not" etc. Initially a tag is indicated to its general relevance as (a) whether or not it has a field in common with the question, and (b) whether or not it has a field in common with any one of the set of all answers for that question. In general, we say that A is a field in common with a field in B if it is the equivalent string, or one is the substring of the other, noting that an effort is made to place the field in a standard canonical format, removing determiners and trivial noun phrase words, and finally regularizing whitespace. If a field of a tag is found in the question it scores a 1 for a particular candidate answer if there is a common field with that candidate answer or with the last tag that had a field in common with that candidate answer. If a field of a tag is found in a candidate answer it scores a further 1 if there is a common field with the question or with the last tag that had a field in common with the question. If a tag is found to have a field in common with the question and also a field that is in common with a specific candidate answer, it scores the additional bonus (here, 10). It will score a second bonus (again, here 10) *both* (a) and (b) above were satisfied. Recall that the score are actually estimates of (positive or negative) mutual information and are all added prior to deducing the probability as an exponential of information, followed by normalization of probabilities of all answers so that the sum to 1. The above is applied to fields as stated, and then to individual worlds, which can be considered as small fields.

GetCrudeKnowledge – This is very similar to *GetCrudeKnowledgePrior*, and is applied after it to *increment* still further the score according to exactly the same scoring rules (at least in the current settings). So, for example, with the current settings the tag could score 10 + 10 in *GetCrudeKnowledgePrior*, and an addittional 10 + 10 in *GetCrudeKnowledge*. The main difference from *GetCrudeKnowledgePrior*, apart from reporting MARPLE'S reasoning in terms of promising knowledge representation tags found, is that as well as *PopperTidy* it also calls the following.

GetIntrinsicKnowledge – The curation step that introduces or curates probabilities assigned to KRS elements. Essentially DiracBuilder [22], introducing the intrinsic probabilities into the calculation but confined to consideration of those tags that can be detected and scored by *GetCrudeKnowledge*. The intrinsic probabilities are in this case are presented as values of the Pfwd and Pbwd attribute on tags in the KRS, and as usual, the default for a probability is 1 if a probability is absent. This routine calls others so that the contextual probability found for each answer is multiplied by the product of the probabilities found in each path composed of tags from question to specific candidate answer, but if there are parallel paths these are added together before the contextual probability associated with each answer is multiplied by the result. Since intrinsic probabilities contribute little to the exam problem, and very similar results are obtained if they are removed, this will be discussed in more detail elsewhere.

GetCrudeHitSCore – Evaluate the common content function. To do this, this routine calls the *Zeta* function. In most of the studies, a prior virtual frequency of 50 was used, meaning that conceptually a prior virtual frequency of 50 is added actual count before entering the zeta function (although in MARPLE it is added within the function).

Zeta - As required for the above, although with settings used in this present study, this is more precisely the z function [22], i.e. for $s = 1$ the value γ n/(n+1) is subtracted from the incomplete zeta function $1 + 1/2 + 1/3 + \ldots + 1/n$ where n is a count such as N_{F+} *plus the prior virtual frequency*. Here $\gamma = 0.5772\ldots$. is the Euler-

Mascheroni constant. Whether the ζ (zeta) or z function is used makes little difference in the present study.

"The Curation Aids" comprise a bundle of routines that relate to automated curation. *PopperTidy* is a simple automatic curation always called when knowledge representation tags are read in to insure they are not severely corrupted and adhere sufficiently to canonical form to be used by the internal working of GetCrudekNowledgePrior (see above). PopperTidy is adequate for use of curated tags already curated via POPPER. It also does basics such as recognizing probability assignment statements in POPPER language code [22], and re-express them as Pfwd and Pbwd attributes as the format preferred in MARPLE KRS elements. *OpenPopper* calls POPPER [22] for automatic curation by metastatements [22] but POPPER HELPER [22] can intercept for manual curation as well as creation of new metastatements. *OpenPopperHelper* calls POPPER HELPER directly, but that is primarily for manual curation, although it does contain contain some automated support tools such that manual component is helped or minimized [22]. *Xtractor* essentially calls XTRACTOR [27] to search the Internet, but XTRACTOR itself does the preliminary curation that allows XTRACTs to be used directly, and *Xtractor* in MARPLE does some further checking tidying. The results remain XTRACTs. In contrast, *XtractTidy* now automatically curates *all* tags found to "well curated KRS element" As XtractTidy is general its name is somewhat a misnomer, but its most dramatic effect is on XTRACTs to STs and LSMs. It is the more sophisticated and slower curation routine that is applied to XTRACTs in the KRS at longer intervals in curation cycles, and so it does not typically act on XTRACTs just extracted. It will reject tags from use if they cannot be fully understood or repaired. *VerifyField* verifies that a field, such as a noun of verb phrase, is correctly constructed. It is used to help curate XTRACT tags. *VerifyOption* verifies that an answer option is correctly constructed and can extract essential content that can be used as a query, but it is also now applied to curate strings that are noun phrase and relationship phrases in XTRACTs. *ValueToHighNormalLow* converts clinical values to standard high, normal, and low ranges. It can be used to help curate both XTRACT and POPPER tags in that respect. *NotNegated* establishes that a relation is at least assured not to be of negative form, such as "is not". It can used to help curate XTRACT as well as POPPER tags. *Detrivialize* curates a relationship field to contain only important words. While "not" is highly significant, nominators presnt and words like "somewhat" etc. are not. It can used to help curate XTRACT as well as POPPER tags, although normally the option to suppress use determiners such as definite and indefinite articles is done by ignoring them, not removing them.

4. Results

4.1. Complexity of solutions and a simple example case

Some 200 questions have been attempted so far with a detailed analysis of performance in regard to 50. Overall performance in terms of percentages of questions answered correctly is described immediately below in Sections 4.2–4 and in Discussion and Conclusions Section 5, but to understand that performance as obtained various conditions, it is helpful to consider first a simple example requiring just one definitional step to answer the question. We now find that about 30% of questions are definitional in the sense of requiring just one knowledge element between question and answer, and 65% clearly two, but that really depends on the run because it is possible to make the connection with one link, or more, depending on what is available. More importantly still, an LSM can be interpreted as a series of fused simple links, i.e. of STs, and the complexity of LSMs or XTRACTs as cruder LSMs begs the question of how many links it really represents, if it is not simply one. However, direct inspection does reveal that some 15% of cases in our study, like the following, are clearly straightforwardly definitional in any reasonable sense, irrespective of the above considerations.

Question 1.

A laboratory has developed a new test for rapid ascertainment of serum parathyroid hormone levels. The test is repeated twenty times on the same sample with a resulting coefficient of variation of one percent. This is a measure of which one of these?

(A) Accuracy
(B) Reliability
(C) Precision
(D) Validity
(E) Mode

The output regarding this question consisted of restating the above, followed by

```
STRONG PRO-CLUE FOR < Reliability > FROM KNOWLEDGE < reliability | is | measure | of |
reproducibility | of | test | in | different conditions >

STRONG PRO-CLUE FOR < Reliability > FROM KNOWLEDGE < coefficient of variation | is |
standardized measure | of | reliability >

STRONG ANTI-CLUE FOR < Accuracy > FROM KNOWLEDGE < coefficient of variation | is not |
standardized measure | of | accuracy >

STRONG ANTI-CLUE FOR < Precision > FROM KNOWLEDGE < coefficient of variation | is not |
standardized measure | of | precision >

STRONG ANTI-CLUE FOR < Validity > FROM KNOWLEDGE < coefficient of variation | is not |
standardized measure | of | validity >

STRONG ANTI-CLUE FOR < Mode > FROM KNOWLEDGE < coefficient of variation | is not |
standardized measure | of | mode >

STRONG PRO-CLUE FOR < Accuracy > FROM KNOWLEDGE < accuracy | is | measure | of | degree |
to which| test | approximates | real value | of that which is | measured >

STRONG PRO-CLUE FOR < Validity > FROM KNOWLEDGE < validity | is | assessment | of | degree | to
which | test | measures | real value | for which it was | designed >
```

(A) (P=17.45%) Accuracy
(B) (P=45.44%) Reliability
(C) (P= 9.82%) Precision
(D) (P=17.45%) Validity
(E) (P= 9.82%) Mode

Question 1. Predicted best answer is B.
According to examiner, the correct answer is B.

Note that a full report of "reasoning" by MARPLE is not usually given because there can potentially be a very large number of tags present that had some relevance to answering the question and which can contribute scores increase or decrease the probability of each possible answer. That is especially so when XTRACT tags are obtained from surfing the Internet as questions are encountered. The stronger of these less strong relevancies are also reported, as WEAK PRO-CLUE and WEAK ANTI-CLUE, when available, though there were no such examples in this case. Whether or not the Internet is to be accessed, the knowledge tags used by MARPLE ultimately come from various sources of which Wikipedia text is prevalent, processed by the XTract subroutine (recall, essentially the earlier XTRACTOR application). A typical extract generated in response to the above question is

< Q-UEL-Marple17 "statistics [0https://en.wikipedia.org/wiki/Mode_(statistics)] |includes| Mode |is| `the `most `common _value |among| `a _group"
| extracted from |
(source:='https://en.wikipedia.org/wiki/Mode' time:='Fri Oct 30 20:02:06 2015' extract:=82) Q-UEL-Marple17 >

Curation via POPPER HELPER [2] would traditionally result in two key semantic triples

< statistics | includes | mode >
< mode | is | the most common value >

and not surprisingly these two tags, and in the absence of too much "noise" from other tags (see below) the second alone, would be sufficient to answer "What is the mode as the term is used in statistics?" with an answer "The most common value". In relation to MARPLE, however, it is common to prepare linear semantic multiples (LSMs) and pay attention to the negative relational forms. This has become extensively automatic in the past few months, including by use of POPPER metastatements as editors [22], which helps avoid the subjective risk of "forcing" the correct answer (Section 4.6). In the following from the KRS used, there is perhaps suspiciously more emphasis on what something is not the case in the following set, because the computer has no initial reason to presume that all alternatives cannot be sets to which 'coefficient of variation' does not belong,

