Ploxoma: Testbed for Uncertain Inference

H. Blau
Department of Computer Science, University of Rochester, Rochester, New York 14627-0226
blau@cs.rochester.edu

Abstract

This paper compares two formalisms for uncertain inference, Kyburg's Combinatorial Semantics and Dempster-Shafer belief function theory, on the basis of an example from the domain of medical diagnosis. I review Shafer's example about the imaginary disease *ploxoma* and show how it would be represented in Combinatorial Semantics. I conclude that belief function theory has a qualitative advantage because it offers greater flexibility of expression, and provides results about more specific classes of patients. Nevertheless, a quantitative comparison reveals that the inferences sanctioned by Combinatorial Semantics are more reliable than those of belief function theory.

1 Introduction

This paper compares Kyburg's Combinatorial Semantics (CS) and Dempster-Shafer belief function theory (BFT), two formalisms for representing uncertain inference from statistical data. I reconsider a medical diagnosis example introduced by Shafer [13] concerning the imaginary disease *ploxoma*. In Section 2, I cite Shafer's example and sketch the steps of its treatment using BFT. Section 3 presents a précis of Combinatorial Semantics[1], and Section 4 shows how the ploxoma example can be handled in the CS framework. Section 5 draws conclusions from this comparison. One of the operations applied in the BFT analysis cannot be translated into CS, and CS groups together two classes of patients that BFT can distinguish. I compute new belief functions after eliminating from the example those elements that are problematic for CS. While BFT appears on qualitative grounds to be more flexible and informative, a quantitative comparison shows that the conclusions sanctioned by CS are more reliable. BFT's results may not conform to frequencies observed in the real world; those of CS do.

Many authors, including [6, 10, 12], have observed that belief functions lack a simple frequency interpretation. The contribution of this paper is a comparison between BFT and CS, discussed here for the first time. In the tradition of Smets [14, 15], I center my investigation on a concrete instance chosen not as a counterexample, nor as a philosophical conundrum like the three prisoner problem [4, 5], but as a basis for contrasting alternative analyses. A good example is a valuable tool to isolate the points of difference among theories, and to identify their strengths and weaknesses. Shafer's ploxoma story suits this purpose well because it is fairly realistic and has a rich structure that calls for a variety of belief function operations.

2 Shafer's Ploxoma Example

Shafer [13, p. 331] uses the example of an imaginary disease *ploxoma* to illustrate how belief functions could be used to model uncertainty in medical diagnosis.

> Imagine a disorder called "ploxoma" which comprises two distinct "diseases": θ_1 = "virulent ploxoma", which is invariably fatal, and θ_2 = "ordinary ploxoma", which varies in severity and can be treated. Virulent ploxoma can be identified unequivocally at the time of a victim's death, but the only way to distinguish between the two diseases in their early stages seems to be a blood test with three possible outcomes, labelled x_1, x_2 and x_3. The following evidence is available:
> (i) Blood tests of a large number of patients dying of virulent ploxoma showed the outcomes x_1, x_2 and x_3 occurring 20, 20 and 60 per cent of the time, respectively.
> (ii) A study of patients whose ploxoma had continued so long as to be almost certainly ordinary ploxoma showed outcome x_1 to occur 85 per cent of the time and outcomes x_2 and x_3 to occur 15 per cent of the time. (The study was made before methods for distinguishing between x_2 and x_3 were perfected.) There is some question whether the patients in the study represent a

[1] Combinatorial Semantics [7] provides the semantic underpinnings for Evidential Probability [8].

	$BEL_{123}(\{\theta_1\}\|x_i)$	$BEL_{123}(\{\theta_2\}\|x_i)$	$BEL_{123}(\{\theta_1,\theta_2\}\|x_i)$
x_1	0.014	0.965	0.021
x_2	0.062	0.918	0.020
x_3	0.165	0.782	0.053

Table 1: BEL_{123} conditioned on results of blood test

fair sample of the population of ordinary ploxoma victims, but experts feel fairly confident (say 75 per cent) that the criteria by which patients were selected for the study should not affect the distribution of test outcomes.

