Computers in Biology and Medicine 56 (2015) 107–123

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

# POPPER, a simple programming language for probabilistic semantic inference in medicine

Barry Robson [a,b,c,1]

[a] St. Matthew's University, School of Medicine, George, Grand Cayman, Cayman Islands
[b] Department of Mathematics Statistics and Computer Science, University of Wisconsin-Stout, WI, USA
[c] The Dirac Foundation, Oxfordshire, UK

ARTICLE INFO

*Article history:*
Received 30 June 2014
Accepted 12 October 2014

*Keywords:*
Medical inference
Decision support system
Expert system
SW
Dirac
Hyperbolic
Complex
Bayes Net
Popper

ABSTRACT

Our previous reports described the use of the Hyperbolic Dirac Net (HDN) as a method for probabilistic inference from medical data, and a proposed probabilistic medical Semantic Web (SW) language Q-UEL to provide that data. Rather like a traditional Bayes Net, that HDN provided estimates of joint and conditional probabilities, and was static, with no need for evolution due to "reasoning". Use of the SW will require, however, (a) at least the semantic triple with more elaborate relations than conditional ones, as seen in use of most verbs and prepositions, and (b) rules for logical, grammatical, and definitional manipulation that can generate changes in the inference net. Here is described the simple POPPER language for medical inference. It can be automatically written by Q-UEL, or by hand. Based on studies with our medical students, it is believed that a tool like this may help in medical education and that a physician unfamiliar with SW science can understand it. It is here used to explore the considerable challenges of assigning probabilities, and not least what the meaning and utility of inference net evolution would be for a physician.



## 1. Introduction

### 1.1. Background

Our suggested Q-UEL universal exchange language for healthcare [1] and for inference [2] is amongst many efforts aimed at a more probabilistic form [3] of the current Semantic Web (SW) [4]. The present report extends the basic Hyperbolic Dirac Net (HDN) [2] that relates to so-called Bayesian approaches [5,6] that are confined to conditional probabilities, i.e., forms like $P(A|B, C)=P(A,B,C) \ / \ P(B,C)$ which is essentially the probability of the statement "*A* if *B* and *C*" being true. The HDN is now extended to encompass other relationships than "if", such as in subject-verb-object, that the SW and Q-UEL provide. However, the methodological emphasis below is on a simple POPPER application and POPPER computer language. In its recent form it is seen in part as a tool (a) to convert Q-UEL tags to a form that makes simpler the building of inference networks and the inference engines that drive them, and, importantly, (b) as an interface by which a medical expert can construct Q-UEL tags (when data mining to obtain probabilistic information is not an easy option). In actuality, POPPER was originally partly developed as a simplest possible "toy system" with which medical students might express and solve inference problems in probabilistic semantics, and the original form goes back to 2009 and is the parent of Q-UEL. The first motivation for Q-UEL was to provide a universal or at least uniform language to extract biomedical information, for automated reasoning systems like POPPER, from the web and the SW in particular. The second motivation was because Q-UEL seemed to meet US Federal demands for a universal exchange language for healthcare in late 2010 [1], including means of securely transporting patient data, and extracting further statistical biomedical information from many millions of consenting patients [1].

### 1.2. Probabilistic conditional inference

It is not obvious even to current SW developers how to assign *plausible* probability values to semantic statements and how a user should interpret them, and perhaps even less clear how one is supposed to interpret and utilize the changes in the inference net that can result from the action of the rules [3]. The diversity of proposed solutions suggests lack of agreement [1]. The popularity of the Bayes Net (BN) [6] is almost certainly because it is easy to understand its quantitative meaning, utility, and mode of use, as what we call a *probabilistic conditional inference* (PCI) approach. PCI in general uses basic building blocks that are the probabilities that one or more aspects are the case given that one or more other aspects are the case. As noted above, a conditional probability like $P(A|B, C)=P(A, B, C)/P(B, C)$ has the force of *P* ("*A* if *B* and *C*"). The aspects *A*, *B*, *C*, etc. are states, events, things,

