Computers in Biology and Medicine 43 (2013) 2297–2310

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Suggestions for a Web based universal exchange and inference language for medicine

Barry Robson [a,b,*], Thomas P. Caruso [a,c], Ulysses G.J. Balis [a,d]

[a] Quantal Semantics Inc, North Carolina, United States
[b] St. Matthew's University School of Medicine, Grand Cayman, The Dirac Foundation, UK, University of Wisconsin-Stout, United States
[c] University of North Carolina, United States
[d] University of Michigan, Michigan, United States

ARTICLE INFO

Article history:
Received 19 November 2011
Received in revised form
6 September 2013
Accepted 11 September 2013

Keywords:
Dirac
Bayes inference
Semantic networks
Probabilistic knowledge representation
Hyperbolic complex algebra
Best practice

ABSTRACT

Mining biomedical and pharmaceutical data generates huge numbers of interacting probabilistic statements for inference, which can be supported by mining Web text sources. This latter can also be probabilistic, in a sense described in this report. However, the diversity of tools for probabilistic inference is troublesome, suggesting a need for a unifying best practice. Physicists often claim that quantum mechanics is the universal best practice for probabilistic reasoning. We discuss how the Dirac notation and algebra suggest the form and algebraic and semantic meaning of XML-like Web tags for a clinical and biomedical universal exchange language formulated to make sense directly to the eye of the physician and biomedical researcher.

1. Introduction

1.1. Inference from mining medical data

In one of the earliest studies involving data mining of a large collection of medical records (667,000), we were able to discover several previously unnoted associations between clinical data [1]. Analogous data mining was also recently applied to 6.7 million chemical compound records obtained by automatic reading of all US patents [2], suggesting scalability of a method used [3]. However, discovery is not the only purpose of such projects. These and many other data mining methods deliver probabilistic statements about relationships that can be brought together to provide medical decision support. There is now a growth of interest in using the Web as a common repository of probabilistic relationships for automated reasoning (e.g. Refs. [4,5]). Such methods are of general application, but medicine has long been recognized as the challenging priority for inference from uncertain knowledge [6]. Our own XML-like Web effort, Q-UEL, comprises knowledge-bearing tags as "probabilistic statements" from relatively structured data sources [1] and specialist text [2], as well as "common sense" and general wisdom from thesauruses and encyclopedias, and automatic surfing of the Internet. Many of these can be rendered probabilistic as described below. This is in pursuit of a probabilistic biomedical Semantic Web [7], and a larger vision of a "Thinking Web" WW4 for healthcare dominated by automated reasoning process on portals and servers. It is not a mainstream effort such as the Ontology for Biomedical Investigations [8], so one may wonder what we see as justifying our effort. We seek to resolve various controversies by unusual methods, so we first review the issues and compare related efforts.

1.2. The need for unification

The lack of unified best practice for the probabilistic Semantic Web is recognized as a significant problem [7]. It is often argued that there are only a few *fundamentally different* inference strategies (induction, deduction, abduction, propositional predicates of order 1 or more, epistemic modal logic, fuzzy, or Bayesian), but "fundamentally different" is a point of view. For many years each new project has typically introduced some new and often controversial features, a tradition set by the pioneering MYCIN Expert System [6]. Some methods (e.g. Ref. [5]) appear to have no special adherence to the Bayes Net (BN) principles that BayesOWL champions [4], whilst others do but still differ significantly (e.g. Ref. [9]). BN philosophy itself shows increasing variations [10,11], and even what the probabilities really are, or should be, troubles

* Corresponding author at: St.Matthew's University School of Medicine, Regatta Park, Grand Cayman, Cayman Islands.
Tel.: +1 345 945 3199x193.
E-mail address: robsonb@aol.com (B. Robson).

0010-4825/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.compbiomed.2013.09.010

Web developers [7]. However, quantum mechanics (QM) [12,13] has been of growing interest to us, because it is the one long-established discipline that at least frequently *claims* to have universal applicability for inference involving probability. Consequently, Q-UEL stands for *Quantum Universal Exchange Language*. This also relates to another need for unification. Admittedly, the Semantic Web has been converging to just a few ways to represent and communicate data and knowledge (i.e. mainly XML, OWL and RDF technology) [14–16]. Yet in December 2010, the President's Council of Advisors on Science and Technology (PCAST)[1] controversially called for a fresh start and an *"XML-like Universal Exchange Language"* for healthcare [17]. We responded with a preliminary proposal [18] following studies [19–24], unusual by being based on Dirac's view of QM [12], but algebra known to A.I. [25–28]. However, there are other differentiators, as follows.

1.3. Comparison of Q-UEL with related efforts

Some offerings stand out to us by having both similarities and differences to Q-UEL. XML itself stands out by having the broadest scope with the efforts discussed below usually expressed in XML, because Q-UEL is *not* expressed in XML, although we do see it as an XML *extension* designed inter-conversion with XML in mind.[2] Q-UEL tags can be used in an XML-like way to structure documents, but unlike XML, Q-UEL tags always have *intrinsic algebraic* meaning and also aspire to simple *linguistic interpretation*. A tag can represent a *statement* in Q-UEL, i.e. a declarative expression that can be assigned a probability as to its truth or scope, by observations and inference.[3] XML can of course "do anything" in terms of specific embodiments. Comparable efforts at the Q-UEL-like level are:

1.3.1. pDAML-OIL [5]

like Q-UEL, focuses on three essentially probabilistic values. However, pDAML-OIL's three values are the degree of truth, the degree of falsity, and the degree of inconsistency with the knowledge. In pDAML-OIL expressed in XML, we have comparable constructions like ⟨Person rdf:ID="#Bob" pdaml:prob= "0.7/0.1/0.1"/⟩. It "*describes an uncertain rdf:type property* (*it is unclear if Bob is a person*)" [5]. It is hard to imagine how this could describe the reliability of a measurement of, say, blood pressure in a way that physicians (and many application developers) could understand directly. Q-UEL achieves a similar and (if required) same effect[4] using two attributes *Pfwd* and *Pbwd* as "directional" probabilities such as the conditional probabilities $P(A|B)$ and $P(B|A)$, supplemented by an attribute *assoc* for association constant $K(A; B)=P(A, B)/P(A)P(B)$. Just from these $P(A|B)$, $P(B|A)$ and $K(A; B)$, one can readily compute[5] all other probabilities involving A and B, like $P(A)$, $P(A \text{ \& not } B)$, and so

1.3.2. BayesOWL [4]

like Q-UEL, emphasizes familiar conditional probabilities such as $P(A|B)$ that can often be interpreted categorically as P("All B are A") but also relationships with the force of P("B causes A"). It expresses the Bayes Net and provides features to aid its construction. Whilst BayesOWL and Q-UEL can convey the same familiar probabilities, BayesOWL avoids combinations that imply cyclic paths that Q-UEL allows (see Section 1.4). The traditional Bayes Net is also constrained to use of AND logic as multiplication of conditional probabilities as the only relationship. It does not evolve by reasoning new probabilistic statements from those already available, and is confined to monotonic logic. Q-UEL addresses all these in various ways. BayesOWL is said to define probabilistic relatedness of *distinct classes*. From Q-UEL's perspective the term "distinct classes" is either restrictive or misleading, because it wants to convey, for example, to what extent obese patients and type 2 diabetics are not distinct classes.

1.3.3. PR-OWL [9]

would appear to agree. It follows BayesOWL but considers the Bayes Net, albeit as the Multi-Entity Bayesian Network (MEBN) in a way that does not suggest a demand for distinct classes. Although Q-UEL can, like PR-OWL, support probability distributions relating to MEBN fragments (which Q-UEL calls *relevancy sets*), an important similarity is that both Q-UEL and PR-OWL recognize first order predicate logic, albeit in a different way. Q-UEL states that all of universal ("all") and of existential ("some") degrees of character, and the concept of negation, can be captured in a single scalar value. Normally, $P(A|B)$ as P("All B are A") and $P(B|A)$ as P("All A are B") are purely symbolic adjoints, requiring Bayes theorem $P(A|B)P(B)=P(B|A)P(A)$ to evaluate one from the other. In Q-UEL the *dual probability* ($P(A|B)$, $P(B|A)$) is a special kind of *complex* value of broad importance.

1.3.4. The OBI (the Ontology for biomedical investigations project) [8]

develops an integrated ontology to support specifically biological and clinical investigations. Its goal is a crisp definition of a set of terms applicable across various domains, as well as domain-specific terms. Q-UEL understands that importance and extends the use of RDF-like methods for clear definitions, and pursues an unambiguous formal language (Section 5). However, "obese patients are type 2 diabetics" can be reinterpreted as a probabilistic ontology, and constructing one's own ontology and trying to work out the possibly new ontological relationships implied in e.g. natural language text, are very different matters (Sections 2.2 and 3.9). These require management of uncertainty or degree of scope in interpretation. Nonetheless the OBI approach can support efforts like the following that allow uncertainty for statements corresponding to hypothesis, evidence, and conclusion. They place probability at the level of classical statistics, rather than of the ontology itself.

1.3.5. HELO [27] defines a hierarchy of research statements

It can combine research hypotheses, negative and alternative hypotheses, assumption, conclusion, and scientific laws. Most importantly, these appear in *production rules* (statement A→statement B). In Q-UEL, *metastatements* (Section 2.6) transform (statement A, statement B, …)→(statement R, statement S,…) by a partial-match and edit approach with binding variables. Q-UEL thus supports deductive or inductive reasoning (although deductive is commonest). Compare SWRL [28] HELO recognizes Bayes Theorem as a rule, but it stands

risk, incidence, prevalence, mortality rate, fatality rate, accuracy, sensitivity and specificity, and also odds such as relative risk, predictive odds, LR+, and LR−, and the odds ratio, and number needed to treat/harm, etc.

---

[1] This is because of their observation that "In other sectors, universal exchange standards have resulted in new products that knit together fragmented systems into a unified infrastructure. By contrast, health IT has not made this transition" [17]. The controversy, however, is more in regard to a fresh start.

[2] Conversion to Q-UEL of XML documents as complex as health records is not particularly easy, and indeed Q-UEL was also developed as an alternative health record representation for more readily extractable data. But at least Q-UEL could diplomatically provide a universal "second language".

[3] Dictionaries also define "statement" as a sentence of words grouped meaningfully to convey knowledge, and as an element of an imperative and sometimes declarative programming language. While Q-UEL alone is a very basic programming language for calculations for inference, Q-UEL tags as algebraic objects, i.e. constant and variable scalars, vectors and operators, may be embedded in a host programming language (Section 3.1).

[4] This must be described elsewhere, but n-valued probability systems imply $(n+1)$-valued logic, and the conversion tools needed are Bayesian belief, odds ratio, and the logical law of the contrapositive.

[5] $P(A)=P(A|B)/K(A; B)$, $P(B)=P(B|A)/K(A; B)$, $P(\text{not } A)=1-P(A)$, $P(\text{not } B)=1-P(B)$, $P(A \text{ \& } B)=P(A|B)P(B)$, $P(A \text{ \& not } B)=P(A)-P(A \text{ \& } B)$, $P(\text{not } A \text{ \& } B)=P(B)-P(A \text{ \& } B)$, $P(\text{not } A \text{ \& not } B)=1-P(A \text{ \& } B)-P(A \text{ \& not } B)-P(\text{not } A \text{ \& } B)$.

as one rule alongside many others, whilst in Q-UEL it has a deeper importance, as follows.

