U.S. DEPARTMENT OF
ENERGY

PNNL-22126

Prepared for the U.S. Department of Energy
under Contract DE-AC05-76RL01830

An Approach for Assessing the Signature Quality of Various Chemical Assays when Predicting the Culture Media Used to Grow Microorganisms

AE Holmes
LH Sego
BM Webb-Robertson
HW Kreuzer
RM Anderson
SD Unwin
MR Weimar
MF Tardiff
CD Corley

February 2013

Pacific Northwest
NATIONAL LABORATORY

Proudly Operated by **Battelle** *Since 1965*

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor Battelle Memorial Institute, nor any of their employees, makes **any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights**. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or Battelle Memorial Institute. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

PACIFIC NORTHWEST NATIONAL LABORATORY
operated by
BATTELLE
for the
UNITED STATES DEPARTMENT OF ENERGY
under Contract DE-AC05-76RL01830

Printed in the United States of America

**Available to DOE and DOE contractors from the
Office of Scientific and Technical Information,
P.O. Box 62, Oak Ridge, TN 37831-0062;
ph: (865) 576-8401
fax: (865) 576-5728
email: reports@adonis.osti.gov**

**Available to the public from the National Technical Information Service,
U.S. Department of Commerce, 5285 Port Royal Rd., Springfield, VA 22161
ph: (800) 553-6847
fax: (703) 605-6900
email: orders@ntis.fedworld.gov
online ordering: http://www.ntis.gov/ordering.htm**

This document was printed on recycled paper.
(9/2003)

An Approach for Assessing the Signature Quality of Various Chemical Assays when Predicting the Culture Media Used to Grow Microorganisms

AE Holmes[1]
LH Sego[1]
BM Webb-Robertson[1]
HW Kreuzer[1]
RM Anderson[2]

SD Unwin[1]
MR Weimar[1]
MF Tardiff[1]
CD Corley[1]

February 2013

1. Pacific Northwest National Laboratory
2. University of Washington Tacoma

The research described in this report is part of the Signature Discovery Initiative at Pacific Northwest National Laboratory. It was conducted under the Laboratory Directed Research and Development Program at PNNL, a multiprogram national laboratory operated by Battelle for the U.S. Department of Energy.

Pacific Northwest National Laboratory
Richland, Washington 99352

Table of Contents

1.0 Summary

In this report, we present a mathematical framework for assessing the quality of signature systems in terms of fidelity, cost, risk, and utility—a method we refer to as Signature Quality Metrics (SQM). We demonstrate the SQM approach by assessing the quality of a signature system designed to predict the culture medium used to grow a microorganism. The system consisted of four chemical assays designed to identify various ingredients that could be used to produce the culture medium. The analytical measurements resulting from any combination of these four assays can be used in a Bayesian network to predict the probabilities that the microorganism was grown using one of eleven culture media. We evaluated combinations of the signature system by removing one or more of the assays from the Bayes network. We demonstrated that SQM can be used to distinguish between the various combinations in terms of fidelity, cost, risk, and utility—and to account for tradeoffs between these attributes. The approach assisted in clearly identifying assays that were least informative, largely in part because they only could discriminate between very few culture media, and in particular, culture media that are rarely used. There are limitations associated with the data that were used to train and test the signature system. Consequently, our intent is not to draw formal conclusions regarding this particular system, but rather to illustrate an analytical approach that could be useful in comparing one system to another.

2.0 Introduction

We present a mathematical framework for assessing the quality of signature systems in terms of fidelity, cost, risk, and utility. The objective is to provide researchers, developers, and decision makers with a holistic approach for assessing the quality of a signature detection system as it moves through the stages of research and development, construction and evaluation of prototypes, and eventual deployment in an operational environment. We call the methodology Signature Quality Metrics (SQM). It may be used to compare the quality of two (or more) different signature detection systems—or to compare the quality of various implementations of a single system. SQM is based on statistical decision theory (Berger 1985) and the application of that theory, which is often referred to as decision science (Edwards et al. 2007). We demonstrate the SQM approach in the context of assessing the signature quality of various chemical assays when predicting the culture media used to grow microorganisms.

The anthrax incidents of 2001 in the United States made painfully clear the vulnerability of populations to the intentional release of deadly microorganisms. While medical and emergency personnel are essential for dealing with the medical side-effects of such an event, chemical and biological forensics assist in determining the source of contagion and where/how the microorganism was grown. Such details aid law enforcement personnel in identifying suspects, and ultimately, the perpetrator—thereby preventing future acts of terror. Many pieces of information taken together can lead investigators to the suspect. The components of the culture medium used to grow the microorganisms are one such example. The culture medium information can be used to narrow the suspect pool and otherwise inform the investigation. *Bacillus anthracis* (*B. anthracis*) is just one example of a number of microorganisms that could be used in such an attack.

Jarman et al. (2008) present a Bayesian network that leverages the results of three different assays to predict the culture medium used to grow a given batch of anthrax spores. The three assays include: 1) electrospray ionization mass spectrometry (ESI-MS) for identification of agar, 2) isotope ratio mass spectrometry (IRMS) to determine the isotope ratios of carbon and nitrogen and the presence of agar, and 3) secondary-ion mass spectrometry (SIMS) to determine metals content (Zn^{2+} and Cu^{2+} in particular). Continuing this work, we added a fourth assay to the Bayes net, matrix-assisted laser desorption ionization mass spectrometry (MALDI-MS), to determine the presence or absence of heme.

A Bayesian Network is a model that provides a link between data or beliefs and outcomes. It is often expressed as a directed acyclic graph that defines causal relationships in a probabilistic setting. In this case the Bayesian network expresses a relationship between analytical measurements and culture media. For each assay, the "raw" data from the assay are mapped to conditional probabilities related to the heme, agar, metal, or carbon/nitrogen content of the assay and then mapped to a prediction of the probability of each culture medium. The Bayesian network is illustrated in Figure 1. For each of the eleven culture media, the Bayes net estimates the posterior probability that a particular culture medium was used to grow the microorganism, given the results of the four assays. Note that IRMS is used to predict both the presence of agar and the type of C/N source.

The Bayesian net developed by Jarman et al. (2008) is just one example of what we call a *signature system*. A signature system is the collection of measurement techniques, data processing, and algorithms that are collectively used to measure and extract features for the purpose of detecting, predicting, or characterizing a phenomenon of interest. For the Bayes net, the features are the test results from SIMS, MALDI-MS, ESI-MS, and IRMS which test for the presence of metals, heme, agar, and the isotopic ratios of carbon and nitrogen. The phenomenon of interest is the culture medium used to grow the microorganism being tested. For the Bayes net signature system, we are interested in how the removal of one (or more) of the assays from the net affects its performance. Specifically, we evaluate the 15 possible combinations[1] that result from including 1, 2, 3, or 4 assays in the net (irrespective of order). The combinations of assays we investigate are given in Table 1.

