2011 IEEE 13th International Conference on e-Health Networking, Applications and Services

Considerations for a Universal Exchange Language for Healthcare

Barry Robson[1,2], Ulysses G.J. Balis[1,3] and Thomas P. Caruso[1,4]
[1]Member, Biomedical Informatics Think Tank, a Division of Projectivity, Inc., Vienna, VA, USA
[2] St. Matthew's University School of Medicine, Florida and Grand Cayman, Cayman Islands and Department of Mathematics & Computer Science, University of Wisconsin-Stout, Stout, WI, USA
[3]Associate Professor and Director, Division of Pathology Informatics, University of Michigan,
[4]Business Ambitions, LLC, Vienna, VA, USA
E-mail: tcaruso@gmail.com

Abstract—**The US Government has called for a Universal Exchange Language for healthcare. This may be an opportunity to start from fundamental principles to deliver medical information in a uniform way that many stakeholders in the use of the system, not just programmers, IT architects, and standards organizations, but physicians, public health analysts, epidemiologists, and clinical decision support system designers, would like. To be truly universal, several aspects of a consistent formalism need to be considered. Some of these with examples are provided here.**

Keywords-UEL; standards; PCAST; inference; CER; CDSS

I. INTRODUCTION

The Office of the National Coordinator (ONC) for Health IT in the US is developing the technical guidance for a Universal Exchange Language (UEL), in response to the report by the President's Council of Advisors on Science and Technology (PCAST) [1]. Patient data is not to be confined to "home base" archives but rather, should be widely available for the patient in emergencies, and to facilitate extraction of knowledge for Comparative Effectiveness Research (CER), Evidence Based Medicine (EBM), epidemiological and public health analysis, and Clinical Decision Support Systems (CDSS).

The general requirements for UEL as framed by PCAST have been controversial for several reasons. For added protection and selective access to parts of the record only on a need-to-know basis, patient data is to travel in chunks, which then requires re-aggregation to once again render an intact full version when required. Primary care providers, long-used to ready access of patient's record, contest the worthiness of this, but an appropriate system can assign an adjustable scope of authority, and a system that is data segregated can be treated as aggregated, while the converse is not true. Ideally the architecture should avoid a unique identifier for a patient not only for added security, but also because attempts to implement such have caused much controversy, especially in the US Healthcare System [2]. Furthermore, such a UEL should overcome or *meaningfully reconcile* a plethora of standards such as HL7 [3], ISO [4], IHE [5], DICOM [6], and CDISC [7]. The UEL appears to be a "green field" opportunity for an architecture based on a self-consistent set of language principles in a unifying theoretical formalism. Here we discuss what this might look like.

II. UNIFORMITY

We describe our approaches within an XML-based format, which does not appear to be a controversial issue. It makes sense that any kind of "universal exchange language" should be applied in a way that is as *uniform* as possible, whether it is used for conveying patient data or for generating the measures intrinsic to CER, etc. In XML terms, this means that emphasis is on a grammar of *markup* relating to *attributes* in XML tags. *Content* wrapped by tags we consider as more arbitrary and not necessarily structured information, comprising supporting textual comments less crucial to rapid medical use and analysis. *Uniformity* more crucially implies as few attributes as possible that can serve in a reusable and consistent fashion for the vast number of extant data types, with the added feature of allowing for omitted attributes, owing to default meaning. We seek a UEL with the same tags and grammar that applies to (a) individual patient data, (b) statistical summaries and metrics of medicine drawn from data analysis of many patients, (c) semantic triples for the Semantic Web, and (d) probabilistic rules for inference.

To achieve this we think of XML tags or analogous objects as being rules, and of tags that relate to an individual patient as being *incidence rules* that portray that patient as the limit-case of a population of one, with this conveyed by an attribute *patients= '1'*. Rules from many patients are *summary rules*.