(iii) It seems that most people who seek medical help for ploxoma are suffering from ordinary ploxoma. There have been no careful statistical studies, but physicians are convinced that only 5–15 per cent of ploxoma patients suffer from virulent ploxoma.

The three items of evidence are represented by three belief functions (BEL_1, BEL_2, BEL_3) whose frame of discernment is the cross product of disease and blood test result: $\Lambda = \{\theta_1, \theta_2\} \times \{x_1, x_2, x_3\}$. The study of virulent ploxoma can be represented by a belief function BEL_α defined over $\Omega = \{(\theta_1, x_1), (\theta_1, x_2), (\theta_1, x_3)\}$. But we need a belief function defined over all of Λ, not just a subset of Λ. We *conditionally embed* BEL_α in Λ to obtain BEL_1. The study of ordinary ploxoma can be represented by a belief function BEL_β defined over $\Omega = \{(\theta_2, x_1), (\theta_2, x_2), (\theta_2, x_3)\}$. To capture the experts' misgivings about the sample selection in this study, we *discount* BEL_β at the rate of 0.25, yielding BEL_γ. We conditionally embed BEL_γ in Λ to obtain BEL_2. The third item of information can be represented by a belief function BEL_δ defined over $\Omega = \{\theta_1, \theta_2\}$. We *minimally extend* BEL_δ to obtain BEL_3 defined over Λ. The belief functions for the three pieces of evidence are then combined using Dempster's rule of combination, yielding BEL_{123}. The last step is to condition BEL_{123} on the results of the patient's blood test (Table 1, corresponds to Shafer's Table 3 [13, p. 332]). Complete details of this derivation may be found in [1].

3 Précis of Combinatorial Semantics

When we have data that are not entirely conclusive, but that support a particular hypothesis to some degree, we may choose to accept the the hypothesis, provided the uncertainty does not seem unacceptably high. Kyburg's Combinatorial Semantics [7] seeks to formalize inference from a set of sentences Γ to a sentence ϕ, where Γ does not *entail* ϕ, but to some extent *justifies* ϕ.

3.1 Syntax of the Language

Kyburg's logic is two-sorted. The *empirical* sort is used to describe the domain of discourse. The *mathematical* sort is intended to represent real numbers. Predicate symbols, function symbols, and variables for the empirical part of the language are as in standard first order logic. The mathematical sort includes mathematical variables, the constants 0 and 1, and the standard operations and relations for real numbers. Kyburg introduces a special variable binding operator, denoted by '%', that relates formulas and real numbers. '$\%(\psi, \phi, p, q)$' is a well-formed formula if (i) ψ and ϕ are open formulas having no free mathematical variables, and the free empirical variables of ϕ include all those of ψ;[2] (ii) p and q are mathematical variables or rigid real number designators. This type of formula will be called *statistical*, and we will refer to '%' as the *statistical operator*. This sentence means: among those objects in the empirical domain that satisfy ϕ, the proportion of objects that also satisfy ψ lies between p and q, inclusive. For example, the formula '$\%(\texttt{practices-sports}(x), \texttt{resident-of}(x, \texttt{Colorado}), 0.70, 0.80)$' says that somewhere between 70% and 80% of Colorado residents practice sports.

3.2 Semantics for the Language

A model $\mathcal{M}$ for this language is a tuple $\langle D_m, D_e, \xi_m, \xi_e \rangle$ where D_m and ξ_m (D_e and ξ_e, respectively) are the domain and interpretation function for the mathematical (empirical) part of the language. The empirical domain D_e is constrained to be *finite* so that the "proportion" involved in the semantics of the statistical operator is well-defined. The definition of truth for statistical statements involves a new

[2] In addition, ψ must belong to the class of target formulas and ϕ to the class of reference formulas—see [7] for details.