< the 'coefficient of variation' | is | a standardized measure | of | dispersion | of | a 'probability distribution' >
< the 'coefficient of variation' | is | a standardized measure | of | dispersion | of | a 'frequency distribution' >
< the 'coefficient of variation' | is | a standardized measure | of | reliability >
< the 'coefficient of variation' | 'is not' | a standardized measure | of | accuracy >
< the 'coefficient of variation' | 'is not' | a standardized measure | of | precision >
< the 'coefficient of variation' | 'is not' | a standardized measure | of | validity >
< the 'coefficient of variation' | 'is not' | a standardized measure | of | mode >
< the 'coefficient of variation' | 'is not' | a standardized measure | of | median >
< the 'coefficient of variation' | 'is not' | a standardized measure | of |'central tendency' >
< the 'coefficient of variation' | 'is not' | a standardized measure | of | range >
< the 'relative standard deviation' | is | a standardized measure | of | dispersion | of | a 'probability distribution' >
< the 'relative standard deviation' | is | a standardized measure | of | dispersion | of | a 'frequency distribution' >
< 'inter-observer reliability' | is | reliability >
< plit-sample reliability' | is | reliability >
< 'repeat testing reliability' | is | reliability >
< accuracy | is | a measure | of | the degree | 'to which'| a test | approximates | the real value | 'of that which is' | measured >
< a test | 'is measured against' | the gold standard >
< validity | is | the assessment | of | the degree | 'to which' | a test | measures | the real value | 'for which it was' | designed >
< precision | is | the degree | 'to which' | a measurement | 'is not subject to' | 'random variation' >

However, the importance of negative 'is not' form is in this case misleading. In fact, the negative forms are there because they are important to help answer other questions of similar type, not drive the correct answer to a particular questions (Question 1 above). If the tags containing 'is not' are removed, are removed, the following output is now obtained in which the chosen answer is still correct.

```
STRONG PRO-CLUE FOR < Reliability > FROM KNOWLEDGE < reliability | is | measure
| of | reproducibility | of | a test | in | different conditions >

STRONG PRO-CLUE FOR < Reliability > FROM KNOWLEDGE < coefficient of variation |
is | standardized measure | of | reliability >

STRONG PRO-CLUE FOR < Accuracy > FROM KNOWLEDGE < accuracy | is | measure | of |
 degree | to which| test | approximates | real value | of that which is | measured >

STRONG PRO-CLUE FOR < Validity > FROM KNOWLEDGE < validity | is | assessment | of | degree | to
which | test | measures | real value | for which it was | designed >
```

(A) (P=21.62%) Accuracy
(B) (P=32.42%) Reliability
(C) (P=12.17%) Precision
(D) (P=21.62%) Validity
(E) (P=12.17%) Mode

Question 1. Predicted best answer is B.
According to examiner, the correct answer is B.

4.2. Summary of overall performance

The ongoing evolution of the KRS and the number of combinations of algorithm variation, input, and time limitations on searching the Internet, all lead to a very large number of potential results. They are best presented as they were obtained, as the results of "computer experiments" addressing specific issues of particular interest. Relevant computer experiments are discussed in Sections 4.3–4.11. The most important areas are as follows.

4.2.1. Curation studies

The most important results as far as curation for CDS purposes is concerned are as follows. 100% is now readily obtained by MARPLE 2 providing they are based on qualitative knowledge and not of calculation type, compared with 80% obtained by MARPLE 1 in similar conditions [3]). We consider the knowledge elements to be well curated as far as MARPLE and the exam is concerned. In the last run at the time of writing, the above100% was obtained using 2435 well curated KRS elements with, and without, previously obtained XTRACTs, and with, *or without*, new XTRACTs obtained by surfing the Internet during the exam, e.g. 11,500 XTRACTs in the last run. In that particular run the number of well curated KRS elements is about 12-13 on average per question and when Internet searching is allowed in addition, the number of XTRACTs is on average about 60 per question. The 12-13 is notably close to the average number of answers per question (13), but in a few cases just one may answer a question, and averaged over some 50 runs in the recent past required a minimum average of 16 per question. However, in some runs some questions appear more recalcitrant for no immediately obvious reason that is intrinsic to the question, and an average of 122 well curated elements per question has been needed to guarantee that exactly 100% is always obtained. The actual results will depend on the number of XTRACTs obtained and hence on the duration of Internet searching. In normal operation, indefinite time is allowed to gather knowledge, and the limiting factor is either when the questions so far are all answered correctly, or because overall performance makes worse the "noise" as a consequence of too many elements and inability of curation to keep up as discussed in more detail in Section 1.6 and Section 5, but for example, 7,200,000 XTRACTs drops the 100% to 65%. For purposes of discussing exam performance as a kind of "exercise in AI", 30 seconds "surfing" is allowed to consider each candidate answer, comparable to an exam for a human student, typically of 50 questions each of 5 answers to be answered in about one hour. See Section 4.3 for an overview of "timing" benchmarks and how they relate to XTRACTs. Noise is still seen as arising mainly from XTRACTs because 96% to 100% is obtained if approximately 1 million well curated (or at least better curated) *legacy knowledge elements* from clinical data mining (Section 3.1) are included on the KRS. The set of 2435 well curated elements used to obtain 100% at the time of writing appears fairly typical in the character of the elements and the number of them, and appear to be close to a *minimal* set required because of the following preliminary results. The "computer experiment" of removing some elements at random to produce a reduced set of 2000 drops the score to 90%, and to 1000 gives 55%. Removing all content from the KRS, i.e. well curated KRS elements plus XTRACTs, gives 8% which is essentially what is expected on a random basis given an average of 13 candidate answers per question. If MARPLE 2 is then allowed to search the Internet during the exam, then 40% is obtained. Note that in this case the only sources of knowledge are the freshly obtained rougher XTRACTs, as yet without fuller curation.

4.2.2. Examination "Blind" studies and jackknifing

The use of "Blind Studies" in contrast with "Curation Studies" above is somewhat misleading because even there the official correct answer is not inspected and taken into account until after an attempt has been made to predict the answer. Curation Studies are really best considered as relating to performance after extensive good curation has been applied, while "Blind Studies" are in practice the same but with performance recorded at the moment that a new question is introduced. Blind Studies are also those in which emphasis is put on testing the "smartness" or otherwise of MARPLE, done by somehow setting the "test" questions apart from the set of "development" questions. That is, the questions previously used in learning are removed from training set for each relevant question. Under those conditions, an exam score of 73% is obtained. That is, providing the Internet can be searched. Recall that we consider that fair because it is the ability to use the use the Internet as a "memory extension" when needed allows MARPLE to run on a standard personal computer. The precise meaning of this result should be explained, due to the number of possible ways, all fair tests, by which it could have been achieved (e.g. the learning machine approach in Ref. [20]). So that we can work with a stable fixed set of questions for the purposes of the present paper, we used *jackknifing* in the present study. One question is taken out at a time and the system is retrained in its absence, and then that question is placed back into the exam and tested as if seen for the first time. Many changes in conditions can affect the score. The above 73% is more precisely a specific if typical last run at the time of writing, using 1641 well curated KRS elements plus 14,612 new XTRACTs gathered from the Internet during the exam, and no older XTRACTs. This kind of result is only slightly more sensitive to minor variations as discussed for Curation Studies, although in earlier studies with fewer questions and fresh Internet search for each exam variously 55%-86% was obtained in otherwise similar conditions (there was a negative skew and a mode at about 77%). The fall from well curated KRS elements was in this case from the above 2435 of the Curation Studies to 1641, because we removed those "well curated KRS elements" that arose by the curation of XTRACTs obtained from the Internet in earlier response to each question that is removed.

4.2.3. Effects of stronger forms of jackknifing

In the above jackknife tests there are always many well curated KRS elements present, although none that were curated to answer the question now being tested. When Internet searching was not permitted reliance is totally upon those and these well curated KRS elements above 73% drops to 44%, recalling that this is in jackknifing conditions. We call this *strong jackknifing*, since is rather like the medical student who neglected to study the topic that came up, and has no way, during the exam, to get the knowledge needed. There is not an obvious common code that is transferable form one topic to another, as there can be in some prediction problems [20]. However, it is a fuller curation of XTRACTs that ultimately produces most of the well curated KRS elements that are present, and indeed the above has recently increased to 47% with ongoing curation of the XTRACTs. Clearly some relevant knowledge has been got, but this is essentially unsupervised learning that has picked up relevant information, but not by being directed to it by MARPLE in tackling a question. Earlier, when repeating with slightly varying conditions we obtained an average with a mean of 45% and standard deviation of about 10%, indicating more sensitivity dependent on the number and relevance of the KRS elements. Conversely, if all KRS elements are removed, and only the XTRACTs newly found in the exam are allowed, the 73% drops to 40%, the result that was already quoted at the end of Section 4.2.1 on "Curation Studies". It is said to express *extreme jackknifing*, as there is no opportunity to learn by searching for knowledge directly related to the question. See also Sections 4.4 and 4.5 for other conditions that reduce this 73% score.

4.3. Speed benchmarks

MARPLE is already significantly faster than a human taking the examination. For conditions of the first row of the Table, using an older Dell Vostro 320, MARPLE tacked questions correctly at a rate of 1.65 seconds per question, and 0.2 to 1.2 seconds per question on a variety of other personal computers. A very large piece of text rendered as a question, say 1-3 standard pages of text from the start of a published case study report, takes about 3 seconds. There is no well-defined distinction as yet between the time required to give the correct answer and the time taken when giving an incorrect answer when only well curated KRS elements are used, rather than Internet searching which along with curation is the slowest component function at present. When not in Popper mode, which applies to all rows below the first except the last, Internet searching is allowed (and obligated) and can take 2-3 minutes a question, of which about half is due to practical Internet access rate issues and half to the above basic degree of automatic curation. As of recently in this study, some 50,000 XTRACT tags can be readily be assembled and subject to some limited degree of automatic levels of curation in about an hour, and in practice roughly 20,000 such new tags can so be generated in a typical exam. One can potentially obtain some 1,200,000 XTRACTs per day allowing or forcing MARPLE to continuously search the Internet, but that number can vary considerably for two main reasons. First, on some occasions, a new questions can prove more troublesome and may take about 3-4 attempts before suitable XTRACTs or KRS elements curated from them are found, and these 3-4 attempts are together typically associated with generating some 60,000 tags taking about 15 minutes per question in very troublesome cases. In very incalcitrant cases a search might precede for an hour to answer a question, but that introduces a practical limit because at around 180,000 XTRACTs the score with jackknifing starts to fall from 73% to 71% and then continues to deteriorate. Straightforward learning without jackknifing, however, requires about 700,000 XTRACTs to fall to a similar score. Note that continued searching even if exam performance deteriorates still produces XTRACTs that can be put aside for later curation. Second, there are always practical considerations for time of day, location, and trace route. Notably, pinging Wikipedia in studies performed from a domestic residence in the Cayman Islands is not always so fast and significantly slower than pinging Google, on average some 200ms per packet response compared with 20ms for Google.