1.4. Cleaner knowledge models and coherence

We were challenged by a referee's view that "cycles indicate bad knowledge representation, and not the feature of the knowledge itself. The authors should advocate for cleaner knowledge models, and not to develop methods for dealing with bad practices". This elegantly and crisply describes a school of thought worthy of discussion, but it is certainly not the only one (e.g. Ref. [10]). What is disliked about cyclic paths in knowledge representation is that a node can represent a cause that has effects which can ultimately affect thecause, or ontologically seems to imply a probability of nodes in a cycle being equivalent. For those not so deterred, "The fundamental problem with reasoning in cyclic Bayesian networks is that the joint probability distribution is not given explicitly, but is rather an asymptotic limit of a multi-stage process" [10]. However, Q-UEL inference does not usually require multistage iteration, and health workers happily consider risk factors *P*(type 2 diabetes|obesity), *P*(obesity|overeating), and *P*(overeating|type 2 diabetes) by sampling a population, and explore causality in both directions. The conceptual or probability-bookkeeping problem that researchers have with cyclic paths is in large part because in going from A to B to C and so on and back again to *A*, the situation should necessarily be seen as bidirectional, hence Q-UEL's Pfwd and Pbwd. Bidirectional systems can lack cycles, but cycles require bidirectional systems. If, however, we change or provide new probabilities in an inference network derived from such probabilities, then we need to ensure coherence [11]. We certainly concur with the desire for those "cleaner knowledge models", but in our view, coherence, not insisting on acyclic character, is the issue, and a more general one. Since the global market crash, financial risk analysts using acyclic Bayes Nets have seen that they have typically been implemented in a way that has little to do with Bayes, and that they should be enforced to be coherent internally and with the data, by actually satisfying Bayes Theorem [11]. Depending on its precise definition, coherence is only partly a matter of marginal sums.[6] Coherence is primarily the need to satisfy Bayes Theorem $P(A|B)P(B)=P(B|A)P(A)$ [11]. Ironically, Bayes Nets cannot by themselves satisfy that requirement because they see only one direction of conditionality, $P(A|B)$ [11]. The importance is discussed in Section 3.6.

2. Theory

2.1. The Dirac basis of Q-UEL

The basis of Q-UEL is the probabilistic language and inference system known as the *Dirac notation* [12] including the *Clifford–Dirac algebra* [13], founded in the 1920s–1930s. The algebra implies, amongst other things, the dual probabilities interpretable as complex values. Dirac notation already looks like the required format extension capturing the Semantic Web's RDFs and semantic triples generally [14–16] in an elegant way (Section 2.3). It enables a quantitative or *probabilistic semantics* that Dirac himself appears to have envisaged.[7] However, Dirac's complex number of interest to us as behaving classically is not i where $ii=-1$ responsible for a wave description, but the *hyperbolic number* $\boldsymbol{h}$ such that $\boldsymbol{hh}=+1$. Dirac did not call it that, and it has gone under many guises that until fairly recently obscured a persuasive convergence of disciplines and uses relevant to our argument.[8] The dual (Pfwd, Pbwd) as our hyperbolic-complex probability value or *hyperbolic probability amplitude* can be represented as the *Hermitian commutator.*

$$(\text{Pfwd},\ \text{Pbwd}) = \frac{1}{2}[\text{Pfwd}+\text{Pbwd}]+\frac{1}{2}\boldsymbol{h}[\text{Pfwd}-\text{Pbwd}] \tag{1}$$

Pfwd and Pbwd are easy to understand as P(A|B) and P(B|A) respectively, but more generally, it is as if the vertical bar "|" were replaced by an *operator*. QM in this kind of way occupies a Hilbert space, essentially a multidimensional space of complex values [13]. It is normally seen as i-complex [13], but the Lorentz rotation $i\rightarrow\boldsymbol{h}$ gives us the $\boldsymbol{h}$-complex description [19]. Eq. (1) describes a point in that new space.

2.2. Implications for Q-UEL

Q-UEL's commonest tag type is

⟨*subject expression*|*relationship expression*|*object expression*⟩

This is Dirac's *"bra-operator-ket"*. If anything, XML implies logical AND between its attributes, but whilst this implied logic is common in Q-UEL and so the absence of AND implies AND by default, "expression" above can mean a much more elaborate logical or other expression.[9] A "simple" and typical example tag[10] of this type is

⟨Q-UEL-MOLECULE Ampicillin | **means**:=www.qexl.org/means_2 | code:=IUPAC:= '(2S,5R,6R)-6-{[(2R)-2-Amino-2-phenylacetyl]amino}-3,3-dimethyl-7-oxo-4-thia-1-azabicyclo[3.2.0]heptane-2-carboxylic acid' **or** code:=SMILES:=O=C(O)[C@@H]2N3C(=O)[C@@H](NC-(=O)[C@@H](c1ccccc1)N)[C@H]3SC2(C)C **or** code:=InChI:=1S/C16H19N3O4S/c1-16(2)11(15(22)23)19-13(21)10(14(19)24-16)18-

[6] That is, $\Sigma_X P(X)=1$ for all mutually exclusive alternatives, and similarly $\Sigma_X P(A, X)=P(A)$ and so on. More correctly, that is the case if we have all the X, but at least we must have $\Sigma_X P(X)\leq 1$ and $\Sigma_X P(A, X)\leq P(A)$ if all X are mutually exclusive.

[7] As Dirac noted in his Nobel Prize Banquet Speech (1933), "*the methods of theoretical physics should be applicable to all those branches of thought in which the essential features are expressible with numbers*". Numbers include probabilities, of course. But that Dirac was modestly referring to his extensions to physics (via his notation and algebra) as also extensible to human language and thought, i.e. as a quantitative semantics, seems plausible because he frequently excluded poetry and (more controversially) economics as subjective.

[8] This needs some brief review. In physics $\boldsymbol{h}$ appears diversely with general application as Dirac's linear operator σ [12], and as γ_{time} (γ_0 or γ_1) of the Dirac–Clifford algebra, and γ_5 of Quantum Field Theory and particle physics [13]. Dirac himself seems have thought that he discovered it, but he was preempted by Cockle and Clifford by some 70 years [13]. Dirac may be forgiven because it appears in the literature in at least twenty other names such as Cockle, Lorentz, motor, or split-complex number. Importantly, it is a generalization of the Wick rotation that renders QM classical [13], and split-complex multiplication has also been seen as a Lorentz boost of a space-time plane [13], where number $z=x+\boldsymbol{h}y$ represents an event in that plane, as does Eq. (1), justifying a "calculus" of bidirectional causality. $\boldsymbol{h}$ *also* appears in the guise of a variety of non-scalar matrices and algebras exhibiting the same essential hyperbolic properties, and like i it can adopt different algebraic variants or "flavors". In the course of our project, we came across the increasing focus (since the 1990s) on use of the same $\boldsymbol{h}$-complex algebra in neural networks (e.g. Ref. [25]). There, one main advantage is that the XOR problem resolves naturally and more computation can be done with a single neuron. It may similarly be an important method of computation in the natural neural networks of the brain [26]. Both these use "hyperbolic", and it seems increasingly common ($e^{\boldsymbol{h}x}=\cosh(x)+\boldsymbol{h}\sinh(x)$ being a hyperbolic function). It is adopted here.

[9] Formally, each expression should be some description that can in principle be observed and counted, so that a probability can be assigned to it. Q-UEL follows Dirac notation in that entities inside tags can be scalar, vector, or matrix or operator and can be moved around the tag, providing Dirac's rules of adjoint algebra are followed for movement in and out of a bra [12,13]. An expression should still, however, describe a countable situation.

[10] Note that QM textbooks are more lax as to QM's "attributes", using say p where Q-UEL would write momentum(eV-sec/Angstrom):=0.2. That is because p etc. are often variables for the purpose of general discussion, although something very similar to Q-UEL is often seen in textbook examples with set values.

12(20)9(17)8-6-4-3-5-7-8/h3-7,9-11,14H,17H21-2H3,(H,18,20)(H,22,23)/t9−,10−,11+,14−/m1/s1 **and** 'empirical formula':=C16H19N3O4S **and** 'monoisotopic mass':=349.109619 **and** 'average mass (Da)':=349.404785 Q-UEL-MOLECULE⟩.

The tag-name is considered a special attribute; other metadata or RDF-links can be appended to it as for any Q-UEL attribute. It is a stand-alone tag. There is an *XML Q-UEL miscible format* or *XQMF* that is valid XML: for the above tag type, it is ⟨*subject expression* relationship="*relationship expression*" *object expression*/⟩. It is to be understood throughout that the expression in a XQMF will for XML include the tag name where appropriate. Note the RDF-like link on **means**, and the convention of showing relationship operators including logical operators etc. in bold font. We say "simple" above in quotes because the above tag is obviously rich in information content, but the structure is relatively simple. Not least, probabilities Pfwd and Pbwd as tag value attributes are lacking. The tag thus has a value of 1, which is the default. More generally, tags values (Pfwd, Pbwd) should have a real part that lies on the interval 0…1, and an imaginary part, if any, that lies on the $-1/2\ldots+1/2$ interval. An exception is allowed for general utility when the bra-operator-ket is a *dyadic function* (Section 2.4). Following Dirac notation, there are also the following less common tag forms, with their normally allowed value ranges.

⟨***expression***⟩ is called the *expectation*. Its XQMF is ⟨*expression*/⟩. Whilst commonly used for averages, it can have any *real* value carried as the value of an attribute called "value", but if called *P* it is a probability with a value on the interval 0…1, being Q-UEL's way of writing a *self probability P*(*A*).

⟨***subject expression*** **|** ***object expression***⟩ called the *braket* is used for a purely conditional relationship. The preferred form of ⟨A|B⟩ as a semantic triple is the synonymous ⟨A| **if** |B⟩. Its XQMF is ⟨*subject expression* relationship="**if**" *object expression*/⟩. As for the bra-relator-ket, its complex value, carried as the values of two attributes Pfwd and Pbwd, must have a real part that lies on the interval 0…1, and an imaginary part, if any, that lies on the $-1/2\ldots+1/2$ interval. It is Q-UEL's way of writing a *conditional probability P*(*A*|*B*), or more correctly the ***h***-complex dual *P*((*PA*|*B*), *P*(*B*|*A*)).

⟨***expression***| called the *bra*, can also appear as ⟨*expression*|⟩ in Q-UEL. It is a row vector with values that can be carried as attribute values (see Section 3.2). Its XQMF is ⟨*expression*⟩. The values of the vector array elements (carried as attribute values) are as for the braket.

|***expression***⟩ called the *ket*, can also appear as ⟨|*expression*⟩ in Q-UEL. It is a column vector. Its XQMF is ⟨/*expression*⟩. The values are as for the bra.

|***expression***⟩⟨**expression**| called the *ketbra*, can also appear as ⟨|*expression*⟩⟨expression|⟩ in Q-UEL. The XQMF is⟨*expression* relationship="***operator***" *expression*/⟩, where operator is the characters "matrix" or a specified operator. It can be operator as described by an algorithm, either (a) intrinsic to standard Q-UEL applications or (b) extrinsic and downloadable from an RDF-like reference, e.g. in an attribute **operator:=join**:=download:=www.qexl.org/join_operator_3/ in the bra part. If it is a matrix, its elements are as for bra and ket carried as described in Section 3.2.

As in Dirac notation, Bra-relator-kets, brakets, and ketbras are actually themselves expressions as bra and ket products, but can be algebraically manipulated as integral objects, and in Q-UEL's case as tags. The braket and bra-relator-ket are mathematically correct as being real or complex *scalar* objects. The ketbra is mathematically correct as being a real or complex operator. As in XML, Q-UEL tags can be used to impose an ontological structure to a document. When bras and kets do function like XML's opening and closing tags, they *still have algebraic force. Content* between Q-UEL tags is an *identity operator* "variable by name", with real scalar value 1. Identity operators are placed in double quotes "….", and construction ⟨…| "…" |…⟩ is read as one tag. To avoid any ambiguity paired double quotes ""…"" are preferred here; single quotes are preferred in attributes. However, the following approach is preferred to impose structure.

2.3. Q-UEL as an Extended EAV model

The *event-attribute-value* model or *EAV* model (also known as the *vertical database model* or *open schema*) supports the sparse matrix model for data [29]. A major purpose is data mining scalability.[11] Q-UEL extends the EAV approach by expressing data structure through its *Attribute Metadata Language* (AML) allowing attributes to be extended as larger linear, tree, or general graph structures, e.g. 'cardiovascular':='blood pressure (mmHg)':=(systolic:=140, diastolic:=90), and 'cardiovascular':= 'blood pressure (mmHg)':= (systolic:=140:=pbwd:=0.95, diastolic:=pbwd:=0.9). The *metatadata operator* ":=" can be replaced by its XML counterpart "=" and understood, but technically it means that everything to the right of last ":=" in that branch is attribute value (not metadata), and Q-UEL encryption tools encrypt attribute values. Q-UEL uses AML for vectors and matrices in various ways, e.g. Pbwd:=(0.3,02,0.15,0.01) or Pbwd:=((0.1,0.2), (0.3,0.2)).[12] Q-UEL can host, via the *code attribute* e.g. (code:=SMILES:=), external standard notations that may sometimes imply a graph which Q-UEL can use for broader purposes than the original domain of application.[13] Units are considered as part of metadata, but Q-UEL has several medically important features that can extend the sense of the attribute's value.[14]

2.4. Relator as Dyadic function

An inference network is seen as an algebraic expression or program with variables as Q-UEL tags or simplified forms of them. The bra-relator-ket may be written as ⟨*A*|***R***|*B*⟩. ***R*** is the *relationship operator* or *relator*, such as a verb or preposition in linguistic use, and/or an operator that may often be seen as a function. ***R'***s action can be defined by downloadable code by attaching an RDF-like link. But in absence of such a link, an application as a compiler will look for a definition of ***R*** as algorithm defined in a function

[11] Data miners have frequently used Semantic Web text analytics methods to extract information from XML-based medical records and re-represent the data in EAV format, because requisite information is often more dispersed or elaborately and diversely structured than structured data mining would like. It is unlikely that even this can scale to the future's hundreds of millions of electronic health records changing dynamically. Q-UEL would solve that by having records written in Q-UEL from the outset, but is more diplomatically positioned as a "universal second language".