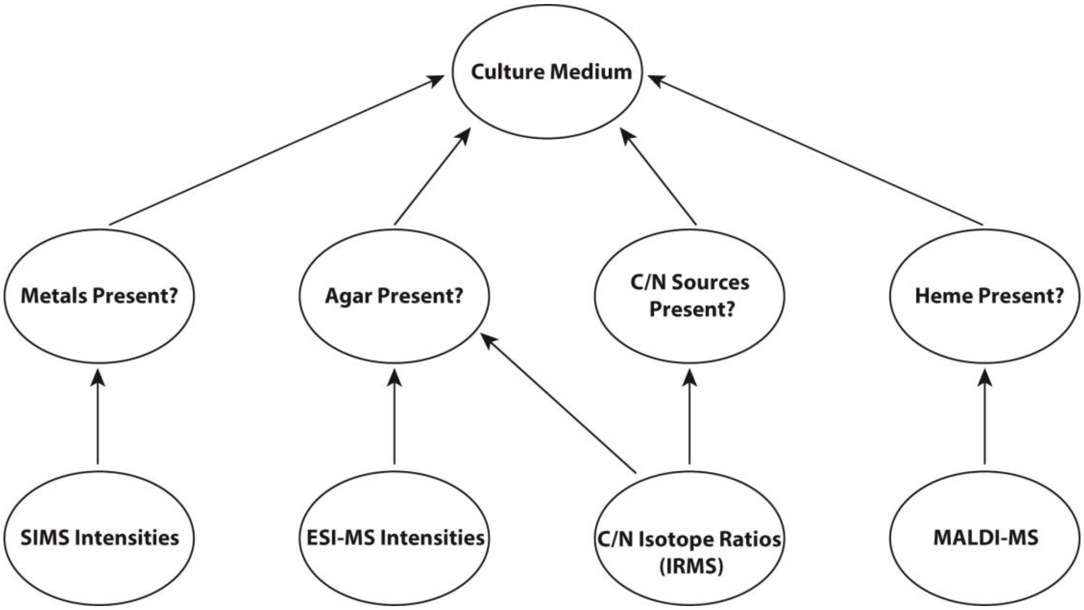

Figure 1: Bayesian network representation of the bioforensic signature system which uses secondary-ion mass spectrometry (SIMS), electrospray ionization mass spectrometry (ESI-MS), isotope ratio mass spectrometry (IRMS), and matrix-assisted laser desorption ionization mass spectrometry (MALDI-MS) to predict the culture medium.

[1] The number of combinations is given by $\sum_{i=1}^{4}\binom{4}{i} = 15$.

Table 1. Combinations of the assays included in the 15 implementations of the Bayes net. An asterisk indicates the test was included, and a blank cell indicates the test was excluded.

Combination	Tests Performed			
Label	SIMS	MALDI	ESI-MS	IRMS
c1	*	*	*	*
c2	*	*	*	
c3		*	*	*
c4	*		*	*
c5	*	*		*
c6	*	*		
c7	*		*	
c8	*			*
c9		*	*	
c10		*		*
c11			*	*
c12	*			
c13		*		
c14			*	
c15				*

We are especially interested in the tradeoffs between fidelity and cost that arise when including (or excluding) various assays in the net, and how principles of decision science (Edwards et al. 2007) can be used to decide between variations of the Bayes net. We define each of the SQM terms below:

- **Fidelity** refers to how well the signature system detects or characterizes the phenomenon of interest. This includes metrics like sensitivity, specificity, accuracy, receiver operating characteristic (ROC) curves, F-score, etc. For the Bayes net signature system, fidelity refers to how well the net predicts the culture media.
- **Cost** refers to the resources expended to develop, deploy, and/or utilize the signature system. Examples include the cost of measurement systems (e.g. SIMS, MALDI-MS, ESI-MS, and IRMS), the cost of consumable reagents, and/or associated labor costs.
- **Risk** refers to the likelihood and consequences associated with decision errors that may result by employing the signature system, helping investigators balance the tradeoff between false positives and false negatives.
- **Utility** provides an overall, relative measure of usefulness, or quality, of a signature system. It is a function of two (or more) of the attributes of fidelity, cost, and risk. It may also include *any* other relevant attribute(s) which may distinguish the value or quality of a signature system from another that are not already accounted for by fidelity, cost, or risk. Examples of other attributes include the time required to collect, process, and analyze samples, human safety, ease of use, etc. For our analysis of the Bayes net signature system, the sample size, i.e., the amount of microorganisms consumed by a particular assay, is of particular interest because the limited forensic sample is arguably the most precious commodity in the analysis.

We examined each of the Bayes net combinations in terms of fidelity, cost, risk, and utility. As is often the case when deciding among alternatives on the basis of multiple attributes (Keeney & Raiffa 1976), a single Bayes net combination that performed the best with respect to each attribute did not exist. Rather, we observed a number of combinations that were ideal in terms of one attribute (such as fidelity), but less ideal in terms of another, such as cost or sample size. In situations like these, multiattribute utility functions (Keeney & Raiffa 1976) provide a useful approach for distinguishing various signature systems by simultaneously accounting for fidelity, cost, risk, and/or other attributes.

3.0 Methods

3.1 Data Sources

The data from the SIMS, MALDI, and ESI-MS tests used for this investigation are described in Jarman et al. (2008). The http://dotearth.blogs.nytimes.com/2013/03/21/scientists-propose-a-new-architecture-for-sustainable-development/?src=recg data for the IRMS test is described in Kreuzer-Martin and Jarman (2007). The dataset used to develop the original Bayes net signature system with all four assays (hereafter referred to as the *training dataset*) included 290 examples of SIMS analysis of *Bacillus* spores grown in various culture media, 50 examples of MALDI-MS analysis, 132 examples of ESI-MS analysis, and 216 examples of IRMS analysis of spores. Training entails using the training dataset to estimate the conditional probabilities associated with the nodes of the Bayes net. There are 11 culture media considered in this study. Not all culture media were represented in each assay type discussed above because the data were, in part, combined from the results of various experiments available in the literature, as discussed by Jarman et al. (2008). The 11 culture media, along with their expected content in terms of metal, heme, agar, and culture media food sources (Atlas 2010, Cliff et al. 2005), are shown in Table 2. The observed carbon/nitrogen ratio (R) for each sample was mapped to one of five categories of culture medium food sources: BP (beef extract/peptone), BT (beef extract/tryptone), ST (soy/tryptone), YS (yeast/sugar), and YT (yeast/tryptone).

Each of the 15 Bayes net combinations was trained using the training dataset. To evaluate the performance of the Bayes net combinations, we constructed an *original test dataset* by randomly sampling the training dataset with replacement based on what should be present for a particular culture medium (as described in Table 2). To illustrate, suppose the culture medium for a single test observation contained heme. Then a random draw of MALDI assay results from samples known to have heme was included in that observation. To complete the data for the observation, a similar process was followed as necessary for the remaining Bayes net components: presence of agar, metal content, and isotope ratios of C and N. Thirty such observations were constructed for each of the eleven culture media, and thus the original test dataset contained 330 observations.

Table 2. Culture media and their constituents, (Atlas 2010, Cliff et al. 2005)

Culture Label	Description	Metal	Heme	Agar	C/N ratio
BA	Blood Agar	None	Positive	Present	BP
GA	Glucose Agar	Cu^{2+}, Zn^{2+}	Unknown	Present	YS
GB	Glucose Broth	Cu^{2+}, Zn^{2+}	Unknown	Absent	YS
NA	Nutrient Agar	None	Unknown	Present	BP
NB	Nutrient Broth	None	Unknown	Absent	BP
LBA	Luria-Bertaini Agar	None	Unknown	Present	YT
LBB	Luria-Bertaini Broth	None	Unknown	Absent	YT
NSMA	Nutrient Sporulating Medium Agar	None	Unknown	Present	BT
NSMB	Nutrient Sporulating Medium Broth	None	Unknown	Absent	BT
TSA	Tryptic Soy Agar	None	Unknown	Present	ST
TSB	Tryptic Soy Broth	None	Unknown	Absent	ST

In practice, multiple assays would be performed on the same biological sample. Unlike what would happen in practice, the training data available to us resulted from convenience samples with assays performed on separate biological samples. Similarly, the assay results were each sampled separately in the test dataset construction. When possible, it is ideal to use separate training and test datasets to construct and evaluate classification methods, such as the Bayes nets under consideration. Often this is done through bootstrapping or k-fold cross-validation (Kohavi 1995). In order to assess the error associated with the estimates of fidelity, risk, and utility calculated from the original test dataset, one thousand test datasets were constructed from bootstrapped samples of the original test dataset, i.e., each bootstrapped test data set had 330 observations and was obtained by sampling the original test data set with replacement. Using test datasets derived from the training dataset has drawbacks. First, it is likely that error estimates obtained from the bootstrapped test datasets will underestimate the true error associated with the measures of fidelity, risk, and utility for each combination. Second, it is possible that estimates of the posterior probabilities of the culture media are biased. While the SQM analysis techniques we present are valid, the particular numerical results computed from these data should be interpreted with caution, in light of the limitations of the data discussed in this paragraph.