*E-mail address:* robsonb@aol.com
[1] Tel.: +1 345 945 3199x193; fax: +1 345 945 3130.

http://dx.doi.org/10.1016/j.compbiomed.2014.10.011
0010-4825/

observations, measurements, descriptions, etc. that we call *attributes* by reference to the attributes of the XML-like Q-UEL language. *A* PCI net is an estimate of a joint or conditional probability with many attributes when there is insufficient data to calculate the probability directly. Primarily, conditional probabilities with fewer attributes such as $P(A|B, C)$ are multiplied together. For a traditional BN this is according to the rule that the attributes as nodes and the conditional relationships between them as connections must describe a *directed acyclic graph* (DAG) [6]. The HDN as described in Ref. [2] is also a PCI but uses *dual probabilities* such as $(P(A \mid B, C), P(B, C \mid A))$, considering two directions of conditionality at the same time, say (0.9, 0.6) [1,2]. This is the *value* of a POPPER tag $\langle A \mid B, C\rangle$. The perhaps unobvious mathematical consequences are that the HDN is no longer constrained to a DAG. The dual itself is, nonetheless, a notion that should be obvious to physicians. For example, there is the risk factor that obesity presents for type 2 diabetes, and the different risk factor that type 2 diabetes presents for obesity. These are conditional probabilities *P*(type 2 diabetes | obesity) and *P*(obesity | type 2 diabetes) though typically expressed on, e.g., a percentage or per millum basis.

### 1.3. Association constants and mutual information

While emphasis here is on assigning probabilities to statements subjectively by human experts, it can sometimes be done more objectively and/or automatically by data mining and text analytics. By such means one may often obtain association constants such as $K(A;\ B)=P(A,\ B)/P(A)P(B)=P(A|B)/P(A)=P(B|A)/P(B)$, or $K(A;\ B;\ C)=P(A,\ B,\ C)/P(A)P(B)P(C)$ [2]. The effect of a verb as the relationship between subject and object expressions can often be quantified by an association constant. Association constants form the basis of our *K* method for building an HDN [2], and $K(A;\ B)$ appears on Q-UEL tags along with forms such as $P(A|B)$ and $P(B|A)$ to allow many clinical and epidemiological metrics to be calculated [1]. It relates to Bayes Eqs. (1) and (2) in that inference can be framed in terms of $P(A)$ as prior probability and $P(A|B)$ as posterior probability in $P(A|B)=K(A;\ B)P(A)$. Reference will also be made below to information $I(A;\ B)=\ln K(A;\ B)$, which the author refers to as *Fano mutual information*, after Fano [7]. Many authors speak *of Fano's inequality mutual information* but, strictly speaking, the only inequality for present purposes is that $P(A|B)$ is not in general equal to $P(A)$. It can be, but then $I(A;\ B)=0$ and $K(A;\ B)=1$, and *A* carries no information about *B* and *vice versa*. Note that, unlike some other kinds of mutual information, it is important to our use of $I(A;\ B)$ that it may be greater or less than 0 (and hence *K* greater or less than 1).