[12] In Q-UEL this represents a *tensor system*. Vectors and matrices that are brought together in operations that normally require matching numbers of elements are rationalized by increasing the number of elements to match, combined with down-weighting their element values accordingly. The scalar results of vector–vector and vector–matrix–vector multiplications are readily shown to be unaffected. The method is important because it is particularly valuable for vectors as probability distributions.

[13] For example, SMILES code developed for chemistry used in Ref. [2], can also be used to represent *any* graphs succinctly, by a slight modification (esp. that atom names like Cl, chlorine, are replaced by words such as Cat). But then Q-UEL would say e.g. taxonomy:=code:=SMILES:='Q-UEL modification' 2:=…..

[14] Q-UEL commonly extends attribute values by three important dimensions, date/time stamp (GMT), description of degree of consent e.g. (consented to nearest 10 and year), and reliability of the measure, e.g. 140+/−8CI where CI is a confidence interval (see also Section 2.7). Standard deviation SD and standard error SE may also be used, with their usual statistical meanings. Q-UEL allows 150∼, i.e. "approximately 150". Note the use of "?", (Section 3.5) in e.g. bloodwork:=fasted:=glucose (mg/dL):=150?. The value in glucose(mg/dL):=150!!? is made individually meaningless but in such a way that a data miner can determine a meaningful *average value* from many; for safety, special tag names are used.

subroutine. In these ways, ⟨*A*|***R***|*B*⟩ is also a *dyadic function* with arguments A and B. If it is provided, any value may be returned. For example, *plus* might be defined so that ⟨*3*|*plus*|*4*⟩ has the value 7, or ⟨*$x*|*plus*|*$y*⟩ has the value dependent on the variables. If not provided, it does not prohibit probability calculations and symbolic operations with ⟨*A*|***R***|*B*⟩ in reasoning.

2.5. The Twistor construct

In Dirac's algebra, arguments in expressions in bras and kets can be any scalar, vector, or matrix, subject to certain rules [12]. By that license, *tags can stand as attributes in tags*. It is also consistent with the following. ⟨A|B⟩ can be seen as the Dirac dual or 2-*spinor* [13], and the physicist's *twistor* [13] such as the construct ⟨⟨*A*|B⟩⟨*C*|*D*⟩|⟨*E*|*F*⟩⟨*G*|*H*⟩⟩. However, physics texts usually use different delimiters for a twistor, e.g. {…} as the brackets. We refer to "extended twistor notation", which also includes extending the usual twistor seen as 4-spinor by often replacing the vertical conditioning bar by "|***R***|" and allowing expressions of ⟨*X*|***R***|*Y*⟩ in bra and ket parts, and indefinite embedding. It is an alternative to XML's structured document use. It is also convenient when we want to represent the parsed structure of a sentence as a *semantic multiple*, or represent the sub-graph or even whole of a knowledge network as a graph. We could exploit here several kinds of imaginary number, but Q-UEL currently focuses on just ***h***.

2.6. Metastatements as twistors

Q-UEL algebra has *statements* and *meta-statements*. The distinction is clear in Q-UEL because a metastatement contains at least one *binding variable* $*A*, $*B* etc. (with a name starting in upper case). A binding variable allows metastatements partially to match and act [2] on one or more statements or other metastatements, say generating a new statement from them, as in syllogistic reasoning.[15] Metastatements dynamically evolve a network. Metastatements allow users to define and enforce laws of basic and categorical logic, and of semantics and language definition. Extended twistor notation is also the formal way of expressing and transmitting a Q-UEL metastatement. The principle is easy to understand from the following syllogism: ⟨⟨$*A*| **are** |$*C*⟩ |=| ⟨$A| **are** |$*B*⟩⟨$*B*| **are** |$*C*⟩⟩, which can evolve an inference network by hunting for each of the two bra-relator-kets to the right, and (in the network mode) replacing them by the equivalent bra-relator-ket to the left. Linguistic and semantic examples are ⟨⟨$*A*| **is** |$*B*⟩|=| ⟨$*A*s| **are** |$*B*s⟩⟩, ⟨⟨$*A*| **"are eaten by"** |$*B*⟩ |=| ⟨$*B*| **eat** |$*A*⟩⟩, ⟨⟨$*A*| **marries** |$*B*⟩ |=| ⟨$*B* |**marries**|$*A*⟩⟩, ⟨⟨$*A*| **as** |$*B*⟩ |=| ⟨$*A*| **is** |$*B*⟩⟩, ⟨⟨*A*| **$*R*** |$*B*⟩ |=| ⟨$*A*| **are** |$*B*–**$*R***ers⟩⟩, ⟨⟨$*A*| **pays** |$*B*⟩ |=| ⟨$*A*| **gives** |money⟩ ⟨money| **to** |$*B*⟩⟩. Note ⟨⟨$*A*| **said** |"⟨$*B*| **are** |$*C*⟩" |=| ⟨$*A* | **said that** | ⟨$*B*| **were** |$*C*⟩⟩ and the gerund conversion ⟨a **$R**ing $*A* |=| a $*A* **$*R***s⟩.[16] It will be evident that these can be used to express languages other than English including Q-UEL's own artificial QUELIC (Section 5). Q-UEL also allows partial match using regular expressions [2], so that can cover a lot of cases. Like any statement they can have probabilities, to express confidence or scope of the metastatement. Some conversions depend on probability values: ⟨⟨$A| are |$B⟩ |=| ⟨**non**-$B| **are not** |$A⟩⟩ is only valid as to probability value if Pfwd ≈ 1. There is an extended notation that allows metastatements with e.g. $$A to act on metastatements with $A, and elaborations on this will be described elsewhere.

2.7. Medical interpretations of dual probabilities

Consider a summary statement ⟨Q-UEL-EBM 'Systolic BP(nearest 10)':= 'over 140 (2012)' Pfwd:=0.875 | **when**:=www.qexl.org:=/when_1/:=assoc:=2.438 | Age:='at least 50 (2012)' **and** BMI:='30 (2012)' **and** 'Fat(%)(nearest 5)':='over 30 (2012)' **and** female **and** club:=1 Pbwd:=0.703 observed:=17892 expected:=7214.516 time:=Fri May 3 12:00:16 2013' Q-UEL-EBM⟩.

Note that standard nominal-categorical values like male need not have metadata. Pfwd and Pbwd are respectively the probabilities P("*A* ***R*** *B*") and P("*B* ***R*** *A*"), first as we would first read the statements i.e. linguistically, and then after subject and object expressions are switched. For example,⟨obesity|**causes**|type 2 diabetes⟩→⟨type 2 diabetes|**causes**|obesity⟩. Medically, neither of these are certainly (or certainly not) the case for any individual, but both occur and the probabilities each way are of medical interest. In general, our entities satisfy the following where the * indicates the complex conjugate, changing the sign of any imaginary part. Usually in Q-UEL, ***R*** is Hermitian, meaning

$$\langle \mathrm{A}|\boldsymbol{R}|\mathrm{B}\rangle = (\mathrm{Pfwd},\ \mathrm{Pbwd}) = \langle \mathrm{B}|\boldsymbol{R}|A\rangle^{*} = (\mathrm{Pbwd},\ \mathrm{Pfwd})^{*} \tag{2}$$

The interpretation only *seems* different for an *incidence statement* (not to be confused with epidemiological incidence rate), say for a clinical message or record entry:-

⟨Q-UEL-PATIENT:=#5473 Pfwd:=0.95 | **has**:=www.qexl.org/has_6/ | cardiovascular:='blood pressure (mmHg)':=www.qexl.org/ blood_pressure/:=systolic:='140+/−8CI (Wed Oct 3 14:02 2012 GMT) (consented visible cardiovascular)' club:=Virginia Q-UEL-PATIENT⟩.

Note the tag-name-as-attribute usage, the dispersion, time, and consent dimensions seen as fundamental to a value, and that more typically we would use the attribute metadata language to include systolic BP and maybe pulse etc. in the attribute. It is still probabilistic because observations are prone to error, and at least we know that there is a typical uncertainty associated with such a measurement[17]. Pfwd is interpreted as P(value lies within range specified | value lies within the fullest possible range)=0.95 by definition of the confidence interval CI of a one dimensional normal distribution[18]. Hence Pbwd=P(value lies within the fullest possible range|value lies within range specified)=1. If we wish to mean otherwise, it is a different relator, e.g. indicating prevalence of the patient's range of systolic blood pressure in that population (Pbwd:=0.34), or the chance that any such measurement encountered is of patient #5473 (Pbwd:=1.2E−7). The above CI theoretically implies repeated measurements on one patient around that time with same conditions, but measurements giving the normal range over a population have a similar interpretation for Pfwd and Pbwd:-

⟨Q-UEL-POPULATION:=#23 Pfwd:=0.95 | **has**:=www.qexl.org/has_7/| cardiovascular:='blood pressure (mmHg)':=www.qexl.org/blood_pressure/:=systolic:='125+/−30CI(2012)(consented visible cardiovascular and year)' club:=Virginia Q-UEL-POPULATION⟩.

[15] Such metastatements are thus technically operators. Contrast *calculation variables* $*x* and $*y* (name starting lower case) in tags, with constant values substituted at execution time.

[16] Note also that irregularities in spelling in converting grammatical status (inter-conversion involving –ing, -ed, -s, -ly,) and semantic exceptions generally are coped with by being less general and having a variety of more specific cases with less binding variables. Surprisingly there are not so many specifics in English, as speakers subconsciously use a limited set of English conversion rules for spelling.

[17] Indeed there are characteristic rates of recording male/female incorrectly (albeit one in millions), of an assigned ethnicity being genetically and genealogically inappropriate, and of a lifestyle change.

[18] It may not be 0.95 for more than one medical value collectively, or if it is not a normal distribution. If it is far from being a normal distribution (and particularly if researchers are involved) then the 140+/−8CI, and Pfwd and Pbwd probability values can replaced by vector that describes the possible values and the corresponding probability distribution, though there is a condensed notation for unimodal distributions.

2.8. Physical interpretations of dual probabilities

Classically, $P(A|B)=P(A)e^{x}$ where e^{x} is the assoc attribute value with x as mutual information $I(A;B)$ (Section 2.12). QM probability amplitudes are similarly proportional to $e^{\boldsymbol{i}x}=\cos(x)+i\sin(x)$, and to $e^{\boldsymbol{h}ix}=e^{-i\boldsymbol{h}x}=\cos(x)+\boldsymbol{h}i\sin(x)$ in a bigger Dirac picture. Hence they are both wave functions, but those proportional to $e^{\boldsymbol{h}x}=\cosh(x)+\boldsymbol{h}\sinh(x)$ are not, yielding classical results [19–24] re-expressible as normal distributions [20]. In contrast to i, $\boldsymbol{h}$ has real eigenvalues $+1$ and -1 respectively (see Dirac's discussion of his linear operator σ as $\boldsymbol{h}$ around his Eqn. 23 Chapter 10 [12]). Consider a simple inference problem ⟨overeating|**causes**|obesity⟩ ⟨obesity|**causes**|type 2 diabetes⟩. The eigensolutions for $\boldsymbol{h}\to+1$, $\boldsymbol{h}\to-1$, are the chain rule estimates as P("obesity causes type 2 diabetes") and P("type 2 diabetes causes obesity"). QM brakets etc. and Q-UEL tags are *poised prior to Dirac's normalization recipe* [12] for observable probabilities, and still complex. The recipe starts with ket normalization that sets (Pfwd, 1) because we prepare B and observe A conditional upon it. It is followed by the Born rule as (Pfwd, 1)(Pfwd, 1)*=(Pfwd, 1)(1, Pfwd)=Pfwd. This may be proven using Eqs. (1) and (2). We are not usually interested in A and B in biomedicine as *conjugate variables* where $A=f(B)$, $B=f^{-1}(A)$ so that Pfwd=Pbwd (exceptions include Boyle's gas law). Hence we can also proceed directly with bra normalization (1, Pbwd) to obtain Pbwd in the same way. Some familiarity with QM might suggest that we should see or use Pfwd, Pwd, and assoc values as their square roots, but all the above does not indicate this, except to remain as an optional *hedge interpretation* [19] (Section 3.4).