concept, the *satisfaction set* of a formula ϕ with respect to a model $\mathcal{M}$: $SS_{\mathcal{M}}(\phi)$. If ϕ is a closed formula, then $SS_{\mathcal{M}}(\phi) = D_e$ iff ϕ is true in $\mathcal{M}$, and $SS_{\mathcal{M}}(\phi) = \emptyset$ otherwise. If ϕ is an open formula having n free empirical variables and no free mathematical variables, the satisfaction set of ϕ is the subset of D_e^n for which $\mathcal{M}$ makes ϕ true. $SS_{\mathcal{M}}(\phi)$ is undefined when ϕ has free mathematical variables. If ϕ consists of an n-ary empirical predicate symbol 'pred' followed by n empirical variables, then $SS_{\mathcal{M}}(\phi)$ is simply ξ_e('pred'). For example, $SS_{\mathcal{M}}$('plays-sport(p, s)') is the set of pairs $\langle p, s\rangle$ of elements in the empirical domain of $\mathcal{M}$ such that person p plays the sport s. But not all the terms following the predicate symbol need be variables. $SS_{\mathcal{M}}$('plays-sport$(p,$ tennis$)$') is the set of tennis players. Note that the free variables in ϕ may occur as arguments to function symbols. $SS_{\mathcal{M}}$('plays-sport(brother-of$(p,$ john$)$, tennis$)$') is the set of tennis-playing brothers of John.

A statistical formula '$\%(\psi, \phi, p, q)$' is true with respect to a model $\mathcal{M}$ and a variable assignment v exactly when (i) $SS_{\mathcal{M}}(\phi)$ is non-empty, and (ii) $p \leq \frac{|SS_{\mathcal{M}}(\phi \wedge \psi)|}{|SS_{\mathcal{M}}(\phi)|} \leq q$. If either condition fails to hold, the statistical formula is false in $\mathcal{M}$. Since the empirical domain D_e is constrained to be finite, $SS_{\mathcal{M}}(\phi)$ and $SS_{\mathcal{M}}(\phi \wedge \psi)$ are likewise finite and the ratio in condition (ii) is well-defined (so long as $SS_{\mathcal{M}}(\phi)$ is non-empty).

3.3 Partial Entailment

The criterion of validity for *deductive* inference is: Γ entails ϕ if and only if every model that makes the sentences of Γ true also makes ϕ true. This criterion is too strong for *uncertain* inference. Kyburg proposes the criterion of *partial entailment*, which captures the idea that the inference from Γ to ϕ is justified (although not valid) if ϕ is true in some proportion (almost all) of the models in which Γ is true. We say that Γ *partially entails, to the degree* $[p, q]$, the formula ϕ (written $\Gamma \models_{[p,q]} \phi$) if, among the models that make Γ true, the proportion that also make ϕ true lies between p and q, inclusive.[3] The proportion is well-defined: there can only be a finite number of models for the empirical part of the language, because the language itself is finite and the empirical domain is finite.

What is the relation between partial entailment and uncertain inference? If $\Gamma \models_{[p,q]} \phi$ and p is sufficiently close to 1, then we regard the inference from Γ to ϕ as warranted. Γ justifies the acceptance of ϕ. How do we decide if p is "sufficiently close" to 1? That depends on the situation: we set the threshold of acceptance at a level appropriate to the context in which we are reasoning.

3.4 Statistical Formulas and Partial Entailment

Suppose we know that between 70% and 90% of all lap swimmers wear swim goggles, and Jane is a lap swimmer.

$$\Gamma_1 = \left\{ \begin{array}{l} \%(\texttt{wears-goggles}(x), \texttt{lap-swimmer}(x), 0.70, 0.90) \\ \texttt{lap-swimmer(jane)} \end{array} \right\}$$

What can we conclude about whether Jane wears swim goggles? Theorem 1 tells us that Γ_1 partially entails, to the degree [0.70,0.90], the sentence '**wears-goggles(jane)**'. We may or may not want to accept this sentence based on our knowledge of the two premises. It depends whether we think 0.70 is sufficiently close to 1 for us to consider the inference warranted.