4.4. "Noise", negation, determiners, and intrinsic probabilities

In the most recent example run, with jackknifing to remove test questions from the training set (Section 4.2), the 73% score 1641 KRS elements fell to 71% with 183,000 XTRACTs and then to 65% with 7,200,000 XTRACTs, though it has subsequently stayed stable at 65%. Of greater concern from an automated curation perspective is that without jackknifing the performance of 100% with based on at least 2435 well curated KRS elements, and 0 to 11,500 XTRACTs based on at least 2435 well curated KRS elements, falls to 90% at about 14,000-15,000 XTRACTs, and to 80% at 173,200 XTRACT tags. This value has persisted at time of first writing for several million XTRACTs, but recently fell to score of 66% when using 7,200,000 XTRACTs, beyond which now appears to be stable. These recent results seem fairly typical in reruns with minor variations. The finer information-theoretic implications of this are still unclear and require further "experiments", but in drilling down to identify XTRACTs that appear responsible for forcing wrong answers, some 20% of the problematic statements that were not explicitly negative have so far been found to be negative or low probability by implication or context. For example, XTRACTOR and MARPLE at the present time will still inevitably have trouble with constructions like "Contrary to the evidence accumulated since the late seventeenth century, it was long believed that vision with one eye was superior to that with two". If KRS element tags representing negative statements are discarded from the curated KRS, the exam performance falls to 56%, so determiners are important through curation when a relationship is of a negative nature. Preliminary work has been in assigning words or phrases such as "very few" as negative, and in general using an extension of the determiners associated with negative information as shown in the screenshot of POPER HELPER in Fig. 1 of Ref. [22]. The direct beneficial effect is on noise. At 7,200,000 XTRACTS the noise effect previously led to 50% exam score even in the absence of jackknifing, but rose to 65% with the above greater care over negatives. Intrinsic probabilities can be taken into account by multiplication with contextual probabilities; the finding is that any reasonable implementation has little effect on exam performance except for negative forms, i.e. providing that reasonable values reflecting the positive and negative sense of a statement were assigned. Paying attention to purely positive qualifiers and ensuring that such forms result in positive statements, has improved exam scores when "noise" is present from many extracts. However it is only by a modest *circa* 2% on average. Drilldown to identify difficulties encountered with intrinsic probabilities and estimates of them based on determiners shows that better treatment is required in expressing relationships using intrinsic probabilities based on scope. For example, that no Americans have eye cancer would seem a good approximation of the probability 0.000008 based on prevalence, and injudicious use of such knowledge in inference risks eliminating such a diagnosis as ever being possible in the exam context where contextual probabilities dominate and "life is relatively simple", i.e. of a more binary nature.

4.5. Changes that impair exam performance

The current approach seems more-or-less optimal in the sense that significant changes in methodology impair exam performance. Contrary evidence is important, and at first we suspected that this was due to better managing the noise arising from use many XTRACTs, because the structure of the formula was designed to focus on cases of more extensive data and focusing on the balance of evidence from it. Indeed, deleting all negative well curated KRS elements (e.g., containing relationships such as "is not associated with"), with 7,200,000 XTRACTs, recently dropped the score from 66% to 56%. Instead, simplifying the common content function by dropping the last two terms gave a similar 55%. Keeping the last two terms but dropping the prior frequency used in the zeta function to 10 or less gave 50%. However, in the first two cases, similar final values are obtained even when dramatically decreasing the number of XTRACTs, suggesting that a very significant effect is on the well curated KRS. In contrast, using a prior of more than 10 uniformly for the prior frequencies in the zeta function had relatively little effect on exam performance when no XTRACTs were used, suggesting that the choice of prior frequency, without other changes, has an important role in damping the "noise' from many XTACTs. Somewhat surprisingly, without jackknifing for 2310 well curated elements and 7,200,000 XTRACTs, dropping the last two common word terms from the common content function *and* dropping prior frequency in zeta function to 10 or less at the same time dropped the exam score to 24%. There appears to be some clue to this in the result of 24% that is also obtained with all well curated KRS elements removed and deleting XTRACTs containing significant words (primarily meaning other than determiners “a”, “the”, “some” etc.) that are found in both the question and answer set. This involved 6,110,433 XTRACTs, the drop in number being due to the above deletions. While some findings still beg detailed explanation, all this would seem to confirm that the full form of the common content equation combined with a strong damping effect from a uniform prior frequency is generally important.

4.6. Forced correctness

Even when obtaining 100% by full use of curated KRS elements, we try to avoid excessive introduction of KRS elements that “force” the correct answer, not least because it does not ensure good curation for CDS purposes. This is most easily policed by only allowing ourselves to curate elements of knowledge that show up as XTRACTs or preexisted in the Q-UEL system prior to use of the current exam set, except in the most recalcitrant cases (see below). We would estimate from “computer experiments” that some 25% of new questions required some unusually high degree of attention to refinement of existing KRS element tags and importantly there was in those cases usually the need to include others with more radical modifications that might suggest a degree of forcing. This represented 40% of cases that have separate questions but an answer set in common (See Section 5.2). The above “recalcitrant cases” are extreme or controversial examples, about 6% of cases. This is being reappraised, however, because it is now noted that many corresponded to introduction of negative statements such as < X | ‘is not associated with’ | Y > to ensure that wrong answers Y are avoided, and this turned out not to have as much impact of scores as might be expected (see Section 5.2). It is, of course, a well justified inclusion if X is indeed not associated with Y in real world medicine, which is the ultimate whole point of curation. This normally turns out to be the case because the exam questions typically thoughtfully and fairly constructed by their authors, and reflect real world medicine. When curation does look like forcing, it most often naturally arises in those cases in which the exam question is not correct, or more likely misleading or ambiguous. Examiners preparing questions in the style of licensing examinations, and medical licensing boards, can appear to adopt definitions and classifications that are to some degree consensus in their community but nonetheless reflect an arbitrary rather than fundamental justification. For example, the question

QUESTION 2:

You are doing a study on the distribution of IQ scores in 15-year-old adolescent males in a standard high school classroom. You have chosen one school from Los Angeles, Seattle, Dallas, Miami, Chicago and New York. The WISC-III is administered to all 15-year-olds in the schools selected. After all tests have been administered, the scores are collected and the distribution of the scores is analyzed. The IQ scores represent what type of statistical measurement scale?

ANSWERS:

(A) Nominal
(B) Ordinal
(C) Interval
(D) Ratio
(E) Correlational

required that “(C) Interval” is the correct answer according to the original question setter, but XTRACT tags often indicated otherwise in the opinion of the Web, and there was preexistence of older KRS elements such as < a negative measure | ‘is not’ | ratio scale >, < IQ | is | a positive measure >, and < IQ | ‘is not’ | a negative measure >. Consequently the following more specific definitions were required to override them.

< IQ | is | an interval scale >
< IQ | 'is not' | an ordinal scale >
< IQ | 'is not' | a nominal scale >
< IQ | 'is not' | a ratio scale >
< IQ | 'is not' | a categorical scale >
< IQ | 'is not' | a correlational scale >

The adding of layers of *specificity* as negation, over the top of general statements of knowledge that provide *sensitivity* that is typically (but as above not always) characterized by positive statements, does not in general terms seems so different from the way a human student is taught, but it is does not necessarily guarantee the truly correct knowledge in some more absolute sense, in every case. However, it also remains that teaching human students in lecture and text tends to emphasize what *is* the case. What is not the case is often tacit, typically requiring the appearance of clearly negative statements as well as positive ones for KRS elements. Apart from strong cases like the above, any extra KRS elements required almost always passed a test of reasonableness. They represented what would, in effect, be taught to a human student.

4.7. Quantitative questions and answers: normal ranges

In MARPLE 2, a subroutine is present that has knowledge of normal ranges of standard clinical values (it is currently being extended). Consider for example the following.

Question 19.

A 22-year-old man with a 3-week history of polyuria and polydipsia has had nausea, vomiting, and decreased responsiveness for the past 12 hours. Urinalysis (dipstick) shows 4+ glucose and 4+ ketones.

A. pH 7.15 PO2 mmHg 98 PCO2 mmHg 33 HCO3 mEq/L 11
B. pH 7.15 PO2 mmHg 98 PCO2 mmHg 24 HCO3 mEq/L 8
C. pH 7.30 PO2 mmHg 56 PCO2 mmHg 80 HCO3 mEq/L 38
D. pH 7.40 PO2 mmHg 100 PCO2 mmHg 40 HCO3 mEq/L 25
E. pH 7.50 PO2 mmHg 100 PCO2 mmHg 33 HCO3 mEq/L 25
F. pH 7.50 PO2 mmHg 100 PCO2 mmHg 24 HCO3 mEq/L 18
G. pH 7.50 PO2 mmHg 56 PCO2 mmHg 33 HCO3 mEq/L 25
....
A. (P=23.18%) pH 7.15 PO2 mmHg 98 PCO2 mmHg 33 HCO3 mEq/L 11
B. (P=47.97%) pH 7.15 PO2 mmHg 98 PCO2 mmHg 24 HCO3 mEq/L 8
C. (P= 3.49%) pH 7.30 PO2 mmHg 56 PCO2 mmHg 80 HCO3 mEq/L 38
D. (P= 3.21%) pH 7.40 PO2 mmHg 100 PCO2 mmHg 40 HCO3 mEq/L 25
E. (P= 3.21%) pH 7.50 PO2 mmHg 100 PCO2 mmHg 33 HCO3 mEq/L 25
F. (P=15.72%) pH 7.50 PO2 mmHg 100 PCO2 mmHg 24 HCO3 mEq/L 18
G. (P= 3.21%) pH 7.50 PO2 mmHg 56 PCO2 mmHg 33 HCO3 mEq/L 25

To answer this question, value conversions occurred as follows.
A. (P=23.18%) ph low po2 mmhg normal pco2 mmhg normal hco3 meq/l low
B. (P=47.97%) ph low po2 mmhg normal pco2 mmhg low hco3 meq/l low
C. (P= 3.49%) ph normal po2 mmhg low pco2 mmhg high hco3 meq/l high

D. (P= 3.21%) ph normal po2 mmhg high pco2 mmhg normal hco3 meq/l high
E. (P= 3.21%) ph high po2 mmhg high pco2 mmhg normal hco3 meq/l high
F. (P=15.72%) ph high po2 mmhg high pco2 mmhg low hco3 meq/l low
G. (P= 3.21%) ph high po2 mmhg low pco2 mmhg normal hco3 meq/l high
Question 19. Predicted best answer is B.
According to examiner, the correct answer is B.
(14,763 conversions of values to high/normal/low ranges occurred)

The program reports conversions as above. With the same answer set another question was as follows. A 25-year-old woman is brought to the emergency department 12 hours after a suicide attempt. She took approximately 100 500-mg aspirin tablets. According to the examiner, the correct question is then F, which is as the answer given by MARPLE. We have occasionally noted, however, some slight differences between examiners as to the precise boundaries of a normal range of clinical value, which could cause problems. Simply implementing a fuzzier or probabilistic notion of boundary has not yet resolved this, perhaps suggesting a problem with those questions.