Elsewhere we shall describe a prototype implementing all the features described in this report, and others. Briefly, the prototype considers attributes *subj*, *verb*, *obj* on the understanding that the attribute *verb* means any *predication* or *relator*. An example is *subj='systolic_BP:=140+/-10'*, which is in metadata:= orthodata format with an optional confidence interval.

For simplicity, consider first the categorical relationship: *verb='are'* or *'all are'*. Probabilistic quantification is by three key attributes *p* or optionally *pfwd* (P-forward) with value of P("All A are B") = P(B|A), and *pbwd* (P-backward) with the value of P("All B are A") = P(A|B), plus *assoc* (association) with the value of K(A; B) = P(A & B) / P(A) x P(B)]. Section III describes the formalism of inverse relationships. Note that, P("All pregnant patients are females") and P("All females are pregnant patients") depends on the fraction that are pregnant. *Pfwd*, *pbwd* and *assoc* are sufficient for the general categorical case in medicine on the understanding that A and B can also be joint or compound events or coincidences, such as (V & W) and (X & Y & Z), respectively. All required probabilities can be generated from *pfwd*, *pbwd*, and *assoc* by P(A) = P(A|B)/K(A; B), P(A & ~B) = P(A) − P(A & B) and so on. From these we may obtain all familiar medical odds and odds ratios, and *number needed to treat/harm* of form 1/[P(A|B) − P(A|~B)].

How are these probabilities obtained? *Associative data mining* can be used to obtain K(A; B) or its logarithm as (Fano) *mutual information* I(A; B). Mining I(A), I(A; B), and generally I(A; B; C;…) = ln(A, B, C, ..) − ln(A) − ln(B) − ln(C)… allows *all* required probabilities to be generated from a list of such, e.g. $P(A|B) = e^{I(A;\, B) - I(A)} = P(A)e^{I(A;\, B)}$, this being also *an analogue of the Bayes relationship*. Simple Bayes-style inference for diagnosis and best therapy may use I(A: ~A | B, C,….) = I(A | B) + I(B | A, C) + … − I(~A | B) − I(B | ~A, C)…, or mutual information I(A: ~A; B, C, ….) in absence of prior information. Our expansion terms are in practice Bayes expected information estimates using all the data available for terms with different numbers of arguments A, B, C... Probabilistic UEL objects could be derived from these terms, and conversely used to reconstruct them for inference in CDSS. Note that *pbwd* allows an inverse form of inference using, for example I(B, C | A: ~A) to study etiology of disease.

For *non-categorical* ("action verbs", etc.), the situation is a more complicated when probabilities are assigned by data analytics. It best uses the considerations of Section III for a formal treatment of a verb as operator that is less familiar in probabilistic semantics. In interpreting it as an extension to the categorical case, there is a need to include further (but calculable) probabilities. But subjective assignments by an expert seem natural and self-evident. We have simply to assign intuitive analogues of conditional probabilities to *pfwd* and *pbwd* to represent as e.g. P("overeating causes obesity") and P("obesity causes overeating").

What meaning can be attached to patients='1', the case of the single patient? Nothing can be certain. The loophole is that probabilities in medicine may relate to the reliability of clinical measure methods determined by repeated measurements on one sample, or repeated measurements on one patient, or measurements on different patients in a cohort or population, or (in systematic reviews, in principal at least) a mix of these. The simplest case for one patient is possible because attributes are optional although their absence may imply defaults: for *subj* and *obj* the default is a probability of one. In that case *subj='systolic_BP:= 140+/-10'* and *p*= '0.95' means the 95% confidence interval, i.e. 95% of repeated measurements will lie in that range. The presence of several *subj*'s implies a joint probability with the confidence interval applying to a *two dimensional* distribution. At least one *obj* implies a conditional probability, so *subj='systolic_ BP:=140+/-10'*, *obj='BMI:=30+/-5'* implies that 95% of systolic BP measurement lie in the 130 to 150 range conditional on the BMI being on its specified range 25 to 35 range. Such measures are much rarer, but meaningful.

