Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Implementation of a web based universal exchange and inference language for medicine: Sparse data, probabilities and inference in data mining of clinical data repositories

Q3

Q1 Barry Robson[a,b,1], Srinidhi Boray[a,b,c]

[a] The Dirac Foundation clg, Oxfordshire, UK
[b] St. Matthew's University School of Medicine, Cayman Islands
[c] Ingine Inc., Potomac Falls, VA 20165, USA

ARTICLE INFO

Article history:
Received 6 April 2015
Accepted 17 July 2015

Keywords:
Universal exchange language
Dirac notation
Bayes
Watson
Electronic health record
Public health reporting
Probability theory
Zeta function
Cognitive computing

ABSTRACT

Q4 Q5 We extend Q-UEL, our universal exchange language for interoperability and inference in healthcare and biomedicine, to the more traditional fields of public health surveys. These are the type associated with screening, epidemiological and cross-sectional studies, and cohort studies in some cases similar to clinical trials. There is the challenge that there is some degree of split between frequentist notions of probability as (a) classical measures based only on the idea of counting and proportion and on classical biostatistics as used in the above conservative disciplines, and (b) more subjectivist notions of uncertainty, belief, reliability, or confidence often used in automated inference and decision support systems. Samples in the above kind of public health survey are typically small compared with our earlier "Big Data" mining efforts. An issue addressed here is how much impact on decisions should sparse data have. We describe a new Q-UEL compatible toolkit including data analytics application (DiracMiner) that also delivers more standard biostatistical results, DiracBuilder that uses its output to build Hyperbolic Dirac Nets (HDN) for decision support, and HDNcoherer that ensures that probabilities are mutually consistent. Use is exemplified by participating in a real word health-screening project, and also by deployment in a industrial platform called the BioIngine, a cognitive computing platform for health management.

1. Introduction

1.1. Background

We previously proposed Q-UEL, a web-based universal exchange and inference language for healthcare and biomedicine [1,2]. Here is it extended to the more traditional domain of public health analysis including general population health sampling, healthcare quality surveys, and screenings. The techniques used can include or extend to cross-sectional studies, cohort studies, and other similar investigations including, to some extent, clinical trials. Q-UEL's capabilities are already quite broad. Unhampered by any obligations due to existing clinical implementations (but see Section 1.2) and other legacy matters, and free of rigorous and laborious systems of voting for change that characterize the major medical standards efforts [2], we have been able to go back to a "green field", exploring fundamental unifying principles and to build Q-UEL extremely rapidly. Q-UEL communication artifacts are XML-like tags that mostly represent probabilistic *statements* as captured clinical data and medical knowledge, derived by mining both structured data records and natural language text [1]. Different Q-UEL tags are concerned with both content and management of data representation including electronic health records and their interoperability and inter-conversion, security and disaggregation [2], knowledge extraction and representation, machine learning [1,3], and hypothesis generation, automated reasoning and decision support [3,4]. The Q-UEL system might reasonably be described as a multivariate architecture for cognitive computing in medicine. Its most basic inference systems are like the *Bayes Net* (BN) [5], except that Q-UEL's counterparts are *general graphs* (Section 1.6). The most advanced resemble IBM's *Watson* [6], except that Q-UEL intrinsically uses *probabilistic semantics* [1,4] and seeks to distribute the required computing power parallelized across a larger number of more standard processes on the Internet rather than on a limited number of very

E-mail address: robsonb@aol.com (B. Robson).
URLS: http://www.diractfoundation.org (B. Robson), http://www.ingine.com (S. Boray).

[1] Tel.: +1 345 928 7242.

http://dx.doi.org/10.1016/j.compbiomed.2015.07.015
0010-4825/© 2015 Elsevier Ltd. All rights reserved.

high performance machines, taking advantage of Cloud Computing, grid architecture and in-memory processing. This all seems far removed from traditional biostatistics, epidemiology, and evidence based medicine. But despite the stark contrast between the familiar and conservative nature of public health analysis and many modern efforts in medical inference (Section 1.13), interoperability will expose each one to the other, raising deeper issues such about how to treat sparse data, how to include well found prior knowledge as if it were virtual data, how to reconcile many different notions about probabilities, and how best to do inference from them. To address these matters, a number of software applications with several novel algorithms were developed, using Q-UEL tags with features suited to public health analysis.