4.8. More "sophisticated" questions

The examination question itself needs little natural language processing in order that its relationship to the answers be assessed. It is only undertaken if the resulting probabilities or the candidate answers is not very distinguishing between at least two answers. However, by use of XTRACTOR and other MARPLE routines, it is fairly easy to process it into form analogous to curated tags in the KRS. Consider the following question as reported in MARPLE 2 output.

Question 3.

A 26-year-old man has insidious onset of low back pain and early morning stiffness. The pain alternates from side to side and occasionally radiates into the buttocks and back of the thighs, but not below the knees. The patient has acute anterior uveitis, diffuse low back and sacroiliac tenderness, and restricted range of motion at the hips. His erythrocyte sedimentation rate is 40mm/h; latex fixation test is negative; and mild hypoproliferative anemia is present.

(A) (P=34%) Ankylosing spondylitis
(B) (P= 8%) Intervertebral disc infection
(C) (P= 8%) Multiple myeloma
(D) (P= 8%) Myofascial pain
(E) (P= 7%) Osteoporosis
(F) (P= 8%) Spinal stenosis
(G) (P=14%) Spondylolysis
(H) (P=14%) Tuberculosis of the spine

Question 3. Predicted best answer is A.
According to examiner, the correct answer is A.

Using XTRACTOR [26], along with POPPER HELPER [21] even in automatic mode, we can break the question itself into the following elements.

< the patient | **has** | age(years):=21-30 >
< the patient | **has** | pain(location):=('low back', diffuse) >
< pain(location):='low back' | **alternates from** | 'side to side' >
< pain(location):='low back' | **radiates into** | the buttocks >
< pain(location):='low back' | **radiates into** | the 'back of the thigh's >
< pain(location):='low back' | **does not radiate into** | 'below the knees' >
< the patient | **has** | tenderness(location):= sacroiliac >
< the patient | **has** | stiffness(time):='early morning' >
< the patient | **has** | 'restricted range of motion (location)':= hips >
< the patient | **has** | 'acute anterior uveitis' >
< the patient | **has** | erythrocyte sedimentation rate (mm/h)':= high >
< the patient | **has** | 'latex fixation test':=negative >
< the patient | **has** | 'hypoproliferative anemia':=mild >.

Note the range conversions on age and erythrocyte sedimentation rate which are in the recent clinical range features in MARPLE 2. In this example it is noteworthy that his erythrocyte sedimentation rate is 40mm/h, which is often considered technically "high" and so indicative of heart failure, but in the USA is actually on the borderline between normal and abnormal. Actually, this question parsing is not absolutely required even in the case of the above question, but detailed discussion is extensive and will be deferred to elsewhere. It may be noted briefly, however, that the KRS may contain relevant statements that have been assigned intrinsic probabilities, i.e. values other than default 1, either from data mining by DiracMiner [20] or a human expert via POPPER [21] working with or without XTRACT tags, or both. Important in this case was that the KRS contained

< 'acute anterior uveitis' Pfwd:=0.3 | **if** | 'ankylosing spondylitis' Pbwd:=0.2 >

It turns out not to be the only path, however, as there is a less direct one. The KRS also contains

< HLA-B27 Pfwd:=0.8 | **if** | 'ankylosing spondylitis' Pbwd:= 0.2 >

Here HLA-B27 is a new feature (an HLA gene variant of the T–system of immunity) does not appear in the question, but there is a tag that contains

< HLA-B27 Pfwd:=0.6 | **if** | 'acute anterior uveitis' Pbwd:= 0.1 >

When the above question was not in parsed form, the above were all needed, but then in addition the following further KRS elements became important.

< osteoporosis Pfwd:=0.06 | **if** | 'ankylosing spondylitis' Pbwd:=0.04 >
< osteoporosis | **'does not indicate'** | 'ankylosing spondylitis' >
< osteoporosis | **'does not necessarily indicate'** | 'ankylosing spondylitis' >

Should this suggest strong forcing, the actual situation is that without these latter three the question was still answered correctly. However, it did tip to second best answer if a large number of uncurated XTRACT tags related to the question topic, but not immediately related to the question, were added to the KRS, in the hybrid mode (ii). For such reasons the answers to standard questions sets are constantly checked in order to maintain a stable KRS that is clear cut enough to avoid to avoid turning to new XTRACTs except as a last resort.

4.9. XTRACT tag errors and "noise" elimination by paraphrasing

The impairment of performance by including many XTRACT tags is of considerable concern. Depending on how one views it, the following, as a means of resolving that, is either an aspect of curation or an easy way of greatly reducing the need for curation, although paradoxically it requires increasing the number of "raw" XTRACTs in a particular way. Note first that to human inspection, some 40% of XTRACT tags seem already to be already in adequate usable form by MARPLE without further curation, certainly at least as far as the rather tolerant common content function is concerned. Unfortunately, some 10%-20% of the remaining 60% of XTRACT tags have some kind of errors that appear of sufficient severity to yield noise and reduce overall performance dramatically. These include parsing ambiguities and misinterpretation of the role of words. As far as Q-UEL's THESAURUS that helps XTRACT curation indicates that some 500-600 common words in English may have 20-30 different meanings with different probabilities calculable *a priori* [26]. Nonetheless, there is significant preliminary evidence, from studies on smaller sets of tags as follows, that the bad effect of increasing the number of XTRACT tags is greatly diminished if *different text sources are used that express the same knowledge content in somewhat different ways.* In that case, many different XTRACTs are produced of essentially same knowledge content that paraphrase each other. It is not hard to do that in practice, because in some instances it occurs naturally on the Web. It does however demand that XTRACTs are being generated over many years. To see this, note the following transcript from the source web page.

"The human brain is the center of the human nervous system [#1]. It has the same general structure as the brains of other mammals [#2], but is larger than expected on the basis of body size among other primates [#3].[#4][#5]"

In most examples below prior to the XTRACT tag, original source text of the above kind is displayed, as it is by XTRACTOR, with index references to the links and citations that were in it.

<QSDFXtractor26 "`The human _brain |^is `the center of| `the human nervous _system

[0http://en.wikipedia.org/wiki/Nervous_system]; `The human _brain |^has `the `same `general _structure as| `the _brains |of| `other mammals [0http://en.wikipedia.org/wiki/Mammal]; `The human _brain |^is larger than ^expected on `the basis of| _body _size |among| `other primates

[0http://en.wikipedia.org/wiki/Primate] [1(0)http://www.ncbi.nlm.nih.gov/pubmed/17148188]

[2file:input.txt#cite_note-Brain-num-1]" (source:='http://en.wikipedia.org/wiki/Human_brain' time:='Wed Oct 3 14:02:19 2012' extract:=0) QSDFXtractor26>

Note above the early date, 2012, to be compared below with XTRACTS generated from the same source in 2015 and 2016 XTRACTs below. Alternative descriptions are important in providing greater coverage of the topic by subsequent curated tags on the KRS. In the following examples 2014 and 2015 examples, one can detect a kind of evolution of the text.

<QSDFxtractor27 "`The human _brain |^has| `the `same `general _structure |as| `the _brains

[0http://en.wikipedia.org/wiki/Brain] |of| `other mammals [0http://en.wikipedia.org/wiki/Mammal]; `The human _brain |^has| `a more |^developed| cortex than |as| `any other" | extracted from |

(source:='http://en.wikipedia.org/wiki/Human_brain'
time:='Thu Jun 26 13:02:03 2014' extract:=0) QSDFxtractor27>

<QSDFXtractor27 "`The human _brain |^is| `the main organ |of| `the human nervous _system

[0http://en.wikipedia.org/wiki/Nervous_system]; `The human _brain |^is ^located in| `the _head

[0http://en.wikipedia.org/wiki/Human_head] |^protected by| `the skull

[0http://en.wikipedia.org/wiki/Human_skull], (?_it) |^has| `the `same `general _structure |as| `the _brains |of| `other mammals

[0http://en.wikipedia.org/wiki/Mammal]; `The human _brain | with| `a more |^developed| cerebral cortex

[0http://en.wikipedia.org/wiki/Cerebral_cortex]" | extracted from | (source:='http://en.wikipedia.org/wiki/Human_brain' time:='Tue Sep 29 13:41:20 2015' extract:=0) QSDFXtractor27>

Note that in using the 2015 source XTRACTOR had trouble (see string (?_it)) with replacing a preposition by the noun or noun phrase referred to, but that it could be resolved in automatic curation by reference to the 2012 version. The second entry in the 2012 run was as follows. Note near the end of the tags the attribute extract:=0 which is 0 for the opening tag, and 1, 2, etc in original sequence, subsequently, so that tags with the same extract value will occur in the some ordinal position as in the source text.