Theorem 1 $\left\{ \begin{array}{l} \text{'}\%(\texttt{A}(x), \texttt{B}(x), p, q)\text{'} \\ \text{'}\texttt{B(a)}\text{'} \end{array} \right\} \models_{[p,q]} \text{'}\texttt{A(a)}\text{'}$

It turns out that lap swimmers who prefer the backstroke do not need swim goggles as much as other lap swimmers do. So among backstroke lap swimmers, only 50% to 60% wear swim goggles. Furthermore, we learn that Jane is a lap swimmer who prefers the backstroke. Our database of knowledge now contains:

$$\Gamma_2 = \left\{ \begin{array}{l} \%(\texttt{wears-goggles}(x), \texttt{lap-swimmer}(x), 0.70, 0.90) \\ \%(\texttt{wears-goggles}(x), \texttt{lap-swimmer}(x) \wedge \texttt{backstroke-swimmer}(x), 0.50, 0.60) \\ \texttt{lap-swimmer(jane)} \\ \texttt{backstroke-swimmer(jane)} \end{array} \right\}$$

Of course, we know that all backstroke lap swimmers are lap swimmers.

$$\forall x((\texttt{lap-swimmer}(x) \wedge \texttt{backstroke-swimmer}(x)) \rightarrow \texttt{lap-swimmer}(x))$$

[3]The idea of partial entailment explained here is a simplification: the full definition of this concept is much more complex (see [7, Section 5]). But this simplified version will be adequate for our purposes.

What do we now conclude about whether Jane wears swim goggles? We have conflicting statistics for lap swimmers in general and for lap swimmers who prefer the backstroke. But since we know that Jane herself prefers the backstroke, we select the more specific reference class instead of the broader one. Theorem 2 tells us that Γ_2 partially entails, to the degree [0.50,0.60], the sentence '`wears-goggles(jane)`'.

Theorem 2, the principle of specificity, says that when we have conflicting statistical information, we guide our judgment by the most specific reference class. What is meant by "conflicting" statistical information? We say that two intervals $[p, q]$ and $[r, s]$ *differ* when neither is included in the other (they may partially overlap, or they may be disjoint).

Theorem 2 If DIF($[p, q], [r, s]$), then $\left\{ \begin{array}{ll} \text{`\%(A}(x), \text{B}(x), p, q)\text{'} & \text{`B(a)'} \\ \text{`\%(A}(x), \text{C}(x), r, s)\text{'} & \text{`C(a)'} \\ \text{`}\forall x(\text{C}(x) \rightarrow \text{B}(x))\text{'} & \end{array} \right\} \models_{[r,s]} \text{`A(a)'}$

The case of two statistical statements in which the interval of one is wholly included in the interval of the other falls under a different principle, that of strength [7, Section 8][8, pp. 380–381]. The principle of strength requires a more complex definition of partial entailment than the one given above. I have summarized here only the concepts of CS that are necessary for the analysis of the ploxoma example.

4 Combinatorial Semantics Analysis of Ploxoma Example

To compute interval estimates, we need to know how many patients participated in the epidemiological studies. Suppose we were told that the study of virulent ploxoma involved a total of 200 patients, of whom 40 had blood test result x_1, 40 had test result x_2, and 120 had test result x_3. The study of ordinary ploxoma involved a total of 1000 patients, of whom 850 had test result x_1, and 150 had test result x_2 or x_3. Since ordinary ploxoma is more common than virulent ploxoma, it is not surprising that the second study had a much larger sample size than the first. From these data we calculate intervals at the 99% confidence level, which we then incorporate into the statistical formulas (1)–(5).

(1) %(`blood-x1`(y), `virulent`(y), 0.137, 0.282)
(2) %(`blood-x2`(y), `virulent`(y), 0.137, 0.282)
(3) %(`blood-x3`(y), `virulent`(y), 0.509, 0.685)
(4) %(`blood-x1`(y), `ordinary`(y), 0.819, 0.877)
(5) %((`blood-x2`(y) $\vee$ `blood-x3`(y)), `ordinary`(y), 0.123, 0.181)

Formulas (6) and (7) capture the opinion of doctors concerning the prevalence of virulent ploxoma.