3.2 Fidelity

The principal objective of the Bayes net signature system is to predict which one of 11 different culture media was used to grow *B. anthracis* based on four assays (SIMS, MALDI-MS, ESI-MS, and IRMS) performed on the spores of the microorganism. Because the output of the Bayes net is a vector of posterior probabilities (i.e. one probability value for each culture media such that the sum of the 11 probabilities is unity), scoring rules are an appropriate means of assessing the accuracy of the 15 combinations of Bayesian networks. A scoring rule (Bickel 2007) is a measure of the performance of the estimated probabilities of a finite set of outcomes, or categories, in light of the eventual outcomes. These estimated probabilities may be the result of some type of mathematical or statistical model (e.g. a Bayes net), or they could simply be the informed guesses of experts. A scoring rule can be any function of these estimated probabilities and the true probabilities. In practice, the true probabilities are often estimated by the frequencies of observed outcomes. Three common scoring rules which also have the property of being *strictly proper* include the logarithmic, Brier, and spherical scoring rules (Winkler 1996). Strictly proper scoring rules have the property that they can be decomposed into the sum of two components (Bickel 2010), *calibration* and *sharpness*. Calibration, or accuracy, is a measure of how well the probability estimates match the true values of the probabilities. Sharpness refers to the degree to which the probability estimates place most of the probability on a single class (or very few classes).

3.2.1 Definition of Scoring Rules

We now define the logarithmic and Brier scoring rules. To simplify the notation, we omit distinguishing subscripts for the 15 combinations of the Bayes net. Likewise, we omit subscripts that distinguish the 1000 bootstrapped test datasets. Thus, the following definitions apply to a particular combination of assays and a single bootstrapped test dataset. Let $\boldsymbol{x}_k,\ k = 1, \ldots, n,$ represent the measured data from the assays, where each $\boldsymbol{x}_k$ represents the data from a single "sample" in a test dataset. In this case, $n = 330$ indicates the size of a bootstrapped test dataset. Note that $\boldsymbol{x}_k$ is a vector of dimension 1, 2, 3, or 4, depending on the number of assays that are included for the Bayes net combination of interest. Let $j = 1, \ldots J$ index the culture media, where, for this example, $J = 11$. Let the Kronecker delta be defined as $\delta_j(\boldsymbol{x}_k) = 1$ if $\boldsymbol{x}_k$ truly belongs to culture media j and 0 otherwise. Let $\hat{f}(j|\boldsymbol{x}_k)$ denote the posterior probability (estimated by the Bayes net) that the culture medium used to grow *B. anthracis* sample is j, given the value of the observed data, $\boldsymbol{x}_k$. The logarithmic score (Hand 1997) is simply the logarithm of the probability estimate assigned to the correct class. It is given by

$$L = \frac{1}{n}\sum_{k=1}^{n}\sum_{j=1}^{J}\delta_j(x_k)\,log\big[\hat{f}(j|x_k)\big], \tag{1}$$

and takes values in $(-\infty, 0]$. Here, "log" denotes the natural logarithm (although any base may be used). The Brier score (Hand 1997) is an average of the squared deviation between the estimated probability of culture media and the binary outcome of each culture media, and is given by

$$B = \frac{1}{n}\sum_{k=1}^{n}\sum_{j=1}^{J}\big(\delta_j(x_k) - \hat{f}(j|x_k)\big)^2, \tag{2}$$

and takes values in $[0,2]$. Larger scores are better for the logarithmic score, while smaller scores are better for the Brier score.

To facilitate comparison of the Bayes net combinations using the log and Brier score, we follow the approach of Bickel (2010) and map the two scores to a common scale and orientation via a linear transformation.[2] In the rescaled versions of the Brier and logarithmic scores, 0 indicates the score that would be obtained by uninformed guessing, i.e., the score resulting from uniform probability assignments, where each class is assigned a probability of J^{-1}, and 1 indicates the largest possible (and best) score. Note that worse-than-uniform probability estimates (which occurs, for example, when estimates greater than J^{-1} are placed on incorrect classes, and less than J^{-1} is placed on the correct class) can give rise to negative scores. Furthermore, the logarithmic score has no minimum possible score, since assigning a probability of 0 to the correct class gives a logarithmic score of negative infinity.

For the logarithmic score, uniform probability assignments (i.e. setting $\hat{f}(j|\boldsymbol{x}_k) = J^{-1}$ for each j in (1)) gives $L = -\log J$. Thus, the linear transformation of the logarithmic score in (1) that satisfies the criteria stated above is given by

$$L^R = 1 + \left(\frac{1}{log\,J}\right)L \tag{3}$$

and takes values in $(-\infty, 1]$. For the Brier score, uniform probability assignments in (2) gives $B = (J-1)/J$, and the corresponding linear transformation is given by

$$B^R = 1 + \left(\frac{J}{1-J}\right)B, \tag{4}$$

taking values in $\left[\frac{1+J}{1-J}, 1\right]$. Thus, for both L^R and B^R, larger is better, the maximum possible score is 1, and a score below 0 indicates performance that is worse than assigning equal probabilities to each of the culture media.

We now present two approaches for measuring the sharpness of the probabilities estimated by the Bayes net. Sharpness increases as probability estimates approach unity for a particular class (but not necessarily the correct class) and as the estimates for the remaining classes go to zero. In the decomposition of the logarithmic score, the measure of sharpness is given by the negative entropy, or *information*, of the estimated probabilities (Bickel 2010), averaged across the n test cases:

$$I = \frac{1}{n}\sum_{k=1}^{n}\Big(\sum_{j=1}^{J}\hat{f}(j|x_k)\ \ log\,\hat{f}(j|x_k)\Big), \tag{5}$$

which takes values in [-log$J$, 0]. This interval arises because I is minimized for uniform probability assignments and maximized when $\hat{f}(j'|\boldsymbol{x}_k) = 1$ and $\hat{f}(j|\boldsymbol{x}_k) = 0$ for $j \neq j'$. Note that $0 \log 0$ is defined to be 0, which makes sense because $\lim_{x\to 0^+} x \log x = 0$.

For the Brier score, a useful measure of sharpness (Murphy 1973) resembles the variance of a Bernoulli random variable (again, averaged across the n test cases):

[2] Any linear transformation of a strictly proper scoring rule is also strictly proper (Toda 1963).

$$S = -\frac{1}{n}\sum_{k=1}^{n}\left(\sum_{j=1}^{J}\hat{f}(j|x_k)\left(1-\hat{f}(j|x_k)\right)\right), \tag{6}$$

which takes values in $[J^{-1}-1, 0]$. The endpoints of this interval arise under the same two conditions that produce the bounds for (5), i.e. the lower bound resulting from uniform probability assignments and the upper bound occurring when all the probability is assigned to a single class. Larger is better for both I and S. As with the Brier and logarithmic scores, we can rescale I and S with nearly the same[3] linear transformations used in (3) and (4) so that they range from 0 to 1:

$$I^R = 1 + \left(\frac{1}{log\, J}\right) I \tag{7}$$

$$S^R = 1 + \left(\frac{J}{J-1}\right) S \tag{8}$$

The information, $I,$ is typically associated with the logarithmic score because it estimates the true (negative) entropy that is a component of the decomposition of that score (Bickel 2010). Likewise, the sharpness, $S,$ is usually associated with the Brier score because it estimates the sharpness component of the decomposition of the Brier score (Murphy 1973).