"Estimates for the number of neurons [#6] (nerve cells) in the human brain range from 80 to 120 billion.[#7][#8]"

<QSDFXtractor26 "`the number |of| neurons[0http://en.wikipedia.org/wiki/Neuron] |Estimates for| `the number, (`the number) | as| _nerve cells |in| `the human _brain |with _value _range from| 80-120 billion

[2file:input.txt#cite_note-Brain-num-1] [3(2)http://www.ncbi.nlm.nih.gov/pubmed/19226510]" | extracted from | source:='file:input.txt' time:='Wed Oct 3 14:02:19 2012' extract:=1 QSDFXtractor26>

For higher "extract:=" values, the content will quickly get out of alignment as historical sources progressively differ. By 2015, extract:=1 above became.

<QSDFXtractor27 "Large _animals |such as| whales >AND< elephants have larger _brains |in| absolute terms |but| when |^measured| (using) 'a (^measure) |of| relative _brain _size

[0http://en.wikipedia.org/wiki/Encephalization_quotient] compensates |for| _body _size, `the quotient |for| `the human _brain |^is| almost twice | as large as that of| `a bottlenose dolphin

[0http://en.wikipedia.org/wiki/Bottlenose_dolphin] &and 'three (^times) |as large as | `a chimpanzee

[0http://en.wikipedia.org/wiki/Chimpanzee]" | extracted from |

(source:='http://en.wikipedia.org/wiki/Human_brain' time:-='Tue Sep 29 13:41:20 2015' extract:=1) QSDFXtractor27>

With the above considerations and need for brevity in mind, the 2012 extraction is continued here without further comment but it contains good examples of "noise" in terms of language processing difficulties that can be encountered. It also illustrates how the text can become stilted and sometimes duplicated in order to create linearity appropriate for LSM generation, although this only makes reading irksome to an English-speaking human in this case, and does not typically mean poor performance in MARPLE. As will be described in more detail elsewhere, some logical connectives like AND, OR, IF, and THEN can start in brackets such as {OR} and used either to become >OR< and ultimately break up into separate final curated KSR elements, or be retained as an operator &or in the basic sentence that will ultimately become an **or** operator within a well curated KRS element.

"Most of the expansion comes from the cerebral cortex [#9], especially the frontal lobes [#10], which are associated with executive functions [#11] such as self-control, planning, reasoning, and abstract thought."

<QSDFXtractor26 "`Most |of| `the _expansion |^comes from| `the cerebral cortex

[0http://en.wikipedia.org/wiki/Cerebral_cortex], `the _expansion |especially by| `the frontal lobes

[0http://en.wikipedia.org/wiki/Frontal_lobe] |^are associated with| executive functions

[0http://en.wikipedia.org/wiki/Executive_functions] |such as| self-control (planning) &and (reasoning) &and abstract (thought)" | extracted from | source:='file:input.txt' time:='Wed Oct 3 14:02:19 2012' extract:=2 QSDFXtractor26>

"The portion of the cerebral cortex devoted to vision is also greatly enlarged in human beings, and several cortical areas play specific roles in language, a skill that is unique to humans."

< QSDFXtractor26 "ˆ`The portion |of| `the cerebral cortex ⌈ devoted to| vision; The portion ⌈is also greatly ˆenlarged in| human beings {AND} several cortical areas ⌈play| specific roles |in| _language |as| `a skill ⌈is| unique |to| humans" | extracted from | source:='file:input.txt' time:='Wed Oct 3 14:02:19 2012' extract:=3 QSDFXtractor26 >

"Despite being protected by the thick bones of the skull, suspended in cerebrospinal fluid [#12], and isolated from the bloodstream by the blood–brain barrier [#13], the human brain is susceptible to many types of damage and disease."

< QSDFXtractor26 "(`the human _brain) |Despite ˆbeing ˆprotected by| `the `thick _bones |of| `the skull > AND < (`the human _brain) |Despite ˆbeing ˆsuspended in| cerebrospinal fluid [0http://en.wikipedia.org/wiki/Cerebrospinal_fluid] > AND < (`the human _brain) |Despite ˆbeing ˆisolated from| `the bloodstream |by| `the blood-_brain barrier [0http://en.wikipedia.org/wiki/Blood%E2%80%93brain_barrier] > THEN < `the human _brain |ˆis| susceptible |to `many types of| (ˆdamage) &and _disease" | extracted from | source:='file:input.txt' time:='Wed Oct 3 14:02:19 2012' extract:=4 QSDFXtractor26 >

"The most common forms of physical damage are closed head injuries [#14] such as a blow to the head [#15], a stroke [#16], or poisoning by a variety of chemicals that can act as neurotoxins [#17]."

< QSDFXtractor26 "ˆ`The `most 'common (ˆforms) |of| _damage ⌈ are| (ˆclosed) |kinds of| _head injuries [0http://en.wikipedia.org/wiki/Closed_head_injury] |such as| 'a (ˆblow) |to| `the _head [0http://en.wikipedia.org/wiki/Human_-head] &or `a (ˆstroke) [0http://en.wikipedia.org/wiki/Stroke] &or (ˆpoisoning) |by `a variety of| chemicals ⌈can ˆact as| neurotoxins [0http://en.wikipedia.org/wiki/Neurotoxin]" | extracted from | source:='file:input.txt' time:='Wed Oct 3 14:02:19 2012' extract:=5 QSDFXtractor26 >

"Infection of the brain, though serious, is rare due to the biological barriers which protect it. The human brain is also susceptible to degenerative disorders, such as Parkinson's disease [#18], multiple sclerosis [#19], and Alzheimer's disease [#20]. A number of psychiatric conditions, such as schizophrenia [#21] and depression [#22], are thought to be associated with brain dysfunctions, although the nature of such brain anomalies is not well understood.[#23]".

< QSDFXtractor26 "Infection |of| `the _brain, Infection |though| `serious, Infection ⌈is| rare |due to| `the biological barriers |ˆprotect| (?_it)" | from source | (source:='file:input.txt' time:='Wed Oct 3 14:02:19 2012' extract:=6) QSDFXtractor26 >

< QSDFXtractor26 "ˆ`The human _brain |ˆis also| susceptible |to| degenerative disorders |such as| Parkinson's _disease [0http://en.wikipedia.org/wiki/Parkinson%27s_disease] &and multiple sclerosis [0http://en.wikipedia.org/wiki/Multiple_sclerosis] &and Alzheimer's _disease [0http://en.wikipedia.org/wiki/Alzheimer%27s_disease]" | extracted from | source:='file:input.txt' time:='Wed Oct 3 14:02:19 2012' extract:=7 QSDFXtractor26 >

< QSDFXtractor26 "ˆ`A _number |of| psychiatric _conditions | such as| schizophrenia [0http://en.wikipedia.org/wiki/Schizophrenia] &and depression [0http://en.wikipedia.org/wiki/Major_depressive_disorder] |ˆare ˆthought to ˆbe associated with| _brain dysfunctions |although| `the nature |of such| _brain anomalies |ˆis not well ˆunderstood| [4(3)http://www.ncbi.nlm.nih.gov/pubmed/15514638] (token object)" || extracted from | source:='file:input.txt' time:='Wed Oct 3 14:02:19 2012' extract:=8 QSDFXtractor26 >

4.10. Accumulation of weight of evidence versus importance of very few KRS elements

The reviewers invited us to emphasize whether there is an accumulation of weight of evidence from many KRS elements in pathways of a network linking question to each answer, or whether a few elements dominate. The answer is that is usually the former, but in it depends on the question and knowledge elements available to answer it. The first two examples below demonstrate the need for greater resolving power when two or more questions have the same set of candidate answers.

Question 4.

A 25-year-old woman has sudden onset of persistent right lower abdominal pain that is increasing in severity. She has nausea without vomiting. She had a normal bowel movement just before onset of pain. Examination shows exquisite deep tenderness to palpation in right lower abdomen with guarding but no rebound; bowel sounds are present. Pelvic examination shows a 7-cm, exquisitely tender right-sided mass. Hematocrit is 32%. Leukocyte count is 18,000/mm3. Serum amylase activity is within normal limits. Test of the stool for occult blood is negative.

A. (P= 5%) Abdominal aneurysm
B. (P=11%) Appendicitis
C. (P= 6%) Bowel obstruction
D. (P= 6%) Cholecystitis
E. (P= 2%) Colon cancer
F. (P= 6%) Constipation
G. (P= 6%) Diverticulitis
H. (P= 3%) Ectopic pregnancy (ruptured)
I. (P= 6%) Endometriosis
J. (P= 6%) Hernia
K. (P= 6%) Kidney stone
L. (P= 5%) Mesenteric adenitis
M. (P= 3%) Mesenteric artery thrombosis
N. (P= 3%) Ovarian cyst (ruptured)
O. (P= 6%) Pancreatitis
P. (P= 5%) Pelvic inflammatory disease
Q. (P= 2%) Peptic ulcer disease
R. (P= 2%) Perforated peptic ulcer
S. (P= 6%) Pyelonephritis
T. (P= 6%) Torsion

Question 4. Predicted best answer is B.
According to examiner, the correct answer is B.

Consistent with the fact that there are a great many diagnoses to eliminate, this kind of question builds up a fairly complicated balance of weights (rather like as in Neural Net learning approach) that are difficult to interpret for discussion purposes and involve a very large number of tags on the KRS. In the above case, however, it was clear the searching of the Internet was successful at helping resolve the issue. There were ample texts on the web which referred to appendicitis and the location of tenderness that provided corresponding KRS tags as the strongest clue, subject particularly by the following contrary evidence. Serum amylase testing was clearly linked to information -testing for pancreatitis, and a test of the stool for occult blood as negative is required to indicate a lesion in the alimentary tract, notably due to say colon cancer, regarding which there was ample information about fecal blood testing on the web.

Question 5.

An 84-year-old man in a nursing home has increasing poorly localized lower abdominal pain recurring every 3-4 hours over the past 3 days. He has no nausea or vomiting; the last bowel movement was not recorded. Examination shows a soft abdomen with a palpable, slightly tender, lower left abdominal mass. Hematocrit is 28%. Leukocyte count is 10,000/mm3. Serum amylase activity is within normal limits. Test of the stool for occult blood is positive.