2.9. Comparison with Semantic grammar

The Semantic Web can consider relationships in basic grammatical terms of symmetry, transitivity, and functionality. So can natural language. So also can QM, allowing us to express these ideas in an algebraic and *quantifiable* way. Eq. (2) with $\langle A|\boldsymbol{R}|B\rangle=\langle B|\boldsymbol{R}|A\rangle^{*}$ defines $\boldsymbol{R}$ as *Hermitian* [13], as are the important QM operators. Formally this rests on the *adjoint operator* $^{\dagger}={}^{T}*={}^{*T}$, where ^{T} indicates taking the transpose of rows and columns, and $\boldsymbol{R}$ has symmetry of its elements such that $\boldsymbol{R}=\boldsymbol{R}^{\dagger}$. It is *trivially Hermitian* if $\boldsymbol{R}=\boldsymbol{R}^{*}=\boldsymbol{R}^{T}$, and is *non-trivially Hermitian* when this is not so. When the latter, we can distinguish $\boldsymbol{R}$ and $\boldsymbol{R}^{*}=\boldsymbol{R}^{T}$ as active-passive inversions of a verb (e.g. eat/are eaten by), or analogously for a preposition (e.g. on/under) and so on. This gives the semantic equivalence $\langle A|\boldsymbol{R}|B\rangle=\langle B|\boldsymbol{R}^{*}|A\rangle$, as in "dogs chase cats", and "cats are chased by dogs". The most basic relator is the conditional interpretation of the bra-ket product $\langle A||B\rangle$ as the braket $\langle A|B\rangle$, which may be rendered as the triple ⟨A| **if** |B⟩=⟨B| **therefore** |A⟩. The important categorical and set theoretic interpretation is ⟨A| **include** |B⟩=⟨B| **are** |A⟩. In that case, $\text{Pfwd}=P(A|B)=P(A,B)/P(B)$ and $\text{Pbwd}=P(B|A)=(A,B)/P(A)$. For non-categorical relationships, the QM concept of *projection of operator values* leads e.g. ⟨dogs|chase|cats⟩=(P(dogs:=chase | cats:=chase*), P(cats:=chase| dogs:=chase*)), although normalization by inclusion of required linking probabilities in an inference network gives (P(dogs:=chase, cats:=chase*), P(cats:=chase, dogs:=chase*)). It remains $\boldsymbol{h}$-complex. We can design unambiguous formal languages (Section 5) by Dirac's rules. The main rule is that if we move entities or expressions of them into or out of the bra ⟨…|, we use their adjoint. For example, in the twistor form this means $\langle\langle A|B\rangle|\langle C|D\rangle\rangle^{*}=\langle\langle D|C\rangle|\langle B|A\rangle\rangle$, as it does in physics [13].

2.10. Comparison with classical categorical logic

Q-UEL seems unlikely to have much to compare and contrast with a project some 2380 years old. Nonetheless, classical categorical logic implies the idea of a dual probability as a point in a complex plane that expresses existential "some" and universal "all" content. For example, we may be interested in P("obese patients are type 2 diabetics"), noting that it only differs *quantitatively* from P("cats are mammals") and P("mammals are cats"). We interpret the real value ½[Pfwd+Pbwd] as the extent to which the statement implied by ⟨A| ***are*** |B⟩ is *existentially qualified* on the scale 0…1, and the imaginary part ½[Pfwd−Pbwd] the extent to which the statement is *universally qualified* on the scale $-1/2\ldots+1/2$. We can say that ⟨**some** A| **are** |B⟩=Re⟨A| **are** |B⟩=Re⟨B| **are** |A⟩=½[Pfwd+Pbwd], whilst ½[Pfwd − Pbwd] represents either ⟨**all** A| **are** |B⟩ or ⟨**all** B| **are** |A⟩ according to how much the value approaches $+1/2$ or $-1/2$ respectively. The imaginary part vanishes when Pfwd=Pbwd, and then a fully existentially qualified categorical statement lies on a continuum between ⟨A| **are** |B⟩=1 when A and B are equivalent, i.e. absolutely indistinguishable, and ⟨A| **are** |B⟩=0 when they are absolutely distinguishable, i.e. mutually exclusive. An important point on that continuum is when A and B are indistinguishable except by recurrence, and are independent and so Bernoulli-countable. Here, ⟨A| **are** |B⟩=$P(A)=P(B)$. Evidently we do not have the same interpretations for the non-categorical relator such as in ⟨**some** A| **eat** |B⟩, except through the interpretation of its *categorical semantic equivalent* ⟨some A| **are** |B-eaters⟩. We say that say that ½[Pfwd−Pbwd] governs the extent to which the relationship is non-trivially Hermitian.

2.11. Ontological probabilities

The above implies probabilistic ontology. XML documents that impose ontology can be data-mined [30], but it is still not immediately obvious what an $\boldsymbol{h}$-complex value (Pfwd, Pbwd) is. Q-UEL seems consistent here with the theoretical basis [31] of PR-OWL [9], but implies specific data mining strategies to get the probabilities. Consider that a procedure extracts and canonicalizes (see XTRACT tags in Methods) the following source Wikipedia sentence: *"all strains of canine parvovirus will affect dogs, wolves, and foxes, but only some of them will infect cats"*. Interpretation of "cats" involves computation of probabilities that consider that the context, and in this case the sentence itself, that cats will mean cats as animals, not say a type of boat or whip. The fact that viruses and dogs are mentioned adds strong evidence, because the association constants are significant, but they are not the prior probabilities that may become influential if context evidence is weak. Q-UEL uses a kind of prior probability in which the words or phrases in an indexed subclass for animals of Roget's thesaurus can be the meaning of what is intended with different probabilities.

⟨Q-UEL-THESAURUS cat | **suggests** | '2. Special Vitality':='366. Animal.':=pbwd:=0.03502 **or** 'Section II. PRECURSORY CONDITIONS AND OPERATIONS':='455. [The desire of knowledge.] Curiosity.':=pbwd:=0.03226 **or** '(ii) SPECIFIC SOUNDS':='407. [Repeated and protracted sounds.] Roll.':=pbwd:=0.00667 **or** '(ii) SPECIFIC SOUNDS':='412. [Animal sounds.] Ululation.':=pbwd:=0.00632 **or** 'Present Events':='151. Eventuality.':=pbwd:=0.00571 **or** '(iii) PERCEPTIONS OF LIGHT':='441. Vision.':=pbwd:=0.00448 or 'SECTION III. ORGANIC MATTER 1. VITALITY 1. Vitality in general':='359. Life.':=pbwd:=0.00392 **or** '3. Fluids in Motion':='348. [Water in motion.] River.':=pbwd:=0.00356 **or** '3. PROSPECTIVE AFFECTIONS':='864. Caution.':=pbwd:=0.00223 **or** '5. INSTITUTIONS':='975. [Instrument of punishment.] Scourge.':=pbwd:=0.00151 **or** '3. Contingent Subservience':='668. Warning.':=pbwd:=0.00142 **or** ……(*etc*)…….. Q-UEL-THESAURUS⟩.

This illustrates the use of multiple Pbwd (expressed as pbwd not Pbwd to make the clear distinction), because the above tag may be considered as a composite of logically OR'ed tags each dealing with one interpretation. Pfwd is absent because it is the

default probability 1. The pbwd's are the number of words that actually belong to the sub-subclass of a cat as an animal divided by the number of words and phrases that are any kind of animal. This is the *class density interpretation*. In another context the *population density interpretation* might be important. For example, what is the Pbwd for ⟨cats| **are** |mammals⟩ that is the Pfwd for⟨mammals| **are** |cats⟩, in the sense that any mammal sampled will turn out to be a cat? The taxonomic sets are not of equal size. Priors can be based on Zipf's law which predicts that out of a population M of N elements, the probability $P(e(k)|M)$ of elements of rank k is $P(e(k)|M)=(\zeta(s,\ k)-\zeta(s,\ k-1))/\zeta(s,\ N)$; here Riemann's partially summated zeta function is defined as $\zeta(s,\ n)=1+2^{-s}+3^{-s}+\ldots+n^{-s})$ [3], and plays other important roles in Q-UEL (see Section 3.3).

2.12. Proofs and Iota algebra

While space does not permit extensive proofs (see however Refs [19–24]), proofs become almost trivial when we express $\boldsymbol{h}$-complex algebra in terms of $\mathfrak{ı}=½(1+\boldsymbol{h})$ and $\mathfrak{ı}^*=½(1-\boldsymbol{h})$ analogous to Dirac spinor quantum field operators. In general, $(\alpha,\beta)=\mathfrak{ı}\alpha+\mathfrak{ı}^*\beta$, and $(\alpha,\alpha)=\mathfrak{ı}\alpha+\mathfrak{ı}^*\alpha=\alpha$. This *iota algebra* is easier: of particular importance is the *idempotent property* $\boldsymbol{u}=\mathfrak{ı}$ and $\mathfrak{ı}^*\mathfrak{ı}^*=\mathfrak{ı}^*$ (so $\log(\mathfrak{ı})=\exp(\mathfrak{ı})=1/\mathfrak{ı}=\mathfrak{ı}$ and similarly for $\mathfrak{ı}^*$), *the annihilation property* $\boldsymbol{u}^*=\mathfrak{ı}^*\mathfrak{ı}=0$, and the *normalization property* $\mathfrak{ı}+\mathfrak{ı}^*=\mathfrak{ı}^*+\mathfrak{ı}=1$. They all follow from $\boldsymbol{hh}=+1$. For example, the basic braket ⟨A|B⟩, which can be rendered as the triple ⟨A| **if** |B⟩, can be expanded in the following manner.

$$\begin{aligned}\langle A|B\rangle &= \mathfrak{ı}P(A|B)+\mathfrak{ı}^*P(B|A)=[\mathfrak{ı}P(A)+\mathfrak{ı}^*P(B)]K(A;B)\\ &=[\mathfrak{ı}P(A)+\mathfrak{ı}^*P(B)]e^{I(A;B)}\end{aligned} \tag{3}$$

The association constant $K(A;B)=P(A,\ B)/P(A)P(B)=e^{I(A;B)}$ with $I(A;B)$ as Fano's mutual information is important as it impacts the algebra in the above way. The conceptual eigensolution outcome of ⟨B| **are** |A⟩=⟨A| **include** |B⟩ may be seen as $\langle A|\ e^{I(A;\ B)}\ |B\rangle$, in which case A and B are seen as randomly associated and the vectors ⟨A| and |B⟩ are independent vectors. Consider this example toy case.

$$\begin{aligned}\langle A| &= [\mathfrak{ı}^*P(A),\mathfrak{ı}]=[\mathfrak{ı}^*,\ \mathfrak{ı}](\mathfrak{ı}+\mathfrak{ı}^*P(A))=[\mathfrak{ı}^*,\ \mathfrak{ı}]\langle\boldsymbol{?}|A\rangle\\ |B\rangle &= [\mathfrak{ı}\ P(B),\mathfrak{ı}^*]^T=[\mathfrak{ı},\mathfrak{ı}^*]^T(\mathfrak{ı}\ P(B)+\mathfrak{ı}^*)=[\mathfrak{ı},\mathfrak{ı}^*]^T\langle B|\boldsymbol{?}\rangle\\ \boldsymbol{R} &= [0,\mathfrak{ı}+\mathfrak{ı}^*e^{I(A;B)};\mathfrak{ı}e^{I(A;B)}+\mathfrak{ı}^*,\ 0]\end{aligned} \tag{4}$$

Note the definition of *preparation-observation brakets* with **?** such that $P(\boldsymbol{?})=1$, needed later. It is easy to show by the annihilation property of iota algebra that ⟨A|B⟩=0, meaning that A and B are mutually exclusive and ⟨A| and |B⟩ are orthogonal vectors. That is the general case for Hermitian operators and their vectors, and it is no worse than saying that ⟨dogs| **are** |cats⟩=0 and ⟨dogs| **chase** |cats⟩≠0. As in BayesOWL, network nodes can be distinct classes, and one purpose of a relator is to override this. The result with $\boldsymbol{R}$ is now $\langle A|\boldsymbol{R}|B\rangle=[\mathfrak{ı}^*P(A)+\mathfrak{ı}P(B)]e^{I(A;\ B)}\neq 0$, because the matrix also flipped the elements of the ket so that its contents are not annihilated by the bra. It indicates that ⟨A| and |B⟩ are *randomly associated*. We can say that the eigenket part of the solution is that |B⟩ is transformed to a vector $\boldsymbol{R}|B\rangle$ that is dependent on ⟨A|, and not orthogonal. Note here that $\langle A|\boldsymbol{R}|B\rangle=(\langle A|\boldsymbol{R})\ |B\rangle=\langle A|\ (\boldsymbol{R}|B\rangle)$. The clue to a less toy application is that QM typically considers its vectors as

$$\begin{aligned}\langle A| &= [\langle A|\psi_1\rangle\ ,\langle A|\psi_2\rangle,\ \langle A|\psi_3\rangle,\ldots\langle A|\psi_L\rangle],\\ |B\rangle &= [\langle\psi_1|B\rangle,\ \langle\psi_2|B\rangle,\ \langle\psi_3|B\rangle,\ldots\langle\psi_L|B\rangle]^T\end{aligned} \tag{5}$$

This seems a problem for Q-UEL because ψ in quantum mechanics is held to be the *universal wave function* or *universal quantum state*, and this implies a literally cosmological amount of information. It is not insurmountable[19]; and see Section 2.3 and Footnote 12.