### 1.4. Probabilistic semantic inference

Conditional probabilities such as $P(A|B)$ are often capable of a number of other semantic interpretations, including the causal *P*("*A* is caused by *B*"), the transformational *P*("*B* becomes *A*"), the comparative *P*("*A* is greater than *B*"), and not least the categorical *P*("*B* are *A*") and similar set-theoretic interpretations, or to imply relationships where something is propagated, or a chain of effect (Section 5.3). These remain important as valid *probabilistic semantics*, and are responsible for the remarkably broad range of applicability of a BN. Also, the role of a verb of action as in "physicians treat patients" can almost always be rendered categorically: "physicians are patient-treaters". However, homely examples are often used below of "dogs chase cats" and "cats chase mice" because "chase" is a good example of a verb of action that not fully interpretable in conditional ways and very difficult to use in chains of inference if its role is rendered categorical. For a probabilistic SW and probabilistic semantic inference (PSI), the vertical bar '|' must also often be replaced by many other kinds of relationships, including verbs of action and prepositions, when perceived linguistically. Traditionally there is no *P*(dogs|**chase**|cats) except as one possible canonical format for *P*("dogs chase cats"), yet humans have no difficulty in seeing that a probability can be assigned, at least qualitatively. Few would bet a thousand dollars that you will never find a dog that does not chase cats, or even that you would never see a cat chase a dog. But ⟨dogs|**chase**|cats⟩ does have special meaning encoding a probability dual in *Dirac notation* and corresponding algebra [8] used in quantum mechanics (QM) [9,10], albeit that there words like dogs and cats are replaced by physical states, events, observations and measurements and a verb like chase is replaced by an operator. Indeed, a verb is seen a kind of operator in Q-UEL and POPPER, using the term "relationship operator" or *relator* to describe a relationship in general: hence the **R** in ⟨*A*|**R**|*B*⟩. Note that POPPER does not simply see a verbal relationship as some kind of additional condition, but as with QM, **R** is an operator that can in principle be defined as a matrix.

As well as such a semantic triple, Q-UEL statements can be more elaborate *semantic multiples* [1], such as the nested form ⟨ ⟨A|**R**|*B*⟩ |*S*| ⟨*C*|**T**|*D*⟩⟨*E*|**U**|*F*⟩ ⟩, that can represent a tree graph as (the parsed structure of) a sentence. ⟨*A*|***R***|*B*⟩ relates to the physicist's *spinor* or dual spinor, and the above nested is an extension of a physicst's higher level *twistor* in ⟨⋯⟩ format [1,2,10]. POPPER is a simplified language that does not support or utilize all Q-UEL features, but while it does not currently *explicitly* support the above nested form, it also does not prohibit it. As described in Section 3.7, POPPER metastatements (essentially, match and edit instructions that act on statements to apply logical and definitional rules in inference) derive from such a nested "extended twistor" form. In addition, the POPPER method allows construction of statements with such nested forms, simply by using assignment statements to create them as it would create a new statement as a semantic triple in ⟨*A*|**R**|*B*⟩ format. However, the resulting forms have not as yet been manipulated in any way that would justify their use. In this report, statements are confined to the ⟨*A*|**R**|*B*⟩ format. It remains that *A*, **R** and *B* can *each* be strings of words that can be manipulated by metastatements, and these could have implied nested forms like that above.

### 1.5. Previous work

Although the literature of probabilistic semantics goes back a long way, it has been overshadowed by use of symbolic logic [11], and only recently has there been a significant increase in interest (see Ref. [12] for discussion and review). The problem of what probabilities to assign in POPPER goes back in part to the work of Karl Popper [13] that addressed more generally what it means to make statements about the world. Dirac himself felt that QM and his methods should also be applicable to many aspects of logical and probabilistic human thought [1,2], but he did not say exactly how. Considering the way in which he extended Schrödinger's wave mechanics to particle theory [10], it seems likely that he had in mind (albeit under other guises) the hyperbolic imaginary number $\boldsymbol{h}$ (i.e., $\boldsymbol{hh}=+1$) that allows use of empirical classical probabilities [1,2]. Adding the semantic interpretation, however, draws on artificial intelligence [14], and linguistic theory [15,16], relational data mining [17], as well as on Expert Systems [18,19] which have a tradition of innovation for medical decision support [20,21]. While our approach also appears to be such an innovation, $\boldsymbol{h}$ has been rediscovered in different guises in separate fields many times since Cockle first described it [22], including by Dirac [10]. A largely $\boldsymbol{h}$-complex QM has only been described fairly recently [23]. Significantly, the relevance to mental function was quickly posited [24], and before that $\boldsymbol{h}$-complex algebra has been of considerable interest to the neural network community (e.g. Ref. [25]). Considering the relevance to quantification of relationships, it is not too surprising that the algebra has also been explored for use in on-line dating recommender systems [26]. Such efforts are relatively recent and may yet converge to a unified field, but at present Dirac's work remains the comprehensive resource.