The logarithmic score (hereafter referred to as the log score) is a local rule, in the sense that it accounts only for how well the Bayes net predicts the correct culture medium. The Brier score, on the other hand, gives weight to the fidelity of the predictions for each of the culture media. In other words, for any signature system that produces a vector of probability estimates, the log score rewards the system only for putting high probability on the correct class, while the Brier score rewards the system for putting low probability on all the incorrect classes (and necessarily, a relatively large amount of probability will be assigned to the correct class). In many, if not most, instances, the log score is preferred because it agrees with the likelihood principle and it provides a more consistent basis (Winkler 1996, Bickel 2010) for comparing the performance of several signature systems. We include the Brier score in the discussion primarily because of its common historical use.

3.2.2 Fidelity of the Bayes Net Combinations

We chose to evaluate the 15 combinations of 4 assays using the rescaled Brier and logarithmic scoring rules. Figure 2 illustrates the rescaled log score (L^R) versus the rescaled information (I^R) and the rescaled Brier score (B^R) versus the rescaled sharpness (S^R). The scores obtained from calculating these quantities for each of 1000 bootstrap sample results are displayed as bag plots (Rousseeuw, Ruts, and Tukey, 1999). Bag plots are essentially two-dimensional box plots where, typically, the inner 50% of the data are contained in the innermost bag. The outermost bag (referred to as the loop) is the convex hull that contains all points not deemed to be outliers. The small red dots indicate outliers that are outside the loop. The bag plots provide a sense of the error in the estimates of the various scores. Informally, if the bags of two combinations overlap, this suggests the fidelity of the corresponding combinations may be indistinguishable.

Inspection of Figure 2 reveals the two pairs of scoring rule metrics (L^R, I^R) and (B^R, S^R) both lead to similar conclusions: combinations c1, c3, c4, and c11 appear to be the most promising in terms of both their fidelity (measured by L^R and B^R) and their information or sharpness (I^R and S^R). The overlap of the bagplots suggests these combinations are statistically similar. These results are corroborated by metrics of risk and utility, as discussed below.

[3] The only difference being that the coefficient in (8) is the negative of the coefficient in (4).

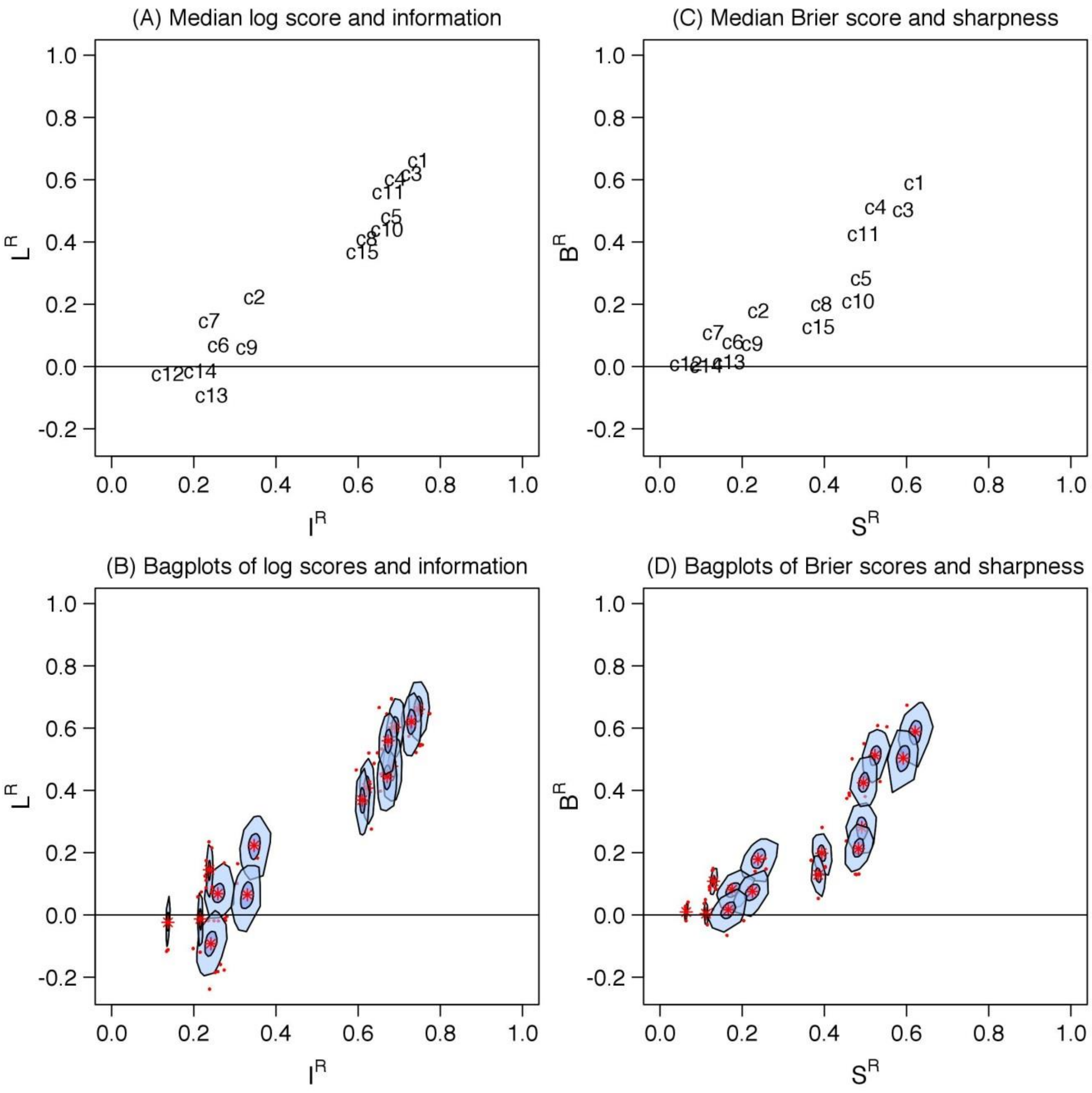

Figure 2: Bayes net models evaluated by rescaled logarithmic score vs. the rescaled information and rescaled Brier score vs. the rescaled sharpness. The median values plotted in panels (A) and (C) were calculated from the 1000 bootstrap samples of each combination. Likewise, the two-dimensional boxplots (known as bagplots) of the 1000 bootstrap samples for each combination are shown in panels (B) and (D).

3.3 Cost and Other Attributes

There are two important considerations with regard to the resources consumed when exercising the Bayes net signature system: 1) the monetary cost to perform the assays, and 2) the sample size required for each assay. The cost of analyzing a single sample (including sample processing) using a given assay was estimated using prices posted on websites by commercial laboratories. In Table 3, we list them as approximate because there may be slight variations from one laboratory to the next. While these prices will change over time, these specific cost estimates serve the purpose to illustrate how we account for cost in the SQM analysis. In what follows, the cost of each Bayes net combination was calculated by summing the costs of the assays in that combination (as shown in Table 1).

In many cases, we anticipate the amount of biological material available for analysis will be small. Many of the assays are destructive, and it will often be desirable to not consume the entire sample so that additional, confirmatory tests may be performed as desired. Consequently, the forensic sample is the most precious resource consumed by the Bayes net signature system. Thus, if two assays provide the same (or similar) forensic intelligence, the one that consumes the least amount of biological sample is preferred. Alternatively, we want to "spend" the sample on assays that provide the greatest insight. The sample size is analogous to a cost, expressed in terms of mg of sample consumed instead of dollars spent. Approximate sample sizes are provided in Table 3.