(6) %(`virulent`(y), `ploxoma`(y), 0.05, 0.15)
(7) %(`ordinary`(y), `ploxoma`(y), 0.85, 0.95)

What about the complication that scientists are only (as Shafer puts it) 75% sure that the results of the second ploxoma study are reliable? There is no good way to represent this information in the CS framework. One could consider two ways to take this aspect into account: (i) broaden the intervals in formulas (4) and (5), to indicate greater ignorance about the true proportion of blood test result x_1 (x_2, x_3) among ordinary ploxoma patients, or (ii) compute intervals at a lower confidence level. The problem with (i) is: by how much do we broaden the interval? What principle would sanction the choice of this number? The problem with (ii) is: the confidence level associated with an interval estimate does not represent a judgment of how well the experiment was designed. The 75% confidence interval would in fact be narrower than the 99% interval. All in all, it seems wise to refrain from adjusting our confidence intervals in ways that have no foundation in statistical theory. We will therefore proceed as if we had no doubts about the quality of the data from the second ploxoma study.

John walks into the doctor's office, suffering from ploxoma.[4] His blood test result is x_1. We cannot infer anything about John from the statistical statements (1)–(5), because we do not know whether John belongs to the reference class of virulent or of ordinary ploxoma patients: indeed, this is precisely what we want to find out. Obviously we need to turn these statistics around to get an estimate of the incidence of virulent ploxoma among ploxoma patients with test result x_1. Bayes' Rule allows us to do this.

$$Pr(\theta_1|x_1) = \frac{Pr(\theta_1)Pr(x_1|\theta_1)}{Pr(\theta_1)Pr(x_1|\theta_1) + Pr(\theta_2)Pr(x_1|\theta_2)}$$

[4]Shafer assumes that we already know the patient is suffering from ploxoma; the task is to estimate the risk of virulent ploxoma based on the result of the blood test. Apparently there is no danger of confusing ploxoma with some other disease.

In this equation we consider only ploxoma patients, so θ_1 (virulent) and θ_2 (ordinary) together partition the sample space. We are interested in an interval, not a point estimate, for $Pr(\theta_1|x_1)$. To calculate the lower bound on this interval, we minimize the numerator of the fraction by taking the lower bound for $Pr(\theta_1)$ and $Pr(x_1|\theta_1)$, which forces us to take the upper bound for $Pr(\theta_2)$ and $Pr(x_1|\theta_2)$. To calculate the upper bound on this interval, we maximize the numerator of the fraction by taking the upper bound for $Pr(\theta_1)$ and $Pr(x_1|\theta_1)$, which forces us to take the lower bound for $Pr(\theta_2)$ and $Pr(x_1|\theta_2)$. Substituting the appropriate values from formulas (1), (4), (6), and (7), we obtain the desired statistical statement.

(8) $\%(\texttt{virulent}(y), (\texttt{ploxoma}(y) \wedge \texttt{blood-x1}(y)), 0.008, 0.057)$

When we try to do the same with formula (2), we run into difficulty. To calculate $Pr(\theta_1|x_2)$, we would have to know $Pr(x_2|\theta_2)$. But we have only $Pr((x_2 \vee x_3)|\theta_2)$, because the x_2 patients and the x_3 patients were grouped together in the study of ordinary ploxoma. Instead of $Pr(\theta_1|x_2)$, we must content ourselves with $Pr(\theta_1|(x_2 \vee x_3))$. We have to re-examine the data from the first ploxoma study to compute a confidence interval for $Pr((x_2 \vee x_3)|\theta_1)$. Of the 200 virulent ploxoma patients who participated in the study, the combined total for x_2 and x_3 test results is 160. This yields the confidence interval in (9). We can now calculate the upper and lower bounds for $Pr(\theta_1|(x_2 \vee x_3))$, as well as $Pr(\theta_2|x_1)$ and $Pr(\theta_2|(x_2 \vee x_3))$.