A. (P= 4%) Abdominal aneurysm
B. (P= 7%) Appendicitis
C. (P= 4%) Bowel obstruction
D. (P= 4%) Cholecystitis
E. (P=11%) Colon cancer
F. (P= 3%) Constipation
G. (P= 3%) Diverticulitis
H. (P= 6%) Ectopic pregnancy (ruptured)
I. (P= 3%) Endometriosis
J. (P= 3%) Hernia
K. (P= 4%) Kidney stone
L. (P= 4%) Mesenteric adenitis
M. (P= 4%) Mesenteric artery thrombosis
N. (P= 6%) Ovarian cyst (ruptured)
O. (P= 3%) Pancreatitis
P. (P= 4%) Pelvic inflammatory disease
Q. (P=10%) Peptic ulcer disease
R. (P=10%) Perforated peptic ulcer
S. (P= 3%) Pyelonephritis
T. (P= 3%) Torsion

Question 5. Predicted best answer is E.
According to examiner, the correct answer is E.

Comments apply as above for Question 5, The fact that an association between colon cancer and about fecal blood testing was prominent on the web was primarily responsible for the corresponding tags on the KRS that provided the correct answer. In the following question, as for the preceding two questions, there are a great many diagnoses to eliminate. Although this kind of question can build up a fairly complicated balance of weights that are difficult to interpret and involve a very large number of tags on the KRS, it appears in some cases that one or very few KRS elements may be crucial in pinpointing a single answer by a marginally higher probability.

Question 6.
A 19-year-old woman has had fatigue, fever, and sore throat for the past week. She has a temperature of 38.3 C (101 F), cervical lymphadenopathy, and splenomegaly. Initial laboratory studies show a leukocyte count of 5000/mm3 (80% lymphocytes, with many lymphocytes exhibiting atypical features). Serum aspartate aminotransferase (AST, GOT) activity is 200 U/L. Serum bilirubin concentration and serum alkaline phosphatase activity are within normal limits.

A. (P= 7%) Acute leukemia
B. (P= 7%) Anemia of chronic disease
C. (P= 7%) Congestive heart failure
D. (P= 7%) Depression
E. (P= 8%) Epstein-Barr virus infection
F. (P= 7%) Folate deficiency
G. (P= 7%) Glucose 6-phosphate dehydrogenase deficiency
H. (P= 7%) Hereditary spherocytosis
I. (P= 7%) Hypothyroidism
J. (P= 7%) Iron deficiency
K. (P= 7%) Lyme disease
L. (P= 7%) Microangiopathic hemolytic anemia
M. (P= 7%) Miliary tuberculosis
N. (P= 7%) Vitamin B12 (cyanocobalamin) deficiency

Question 6. Predicted best answer is E.
According to examiner, the correct answer is E.

Nonetheless, this can sometimes be misleading. The balance of evidence represented by the KRS elements was quite complicated in the above case for several of the answers, effectively canceling out to nearest integer. The 14 main ones found to be influencing the score, are as follows.

< 'Streptococcus pyogenes' | causes | sore throat >
< 'Streptococcus pyogenes' | 'may cause' | malaise >
< 'Streptococcus pneumoniae' | 'may cause' | malaise >
< 'Streptococcus pneumoniae' | 'may cause' | fatigue >
< 'Streptococcus pyogenes' | 'may cause' | fatigue >
< 'Epstein-Barr virus' | causes | sore throat >
< 'Epstein-Barr virus' | 'may cause' | fatigue >
< 'Epstein-Barr virus' | causes | 'cervical lymphadenopathy' >
< 'Epstein-Barr virus' | causes | splenomegaly >
< 'Epstein-Barr virus' | causes | increased 'serum aspartate aminotransferase activity' >
< 'Epstein-Barr virus' | causes | increased 'leukocyte count' >
< 'Epstein-Barr virus' | 'does not cause' | increased 'serum bilirubin' >
< 'Epstein-Barr virus' | 'does not cause' | increased 'serum alkaline phosphatase' activity >
< 'Epstein-Barr virus' | causes | 'atypical leukocyte' >

This is also an example of a question with a partner, as follows, that addresses the same answer set.

Question 7.
A 15-year-old girl has a two-week history of fatigue and back pain. She has widespread bruising, pallor, and tenderness over the vertebrae and both femurs. Complete blood count shows hemoglobin concentration of 7.0 g/dL, leukocyte count of 2000/mm^3, and platelet count of 15,000/mm^3.

A. (P=20%) Acute leukemia
B. (P= 6%) Anemia of chronic disease
C. (P= 6%) Congestive heart failure
D. (P= 6%) Depression
E. (P= 7%) Epstein-Barr virus infection
F. (P= 6%) Folate deficiency
G. (P= 6%) Glucose 6-phosphate dehydrogenase deficiency
H. (P= 6%) Hereditary spherocytosis
I. (P= 6%) Hypothyroidism
J. (P= 6%) Iron deficiency
K. (P= 6%) Lyme disease
L. (P= 6%) Microangiopathic hemolytic anemia
M. (P= 6%) Miliary tuberculosis
N. (P= 6%) Vitamin B12 (cyanocobalamin) deficiency

Question 7. Predicted best answer is A.
According to examiner, the correct answer is A.

The strong score picked up here was in regard several tags relating to bruising and pain in the curated KRS store, such as < 'acute leukemia' | **causes** | bruising >.

Question 8.
A 7-year-old girl has a high fever and a sore throat. There is pharyngeal redness, a swollen right tonsil with creamy exudate, and painful right submandibular lymphadenopathy. Throat culture on blood agar yields numerous small beta-hemolytic colonies that are inhibited by bacitracin.

A. (P= 5%) Adenovirus
B. (P= 5%) Aspergillus fumigatus
C. (P= 5%) Bacillus anthracis
D. (P= 5%) Candida albicans
E. (P= 5%) Chlamydia psittaci
F. (P= 5%) Coccidioides immitis
G. (P= 5%) Coronavirus
H. (P= 5%) Corynebacterium diphtheriae
I. (P= 5%) Coxiella burnetii
J. (P= 5%) Coxsackievirus
K. (P= 7%) Epstein-Barr virus
L. (P= 5%) Haemophilus influenzae
M. (P= 5%) Histoplasma capsulatum
N. (P= 5%) Mycobacterium tuberculosis
O. (P= 5%) Mycoplasma pneumoniae
P. (P= 5%) Neisseria gonorrhoeae
Q. (P= 5%) Neisseria meningitidis
R. (P= 5%) Pneumocystis carinii
S. (P= 5%) Rhinovirus
T. (P= 3%) Streptococcus pneumoniae
U. (P= 8%) Streptococcus pyogenes

Question 8. Predicted best answer is U.
According to examiner, the correct answer is U.

“Beta-hemolytic colonies” in the question was a strong clue, impacting selection of the following KRS tags as of relevance.

```
< 'Streptococcus pyogenes' | is | a group B Streptococcus >
< 'Streptococcus pneumoniae' | 'is not' | a group B Streptococcus >
< 'Streptococcus pyogenes' | 'is not' | a group A Streptococcus >
< 'Streptococcus pneumoniae' | is | a group A Streptococcus >
< 'Streptococcus pyogenes' | is | a group B Streptococcus >
< Group A Streptococcus | is not inhibited by | bacitracin >
< Group B Streptococcus | is inhibited by | bacitracin >
< Group B Streptococcus | is | beta-hemolytic >
< Group D Streptococcus | is | alpha-hemolytic | or | gamma-hemolytic >
< Group B Streptococci | are | beta-hemolytic >
< Group D Streptococci | are | alpha-hemolytic | or | gamma-hemolytic >
< Streptococcus viridans | is | alpha-hemolytic >
————————————————————————————————————————————————————————————————————————
```

Question 17.

Select the drug most likely to have caused the adverse effect. A 56-year-old man with recurrent ventricular arrhythmias began taking an antiarrhythmic drug 5 months ago. He now has progressive dyspnea, cough, and low-grade fever. Erythrocyte sedimentation rate is increased. X-ray film of the chest shows a diffuse interstitial pneumonia. Pulmonary function tests show that diffusing capacity for carbon monoxide is decreased.

A. (P= 5%) Acetaminophen
B. (P=16%) Amiodarone
C. (P= 9%) ACE inhibitors
D. (P= 5%) Aspirin
E. (P= 5%) Atenolol
F. (P= 5%) Bleomycin
G. (P= 5%) Cytosine arabinoside
H. (P= 5%) Furosemide
I. (P= 5%) Metronidazole
J. (P= 5%) Nalidixic acid
K. (P= 5%) Nitrofurantoin
L. (P= 5%) Penicillin
M. (P= 5%) Prednisone
N. (P= 5%) Procainamide
O. (P= 5%) Propranolol
P. (P= 5%) Sulfasalazine
Q. (P= 5%) Tetracycline
R. (P= 5%) Verapamil

Question 17. Predicted best answer is B.
According to examiner, the correct answer is B.

The above question was fairly crisply resolved by KRS tags derived from XTRACT tags concerning contraindications.

4.11. Case study report as examination question

With the ultimate applications also of the above to CDS in mind, a presentation of a patient and lab results from a whole case study [38] can serve as a question, as follows. The original part of the text describing presentation of the patient and initial Specific statements about contraindications for other differential diagnoses were removed, and helped provide the alternative incorrect candidate answers. However, they could in principle confuse and decoy a simple algorithm, so note that the same correct answer was obtained this information was left in. With removal this still left a lengthy “question” of 769 words which has been left out for brevity and respect of copyright, but is the section in Ref. [38] starting “A 46-year-old woman was seen in the emergency department at this hospital because of muscle pain and swelling in her arms and lower legs. The patient had been well until approximately 3 weeks before admission, when a deep ache developed in her left triceps…”, containing “A chest radiograph showed a soft-tissue opacity, 9.0 cm by 9.0 cm by 12.4 cm, in the right lower hemithorax that obscured the right heart border…. suggestive of an anterior mediastinal mass….”, and concluding “Blood levels of parrathyroid hormone, hCG, and alpha-fetoprotein were normal, and testing for antibodies to Ro, La, Sm, RNP, and Jo-1 was negative”.