This begs the question of what such measures mean when *patients='2'* or some other low number. Even putting aside the intrinsic uncertainty in one measurement, a proper treatment really requires the notion of patient-observations analogous to person-hours; say that 3 observations were made of 5 patients and 7 observations on two patients. Within a UEL formalism we could have multiple attributes for a patient each with different values (number of patients), and strictly speaking the recurrence of these attributes in a tag relates to further observations on each group of patients. Note, for Section IV, that this implies an aggregation of data for each patient and an aggregation of data for many patients.

Again formally, at least, there can be many *subj* and many *obj* attributes duplicated by having the same value for an event, state, or measurement, in the same UEL tag, and for repeated measurements under uncertainty, many with the same metadata but different orthodata. If, for example, in evaluating P(A|B) and P(B|A) it is found that these are equal then P(A) = P(B) and $K(A; B) = P(A)^{-1} = P(B)^{-1} > 1$, and this is basically the same as stating that A and B are *indistinguishable*. But counting implies, as is normally required, that observations are probabilistically independent, which means that K(A; B) = 1. In the other extreme, A and B are completely distinguishable, if so they are necessarily *mutually exclusiv*e and K(A; B) = 0. These are consistent with Section III.

While for brevity this leaves some loose ends, it is evident that a formal theory of observations and data analytics (and medical inference from these) can be forged from such considerations. All aspects notwithstanding, a simple practical recommendation is that *patient= '1'* be a special case, albeit the grammar should remain the same, allowing for associated probabilities to be used in inference in exactly the same way as for tags representing many patients.

III. THEORETICAL GUIDELINES

In all the above, and in application to data mining for rules and inference from them, we use quantum mechanics (QM), e.g. [8], as a guideline (e.g. [9, 10]). Physicists regard QM as a universal Best Practice [8, 9], and the methodology of interest here takes ideas from a *branch* of QM. Everything discussed in this Section can be described without resorting to quantum mechanical descriptions. However, QM provides a set of mathematical formalisms and concepts that are already operationally rigorous and well validated. Thus, a UEL could take advantage of these principles.

The UEL tags, in regard to attributes *subj*, *obj*, *verb*, *pfwd*, *pbkd*, *assoc*, are not only useful and beneficial, but also the chosen for their fit with QM. These UEL tags are seen as relating directly to the Dirac's notational *bra-ket* form <A | ***R*** |B> having a complex value [8] of form a + *h*b, where a and b are real and *h* is an imaginary number. A corresponds to our simple or compound subject (*subj)*, and B to our simple or compound object (*obj*). ***R*** is the *verb*, specifically for probabilistic semantic purposes a *verb* is a Hermitian operator such that <A | ***R*** | B>* = <B | ***R*** | A> [8]. But now the imaginary number in a + *h*b is *h* [9,10] not *i* as the square root of minus one. This branch of quantum mechanics describes relativistic, particle, and importantly classical phenomena under a Lorenz-Dirac transformation [8, 10]. This transformation is of *i* to *h*, the square root of *plus* one, i.e. the *hyperbolic number* commonly identified with the particle physicists' γ_5. It is a rotation of the reference frame as a generalization of the Wick rotation that renders quantum mechanics classical [8,10].