(9) $\%((\texttt{blood-x2}(y) \vee \texttt{blood-x3}(y)), \texttt{virulent}(y), 0.718, 0.863)$
(10) $\%(\texttt{virulent}(y), (\texttt{ploxoma}(y) \wedge (\texttt{blood-x2}(y) \vee \texttt{blood-x3}(y))), 0.173, 0.553)$
(11) $\%(\texttt{ordinary}(y), (\texttt{ploxoma}(y) \wedge \texttt{blood-x1}(y)), 0.943, 0.992)$
(12) $\%(\texttt{ordinary}(y), (\texttt{ploxoma}(y) \wedge (\texttt{blood-x2}(y) \vee \texttt{blood-x3}(y))), 0.447, 0.827)$

Now what can we say about John the ploxoma patient with blood test result x_1? There are two pairs of statistical statements that apply to John: (6) and (7), (8) and (11). By the principle of specificity (Theorem 2), (8) takes precedence over (6), and (11) takes precedence over (7). Our knowledge base partially entails, to the degree [0.008,0.057], the sentence '`virulent(john)`'; it partially entails, to the degree [0.943,0.992], the sentence '`ordinary(john)`'. Depending on our threshold for acceptance, we may want to accept the latter sentence. What about Peter, a ploxoma patient whose blood test result is x_2? Again, there are two pairs of statistical statements that apply to Peter: (6) and (7), (10) and (12). By the principle of specificity, (10) takes precedence over (6), and (12) takes precedence over (7). Our knowledge base partially entails, to the degree [0.173,0.553], the sentence '`virulent(peter)`'; it partially entails, to the degree [0.447,0.827], the sentence '`ordinary(peter)`'. We cannot accept either of these, because in each case the lower bound of the confidence interval does not exceed 0.5. If Peter's test result had been x_3, we would have drawn the same conclusion: we do not have enough information to distinguish between the x_2's and the x_3's when it comes to judging their risk of virulent ploxoma. In sum, a blood test result of x_1 strongly suggests that the patient has ordinary ploxoma, but a result of x_2 or x_3 does not give a clear indication either way.

5 Comparison of the Two Approaches

Having represented the ploxoma example using belief functions and Combinatorial Semantics, we can compare these approaches both qualitatively and quantitatively. Qualitatively, BFT seems to be more flexible. The discounting operation (applied in the derivation of BEL_2) allows us to express the doubts we may have concerning the reliability of the evidence at our disposal. As we have seen, there is no counterpart to this operation in CS. With BFT, we can condition BEL_{123} on x_2 or on x_3 individually, even though the second ploxoma study does not distinguish between them. With CS, we have to merge these two classes of patients. We seem to be losing information that the first study gives us about the difference between x_2 and x_3: result x_3 is more common than x_2 among virulent ploxoma patients.

To make a meaningful quantitative comparison of the two techniques, we must be sure the numbers we compare are based on the same problem description. First, we must recompute BEL_2 without discounting the second item of evidence. This in turn forces us to recompute BEL_{123}, giving us BEL'_{123}. Second, we must condition BEL'_{123} on the set $\{x_2, x_3\}$ instead of the singleton sets $\{x_2\}$ and $\{x_3\}$. The results are shown in Table 2. (See [1, Appendix] for the details of this recalculation.) We now consider the interval [belief,plausibility] for virulent and ordinary ploxoma according to the blood test result. Table 3 sets side by side the belief-plausibility intervals and the degrees of partial entailment from CS. In each case, the

	$BEL'_{123}(\{\theta_1\}\|x_i)$	$BEL'_{123}(\{\theta_2\}\|x_i)$	$BEL'_{123}(\{\theta_1,\theta_2\}\|x_i)$
$\{x_1\}$	0.016	0.963	0.021
$\{x_2,x_3\}$	0.431	0.521	0.048

Table 2: BEL'_{123} conditioned on results of blood test

	Belief Functions	Combin. Semant.	Naive Probability
$\theta_1\|x_1$	0.016 0.037	0.008 0.057	0.012 0.040
$\theta_1\|x_2 \vee x_3$	0.431 0.479	0.173 0.553	0.219 0.485
$\theta_2\|x_1$	0.963 0.984	0.943 0.992	0.960 0.988
$\theta_2\|x_2 \vee x_3$	0.521 0.569	0.447 0.827	0.515 0.781

Table 3: Comparison of intervals

belief function intervals are properly included in the intervals derived by statistical methods. How shall we interpret this relationship?