1.2. The origins and nature of Q-UEL

Q-UEL is based on the *Dirac Notation* and associated algebra [7]. The notation was introduced into later editions of Dirac's book to facilitate understanding and use of quantum mechanics (QM) [8] and it has been a standard notation in physics and theoretical chemistry since the 1940s. QM is a system for representing observations and measurements, and drawing probabilistic inference from them. The Q in Q-UEL refers to QM, but a simple mathematical transformation of QM [9,10] gives classical everyday behavior. Q-UEL inherits the machinery of QM by replacing the more familiar imaginary number $\boldsymbol{i}$ (such that $\boldsymbol{ii} = -1$), responsible for QM as wave mechanics, by the *hyperbolic imaginary number* $\boldsymbol{h}$ (such that $\boldsymbol{hh} = +1$). It was rediscovered in various guises by Dirac. Hence our inference net in general is called the *Hyperbolic Dirac Net* (HDN) [3]. The "UEL" in Q relates to the "PCAST 2010" call for a Universal Exchange Language (UEL) for healthcare. The usual qualification "web based" is because Dirac notation by coincidence looks a little like XML and has a bracket structure (it is also called the *braket* notation), and it also represents a semantic system that even maps to natural language [4]. It thus seemed to us the natural choice of XML extension required to render the Semantic Web (SW) [12–14] probabilistic [15], and for using it as input for Clinical Decision Support Systems (CDSS) [16]. Q-UEL's older information-theoretic origins lie in bioinformatics [18,19] and clinical "Big Data" data mining [20–22], thus ensuring Q-UEL's applicability to those disciplines [3]. They bring to Q-UEL the use of "inference probabilities" based on *expected information* [18] now seen as based on the *partially summated Riemann Zeta Function* [20,21] that incorporate sparse data and prior subjective opinion into the inference process. It is of importance in this paper.

1.3. Purpose and content of the present paper

Q-UEL may seem exotic, but the whole field of ES, CDSS, and recently probabilistic reasoning on the web, has been characterized by a tradition of innovation and controversy [1] set by the pioneering MYCIN effort [17]. It seems inevitable that there will be a boundary of tension in healthcare information technology, between classical notions of probability and these seemingly better to suited to inference, if not resolved before full interoperability is achieved. The traditional small sample population studies addressed in this paper conveniently and quickly raise *issues of sparse data and inference from it* that relate to, and allow us to address, the above concern. Although Q-UEL has benefited so far by not being implemented in healthcare data management, the need to address the issue of sparse data in greater detail became pressing because Q-UEL has been deployed in the construction, management, and analysis of patient records in effectively "real time" in an ongoing cardiovascular risk screening and study of a Caribbean population (Section 1.12). The sample size is typical of the kind of studies mentioned in Section 1.1. However, sparse data is also always ultimately encountered in high dimensional (multifactor) data mining, because there are always many probabilities based on large combinations of factors that are seen very few times, if at all [20–22]. Consequently, studies will be performed to demonstrate scalability. Sparse data is important because a great deal of evidence that is individually weak can combine to overthrow a decision made without it. Neglect can be an accumulative error.

1.4. Overall strategy used in the present paper

Though the work described in the present paper is most closely related to the basic BN concept, our general strategy of *observe, evaluate, interpret, and decide* used to build and evaluate a final net looks very like Watson's advertized "cognitive principles" of *observe, interpret, evaluate and decide*. Our use of "evaluate" is different, meaning evaluate Q-UEL's "inference probabilities", but these do appropriately represent weights of evidence from data, sparse as well as ample, and subjective prior opinion. Watson's evaluation involves running hundreds of language analysis algorithms simultaneously and selecting the one most probably giving the right answer [6]. At last inspection, its main use of probability resides there. From Watson as it competed on the TV quiz show "Jeopardy!" [1,6] one might expect similar qualitative answers in a medical setting, e.g. "pulmonary circulation disorder" as a most plausible fit to a long description as the question, based on association. Q-UEL has some such capabilities [1], but in the present paper, the questions are such as "Will my female patient age 50–59 taking diabetes medication and having a body mass index of 30–39 have very high cholesterol if the systolic BP is 130–139 mmHg and HDL is 50–59 mg/dL and non-HDL is 120–129 mg/dL?". This forms a preliminary inference net as the query, which may be refined and to which probabilities must be assigned. The real answers of interest here are not qualitative statements, but the final probabilities. The protocols involved map to what data miners often seem to see as two main options in mining, although we see them as the two ends of a continuum. Method (A) may be recognized as *unsupervised* (or *unrestricted*) *data mining* and *post-filtering*, and is the method mainly used here. In this approach we (1) mine data ("observe"), (2) compute a very large number of the more significant probabilities and render them as tags ("evaluate"), (3) use a proposed inference net as a query to search amongst the probabilities represented by those tags, but only looking for those relevant to complete the net and assign probabilities to it, assessing what is available, and seeing what can be substituted ("interpret"), and (4) compute the overall probability of the final inference net in order to make a decision ("decide"). Unsupervised data mining is preferred because it generates many tags for an SW-like approach, and may uncover new unexpected relationships [22] that could be included in the net. Method (B) uses *supervised* (or *restricted*) *data mining* and *pre-filtering*. Data mining considers only what appears in the net. The downstream user interested in inference always accesses the raw database, while in (A) he or she may never see it. The advantage of (B) is that mining is far less computationally demanding both in terms of processing and memory. It is a persuasive option when using extensive patient records [22]. Indeed it is used with such larger data below. However, even there it is only applied in a limited way rather than as very specific questions, to avoid what is irrelevant at least for the purposes being demonstrated.

1.5. Overview of common types of probability

Common notions about probability or uncertainty used or potentially used by medical information technology developers can differ in many ways, as illustrated by most of the work cited above [1–10,15–28] supplemented by some key works and reviews [23–28]. Most of these share some detectable connection