[19] As do physicists, we can consider a local subsystem reference state, for the subsystem considered, on which things are conditional at a tractable level of precision. In biomedicine, age in years would often be a very suitable choice, or we can have a two element case for male and female. We can refer to the actual choice as the *referential basis*. However, we may need to change that reference basis, and we should state that age, or whatever, was used. Rather than think of ⟨A| we are better to think of say ⟨A|*age*⟩ where *age* and hence⟨A|*age*⟩ is allowed by D-COMPILER (Section 3.1.6) as a vector.

3. Methods

3.1. Software

Q-UEL sees consent of structured or unstructured data for mining, the mining itself, and automated reasoning from it, as an integrated system of applications.

3.1.1. QuantalUEL

is a specific Q-UEL embodiment being developed primarily as open source middleware for cloud implementation, but its architecture shown in Fig. 1 was also the "laboratory bench" for the studies described here. The architecture demonstrates a patient-to-miner model in which data is private, public, or made public for data mining by using a simple fine grained consent language build into source patient records. Fig. 1 distinguishes an *incidence statement tag* (a *consented statement tag*, or a private *"data tag"*) which is data for one patient, from a *summary statement tag* in Fig. 1, which is a statistical summary generated by data mining incidence tags. In practice, Q-UEL sees an incidence as simply meaning a sample size of 1, and hence uses a common format. It means that the patient record or data mining many are all seen as part of the inference process, so QuantalSHRED and QUANTAL UNSHRED and QuantalMINE together form the portal for a physician under Professor Baylis's graphical user interface.

3.1.2. QuantalMINE

has a simple and efficient miner for physicians, but for our studies it accepts the plug-in QFANO for high dimensional mining of *structured and semi-structured data*. QFANO is algorithmically based on FANO [1,3,30] but reads and writes Q-UEL tags like the Q-UEL-EBM tag in Section 2.7, and automates source file joining and mining strategy. In practice, high-dimensional mining usually means generating summary tags containing up to some 5–12 clinical or other data features as attributes, depending on their individual abundance, before data becomes too sparse. QFANO can also run constantly on QuantalMASTER, the Cloud manager, or at remote permitted sites [20], for which purposes it carries its own security tools [30].

3.1.3. Q-ROBOSURFER

is the Web interface and text miner, and contains primarily XTRACTOR, THESAURUS, and QCHEMBIZ which can reside in QuantalMINE, but also QuantalMASTER if running continually to extract knowledge from the Web. They generate tags as in Sections 3.9, 2.9 and 2.11. QCHEMBIZ is not discussed there in detail, but see the description of a similar system in Ref. [2]. Briefly, it manipulates chemical formulae found on the Web. These formulae can be resolved into component parts that are subgraphs, and the degree of association of subgraphs with each other or other features on source texts can be computed [2].

3.1.4. QuantalSHRED and QuantalUNSHRED

A data miner may mine data that a patient consents to be visible, and in return Q-UEL supports optional secure backtrack mechanisms to alert patients if errors or risks are detected by mining. The miner may also be authorized to mine subsets of *encrypted private data*. QuantalUEL implements the President's Council request for a *disaggregation model* for a universal exchange language for healthcare. In this, patient records are disaggregated

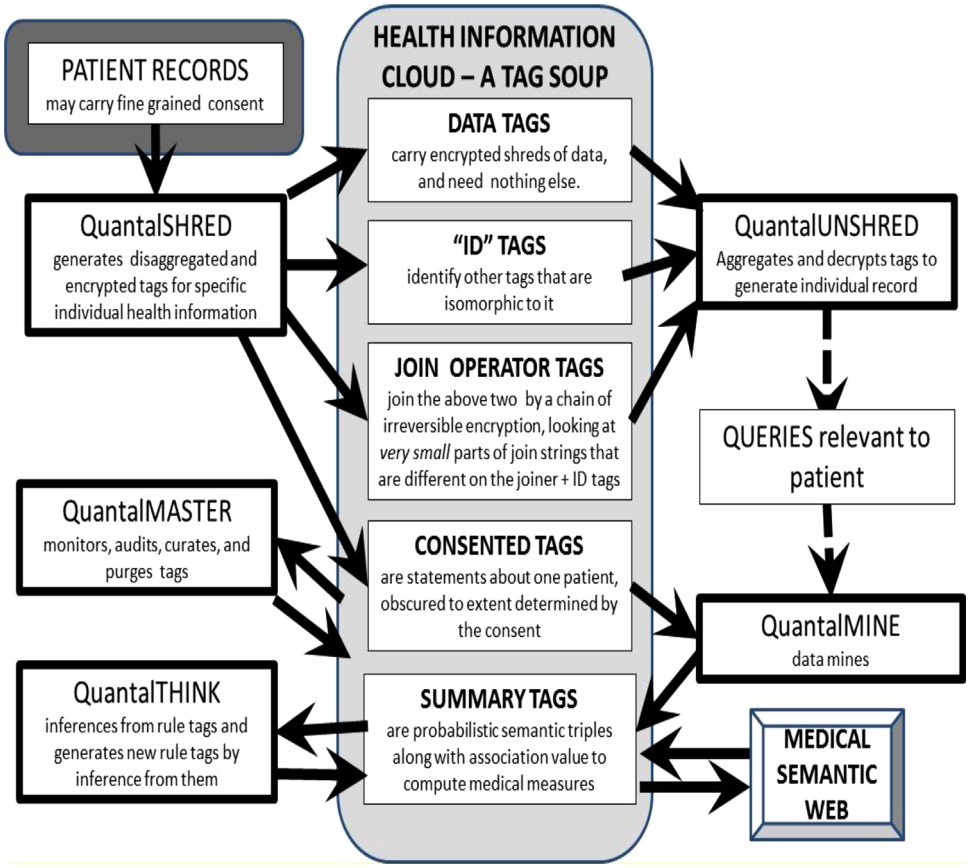

Fig. 1. The current implementation of Q-UEL for medical records, consented data mining of them, and clinical decision support from the probabilities generated.

(shredded) on the Internet and can be recovered anytime, anywhere, via Internet access [17]. The records to be recovered by authorized persons are shredded in a *tag soup* and mixed with shreds from all other patients. Briefly, reaggregation involves multiples of three dissociated tags, the pseudo-ID, join operator, and the data tag that is the only one to carry the reversibly encrypted data. The data miner will often only have access to the third tag, but it may have higher-order metadata with many encrypted clinical observations (see Section 2.3 for how the alternative metadata operator "=" controls visibility). For a physician or appropriately authorized miner, each tag rebuilding a record queries (via QuantalUNSHRED) for the next tag by a system of evolving keys matching strings or substrings on tags, variously using reversible and irreversible decryption.

3.1.5. *QuantalTHINK*

comprises the vehicle for our plug-in experimental automated reasoning technologies such as in Section 3.2 below. That includes D-COMPILER in our most recent experimental version for inference, where the shell is the Perl language interpreter. Our tags are accepted as Dirac's algebraic objects in Perl code but, simply, "Dirac Perl" is preprocessed to give normal valid Perl as input. The (Pfwd, Pbwd) are then held in memory with tags or derivative forms as the variables. Our original and main prototype QPOPPER is not embedded in Perl, and is thus a simpler language. This is compensated because Perl functions and executables can be inserted within tag attributes and pass variables. Metastatements can dynamically evolve the statements, and generate new statements from old in the manner of syllogistic reasoning. The old are usually deleted to shrink the inference net, but the reverse process is possible.

3.2. *The Hyperbolic Dirac Network (Application QuantalTHINK)*

The Hyperbolic Dirac Network (HDN) is most easily introduced by deriving one from a simple example Bayes Net. It therefore neither contains cycles nor does it evolve under the action of metastatements, but the treatment is the same with cycles and for each iteration step in evolving the net by metastatements. As in a Bayes Net, terms now simply multiply. Perhaps the only unexpected feature is the appearance of the preparation-observation brakets (Eq. (4)) as the analogs of prior probabilities. Consider the following derived from test data on tags: ⟨smoker |**?**⟩=(0.3, 1), ⟨not-chronic-bronchitis|smoker⟩=(0.06, 0.11), ⟨lung cancer| smoker⟩=(0.092, 0.897), ⟨breathless|not-chronic-bronchitis & lung cancer⟩=(0.203, 0.145),⟨mass on X-ray **and** biomarker-SMP175 | lung cancer⟩=(0.6, 0.828), ⟨**?**|breathless⟩=(1, 0.089), ⟨**?**|mass seen in X-ray **and** biomarker-SMP175⟩=(1, 0.019). A coherent joint probability (Sections 1.4 and 3.6) is first verified as Pfwd= Pbwd=0.000020 as can be seen by multiplying the Pfwd and Pbwd values. Then division by a single braket term, or division by a comparable smaller HDN after one or more factors such as lung cancer is removed, gives a conditional probability. Removing lung cancer, the joint probability of the net went to Pfwd=Pbwd =0.000094, also coherent in this case. Dividing that net into the former larger net gives ⟨lung cancer|all other considered factors⟩= (0.03, 0.213), i.e. it has an imaginary part ½[0.030–0.213]=−0.0915. Here Pfwd=0.03 is the collective probability of the states or events not removed, conditional upon those removed, Pbwd=0.213 is the converse, the collective probability of those removed as the etiology or cause. The problem is that a shared public repository of probabilistic tags is unlikely to have coherence: it may contain contributions from many sources. Even a

single Bayes Net example manually built by an expert may not allow a sensible coherent solution [11]. Software such as Q-UEL's HDN-CONSTRUCTOR (a QuantalTHINK plug-in) is thus required to enforce coherence (Section 3.6) *at each step above*. It is better to mine one large sample to get coherent data. The remaining problems are then the plethora of probability terms to choose from (Section 3.4), and the fact that many will contain sparse data, as follows.