Classical probability theory provides a common framework in which to evaluate BFT and CS. Kyburg shows in [9] that to any belief function BEL, defined over a frame of discernment Ω, there corresponds a closed, convex set $\mathcal{P}$ of classical probability functions, where

$$\mathcal{P} = \{ Pr : Pr \text{ is a probability function and } \forall A \subseteq \Omega\ (BEL(A) \leq Pr(A) \leq PL(A)) \}.$$

If we understand the belief-plausibility intervals in this way (although Shafer would disapprove), it might seem that they give us more information than CS, because they place tighter bounds on the frequencies that interest us. Unfortunately, in this case the "information" is simply misleading. Consider $Pr(\theta_1|x_2 \vee x_3)$: BEL'_{123} says that this probability is no lower than 0.431 and no higher than 0.479, but the 99% confidence interval extends as low as 0.173 and as high as 0.553. Given the evidence at hand, there is a 1% risk that the true value lies outside [0.173,0.553]. But the risk is greater that the true value lies outside [0.431,0.479]. BEL'_{123} is wrong, and CS right, about the possible worlds in which $Pr(\theta_1|x_2 \vee x_3)$ falls in [0.173,0.431) or in (0.479,0.553]. If we use BFT to guide us in making decisions—for instance, when choosing among various treatment protocols for ploxoma—we will make mistakes more often than if we rely on CS.

Suppose we now ignore the issue of confidence intervals entirely, and accept the observed frequencies of the three blood test results as point estimates for the corresponding probabilities. That is, since 20% of the virulent ploxoma patients tested x_1, we will take $Pr(x_1|\theta_1) = 0.20$, and similarly for the other percentages mentioned in the first and second items of evidence. Doubt remains about the incidence of the virulent form of the disease, which according to doctors strikes 5–15% of all ploxoma patients (the third item of evidence). Again we use Bayes' Rule to calculate probability intervals for $Pr(\theta_1|x_1)$, *etc.*, substituting point estimates for all the conditional probabilities, and the upper and lower limits for $Pr(\theta_1)$ and $Pr(\theta_2)$. The resulting intervals appear in the third column of Table 3. As one would expect, the "naive probability" intervals are narrower than CS's confidence intervals, because we have removed the uncertainty arising from the limited sample sizes in the epidemiological studies. However, the naive probability intervals are wide enough to properly include the belief-plausibility intervals.[5] Even if we had huge sample sizes and could get highly accurate estimates of the proportion of {virulent, ordinary} ploxoma patients with test results $\{x_1,x_2,x_3\}$, we still could not attain the level of certainty that BFT purports to give us.

6 Where's the Beef?

The preceding discussion supports the judgment that Combinatorial Semantics is superior to belief function theory as a formalism for uncertain inference (if we interpret belief functions in terms of convex sets of classical probability functions). CS may be less flexible in its representation of knowledge, may yield less "informative" conclusions, but at least the inferences it does sanction are not misleading. However, the

[5]This comes as no surprise in light of Theorem A.3 in [9, pp. 286-287], which states that the lower and upper bounds on the conditional probability derived using Bayes' Rule are always as wide (and often wider) than the belief-plausibility intervals resulting from Dempster conditioning. This point is mentioned as early as [2].

practical usefulness of CS has yet to be demonstrated. Dempster and Kong [3, p. 33] emphasize that we cannot evaluate a theory of uncertain inference without reference to its practical applications.