Another attribute that might be of interest would be the time required for sample processing and analysis for each of the assays. For this analysis, however, we did not include the attribute of time because the relatively minor differences in processing time among the assays would be insignificant compared to the amount of time that would be required to deliberate and execute a comprehensive forensic investigation.

Table 3. Estimated commercial costs and sample size requirements for a single sample

Assay	Cost ($)	Sample Size (mg)
SIMS: Elemental content	200	0.1
MALDI-MS: Heme analysis	170	0.01
ESI-MS: Agar analysis	250	1.0
IRMS: C,N isotope ratio and Agar analysis	100	0.3

3.4 Risk

While the Bayes net heretofore discussed only makes predictions about the likelihood that a particular culture medium was used to grow the organism of interest, investigators ultimately need to associate the predicted culture medium with potential suspects. As the perpetrator clearly must have the specialized knowledge required to produce *B. anthracis* spores, one way to make this association is to link the culture medium to research institutions that are known to have used that particular medium to culture *B. anthracis* and to focus the investigation on individuals associated with those institutions such as current or former employees, students, or trainees. Consequently, our conceptual model for assessing the risk of the Bayes net signature system is that law enforcement officials will sequentially investigate institutions that are likely to be associated with the culprit, beginning with the most likely institution(s) first, followed by the next most likely, and so on. For convenience, we will use the term *culpable institution* to refer to the institution whose investigation leads to the identification of the culprit. However, this terminology is not meant to infer that an institution itself is culpable.

Thus, we calculate risk as a function of the number of institutions investigated until the culpable individual(s) are found. If a particular combination of assays results in a Bayes net signature system that results, on average, in five institutions being investigated before a culprit is identified, that combination has lower risk than, say, a combination that results in investigating six institutions on average before identifying the culprit. While not perfectly realistic, this measure of risk more closely reflects the context of the use of the signature system than do measures of fidelity, like the log and Brier score.

3.4.1 Extending the Bayes net

Webb-Robertson et al. (2012) extended the Bayes net to predict likely institutions and even geographical regions where suspects may be present. By curating scientific literature, Webb-Robertson et al. (2012) related culture media used to grow *Bacillus* species with those institutions where they were grown. They identified over 2,469 documents with abstracts that mention a *Bacillus* species published between 2001 and 2011. A random sample of 150 documents was taken from the twenty-five journals with the most publications among the identified documents. The text from the title, authors, institutions, abstract, and *Materials and Methods* sections was extracted from these 150 documents. Two microbiology students independently reviewed each of the publications' abstract and Materials and Methods sections following a strict protocol to annotate each citation with the specific culture medium used (e.g., tryptic soy broth). The annotation protocol was as follows: 1) read the abstract and Materials and Methods sections, 2) open spreadsheet, 3) indicate whether the document indeed mentioned cultured *B. anthracis*, 4) note the culture media, 5) look again, and 6) record any additional information. Among the 150 annotated documents 144 were validated to be associated with *Bacillus*, and in 52 of those, they were able to identify the culture medium or media with confidence.

Subject matter experts identified a set of 34 key words (83 including abbreviations and spelling variations) that would be associated with the culture media found within the 52 documents. Identifiers were chosen for the culture media and for institutions at the institution level (e.g., Florida Department of Health, Bureau of Laboratories is simply coded as the Florida Department of Health). Each of these 52 documents has 1) binary vectors of size 83 where a '1' indicated the presence of a particular key word, 2) the institution(s) that were listed as associated with the authors of the manuscript and 3) the culture medium (or media) used to grow *Bacillus*. The remaining 92 documents (144 total *Bacillus* documents less the 52 annotated documents) had the first two sets of descriptions, key words and institutions, but the true culture medium was unknown. A total of 84 institutions were represented among the documents. We subsequently refer to the institution and culture medium data obtained from the literature as the *textual data*.

3.4.2 Risk Model

As with the scoring rules from Section 3.2.1, for the sake of simplicity we omit distinguishing subscripts for the 1000 bootstrapped test datasets. Thus, the following definitions apply to a particular combination of assays and a single bootstrapped test dataset. Let $c = 1, \dots 15$ index the Bayes net combinations defined in Table 1. Let $i = 1, \dots, I$ index the institutions, where, in this case, $I =$ 84 institutions were identified via the curating process described in Section 3.4.1. We use $j = 1, \dots, J$ to index the 11 culture media, and $k_j = 1, \dots, n_j$ to index the measured outcomes of a single "sample" grown in culture medium j in the test dataset. Let $\mathcal{X} \subseteq \mathbb{R}^h$ denote the space of assay measurements, where $h = 1, 2, 3,$ or 4, depending on the number of assays in the particular Bayes net combination, and let $\boldsymbol{X} \in \mathcal{X}$ denote the random vector of assay measurements, and $\boldsymbol{x}_{k_j}$ represent the *observed* assay measurement for sample k_j. Sometimes we will refer simply to an observed datum $\boldsymbol{x}$ when it is not necessary to distinguish between culture media or sample replicates in the test dataset.

Let Y represent the *culpable*[4] institution, the investigation of which will lead to the successful identification of the culprit(s). We define Y as a discrete random variable that takes values in the set $\{1,2, \dots, I\}$. Hence, $Y = i$ will indicate that institution i is culpable, and $P(Y = i)$ will represent the corresponding probability. For simplicity, we assume there is only one institution whose investigation

[4] For convenience, we often refer informally to Y with admittedly imprecise language such as the *culpable* institution. Or we may refer to $P(Y = i)$ as the *culpability* probability of institution i.

will lead to identification of the culpable party. Let M represent the true culture media used to grow the microorganism in the attack, taking values in $\{1,2,\ldots,J\}$. Thus, $M=j$ indicates that culture medium j was used, and $P(M=j)$ represents the corresponding probability of that event. Again, we assume only a single culture medium was used to grow the microorganism. Clearly, the series of possible events leading from the prediction of the culture medium to correctly identifying and prosecuting the culprit are complex. While these simplifying assumptions of the risk model may lack realism, they provide a framework that makes it possible to compare one Bayes net combination to another in terms of risk.

Let $p_j := P(M=j)$ and $P(Y=i)$ represent *a priori* probabilities[5] for the culture medium and institutional culpability probabilities, respectively. Values of p_j may be chosen based on the prevalence of the culture media, as some media may be more common (or preferable to use) than others. However, in the absence of such information, a non-informative prior distribution could be used where each culture medium is equally likely, i.e. $p_j = 1/J$. Values of $P(Y=i)$ may be chosen based on available information, e.g. institution size, location, and/or their known capability to produce certain microorganisms. Or, a non-informative distribution could be used such that $P(Y=i)=1/I$ for all i. Naturally, these probabilities must be chosen such that $\sum_j p_j = 1$ and $\sum_i P(Y=i)=1$.

We estimated the values of p_j using the counts of the number of instances where a particular culture media was mentioned in the textual data. These are shown in Table 4. Of the 11 culture media considered in this work (see Table 2), only 8 were represented in the textual data, i.e., all but GA, GB, and NSMB. To reconcile this discrepancy, we assumed GA, GB, and NSMB are unlikely to be used and, consequently, we set their prior probabilities to be small (0.005) to reflect the fact they were not observed in the textual data, but also not exclude them entirely. The media NA and NSMA were not distinguished from each other in the textual data, i.e., they were given the same label. We chose to set the prior probabilities for NA and NSMA to be equal to half the frequency with which the single label occurred in the textual data, thereby presuming that NA and NSMA are equally likely. The remainder of the probabilities were chosen by dividing the count by 171 (the sum of the counts) and then renormalizing so that $\sum_j p_j = 1$.

Table 4: Assigning the prior probabilities, p_j, of the culture media, using the textual data. Media acronyms are defined in Table 2.