3.3. Sparse Data Mining (Application QuantalMINE)

"Big Data" sources [1,2] may be big, but in high dimensional data mining there is inevitably sparse data for many terms (Section 3.1.2). Argumentation as in a court of law, however, allows weaker evidence to combine to influence a decision or prediction. Ideally we should not discard data, but what does weak data look like as revealed by the limiting case of no data? A default value for (Pfwd, Pbwd) of close to zero is, as for a probability in a BN, a bad idea: the overall joint probability will be zero. In a purely multiplicative HDN or BN, terms seen as providing no new relevant information are neglected, but this is equivalent to including them with probability 1, the Q-UEL tag default value. It is consistent with Popper's *principle of refutation of evidence* [32], and $P=1$ implies information $I=-\ln(P)=-\ln(1)=0$. However, we mean something specific here. The particular preferred approach to data mining is based on the incompletely summated Riemann zeta function (recall Section 2.11) as an expectation of available information [3,30]. For no data, $N=0$ and $\zeta(s, n=0)=0$, and $P=1$ is required as the result:

$$\begin{aligned}\langle A|B\rangle &= [\mathfrak{u}P(A)+\mathfrak{u}^*P(B)]e^{I(A;\,B)}\\ &= [\mathfrak{u}e^{-\zeta(s=1,n[B])}+\mathfrak{u}^*e^{-\zeta(s=1,n[A])}]\,e^{+\zeta(s=1,n[A,\,B])}\\ &\approx [\mathfrak{u}e^{+\zeta(s=1,n[A])}\\ &\quad +\mathfrak{u}^*e^{+\zeta(s=1,n[B])}]\,e^{+\zeta(s=1,n[A,B])]-\zeta(s=1,e[A,B])-\,\zeta(s=1,\,N)}\end{aligned} \tag{6}$$

Here $n[\ ]$ is the observed frequency of occurrence of an event, state, or measurement value, and $e[\ ]$ is an *expected frequency*, calculated on the classical, e.g. chi-square test basis: $e[A, B]=n[A]n[B]/N=n[A]P(B)=P(A)n[B]$. The last line in Eq. (6) using $e[A,B]$ is a convenient close approximation of the second. Both go to real value 1 for no data, and both approach probabilities, as classically determined, for large data, and within a few percent for $N\rangle 20$ (approximately). There are many variations and uses[20] of Eq. (6).

3.4. The Hedge (Application QuantalTHINK)

The remaining problem is that there many different combinatorial ways many of expanding a joint probability like $P(A, B, C, \ldots Z)$ as a production of conditional terms like $P(B|A, F, H)$, i.e. many inference networks. We can include all or many providing we power-weight terms as a geometric mean, e.g. $P(B|A, F, H)^{1/n}$ to remove redundant information.[21] The corresponding braket is $\langle \mathrm{B|A, F, H}\,\rangle^{1/n}=\mathfrak{u}P(B|A, F, H)^{1/n}+\mathfrak{u}^*P(A, F, H|B)^{1/n}$, an equality provable via the idempotent property of $\boldsymbol{h}$. Here n is interpretable as the amount of information that measurement allows us to discern, related to the standard deviation of the square root of the information. We can see this in QM's statistical weight $e^{\boldsymbol{A}/\hbar}=(e^{\boldsymbol{A}})^{1/\hbar}$ where $\boldsymbol{A}$ is information as the physical action and $\hbar$ is the reduced Planck's constant [13]. $1/n$ is called the *hedge*, consistent with the idea that we should increase the probability when we hedge (i.e. weaken) a statement. We have more confidence in the safer or more conservative statement, e.g. in interpreting categorical logic statements from statistics [19]. As to be discussed elsewhere, the hedge appears in a modified form in quantifying "the patient", "a patient", or "six patients", but also in the following.

3.5. Risk mitigation in interpretation (Application QuantalMINE).

consider

⟨Q-UEL-COMPLICATIONS context:='adverse drug reaction' drug:=CODE:=SNOMED-CT:='=373270004/Penicillin – class of antibiotic – (substance)':=www.uel.org/drug/Pfwd:=0.05 | **causes:='causative agent':=www.qexl.org/causes/:=assoc:=1.5** | allergy:=CODE:=SNOMED-CT:=2.16.840.1.113883.6.96:=106190000/Allergy/:246075003:=www.uel.org/allergy/Pbwd:=0.001?:=www.uel.org/allergy_misdiagnosis/Q-UEL-COMPLICATIONS⟩

"Penicillin *causes* an allergy" is evidently correct, but for "allergy causes penicillin", apart from a possible infection associated with an adverse reaction, a penicillin prescription in response to an allergy would depend on the misdiagnosis of the symptoms of, say, sore throat, rash or reddening, or fever as a (gram positive) infection. What we do not wish to risk *conveying* is that any allergy should be treated by penicillin. One could design so that e.g. 0.0316 in 0.0316? is the actual data mined value "but physicians please play safe and consider its square 0.001 instead". However, it is safer to *make visible* the *lower* probability, which can be a hedged value.[22]

3.6. Coherence and Q-UEL normalizations (Application QuantalTHINK)

For reasons of establishing coherence, an HDN is first considered as an estimate of a joint probability, *prior* to divisions which will express the notion of conditionality. An HDN is rendered as a joint probability by the $\boldsymbol{h}$-complex counterparts of prior probabilities. They are the preparation-observation brakets such as $\langle \boldsymbol{?}|A\rangle=\mathfrak{u}+\mathfrak{u}^*P(A)$ of Eq. (4), and similarly $\langle \boldsymbol{?}|B\rangle$,$\langle \boldsymbol{?}|C, D\rangle$, but also say $\langle S|\boldsymbol{?}\rangle\ \mathfrak{u}P(S)+\mathfrak{u}^*$, $\langle T|\boldsymbol{?}\rangle$, $\langle U, V|\boldsymbol{?}\rangle$, i.e. HDN counterparts of prior probabilities, because an HDN is bidirectional. Multiplication or division by them is the common method for all normalizations. The HDN builder HDN-CONSTUCTOR has the task, and it targets first the branch points like $\langle A|B, C\rangle\langle B|D\rangle\langle C|E\rangle$ as hot-spots for the origin of non-coherence. Starting with a Bayes Net and the simple example $P(A|B, C)\ P(B|D)\ P(C|E)\ P(D)\ P(E)$, we have as a first step $\langle \boldsymbol{?}|A\rangle\langle A|B,C\rangle\langle B|D\rangle\ \langle C|E\rangle\langle D|\boldsymbol{?}\rangle\langle E|\boldsymbol{?}\rangle$. We can multiply by $\langle B,\ C|\boldsymbol{?}\rangle/\langle B|\boldsymbol{?}\ \rangle\langle C|\boldsymbol{?}\rangle$, and interpret this as correcting the terms directly involved so that B and C are interdependent in *both* directions.[23]

[20] By adding virtual frequencies to observed frequencies $n[\ ]$ we can include prior belief [30]. With s with real values other than 1 we have various surprise measures other than information [2,3,30], and if s is $\boldsymbol{h}$-complex there is a deep relationship with Q-UEL algebra and QM [20] that is the subject of current research. In semantic use, note that $n[\text{cats, mammals}]=n[\text{cats}]$, since all cats are mammals. By this, we can write $\langle \text{cats|mammals}\rangle=\mathfrak{u}+\mathfrak{u}^*e^{\zeta(s=1,n[\text{mammals}])-\,\zeta(s=1,\,n[\text{cats}])}$ typical of such statements.

[21] For example, if we want to predict concerning C and have all terms for conditional probabilities of that form P(C|…) with r conditioning arguments A,B,D, E,F,… excluding C, we use $1/\mathrm{n}=(\nu-r)!r!/\nu!$.

[22] The hedge comes into it via the recommended precise meaning of? as an *inverse hedge operator* in e.g. 0.001?, 0.001??, and 0.001??? means that if there are $n-1$ "?" in the string, the data mined value was actually the probability shown raised to the power $1/(n)$. Contrast e.g. Pbwd:=? to indicate an as yet unknown probability, and the use of dispersion notation, e.g. Pbwd:=0.001+/−0.0003CI, and also Pbwd:=0.001+/−?CI to indicate a guess by an uncertain confidence interval.

[23] This is still an approximation, though a better one. We ideally need to multiply by $\langle B, C|D, E\rangle/\langle B|D\rangle\langle C|E\rangle$. However, it is possible that we do not have data for the joint occurrence of the four argument (B, C, D, E) that this implies, and it is possible that it is not coherent and still has an imaginary part. We could multiply by $\langle \boldsymbol{?}\ B\rangle\langle \boldsymbol{?}|C\rangle/\langle \boldsymbol{?}|B, C\rangle$. and have coherence, but that loses information $I(B; C)$. Overall, there may be more than one normalization method yielding different networks, and there may sometimes be a choice of discarding less reliable information or retaining reliable information correctly. Nonetheless, whatever terms are needed to

3.7. Distinguishability (Application QuantalTHINK)

When we write P(male | diabetes) P(cancer) do we mean the set of males, or one male, or in fact two distinct males? QFANO by statistics and often Q-ROBOSURFER by context can tell us this, but we focus here on the impact on normalization. Consider B in $\langle A|B\rangle\langle B|C\rangle\langle B|D\rangle$. If $\langle B|B\rangle = P(B)$ meaning that B is distinguishable from B by recurrence, recurs independently, and is Bernoulli countable, then *no renormalization is required*. In contrast, recurrence of A in, say, $\langle A|B\rangle\langle B|C\rangle\langle C|A\rangle$ has a different impact, as follows.

3.8. Cyclic paths (Application QuantalTHINK)

The simplest cyclic path is $\langle A|$ **if** $|A\rangle = \langle A|A\rangle = 1$, or $\langle A|A\rangle = P(A)$ if A can recur and be counted. A more elaborate path is $\langle A|B\rangle\langle B|C\rangle\ldots\langle X|Y\rangle\langle Y|A\rangle$. *Such a cyclic path is readily shown to be a purely real joint probability* $P(A, B, C,\ldots X, Y)$ estimated as $P(A,B)\ P(B,C)\ldots P(Y,A)/P(A)P(B)\ldots P(Y)$. Also, $\langle A|B\rangle\langle B|C\rangle\ \ldots\langle X|Y\rangle\langle Y|Z\rangle$ is purely real if $P(A) = P(Z)$. A coherent network satisfied by all required "**?**" brakets is a joint probability and purely real, and between the "**?**" it is an example of a system of cyclic paths, since **?** can be seen as a common node, and in any case all P(**?**) are equal at value 1. The seeming problem of cyclic paths is the solution. The difficulty arises not with a cyclic path, but in changing a prior probability of a $\langle A|B\rangle$ or $\langle A|\boldsymbol{R}|B\rangle$ that lies in, or at a branch point to, a cyclic path. However, it is a matter of coherence and the above "**?**" normalizations. Cyclic paths are an almost indistinguishable special case. *For a cycle or any net when a properly coherent joint probability or, or in completely existentially qualified categorical descriptions, or in statements in which the Hermitian relationship is trivial (real valued) and prior, marginal, or when self-probabilities are equal, the consequence is that the imaginary part of the complex value vanishes*. Then, Pfwd overall equals Pbwd overall. Note that, unlike ideal ontologies, *knowledge networks* can be rich in cycles, and so when probabilistic require the above.

3.9. Building General Knowledge Graphs (Application XTRACTOR)

To build such knowledge networks automatically, XTRACTOR is an off-screen search engine accessing website HTML, though it also works on pure text files found. It goes from website to website via links, and extracting and parsing (mainly) sentences into a *semantic multiple* form that facilitates production of semantic triples from it.

⟨Q-UEL-XTRACT-BIOLOGY "`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]" | **from** | source:='http://en.wikipedia.org/wiki/Human_brain' time:='Wed Oct 3 14:02:19 2012' extract:=0 Q-UEL-XTRACT-BIOLOGY⟩.

Such "Xtracts" are said to *auto-surf and spawn*. Almost all text on a Web page is converted to tags, and the tags themselves retain (a) links that occurred in the source text at that point, and (b) the links constructed from the citations as reference numbers in the source text. New tags are then generated in the same way from the source text at those links, so the number of tags generated can grow explosively. If the extracted text contains less than a specified density of links (e.g. current PubMed synopses), XTRACTOR can create them by initiate a Google query where words appear that do not appear in everyday speech. In Xtracts, sentences have clauses rearranged to represent as linear a parse structure as possible (as in e.g. "The man at the wheel of the car on the highway to Rome"). The tag format details relate to that.[24] In reverse query like "It is on the kidneys and releases hormones in response to stress by synthesis of corticosteroids" Q-UEL reasoners using X-tracts can reach http://en.wikipedia.org/wiki/Adrenal_gland and deliver "Adrenal gland" as the answer. Xtract tags can be queried to find a tag that directly answers the question, and if one does not exist, new Xtract generation can be launched via a Google query. But even with an intensive focus on improving prediction of cardiovascular events [33,34] there are no examples as yet that convincingly solve a medical mystery. There are, however, test cases of general knowledge that one can imagine included in epidemiological detective work.[25]

4. Results

This paper concerns a specification and its theoretical basis, but use cases, proof of concept, and scalability are important, and some findings should be shared. The current reaggregation rate is remarkably linear in scaling at one metadata item per second per 100,000 shreds of many records in the tag soup, and we anticipate a 100-fold improvement. Some 16000 English words and phrases as THESAURUS tags, and some 260,000 specialist biological, biomedical and chemical terms have been generated and could continue to be generated to the order of a trillion entries that are found a Semantic Web RDF Triplestore [16]. In using tags for inference, metastatements can typically transform nets derived from more than a 100 tags to 50% of their size within 25 iterations at about one iteration per second. The cardiovascular field is a popular test case [33]. Using tags from data in Ref. [1] to predict congestive heart failure (in 10% of records removed from the "training set") is 91% accurate, but his proves little because the power to predict the patient's future (hard task) and diagnostic power (much easier task), and obvious etiologies and comorbidities, were entangled. Indeed, identification of new strongly causative factors in cardiology typically gives more benefit than trying to improve an algorithm (e.g. Ref. [34]). Fortunately, ⟨outcome | input

(*footnote continued*)
provide coherence overall, we know that the product of their values yield −1 times the imaginary value of the network.