Answers:

A. (P= 8.89%) vascular compression
B. (P= 8.89%) thrombosis.

C. (P= 8.89%) farction
D. (P= 8.89%) compartment syndrome
E. (P= 9.23%) infection
F. (P=10.62%) thymoma
G. (P= 8.89%) lymphoma
H. (P= 8.89%) germ cell tumor.
I. (P= 8.89%) thyroid cancer
J. (P= 9.06%) trauma
K. (P= 8.89%) polymyositis

Question 1. Predicted best answer is F.
According to examiner, the correct answer is F.

As reported in output, Marple17 responded to this single "question" in elapsed time 3 seconds. The most relevant KRS elements automatically found and used were

< 'mediastinal mass' | suggests | thymoma | or | 'neurogenic tumor' >
< 'mediastinal mass' Pfwd:=0.2 | is | 'neurogenic tumor' >
< 'mediastinal mass' Pfwd:=0.175 | is | thyoma >
< 'neurogenic tumor' | 'is usually found in ' | 'posterior mediastinum' >
< thymoma | 'is usually found in' | 'anterior mediastinum' >
< thymoma | 'is sometimes associated with' | 'myasthenia gravis' >
< 'mediastinal mass' | suggests | 'paraneoplastic manifestation of thymoma >
< patients | with | thymoma | may be | asymptomatic >
< patients | with | thymoma | may have | 'paraneoplastic autoimmune disease' >
< patients | with | thymoma | may have | 'vascular compression' >
< patients | with | thymoma | may have | 'chest pain' >
< 'vascular compromise' | does not account for | 'regional muscle swelling' >
< 'topical pyomyositis' | may be associated with | 'focal lesion in muscle >
< 'topical pyomyositis' | may be associated with | fever | and | trauma >
< 'topical pyomyositis' | may be associated with | immunocompromise >
< 'topical pyomyositis' | may be associated with | malnourished host >
< thymoma | 'is not associated with' | fever | and | trauma >
< thymoma | 'is not associated with' | immunocompromise >
< thymoma | 'is not associated with' | malnourished host >
< 'asymmetric muscle pain' | and | swelling | 'is not associated with' | 'inflammatory myopathy' >
< symmetric | and | 'proximal-muscle weakness' | 'is associated with | 'inflammatory myopathy' >
< 'diffiulty swallowing' | 'is associated with | 'inflammatory myopathy' >
< dyspnoea | 'is associated with | 'inflammatory myopathy' >
< 'idiopathic inflammatory myopathies' | 'may be' | paraneoplastic >
< 'idiopathic inflammatory myopathies' | 'may be associated with' | 'distal neuromuscular weakness' >
< 'idiopathic inflammatory myopathies' | 'may be associated with' | 'inclusion-body myositis' >
< 'idiopathic inflammatory myopathies' | 'may be associated with' | 'photosensitive rash' >
< 'idiopathic inflammatory myopathies' | 'may be associated with' | dermatomyositis >
< 'idiopathic inflammatory myopathies' | 'may be associated with' | myalgia >
< 'idiopathic inflammatory myopathies' | 'may be associated with' | myalgia >
< 'viral myopathies' | 'are associated with' | myalgia >

5. Discussion and conclusions

5.1. Assessment of overall success

The notion of success depends on the goal and, for our main goal as the curation of the KRS for CDS purposes, that one can achieve 100% score over large question sets with well curated KRS elements is pleasing and means that MARPLE is a very useful tool for auditing and curating knowledge elements. It is not sufficient, not least because intrinsic probabilities are not well tested, except as binary degrees of truth represented by positive and negative statements. As to being a success in AI, it emerges that passing medical exams is not as much an AI challenge as one may think. It continues to looks as if medical licensing examinations are set up so that a student taking the examination can reason as follows: *"The question has a part that has something to do with part of one of my knowledge elements, and part of that knowledge element has something to do with part of one of my other knowledge elements, which has a part that has something to do with one of the answers. So that must be the answer. Well, maybe that is not so in real life, but I bet the examiner wrote the question that way*" [41]. The kind of work described here gives some insight into how examiners set questions and how human students may be tackling them, and into how to structure a fair question.

Nonetheless, the exam remains a test of knowledge because that knowledge must still be there even to use the above elementary reasoning, even though that reasoning and the knowledge that it uses is not necessarily all that would be required outside the exam and in a more practical everyday context. We can see that the above goes part of the way, and that it can be an important heuristic to gather a plausible set of candidate KRS elements as potential solutions to consider. A typical definition of a syllogism, which is a minimal case or part of, and extensible to, the above chain, is "an instance of a form of reasoning in which a conclusion is drawn from two given or assumed propositions, each of which shares a term with the conclusion, and shares a common or middle term not present in the conclusion", but that is not intended to be sufficient as the definition of a *valid* syllogism, just of the general form that syllogistic reasoning must have[5]. We may say that *presyllogistic logic* is being used, and that it represents MARPLE's most effective heuristic algorithm. If one initially has few questions and few curated knowledge elements, the overall approach can look forced, but that feeling diminishes as the number of questions and curated KRS elements increase towards covering the whole syllabus. It is the gathering and curation of KRS elements to obtain a correct answer that looks like supervised learning, while searching the Internet looks like unsupervised learning.

[5] More symbolically, in a syllogistic chain of reasoning, to say something about the truth of, for example, statement $<\mathrm{A}|\,\mathbf{R_{AZ}}\,|\mathrm{Z}>$ requires that we can build a path of KRS elements $<\mathrm{A}|\,\mathbf{R_{AB}}\,|\mathrm{B}><\mathrm{B}|\,\mathbf{R_{AC}}\,|\mathrm{C}>\ldots<\mathrm{Y}|\,\mathbf{R_{AY}}\,|\mathrm{Z}>$. To do that, we first only have to look for KRS elements that contain A and Z, or which contain something directly or indirectly related to them, where "related" means that there exists a KRS element that contains something that is related by occurring along with it in another KRS element, and so on.

5.2. Automatic curation

Capturing and curating knowledge in computers has been regarded as one of the hardest steps and most enduring challenges of AI [42]. Early recognition of this has been widely attributed to a statement by Edward Feigenbaum and subsequently called *the Feigenbaum Bottleneck* [42]. While we feel bound to continue auditing experts of KRS elements by human with CDS applications in real clinical settings in mind, escalating numbers of XTRACTS in particular are making routine manual intervention increasingly impractical. Any automation of curation processes is important. There are a number of ways that a knowledge representation may be wrong, and hence a number of ways to correct it. Some are easier to automate than others. Readily automated were simple detection and correction of obvious canonical structure errors, and the characteristic content of noun and relationship phrases about to be laced in the wrong “slots” of STs and LSMs. The POPPER HELPER interface [22] facilitates the progressive transformation from human expertise capture to its replacement by automation, so that this evolves at the same time as the KRS is growing. This is easy to do by constructing metastatements that act on statements as KRS elements [22], but steps that are essentially curation are being added directly to the code of MARPLE on an ongoing basis. All these approaches were used in the later of studies reported above. A recent implementation in MARPLE 2 which resulted in the later results reported above was automatic creation of negative statements to provide appropriate evidence against incorrect candidate answers appears important, and is readily automated. That is to say, statements like < X | ‘is not associated with’ | Y > are important, and effectively eliminate Y as an answer, but therefore needs to be verified that this is medically the case to avoiding forcing correct answers that will damage performance in later questions on similar topics. It is not as yet applied to XTRACTs, but only in their processing to become well curated KRS elements. Note that we are increasingly making use of multiple questions sharing the same answer sets [35], when sometimes the same answer is the case and more commonly not, so the common content function will tend to increase specificity and weight contributions naturally without much need for human intervention (except for ongoing checks). Probably because of use of such questions, and because MARPLE was already working will a well curated core set of KRS elements, this implementation has only minor effect in the present study. Results are not significantly changed by removing this feature. However we expect it to become important in the future as a large number of more recently acquired XTRACTs are progressively further curated. Studies on these curation aspects and others will be reported elsewhere.

5.3. The relationship between determiners and intrinsic probabilities

A detailed analysis of the role of determiners including those of definiteness, scope and number such as like “a”, “the’, “some”, “two” “few”, “many” will be discussed elsewhere in consideration of various cases when they are, or are not, important. In the exam case, they are found empirically to be much less important. It helps to see that determiners and quantifiers like “most” are essentially “fuzzy” counterparts of the Pfwd and Pbwd attributes carried on Q-UEL tags, i.e. that specify the probabilities that the statement is true, or of a certain degree of scope, e.g. Pfwd:=0.85. The challenge is conversion to a specific value in Pfwd and Pbwd attribute values, though note that Q-UEL allows e.g. Pfwd:=0.85+/-0.05, standard deviations etc., and other qualifications of certainty, as it does for any attribute values [23,27]. Some aspects of our current theoretical thinking on determiners are indicated elsewhere [22,41]. These Pfwd and Pbwd values are, however, precisely the same intrinsic probabilities that we have shown to be less important in the present exam context, except for taking account of negation, so the minor role of determiners in exams seems less surprising in hindsight.

5.4. Comparing MARPLE with human medical students, and IBM’s Watson

While the challenge of gathering enough well curated knowledge for MARPLE to tackle full syllabuses is significant and ongoing, it does not seem to be of Herculean proportions. MARPLE can gather 10^6 XTRACT tags per day from the Internet, and their curation, now the main focus of current efforts, is improving rapidly. With well curated KRS elements, each question in an exam takes the order of a second to answer. An average human medical student may study privately about 500-600 total hours for the USMLE Step 1 and spend 3000-5000 hours overall including lectures and practical classes, and a real comprehensive medical licensing examination can take 4 or 8 hours involving perhaps 200 multiple choice questions each with just 5 answers. In contrast, the TV quiz show “Jeopardy!”, is specifically an interactive competition against humans, and winning is based on the total amount of money won, which is based on wagers by participants that the answer given would be correct. On that basis, Watson was a clear winner. However, Watson only answered 50% of all questions in Round 1, 77% in Round 2, and 0% in the short final according to Ref. [12]. That gives 44% questions answered in this quiz game interpreted as an exam, overall, in which those that are not attempted are counted as fails.