> Belief function methodology does introduce more complexity into the class of available representations of uncertainty ... The important question is whether the added flexibility is necessary in practice to permit satisfactory representation of an analyst's state of uncertainty about the real world. We believe that it is literally impossible to answer the question outside the context of real examples based on attempts to construct formal representations of uncertainty reflecting actual uncertain knowledge of the real world.

I agree with Dempster and Kong that the true test of a formalism for uncertain reasoning is its application to real problems in the real world. Loui's experiment in predicting computer network usage [11] is the only such application of Combinatorial Semantics reported to date.

Acknowledgements

I thank my advisor, Professor H. E. Kyburg, for his guidance in this research, and in particular for his patient explanations of Combinatorial Semantics. I am indebted to I. Macarie and C. M. Teng for their helpful comments, and to B. Murtezaoğlu for the use of his program to calculate the normal approximation of the binomial confidence interval. I am also grateful to Professors L. Schubert and N. Jochnowitz for their unfailing confidence in my work.

References

[1] H. Blau and H. E. Kyburg, Jr. Ploxoma: Testbed for Uncertain Inference. Technical Report 537, Department of Computer Science, University of Rochester, December 1994.

[2] A. P. Dempster. Upper and Lower Probabilities Induced by a Multivalued Mapping. *Annals of Mathematical Statistics*, 38(2):325–339, April 1967.

[3] A. P. Dempster and A. Kong. Commentary on Papers by Shafer, Lindley, and Spiegelhalter. *Statistical Science*, 2(1):32–36, 1987.

[4] P. Diaconis. Review of "A Mathematical Theory of Evidence". *Journal of the American Statistical Association*, 73(363):677–678, September 1978.

[5] P. Diaconis and S. Zabell. Some Alternatives to Bayes's Rule. In B. Grofman and G. Owen, editors, *Information Pooling and Group Decision Making: Proceedings of the Second University of California, Irvine, Conference on Political Economy*, pages 25–38. JAI Press, 1986.

[6] D. Hunter. Dempster-Shafer vs. Probabilistic Logic. In *Proceedings Third AAAI Workshop on Uncertainty in Artificial Intelligence*, pages 22–29, University of Washington, Seattle, July 1987.

[7] H. E. Kyburg, Jr. Combinatorial Semantics. Technical Report, Department of Computer Science, University of Rochester, revised version forthcoming.

[8] H. E. Kyburg, Jr. The Reference Class. *Philosophy of Science*, 50:374–397, 1983.

[9] H. E. Kyburg, Jr. Bayesian and Non-Bayesian Evidential Updating. *Artificial Intelligence*, 31:271–293, March 1987.

[10] J. F. Lemmer. Confidence Factors, Empiricism and the Dempster-Shafer Theory of Evidence. In L. N. Kanal and J. F. Lemmer, editors, *Uncertainty in Artificial Intelligence*, pages 117–125. North-Holland, 1986.

[11] R. P. Loui. Evidential Reasoning Compared in a Network Usage Prediciton Testbed: Preliminary Report. In R. D. Shachter, T. S. Levitt, L. N. Kanal, and J. F. Lemmer, editors, *Uncertainty in Artificial Intelligence 4*, pages 253–269. North-Holland, 1990.

[12] J. Pearl. Reasoning with Belief Functions: An Analysis of Compatibility. Technical Report CSD-910047, Computer Science Department, UCLA, July 1991.

[13] G. Shafer. Belief Functions and Parametric Models. *Journal of the Royal Statistical Society, Series B (Methodological)*, 44(3):322–352, 1982.

[14] P. Smets. About Updating. In B. D. D'Ambrosio, P. Smets, and P. P. Bonissone, editors, *Proceedings of the 7th Conference on Uncertainty in Artificial Intelligence*, pages 378–385. Morgan Kaufmann, 1991.

[15] P. Smets. What is Dempster-Shafer's Model? In R. R. Yager, J. Kacprzyk, and M. Fedrizzi, editors, *Advances in the Dempster-Shafer Theory of Evidence*, chapter 1, pages 5–34. John Wiley & Sons, 1994.