[24] Branches are either shown by semicolons ";" showing connection to the first noun phrase and the first noun phrase is reproduced to make the connection clear, or a comma "," joined to a later noun phrase where the noun phrase is reproduced to make the connection clear. An attempt is made to replace all pronouns etc. by what they stand for. Adjacent sentences with a common subject noun phrase are fused into one tag, and conversely phrases and sentences containing phrases as alternatives as flagged by the appearance of "and" and "or" can be split onto separate tags, and words made explicit. Symbols such as " and ^ are to indicate the grammatical parts of speech as the X-tractor program perceived them, and errors may be corrected in a curation phase (Ctract tags) before final disassembly into RDF-like triples in bra-relator-ket format. The X-tract parse form as semantic multiple can be retained: it can be refined as a twistor tag as in Section 2.3. The main problem is not ungrammatical parsing but ambiguous parsing affecting meaning, i.e. the source of much humor as in "One morning I shot an elephant in my pajamas" (Groucho Marx), but also in construing "I launched the cat" (a kind of boat). The context of text, links, and citations, use of metastatements, and THESAURUS tags are all important to try and remove ambiguity.

[25] One relates to the IBM Watson computer, which beat human champions at Jeopardy but thought O'Hare airport was in Toronto [35]. A Q-UEL metastatement did already know that if A travels to B, then A is not B. A key rule in that process was that below reached by an automatically generated Google query:-⟨Q-UEL-CTRACT 'Chicago-O'Hare International (ORD) |**to**| Toronto; Chicago-O'Hare International (ORD) |(**concerning**)| flights' (source:=http://www.cheapflights.com/flights-to-toronto/chicago-ohare-intl | **from** | Extract:=0 presource:= http://www.google.com/search?hl=en&source=hp&q=O%27Hare+airport+Toronto&gbv=2&oq=O%27Hare+airport+Toronto&gs_l=heirloom-hp.3…12891.24047.0.24750.22.19.0.3.3.0.531.3292.0j15j2j5-1.18.0…0.0…1c.1.nX3-bh6US_AQuery:%3D+O%27Hare Query:='O%27Hare+airport+Toronto' AutoQuery:='O\' Hare airport Toronto' Hits:=2,330,000 'Select (not advert)':=1) Q-UEL-CTRACT⟩.

event 1, input event 2, ….⟩ is quite general and at least the methodology can be tested right now on cleaner "code breaking" problems such as prediction of protein secondary structure from amino acid sequence [36]. Excluding predictions of proteins too similar to those in the training set (but allowing those not recognized as similar *a priori*) correctly assigns one of three states 94% of the time, using tags with up to 5 attributes as input events (amino acid types at relative positions).

5. Discussion and conclusions

5.1. Criticisms of Q-UEL: why reinvent the wheel?

Standards are of course already available to represent appropriate relations and logical properties, but recall that there is no agreed best practice regarding *probabilistic* data and knowledge (Sections 1.2 and 1.3). The choice of which one to build on was not obvious, and even less obvious was which preexisting single probabilistic implementation, that had taken that route, was best to emulate (Section 1.3). All had both interesting and missing features. A natural starting point for us from the medical perspective would be the previous effort of the first author in a biomedical and clinical exchange language that involved co-packaging and manipulating XML, HL7-CDA XML, and various non-XML standards for combined use [37]. This was well received by the XML community (e.g. Ref. [38]), so one may well ask why we did not avoid controversy and extend that effort. Essentially, it was vulnerable to any limitations and criticisms of any the standards that it carried, at least where it affected our particular requirements. There the PCAST report [17] became impactful on our thinking. We therefore explored what PCAST's "XML-like Universal Exchange Language for healthcare" would look like, given a free hand in a fresh start approach that the PCAST report would seem to favor. Had early XML developers referred to established standards of data, knowledge, and inference management in physics, we believe it plausible that XML today would automatically cater elegantly for semantic representations currently treated by OWL and RDF methodologies, look much like Q-UEL, and also rather like what PCAST had in mind. We hoped that this research would generate a body of work with useful ideas of more general value to the medical IT community members whatever their preferred format. Nonetheless, we have reached the conclusion that something very like Q-UEL would be at least a useful *second* language, a common hub. This has the following implications.

5.2. Criticisms of Q-UEL: Q-UEL is disruptive

We would hope so, but this needs serious comment. Q-UEL can make no claim to be a "disruptive technology" as can the first wheel, the microchip, and the Internet. It remains that any effort comparable to Q-UEL, and successful, would be in an interesting position. Once one is able fully to extract medical information from an Electronic Health Record (EHR) in a universal second language, one can rebuild the EHR expressed in that language, so that it might automatically become pervasive and displace established EHR methods. Here, the current EHR is in a particularly vulnerable position. PCAST's call for a Universal Exchange Language arose from the observation that that in healthcare IT there never was a single invented wheel to reinvent, and that just because the wheels worked well historically does not mean that they will provide the best means of moving information in the emerging ecosystem: *"data mining and presentation is something that computers, augmented by communications networks and distributed data storage, are very good at…it is not something that current EHR systems are optimized for."* [17] Any solution as a design by committee almost solely comprising the preexisting medical records standards bodies could result in preserving legacy structure by adding further layers of features. A significant core structural reorganization of the EHR is implied to facilitate data mining and avoid extra steps. Recall the EAV and data mining basis of Q-UEL (Section 2.3). Extra steps or not, such structure is required for sufficient granularity of disaggregation [17] (Section 3.1.4) and consent (Section 3.1.1) mechanisms. These are our opinions, but there clearly needs to be independent exploratory efforts alongside those of the standards communities, and a fair hearing for the "minority report".

5.3. Criticisms of Q-UEL: complexity

To be an attractive candidate in the above scenario and more generally, developers would have to be comfortable with Q-UEL. The commonest criticisms of Q-UEL are that we are dealing with a complex set of algorithms described in terms of difficult mathematics that would make at least some aspects of future work very hard, and that the approach used is very expensive and unsuitable for *public* use as literally being a "Universal Exchange and Inference Language". Such criticisms are typically combined, although really distinct issues, but all of them are misconceptions. These often appear to arise from the fact that the mathematics is unfamiliar. In general, reasons for introducing novel or unfamiliar mathematics include ultimate simplicity, consistency, insight, explanatory and predictive power, and efficiency, and we believe all apply here. As illustrated by our previous work (e.g. Refs. [1–3]), a motivation is to find mathematics that leads to algorithms enabling scalability with "Big Data", including rapid inference from large knowledge networks; that includes also information-theoretic and number-theoretic approaches [3]. Much more than simplicity, efficiency is important to us because of the need for scalability (see Results), but of all the mathematical tools explored, ***h***-complex algebra becomes the simplest with experience, and much simpler than *i*-complex algebra at least when expressed as easily manipulated iota algebra (Section 2.12). Unfamiliarity of the mathematics at the level of individual tags is irrelevant, because the three parameters needed are rendered in Q-UEL tags as intuitive probabilistic quantities from which, for example, medical metrics can be calculated (Section 1.3.1). However, the mathematics is available for a crucial next step: inference from the knowledge that the tags represent. We resist the temptation to say that the mathematics is then irrelevant to the user when "under the hood" of the inference engine, because physicians are ultimately responsible for accepting or rejecting its conclusions. Elsewhere we shall describe how use of Pfwd, Pbwd, and association constants for all or parts of a network can provide everyday intuitive explanations of how conclusions were reached.

5.4. Criticisms of Q-UEL: but why not in XML?

We could still preserve many essential ideas of Q-UEL by implementing it in XML. We could in the usual way implement appropriate relations and logical properties by creating appropriate "fields" and analysis software, essentially as in the previous effort [37]. As it happens, Q-UEL can use data fields but it was designed to focus on the EAV model with its advantages for data mining, and instead extends that EAV approach via the metadata attribute language to present and define data (Section 2.3). Developers cannot be stopped from developed applications to use this data as they wish (but see discussion in Sections 5.4.4 and 5.4.5 below). However, there is a favored option: the general idea of tag analysis software (Section 2.1.5) has been replaced by the D-COMPILER (Dirac complier, see Section 3.1.5) which plays a similar role, and one writes programs using Q-UEL as outlined in Sections 3.1.5, and

5.2.4 below. This is not as dismissively hand-waving as might appear: the most simple and common program intended for routine use of tags is an HDN as a generalized Bayes Net but with the option to evolve it by metastatements. It is merely a collection of Q-UEL tags multiplied in any order with no further code (Sections 3.2 and 5.2.4 below). It remains that this could all still be done in legal, if perhaps unusual-looking, XML. In fact, we have not only outlined the principles of XML and Q-UEL interconversion, but also an XQMF specification (Section 2.2) that will (with a little standard wrapping) work in an XML browser as legal XML. But whilst adoption of Q-UEL as an XML embodiment would be satisfying, our arguments for Q-UEL as an XML *extension* are as follows.

5.4.1. Human-friendliness

It has been emphasized that Q-UEL tags are directly readable by human medical workers and developers. In contrast, the direct translation of Q-UEL into XML (e.g. as XQMF) look distractingly "ugly", for several reasons discussed below. That an underlying communication structure format is human-friendly throughout has important benefits. It reduces errors in application development, testing, and maintenance. Further, most Q-UEL tags are stand-alone pieces of medical information or a fragment of information about a patient (see Section 5.4.4), so an email or telephone text is a sufficient application for Q-UEL usability. It is an important consideration in medicine if an application is not available or fails, or the IT infrastructure fails in a disaster. Such scenarios were a serious design consideration in the earlier standards-based effort [37], but the solution was different. It certainly could not be said that the underlying communication structure format was particularly human-friendly throughout because the code that wrapped the standards like HL7 XML were designed for extreme brevity and error correction [37]. These issue overlap with the following.

5.4.2. Reasonable brevity

The commonest criticism levied against XML is that a large burden of characters is required to communicate small amounts of useful information. In fairness to XML's founders, one should perhaps distinguish XML's elaborations from the informal references of some developers to the more friendly "plain old XML" (POX), but both have a burden of characters, and it is evident that in practice public XML embodiments are seen as needing their elaborations. A physician's addition of a note such as "Chest X-ray normal 2016-1-1" into a traditional paper file expands into a tolerable 8 lines in HL7's original condensed messaging system, but the same information in the screen print out of the XML version occupies some one-and-a-third pages plus some three-and-a-third pages of XML showing how to display it (Ref. [31], pp. 215–220). The usual greater brevity of Q-UEL, especially compared with any XML version of it, is due to fewer tags, its extended attribute metadata language which packs XML bracketed structure into attributes, allowance of nominal categorical data like "male" without metadata, and its logic management. The typical Q-UEL tag by itself comprises not an arbitrary collection of attributes, but rather a human-readable expression in attributes which is typically a logical expression, e.g.⟨*logical expression*|*relationship logical expression*|*logical expression*⟩. Handling this in XML would be much more elaborate, and certainly less directly readable by physicians (Section 5.2.1).

5.4.3. Molecular character

Q-UEL emphasizes and refines the EAV approach for scalability of data mining medical records (Section 2.3 and Footnote), and that includes extending XML's method of attribute representation via its attribute metadata language, so that Q-UEL tags are typically perceivable as self-sufficient "molecules of knowledge" joining attributes as atoms or groups of atoms. Though Q-UEL can structure documents as does XML (Section 2.2), all the tags used in the present study were randomly mixed in the tag soup (Section 3.1.4). They are recovered by querying or reassembled by issuing queries to each other via portal or server. This has advantages. It can be used to enhance robustness (against accidental data losses or fragmentation) or deliberately obscure reassembly as a security feature (data disaggregation or "shredding"). Stand alone XML tags could have this function, but it is highly questionable that they are, or need to be, XML outside a structured XML document.