Media	BA	GA	GB	NA	NB	LBA	LBB	NSMA	NSMB	TSA	TSB
Count	11	0	0	6	15	39	22	6	0	35	10
$\boldsymbol{p_j}$	0.0752	0.0050	0.0050	0.0410	0.1026	0.2668	0.1505	0.0410	0.0050	0.2394	0.0684

In comparing the 15 Bayes net combinations, we proceed with the assumption, for a given culture medium, that institutions capable of producing *Bacillus* are likely to be culpable, *a priori*, based on the frequency with which they discuss using a particular culture medium in scientific publications. Specifically, institution i is deemed capable of obtaining and using culture medium j if they published on having done so in the textual data. We represent institutional capability symbolically as $P(Y=i|M=j)>0$, that is, there is a non-zero prior probability that institution i is culpable, given that culture medium j was used in the attack. Institution i is deemed incapable of using culture medium j if there are no records in the textual data of that institution using that particular medium, in which case $P(Y=i|M=j)=0$. Because our textual data were obtained from a sample of the literature, the prior estimates of institutional culpability and culture media prevalence obtained from the textual data will be

[5] These probabilities reflect our best guess of the culture media prevalence and the probability of institutional culpability before assay data from a forensic sample are available.

subject to sampling bias. In particular, it is likely there are cases where we incorrectly set $P(Y = i|M = j) = 0$ because we simply did not observe any cases where institution i published about using culture medium j. Notwithstanding these limitations, for the purpose of illustrating the methodology of calculating the risk, we proceed with the understanding that conclusions from our example analysis may be inaccurate because of limitations in both the textual and assay data.

We now define a *capability matrix*, $W = \{w_{ij}\}$, with I rows and J columns, where w_{ij} is a count of the number of articles in the textual data where institution i mentions the use of culture medium j. If any of the rows of W are equivalent, the publishing profiles of the corresponding institutions are identical. There were 11 instances where 2 or more institutions had identical publishing profiles, making it difficult for the extended Bayes net model to distinguish the culpability of these institutions. We used W to estimate the prior conditional culpability probability for each institution and the prevalence of the culture medium to arrive at the estimate of the marginal culpability probabilities, as follows:

$$\begin{aligned} P(Y = i) = &\quad \textstyle\sum_{j=1}^{J} P(Y = i|M = j)P(M = j) \\ \approx &\quad \textstyle\sum_{j=1}^{J} \left(\frac{w_{ij}}{\sum_{i=1}^{I} w_{ij}}\right)\left(\frac{\sum_{i=1}^{I} w_{ij}}{\sum_{i=1}^{I}\sum_{j=1}^{J} w_{ij}}\right) = \frac{\sum_{j=1}^{J} w_{ij}}{\sum_{i=1}^{I}\sum_{j=1}^{J} w_{ij}} := q_i \end{aligned} \qquad (9)$$

In the absence of assay data and a prediction of the culture medium from the Bayes net, these *a priori* estimates of the institutional culpability probabilities q_i would guide the order in which an investigation would take place. The institution with the largest q_i would be investigated first, followed by the institution with the next largest q_i, and so on. We make the simplifying assumption that the culprit will successfully be identified after investigating one, several, or, in the worst case scenario, all, of the I institutions. When there are ties in the q_i, we assume institutions with ties are investigated in a randomized order. In what follows, it will be useful to rank the q_i's such that the largest q_i is ranked 1 and the smallest receives rank I, and the sum of the ranks is $\sum_{i=1}^{I} i = I(I + 1)/2$. Ties are handled by averaging the corresponding ranks, as illustrated in Table 5. We define the vector of ranks of *a priori* probabilities as follows: $\boldsymbol{r} = \left(\text{rank}(q_1), \text{rank}(q_2), \ldots, \text{rank}(q_I)\right)^T$.

Table 5: Example of average ranking method for four institutions

q_i	$q_1 = 0.1$	$q_2 = 0.2$	$q_3 = 0.5$	$q_4 = 0.2$
$\boldsymbol{r}$	4	2.5	1	2.5

In the more meaningful situation where assay data from a forensic sample are available, (9) can be modified by replacing the prior probability $p_j = P(M = j)$ with the posterior probability, $P_c(M = j|\boldsymbol{x})$, obtained from Bayes net combination c, as follows:

$$P_c(Y = i|x) = \textstyle\sum_{j=1}^{J} P(Y = i|M = j)P_c(M = j|x) \approx \sum_{j=1}^{J} \left(\frac{w_{ij}}{\sum_{i=1}^{I} w_{ij}}\right) P_c(M = j|x) \qquad (10)$$

Let $\boldsymbol{r}_c(\boldsymbol{x}) = \left(\text{rank}\big(P_c(Y = 1|\boldsymbol{x})\big), \text{rank}\big(P_c(Y = 2|\boldsymbol{x})\big), \ldots, \text{rank}\big(P_c(Y = I|\boldsymbol{x})\big)\right)^T$ represent the ranks of the posterior probability estimates of institutional culpability obtained from Bayes net combination c. Again, we make a simplifying assumption that the cost, C, of investigating a particular institution is the same for all institutions. We now specify a loss function that reflects the cost of the number of institutions investigated until the culprit is found. The posterior loss function is given by

$$L\big(y, \boldsymbol{r}_c(\boldsymbol{x})\big) = C \cdot \text{rank}\big(P_c(Y = y|\boldsymbol{x})\big) \overset{\text{def}}{=} C\boldsymbol{r}_c^{[y]}(\boldsymbol{x}) \qquad (11)$$

respectively, where the value of the superscript $[y]$ indicates the y^{th}element of the posterior rank vector $\boldsymbol{r}_c(\boldsymbol{x})$. For example, if the culpable institution is $Y = 3$, the posterior loss would be $L\big(3, \boldsymbol{r}_c(\boldsymbol{x})\big) =$

$C\boldsymbol{r}_c^{[\mathbf{3}]}(\boldsymbol{x}) = C \cdot \mathrm{rank}\big(P_c(Y=3|\boldsymbol{x})\big)$. If there are ties among the $P_c(Y=i|\boldsymbol{x})$, the average ranking method reflects the assumption that tied institutions are inspected in a random order. Consequently, the loss function in (11) returns the expected, or average, loss for the tied institutions. To illustrate, consider again the ranking in Table 5 and suppose $Y=4$. Because $q_2 = q_4$, we randomly choose to investigate either institution 2 or 4. If institution 2 is investigated before 4, the cost will then be $3C$, since we must investigate institutions 3, 2, and then 4 before finding the culprit. Because the ordering is random, investigating institution 2 before 4 occurs with probability one half. If we randomly choose to investigate institution 4 before 2—which again occurs with probability one half—the cost will be $2C$. Hence, the expected cost, or expected loss, when Y_4 occurs is $0.5(3C) + 0.5(2C) = 2.5C$.