Again for simplicity consider the categorical case, the relation with QM can clearly be seen in metadata:=orthodata format where in Dirac notation:

<momentum(kgm/s):=7654 | position(m):=273>

With a value say 0.6 – *h*0.3, this becomes in UEL notation, for example,

<uel *subj* ='position(m):=273'
obj='momentum(kgm/s):=7654'
pfwd=0.6 *pbwd*=0.3>

As hyperbolic or *h*-complex probabilities, these objects < >, coincidentally similar in Dirac QM notation and XML, satisfy all classical laws of probabilistic inference (such as the law of composition of probabilities and the chain rule) when applied to the values of *pfwd* and *pbwd* separately (see below). The relation between the probabilities and the *h*-complex value is that in the *h*-complex space the bra-ket value with *empirical probabilities* (as opposed to exponentials as statistical weights calculated *ab initio* from physical principles) can be expressed in a simple *spinor* form [9,10]:

$$\langle A|B\rangle = (1-h)\, P(A|B) + (1+h)\, P(B|A)$$
$$= [(1-h)\, P(A) + (1+h)\, P(B)]\, K(A; B) \quad (1)$$

The complex value, e.g. 0.6 – *h*0.3 corresponds to (*pfwd* + *pbwd*) + *h*(*pbwd* − *pfwd*) in UEL attribute notation. Square roots of P's and K's are used; this need not concern us here, but using square roots does have a semantic interpretation [9].

There is an implicit finer structure, in the advocacy for bra-ket notation, with associated empirical probabilities, and alluring analogies. In physics, the use of a Wick-like rotation necessarily imparts the need for scalar fields based upon a Euclidean functional integral. So, with such a field metric, increasing degrees of variability in $\varphi(x)$ in any d-dimensional space result in diminishing component contributions to the overall Euclidean functional integral. The proposed system has properties very similarly to well-formed statistical mechanical systems, which obey Boltzmann distributions. With the availability of a Boltzmann-like fine structure, it may be possible to extend the QM analogy of the UEL-tag encoded data into distinct NP-complete partition sets, defined by a well-formed Hamiltonian that serves as the partition function for the quantum-like encoding system now possible. The consequence is that the resultant UEL spinors can be seen as a means of bridging the limit function of $\varphi(x)$ to all the possible Bose fields implicit under the current encoding map, as derived by Zee [11]. In any event, it can be shown that the *h*-complex counterpart of a wave function illus-trates the required properties of everyday probabilistic inference [9] and, expressed as distributions, shows the locality characteristic of particle behavior [10]. But what is also worth exploring is that the above equivalency implies that the UEL-encoded spinors of a common set, representing an orthogonal basis, and extracted from original un-encoded data, essentially guarantee reverse decodability in settings with continued availability of appropriate keys.

In general, UEL tags may be used in a broad formal system of probabilistic inference [9, 10] that can make use of data mined or of expert rules, wherever one imagines that a bra-ket or bra-relator-ket would play in an analogous quantum mechanical role. There is even a counterpart of confidence interval in both cases. For conjugate variables $\Delta A \Delta B$ = h/4π (where h, non-italicized, is Planck's constant), and literally translates as "standard deviation of…" Usually, the states, events, or measurements of interest clinically are not conjugate variables. But, that said, it is not generally true that K(A; B) ≈ 1, implying a possible need to treat the confidence intervals as interdependent.

More generally <A| and |B> are *h*-complex vectors and the dot product <A|B>·***R*** when present is an *h*-complex Hermitian operator and ***R*** |B> is a vector. This is useful for handling probability distributions and in reducing order of computation (see N^2 problem below). |B><A| is the probability interaction matrix. However, especially once values are assigned to *pfwd* and *pbwd*, we need only be concerned here with the final dot product. The difference between a categorical and non-categorical verb case can be thought of as formally analogous to the Dirac notational quantum mechanical structure <A|B> and <A| ***R*** |B> respectively, but noting that defining a verb "to be" as the canonical basis for verbs implies <A| ***R*** |B> = <B | A>.