5.4.4. Algebraic character

Recall that Q-UEL in its fullest definition builds on Dirac notation and algebra to become a programming language. Consider $-\ln(\langle$A Pwd:=0.6$|\boldsymbol{R}|B$ Pbwd:=0.8$\rangle\langle A$ Pwd:=0.2$|R|B$ Pbwd:=0.6$\rangle)/2$ (which yields the complex value (1.060, 0.367)=$\imath$1.060+$\imath$*0.367=0.7135+$\boldsymbol{h}$0.3465). One can see how to express that expression as an XML document, or similarly insert XML tags into programming code to serve as constants and variables, but most workers would surely agree that this is stretching XML's role a little too far, whereas for Q-UEL it was a principal brief. Be that as it may, the main concern is that to allow a free hand with a "general carrier vehicle" like XML with a variety of potential kinds of probability, and with a variety of applications for diverse interpretation for use, could be dangerous in medical decision making. In our opinion, browsers, portals and decision engines should recognize one or a limited well specified number of finally agreed probability-algebraic structures as innate to the tag language, because of these dangers of freedom of interpretation. Standards, whether or not they finally reflect the principles of this report, are needed. But with no control over developers, the only obvious way to help make other input illegal is as follows.

5.4.5. Safer Q-UEL Universe

XML's designers also recognized that they had no control over application developers, good or bad, but sought to mitigate the risks as much as possible. For example, it is well known that XML tag names containing "-" characters are forbidden in case an application thinks it means that something is to be divided by something. Imperfect software is a fact of life, and whether commercially released, freely circulated executables, or freely circulated source code modified by researchers, it can fall into trusting hands. Even the best applications could accept and act on future input with actions that cannot be foreseen. The much larger universe of XML documents and applications in general, and the variety of XML-based medical efforts, was seen as posing a slight risk for Q-UEL if expressed as XML. To put the above slight risk in more significant perspective, however, it should be appreciated that a design feature of Q-UEL was that it will work against a challenging background. To monitor and audit use, harvest knowledge, explore provenance, recover lost information, and assemble inference networks from sub-graph components, Q-UEL applications can scour the Internet for Q-UEL tags containing relevant information not only in random mixes of Q-UEL tags in a tag soup, but mixed in with text or effectively arbitrary strings of characters including those which contain XML documents and snips of XML. It is much less likely that a Q-UEL (XML extension) application can accidentally use XML or join Q-UEL with XML documents (and vice versa), if XML is illegal Q-UEL (and vice versa). Indeed, to further

ensure that, it is strongly recommended that any public Q-UEL tag should contain the string "⟨Q-UEL" or "Q-UEL⟩" (note the minus sign in the tag name!). We see no huge effort if downloadable Q-UEL (XML extension) browsers are made available, and we see inclusion of a Q-UEL interpreter at the input point to applications as essentially straightforward.

5.5. Current and Future Work

Our aim has also been to point that Dirac's system is a good specification for a Universal Exchange Language. Also, this system maps to semantics. All this seems worthy of further research. In our larger *QEXL Consortium*, Avner Levy's KODAXIL [39] is a universal natural-language-like but natural-language-free semantics for the Web. It is to be integrated with XTRACTOR and Barry Robson's QUELIC that considers "dimensions of meaning". QUELIC constructs unambiguous words from roots that are Dirac objects, based on Roget's Thesaurus. METACLOUD and INGINE will be respectively based on Paul Peters' Fluxology Cloud integration [40] and Srinidhi Boray's enterprise-wide integration software at Ingine Inc. [41], providing an infrastructure for Q-UEL. Quantal Semantics [42] provides a preliminary architecture focusing on medical record security and consent mechanisms for data mining, still using some quantum mechanical principles. Mathematician Steve Deckelman at University of Wisconsin-Stout and Berkeley is exploring novel uses and implications of complex algebra in inference. One basic task ahead, following Results (Section 4), is to ask medical questions as clean as those in bioinformatics. We foresee this as aided by additional metadata in tag attributes, as keywords that draw from the limited set of possible medical question types (e.g. diagnosis, best therapy, etiologies, comorbidities, epidemiology, and also predictions for prognosis, risk, preemption and prevention). Best practice for the inference here will doubtless be an ongoing argument, but our current perception is this: since cycles and oscillations between states abound in QM [12,13], we might paraphrase Einstein and conclude that God may play with dice, but does not seem to play with traditional Bayes Nets. As biology, medicine and healthcare do not fit into a unidirectional acyclic graph, we would like to understand that.

Conflict of Interest

None declared

References

[1] 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.

[2] B. Robson, R. Dettinger, A. Peters, S.K.P. Boyer, Drug discovery using very large numbers of patents: general strategy with extensive use of match and edit operations, J. Comput. Aided Mol. Des. 25 (5) (2011) 427.

[3] B. Robson, Clinical and pharmacogenomic data mining: 3. Zeta theory as a general tactic for clinical bioinformatics, J. Proteome Res. (Am. Chem. Soc.) 4 (2) (2005) 445.

[4] ⟨http://semanticweb.org/wiki/Bayes_OWL⟩ (last accessed 7/3/2513).

[5] H. Nottelmann, N. Fuhr, pDAML+OIL: a probabilistic extension to DAML+OIL, (last accessed 7.28.2013) based on probabilistic Datalog, duepublico.uni-duisburg-essen.de/servlets/…/Nottelmann_Fuhr-04a.pdf (last accessed 7.28.2013).

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

[7] L. Prediou and H. Stuckenschmidt, H. Probabilistic Models for the SW – A Survey. ⟨http://ki.informatik.uni-mannheim.de/fileadmin/publication/Predoiu08Survey.pdf⟩ (last accessed 4.29.2010) (2009).

[8] ⟨http://obi-ontology.org/page/Main_Page⟩ (last accessed 4.29.2012).

[9] ⟨http://www.pr-owl.org/⟩ (last accessed 7.27.2013).

[10] M.A. Klopotek, Cyclic Bayesian Network – Markov Process Approach, Studia Informatica, 1/2(7) Systemy i Technologie Informacyjne (2006).

[11] R. Rebonato, Coherent Stress Testing. A Bayesian Approach to the Analysis of Financial Stress, John Wiley, New York, 2010.

[12] P.A.M. Dirac, The Principles of Quantum Mechanics, Oxford University Press, Oxford, 1930.

[13] R. Penrose, The Road to Reality. A Complete Guide to the Laws of the Universe, Joanthan Cape, Random House, London, 2004.

[14] ⟨http://en.wikipedia.org/wiki/Semantic_Web⟩ (last access 3.30.2013).

[15] ⟨http://en.wikipedia.org/wiki/Resource_Description_Framework⟩ (last accessed 4.10.2013).

[16] ⟨http://en.wikipedia.org/wiki/Triplestore⟩ (last accessed 6.5.2013).

[17] ⟨http://www.whitehouse.gov/sites/default/files/microsites/ostp/pcast-health-it-report.pdf⟩.

[18] B. Robson, U.G.C. Balis, UGC, and T.P. Caruso, Considerations for a universal exchange language for healthcare, in: IEEE Healthcom 11 Conference Proceedings, June 13–15, Columbia, MO 173 (2012).

[19] 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.

[20] B. Robson Towards, Intelligent internet-roaming agents for mining and inference from medical data, Stud. Health. Technol. Inform. 149 (2009) 157.

[21] B. Robson, Links between quantum physics and thought, Stud. Health. Technol. Inform. 149 (2009) 236.

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

[23] B. Robson, Towards new tools for pharmacoepidemiology, Advances in Pharmacoepidemiology and Drug Safety, 1, (6) http://dx.doi.org/10.4172/2167-1052.100012 (2013).

[24] B. Robson, Rethinking Global Interoperability in Healthcare. Reflections and Experiments of an e-Epidemiologist from Clinical Record to Smart Medical Semantic Web Johns Hopkins Grand Rounds Lectures (last accessed 3.14.2013). ⟨http://webcast.jhu.edu/Mediasite/Play/80245ac77f9d4fe0a2a2bbf300caa8be1d⟩.

[25] Y. Kuroe, T. Shinpei, H. Iima, Models of hopfield-type clifford neural networks and their energy functions – hyperbolic and dual valued networks, Lect. Notes. Comput. Sci. 7062 (2011) 560.

[26] A. Khrennikov, On quantum-like probabalistic structure of mental information, Open Systems and Information Dynamics 11 (3) (2004) 267.

[27] N. Soldatova, A. Rzhetsky, A., K. De Grave, K., and R.D. King, Representation of probabilistic scientific knowledge (2012) ⟨https://lirias.kuleuven.be/bitstream/123456789/371072/1/bmc_article.pdf⟩.

[28] ⟨http://en.wikipedia.org/wiki/Semantic_Web_Rule_Language⟩ (last accessed 3.30.2013).

[29] ⟨http://en.wikipedia.org/wiki/Entity%E2%80%93attribute%E2%80%93value_model⟩ (last accessed 7.27.2013).

[30] B. Robson, Clinical and pharmacogenomic data mining: 4. the fano program and command set as an example of tools for biomedical discovery and evidence based medicine, J. Proteome Res. (Am. Chem. Soc.) 7 (9) (2008) 3922.

[31] ⟨http://seor.gmu.edu/~klaskey/papers/Laskey_MEBN_Logic.pdf⟩ (last accessed 7.26.2013).

[32] K. Popper, The Logic of Scientific Discovery, (as Logik der Forschung; English translation 1959, Routledge, London, 1934.

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

[34] W. de Ruijter, W., G.J. Rudi, R.G.J. Westendorp, W.J.J. Assendelft, W.P.J. den Elzen, A.J.M. de Craen, S. le Cessie, S., J. Gusskloo, Use of Framingham risk score and new biomarkers to predict cardiovascular mortality in older people: population based observational cohort study, Br.Med. J. 338 (2009) a3083.

[35] ⟨http://en.wikipedia.org/wiki/Watson_(computer)⟩ (last accessed 6.5.2013).

[36] B. Robson, Analysis of the code relating sequence to conformation in globular proteins: theory and application of expected information, Biochem. J. 1974 (1974) 141.

[37] B. Robson, R. Mushlin, R., Genomic messaging system for information-based personalized medicine with clinical and proteome research applications, J. Proteome Res. (Am. Chem. Soc.) 3 (5) (2004) 930–948.

[38] ⟨http://xml.coverpages.org/ni2004-10-11-a.html⟩ (last accessed 4.2.2013).

[39] ⟨http://www.kodaxil.org/⟩ (last accessed 6.3.2013).

[40] ⟨http://www.fluxology.net/⟩ (last accessed 6.3.2013).

[41] ⟨http://www.ingine.com/⟩ (last accessed 6.3.2013).

[42] ⟨http://www.quantalsemantics.com⟩ (last accessed 6.3.2013).

Barry Robson retired after 11 years (1998–2009) as Chief Scientific Officer IBM Global Healthcare, Pharmaceutical and Life Sciences, IBM Distinguished Engineer, and the Strategic Advisor to IBM Global Research Headquarters at Yorktown Heights, NY. He is currently Director of Research and Professor of Biostatistics Epidemiology and Evidence Based Medicine at St Matthew's University School of Medicine, Grand Cayman, and Distinguished Scientist in the Department of Mathematics and Computer Science in the University of Wisconsin-Stout. Barry is also Chief Scientific Officer of Quantal Semantics Inc., Virginia, and Chair of The Dirac Foundation, Oxfordshire, UK. According to his biography in Nature (389,418–420, 1997) he was a pioneer in bioinformatics, protein modeling, and computer-aided drug design. Prior to joining IBM he was an industry executive serving on the boards of five biopharmaceutical companies. Notably he was scientific founder and Science Director (CSO) of Proteus International which went to the London Stock

Exchange in 1995, currently part of the BTG plc. group. Barry has a BSc Joint Honors in Biochemistry and Physiology, a Ph.D. in medical and protein science, and a higher UK doctorate (DSc) in computational chemical physics. Barry has had several advisory roles to four governments, e.g. via the Council on Competitiveness's white paper "Innovate America". He is the author of "Introduction to Proteins and Protein Engineering" (B. Robson and J. Garnier, 1984, 1988, Elsevier Press), and "The Engines of Hippocrates: From the Dawn of Medicine to Medical and Pharmaceutical Informatics" (B. Robson and OK Baek, 2009, John Wiley & Sons).