To determine the risk we must average the loss over the variety of conditions that are reflected in the assay test data, i.e., averaging over the distribution of $\boldsymbol{x}$. Let $F(\boldsymbol{x}|y)$ denote the conditional distribution of $\boldsymbol{x}$ given that institution y is culpable. The conditional risk of Bayes net combination c, i.e., the expected loss when y is culpable, is given by

$$R(y,c) = \int_{\mathcal{X}} L\big(y, \boldsymbol{r}_c(\boldsymbol{x})\big)dF(\boldsymbol{x}|y), \tag{12}$$

where $dF(\boldsymbol{x}|y)$ is a flexible measure theoretic notation that accommodates any combination of discrete or continuous elements of $\boldsymbol{X}$. For example, if all the elements of $\boldsymbol{X}$ were discrete, then (12) can be written as the finite sum of the loss function multiplied by the corresponding elements of the probability mass function associated with $F(\boldsymbol{x}|y)$. If all the elements of $\boldsymbol{X}$ were continuous, the integral in (12) would still remain, but $dF(\boldsymbol{x}|y)$ would be replaced with $f(\boldsymbol{x}|y)d\boldsymbol{x}$, the corresponding conditional density function and the variable of integration. In practice, we often estimate $F(\boldsymbol{x}|y)$ using the empirical cumulative distribution function calculated from the test data. For our Bayes net problem, applying this approach requires that we first define $\mathcal{V}_i = \{j : w_{ij} > 0\}$, the set of culture media that could have been obtained or used by institution i. We can then define the empirical estimate of $F(\boldsymbol{x}|y)$ by assigning probability mass to each element of the dataset. Specifically, we have

$$P\left(\boldsymbol{X} = \boldsymbol{x}_{k_j}\middle|Y=i\right) = P\left(\boldsymbol{X} = \boldsymbol{x}_{k_j}\middle|M=j, Y=i\right)P(M=j|Y=i) \approx \frac{1}{n_j} \times \frac{p_j}{\sum_{j\in\mathcal{V}_i} p_j} := \pi_{ij} \tag{13}$$

This estimate of $F(\boldsymbol{x}|y)$ assumes that within a culture medium, each observed data point $\boldsymbol{x}_{k_j}$ is equally probable and the set of $\boldsymbol{x}_{k_j}$ (for $k_j = 1, \dots, n_j$) in the dataset are a comprehensive representation of the distribution $F(\boldsymbol{x}|M=j)$. Using (13) we can estimate the conditional risk (12) as follows:

$$\hat{R}(Y=i,c) = \sum_{j\in\mathcal{V}_i} \sum_{k_j=1}^{n_j} L\left(i, \boldsymbol{r}_c\left(\boldsymbol{x}_{k_j}\right)\right)\pi_{ij} \tag{14}$$

When evaluating the various combinations of Bayes nets, it is of interest to average the conditional risk in (12) across the possible institutions, weighted by the *a priori* culpability probabilities of the institutions. This is known as the Bayes risk:

$$\rho(c) = \int_{\mathcal{Y}} R(y,c)dF(y) \tag{15}$$

where $\mathcal{Y}$ denotes the support, or range, of Y, and $F(y)$ is the prior probability measure for Y. Thus, for hypothetical Bayes net combinations A and B, if $\rho(A) < \rho(B)$, then A is preferred to B in terms of risk. To compare the 15 Bayes net combinations, we estimated the Bayes risk of Bayes net combination c by averaging (14) over the prior institutional culpability probabilities as follows:

$$\hat{\rho}(c) = \sum_{i=1}^{I} \hat{R}(Y=i, r_c)q_i \tag{16}$$

We calculated (16) for each of the 1000 bootstrapped test datasets produced from each of the 15 Bayes net combinations. Doing so requires the assumption that dependencies induced by estimating $F(\boldsymbol{x}|y)$ with overlapping (bootstrapped) test datasets are non-consequential (i.e. the estimates will behave as if they had been calculated on independent test datasets). Given the definition of the loss function in (11), equation (16) is an estimate of the expected investigation cost (EIC).

3.4.3 Calculation

In calculating the EIC, we set $C = 1$ in the loss function. Thus, the EIC may be interpreted as the expected number of institutions that would be investigated before identifying the culprit. Keep in mind that these estimates of the EIC only reflect the information available from the Bayes net prediction of the culture medium and the fairly limited sample of the textual data. Boxplots of the estimated Bayes risk (16), calculated for each of 1000 bootstrapped samples of size 330 for each of the 15 test combinations are presented in Figure 3. Corroborating the results from the fidelity metrics, combinations with lower EIC are preferred to those with higher EIC, which, for these data, are combinations 1, 3, 4, and 11. Referring to Table 1, c1 included all four assays, c3 consisted of MALDI, ESI-MS, and IRMS, c4 included SIMS, ESI-MS, and IRMS, and c11 included only ESI-MS and IRMS. It may be of interest that these four combinations always included ESI-MS and IRMS tests, which suggests that ESI-MS and IRMS are most effective in minimizing the expected cost of sequentially investigating institutions until the culprit is identified.

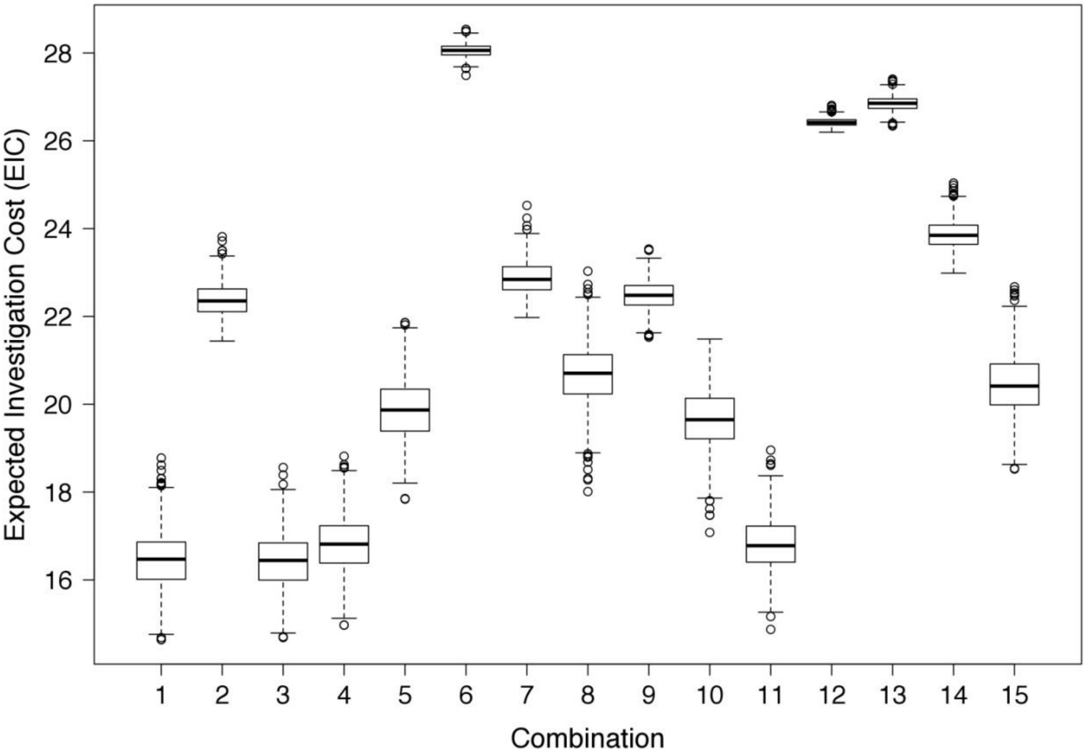

Figure 3: Boxplots of the EIC, $\hat{\rho}(c)$, of 1000 bootstrapped samples for each test combination

3.5 Applying Decision Science

Comparing signature systems on the basis of multiple attributes often leads to tradeoffs for decision makers. As discussed by Edwards et al. (2007) and others, a good place to begin is to list possible outcomes in terms of attributes of interest for each signature system and then compare the various systems (in our case, Bayes net combinations) in terms of their performance with respect to each attribute. For example, suppose some system (call it A) is better than another system (call it B) for at least one attribute, and that A is at least as good as B with respect to the remaining attributes. Then system B would be considered inferior to A and would be removed from future consideration. This process can be repeated for each system until all inferior combinations are identified. Those that remain constitute the Pareto frontier (Henderson and Quandt, 1980), also called the efficient frontier (Keeney and Raifa, 1976).