Recall that values of *pfwd* and *pbwd* considered separately satisfy classical probabilistic laws. For most purposes, therefore, it is sufficient to think of the complex value as a simple vector [*pfwd*, *pbwd*], and by only considering one element at a time, the calculation is purely classical (e.g. in the manner of a Bayes Net). This begs the question of why not, for example, use two Bayes Nets? One reason is having both directions of conditionality forces us to consider semantic repercussions that influence overall inference. For example, "All cats are mammals" implies "All non-cats are non-mammals", and the probabilities may be different (except in the case of certainty). Significant use is lost by breaking down hyperbolic inference into two independent classical systems if we do not constantly compare them. If P(A) = P(B), then <A|B> is real-valued, and this means that it has an existential rather than a universal categorical interpretation. The former lacks directional conditionality. Related to that, a cyclic path of bra-kets can be shown to be real-valued [8]. As an indication of how such guidelines may be useful [8,9], we note the generalization of Bayes Nets to quantum mechanical descriptions that can use all forms or probabilistic logic, and potential applications in the ease of treatment of cyclic paths in graphs representing inference networks. The real value of such paths lies in their freedom from requiring iteration. Also, there is the important N^2 problem. *A priori*, for any set of tags or equivalent objects that describe semantic triplets, memory and calculation demands rise proportionally to N^2 for N "nouns" or nodes A, B, C etc. in the general knowledge network as in a graph. If tags describing triples are simply neglected on the grounds that a relationship is fairly close to random, there can be an accumulation of weak evidence from many tags as rules that dominate conclusions and recommendations reached by inference. One may avoid the N^2 problem by instead making reference to at least locally *universal quantum states* $<A|\Psi>$. Again, many aspects can be considered in more familiar mathematical and algorithmic ways, but whether within a single framework is questionable. There is also a striking relationship with natural semantic structure. For example, though beyond present scope, the physicist's *twistor* could be written in forms such as $<<A|B> \mid <C|D>>$ which, with operators analogous to verbs and prepositions, provides guidelines to a probabilistic semantic grammar even closer to practical human language structure.

IV. DISCUSSION AND CONCLUSIONS

It is not at all evident that the grander vision of PCAST will be met, and one suspects that standards bodies will encourage extensive reuse of the work that has resulted from years of hard effort. To achieve uniformity it is important however, to know the theoretical framework, or at least have in mind *some* theoretical framework. At the very least, reference to a physically proven and well worked out formal framework such as QM is insightful for mathematical and algorithmic development of a UEL. Other considerations, such as secure aggregation, will be introduced in upcoming manuscripts.

REFERENCES

[1] President's Council of Advisors on Science & Technology, "Realizing the Full Potential of Health Information Technology to Improve Healthcare for Americans: The Path Forward.". 2010, http://bit.ly/gV2Qun.

[2] B. Robson and OK Baek, The Engines of Hippocrates. From the Dawn of Medicine to Medical and Pharmaceutical Informatics, John Wiley & Sons, US and Canada (2009),

[3] HL7 Standards, http://bit.ly/gqDm7A.

[4] ISO - ISO Standards - TC 215 - Health informatics, http://bit.ly/gKcDbL.

[5] IHE Technical Frameworks, http://bit.ly/grH8wl.

[6] DICOM Homepage, http://bit.ly/g2QbXu

[7] CDISC BRIDG Model, http://bit.ly/i1xwgH

[8] R. P. Feynman and A. R. Hibbs, *Quantum Mechanics and Path Integrals*, emended D. F. Styer, Dover Publications, Mineloa, NY, 2010.

[9] B. Robson (2007), "The New Physician as Unwitting Quantum Mechanic: Is Adapting Dirac's Inference System Best Practice for Personalized Medicine, Genomics and Proteomics?" J. Proteome Res. (A. Chem. Soc.), Vol. 6, No. 8, 2007, pp. 3114 – 3126.

[10] B. Robson, "Towards Intelligent Internet- Roaming Agents for Mining and Inference from Medical Data", *Studies in Health Technology and Informatics*, 149, IOS Press Amsterdam, Berlin, Tokyo, Washington DC, 2009, 157-177.

[11] A. Zee, *Quantum Field Theory in a Nutshell*. Princeton University Press, Princeton, NJ, 2003, 261-263.