5.5. Ultimate potential

As to the quality that in the long term may be or may not be achievable in quizzes and examinations, and not least in medical decision making, some closing comment might be made in the light of the movie “Slum Dog Millionaire” [43], said to be based in part on a true story. It suffices to state the plot itself, and to think of the possible implications for applying the celebrated Turing Test [44] in such situations, replacing Jamal Malik, the hero of the movie, by a computer system. In the movie, Jamil is one question away from winning a staggering 20 million rupees on India's Kaun Banega Crorepati? (2000) (Who Wants To Be A Millionaire?), but that night, police arrest him. It is on suspicion of cheating: he is an 18 year-old orphan from the slums of Mumbai. How could such a person have so much wisdom and knowledge? To prove his innocence, Jamal tells the story of his life and how it provided key knowledge and wisdom in stages that could be used to provide the answer to one of the game show's seemingly impossible questions. It would seem hard for a computer to pass a quiz or exam if so much depends on subtle human experience. However, while we can live complicated lives, and present a challenge for a guru, spiritual guide, counselor, or even psychiatrist, most of us usually get physically sick in a relatively limited number of ways, as “case” types that allow a physician, and hopefully future CDSS, to function effectively.

References

[1] B. Robson, O.K. Baek, The Engines of Hippocrates. From the Dawn of Medicine to Medical and Pharmaceutical Informatics, Wiley, 2009.

[2] B. Buchanan, E.H. Shortliffe, Rule Based Expert Systems. The MYCIN Experiments of the Stanford Heuristic Programming Project, Addison-Wesley, Reading, Massachusetts, 1982.

[3] B. Robson, S. Boray, Interesting things for computers systems to do: keeping and data mining millions of patient records, guiding patients and physicians, and passing medical licensing exams, in: Proceedings SB204, Bioinformatics

and Biomedicine (BIBM), Proceedings 2015 IEEE International Conference, 1397-1404, IEEE.

[4] J.E. Richardson, J.S. Ash, D.F. Sittig, A. Bunce, J. Carpenter, R.H. Dykstra, K. Guappone, J. McCormack, C.K. McMullen, M. Shapiro, A. Wright, B. Middleton, Multiple perspectives on the meaning of clinical decision support, AMIA Annu. Symp. Proc. 2010 (2010) 1427–1431.

[5] I.M. Mullins, I. M., M.S. Siadaty, J. Lyman, K. Scully, G.T. Garrett, G. Miller, R. Muller, B. Robson, C. Apte, C.S. Weiss, I. Rigoutsos, D. Platt, S. Cohen, Data mining and clinical data repositories: insights from a 667,000 patient data set, Comput. Biol. Med. 36 (12) (2006) 1351.

[6] A.B. McCoy, W. t A. Krousel-Wood M, T. E. J., J.A. McCoy, D.F. Sittig, Validation of a crowdsourcing methodology for developing a knowledge base of related problem-medication pairs eCollection, Appl. Clin. Inform. 6 (2) (2015) 334–344.

[7] D. Ferrucci, E. Brown, J. Chu-Carroll, J. Fan, D. Gondek, A.A. Kalyanpur, A. Lally, J.W. Murdock, E. Nyberg, J. Prager, N. Schlaefer, C.S. Welty, Building Watson: an overview of the DeepQA project, AI Mag. 31 (3) (2010) 59.

[8] Y. Pan, Z.M. Qiu, C. Welt, A framework for merging and ranking of answers in DeepQA Paper 14, IBM Res. Dev. 56 (3–4) (2012).

[9] 〈http://www.cs.uky.edu/~raphael/grad/keepingCurrent/HowWatsonWorks.pdf〉.

[10] 〈http://bits.blogs.nytimes.com/2012/10/30/i-b-m-s-watson-goes-to-medical-school/?_r=0〉 (last accessed 18.12.15).

[11] 〈http://www.forbes.com/sites/bruceupbin/2011/05/25/ibms-watson-now-a-second-year-med-student/〉 (last accessed 15.12.15).

[12] 〈https://docs.google.com/spreadsheets/d/1e0o0R-eOAbnkFEHFFqlcPole77ZR06tLr0EPxGfaqN0/edit?hl=en#gid=0〉.

[13] P.A.M. Dirac, A new notation for quantum mechanics, Math. Proc. Camb. Philos. Soc. 35 (3) (1939) 416.

[14] P.A.M. Dirac, The Principles of Quantum Mechanics first edition, Oxford University Press, 1930), fourth edition, Clarendon Press, 1982.

[15] B. Robson, The new physician as unwitting quantum mechanic: is adapting Dirac's inference system best practice for personalized medicine, genomics and proteomics?", J. Proteome Res. (Am. Chem. Soc.) 6 (8) (2007) 3114.

[16] B. Robson, Links between quantum physics and thought, Future of Health Technology Congress, Technology and Informatics, vol. 149, IOS Press (2009), p. 157.

[17] B. Robson, Towards intelligent internet-roaming agents for mining and inference from medical data, Future of Health Technology Congress, Technology and Informatics, vol. 149, IOS Press (2009), p. 157.

[18] B. Robson, Towards New Tools for Pharmacoepidemiology, Adv. Pharmacoepidemiol. Drug Saf. 1 (2013) 6, http://dx.doi.org/10.4172/2167-1052.100012.

[19] B. Robson, Towards automated reasoning for drug discovery and pharmaceutical business intelligence, Pharm. Technol. Drug Res. 1 (2012) 3.

[20] B. Robson, Hyperbolic Dirac nets for medical decision support. Theory, methods, and comparison with Bayes nets, Comput. Biol. Med. 51 (2014) 1832014.

[21] S. Deckelman, B. Robson, Split-complex numbers and Dirac Bra-Kets, Commun. Inf. Syst. (CIS) 14 (3) (2015) 135–149.

[22] B. Robson, POPPER, a simple programming language for probabilistic semantic inference in medicine, Comput. Biol. Med. 56 (2014) 107.

[23] B. Robson, S. Boray, Implementation of a web based universal exchange and inference language for medicine. Sparse data, probabilities and inference in data mining of clinical data repositories (in press), Comput. Biol. Med. (2016).

[24] 〈http://www.whitehouse.gov/sites/default/files/microsites/ostp/pcast-health-it-report.pdf〉 (last accessed 30.03.13).

[25] B. Robson, U.G.J. Balis, T.P. Caruso, Considerations for a universal exchange language for healthcare, in: Proceedings of the 13th IEEE International Conference on e-Health Networking Applications and Services (IEEE Healthcom '11), Columbia, MO, June 13, 2011, p. 173.

[26] B. Robson, T.P. Caruso, A Universal Exchange Language for Healthcare" MedInfo '13, in: Proceedings of the 14th World Congress on Medical and Health Informatics, Copenhagen, Denmark, Edited by Lehmann, Ammenwerth, and Nohr. IOS Press, Washington, DC, USA, 2013. 〈http://quantalsemantics.com/documents/MedInfo13-RobsonCaruso_V6.pdf〉; 〈http://ebooks.iospress.nl/publication/ 34165〉.

[27] B. Robson, T.P. Caruso, T, U.G. J. Balis, Suggestions for a web based universal exchange and inference language for medicine, Comput. Biol. Med. 1 (12) (2013) 229.

[28] Suggestions for a web based universal exchange and inference language for medicine, continuity of patient care with PCAST disaggregation, Computers in Biology and Medicine, 56, (2014) 51.

[29] R.P. Feynman, Josiah R. Gibbs, Quantum Mechanics and Path Integrals, McGraw Hill, 1965.

[30] 〈https://en.wikipedia.org/wiki/Digital_curation#cite_note-ijdcpw200711-10〉 (last accessed 02.01.16).

[31] P. Watry, Digital preservation theory and application: transcontinental persistent archives testbed activity, Int. J. Digit. Curat. 2 (2) (2007) 41 (last accessed 02.01.16).

[32] A. Gómez-Pérez, M. Fernández-López, O. Corcho, Ontological Engineering: With Examples from the Areas of Knowledge Management, e-Commerce and the Semantic Web, Springer Science & Business Media, 2006.

[33] H. Khan, B. Caruso, J. Corson-Rikert, D. Dietrich, B. Lowe, G. Steinhart, DataStaR: using the semantic web approach for data curation, Int. J. Digit. Curat. 6 (2) (2011) 209.

[34] J.D. Myers, The Background of INTERNIST-I and QMR, in: B.I. Blum, K. Duncan (Eds.), A History of Medical Informatics, 423, ACM Press, 1990.

[35] S.M. Case, D.B. Swanson, Constructing Written Test Questions for the Basic and Clinical Sciences, National Board of Medical Examiners, 〈https://www.uclouvain.be/cps/ucl/doc/adef/documents/EVA_Res_Ext_Questions_type_Appariement.pdf〉.

[36] United States Medical Licensing Examination Board, United States Medical Licensing Examination, 〈http://www.usmle.org/〉 (last accessed 26.09.15).

[37] S. Katz, Week 7 USMLE Step 1 Review: Biostatistics, Behavioral Science, and Nutrition, 〈http://www.google.com/url?url=http://som.uthscsa.edu/AcademicEnhancement/documents/week7ppt.ppt&rct=j&frm=1&q=&esrc=s&sa=U&ved=0CBQQFjAAahUKEwjxkqOjg6LIAhWI_R4KHb64Bpo&usg=AFQjCNEEZooFbzFv95bWXy-MA-Pe1uS8Zw〉.

[38] M. Seton, C.C. Wu., A. Louissant Jr., Case 26-2013: a 46-year-old woman with muscle pain and swelling, case records of the Massachusetts General Hospital, N. Engl. J. Med., Mass. Med. Soc. 769 (2013) 364.

[39] 〈http://lskampe.com/grammar/pronoun.htm〉 (last accessed 26.09.15).

[40] 〈http://usarad.com/pdf/CT/CT%20BRAIN%20W_O%20CONTRAST_2.pdf〉.

[41] B. Robson, S. Boray, The structure of reasoning in answering multiple choice medical licensing examination questions. computer studies towards formal theories of clinical decision support and setting and answering medical licensing examinations, workshop lecture presentation, in: Proceedings of the IEEE International Conference of Bioinformatics and Biomedicine, 9–11 November, 2015, Washington DC.

[42] J.H. Fetzer, Artificial Intelligence: Its Scope and Limits, Studies in Cognitive Systems, 4, Springer Netherlands, Dordrecht, 1990.

[43] 〈http://www.imdb.com/title/tt1010048/plotsummary〉.

[44] 〈https://en.wikipedia.org/wiki/Turing_test〉.