Some attributes are stochastic, (e.g. the EIC), because they are functions of data. When using stochastic attributes to compare combinations, we must approach the comparison with care to ensure that as we attempt to identify inferior solutions, the difference observed between two combinations is statistically meaningful. At this point, we avoid the formal notion of statistical significance, because traditional statistical hypothesis tests (e.g. t-test, Wilcoxon rank-sum test), which determine whether the means of two populations are distinct, require independent observations. Clearly this was not the case for our analysis of the Bayes net combinations because the bootstrapped datasets were all generated from a single test dataset. Nonetheless, we compared the distributions of stochastic attributes in an informal manner, e.g. using boxplots, and if overlap is observed or suspected in their distributions (especially overlap in the inner quartile ranges of two Bayes net combinations), then those combinations may be statistically indistinguishable in terms of that attribute. Such statistical ties may be resolved by using other attributes as a basis for comparison, which can be formally accomplished by employing a utility analysis, presented below.

In Table 6, we compare the 15 combinations on the basis of three attributes: EIC, the sample size, and the total assay cost.[6] Recognizing that it is better to have smaller risk, a smaller sample size, and a smaller cost, we can begin to pare down these fifteen combinations by removing those that are inferior to another combination. Specifically, a combination (call it A) is inferior to another combination (call it B) if 1) all of the attributes of A are equally or less preferable than the corresponding attributes of B, and 2) for at least one attribute, A is strictly less preferable than B.

Table 6: Multiattribute assessment of each assay combination. The estimated Bayes risk, $\hat{\boldsymbol{\rho}}(\boldsymbol{r_c})$, is averaged over the 1000 bootstrap samples. Non-inferior combinations are highlighted with enclosing rectangles.

Combination	EIC (Number of Institutions)	Sample Size (mg)	Assay Cost ($)
c1	16.46	1.41	720
c2	22.37	1.11	620
c3	16.43	1.31	520
c4	16.83	1.4	550
c5	19.89	0.41	470
c6	28.06	0.11	370
c7	22.88	1.1	450
c8	20.67	0.4	300
c9	22.48	1.01	420
c10	19.67	0.31	270
c11	16.82	1.3	350
c12	26.42	0.1	200

[6] We did not include a measure of fidelity in the analysis because they are closely related to risk. However, we could have used the log score instead of the Bayes risk as one of the attributes.

Combination	EIC (Number of Institutions)	Sample Size (mg)	Assay Cost ($)
c13	26.85	0.01	170
c14	23.87	1	250
c15	20.45	0.3	100

Using Figure 3, we can informally identify combinations that are statistically indistinguishable in terms of their EIC if the box (defined by the inner quartile range) of one combination overlaps the median (defined by the solid line inside the box) of another combination. We will consider the following pairs enclosed in parentheses as having the same EIC: (c1, c3), (c4, c11), (c10, c5), (c8, c15), and (c2, c9). Note that c1 has a larger sample size and cost than c3, while the EIC is statistically equivalent. Thus, c1 is inferior to c3. The same is true for c4 versus c11. Following this process of elimination, the inferior combinations can be identified as follows:

- c1 is inferior to c3,
- c4 is inferior to c11,
- c5 is inferior to c10,
- c8, c9, and c14 are inferior to c15,
- c2, c7 are inferior to c9, and
- c6 is inferior to c13.

This leaves c3, c10, c11, c12, c13, and c15 as the combinations that comprise the Pareto frontier with respect to risk, sample size, and cost. The frontier, which has the appearance of a convex hull, is illustrated by the red points in Figure 4.

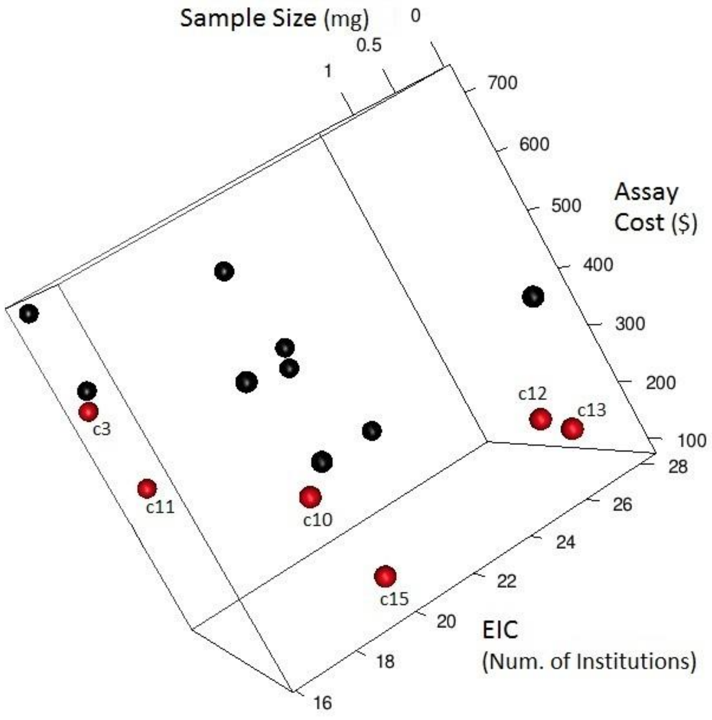

Figure 4: Three-dimensional scatter plot of attributes listed in Table 6. The red points indicate the Pareto frontier.

The only way to decide between combinations on the frontier is to determine the value we place on each of the attributes. Which of these matters the most: EIC, sample size, or assay cost? Which matters least? And by how much? A multiattribute decision analysis (Keeney & Raiffa, 1976, Edwards et al. 2007) addresses these questions, and ultimately, helped us identify the preferred Bayes net combination(s). This involves constructing a single attribute utility function for each of the three attributes, and then aggregating them together in an additive or multiplicative fashion, as we discuss next.

3.5.1 Utility Analysis

Following the approach of Keeney (1972), we construct a multiattribute utility function of three attributes: 1) the investigation cost incurred to identify the culprit, 2) sample size, and 3) assay cost. The expectation of the multiattribute utility function will serve to compare all fifteen combinations, and in particular, the set of combinations that are on the Pareto frontier. Hence, the multiattribute utility function will replace the loss function in (12), (14), and by extension, (16).

The first step is to elicit a single attribute utility function, $u_m: [z', z''] \rightarrow [0,1]$, for each attribute m that reflects the utility we ascribe to the attribute for a given value of z. While u_m is usually monotonic, it does not have to be. An increasing u_m is used when larger values of the attribute are preferred to lower values. A decreasing u_m is used when smaller values of the attribute are preferred to larger ones. Thus, larger utility is always preferred to lower utility. A linear u_m represents a constant rate of change in the utility as the value of the attribute, z, increases. Exponential (or other curved shapes) model an increasing or decreasing rate of change in the utility, corresponding to convex and concave shapes, respectively. Nonlinear utility functions are appropriate when the change in utility resulting from a unit increase in z depends on the value of z. For example, if increasing z from a to $a + \Delta$ results in a greater (or smaller) change in utility than we obtain by increasing z from $b - \Delta$ to b, the utility function would be nonlinear. A variety of techniques exist for helping investigators construct single attribute utility functions that best suit their values and preferences (Berger 1985, Keeney & Raifa 1976 , Edwards et al., 2007). These often include using a gamble (or lottery) heuristic (Berger 1985).

Let $m = 1, \ldots, M$ index the attributes we will use to compare two or more signature systems, let z_m denote the value of attribute m, and let u_m denote the corresponding single attribute utility function. Also, let $\boldsymbol{z} = (z_1, z_2, \ldots, z_M)^T$ represent the vector of attributes measured for a particular signature system. Under reasonable independence conditions (Keeney & Raiffa 1976), a common choice for the multiattribute utility function is given by

$$u(\boldsymbol{z}) = \sum_{m=1}^{M} \alpha_m u_m(z_m) \tag{17}$$

or

$$u(\boldsymbol{z}) = \frac{1}{\beta}\left(\prod_{m=1}^{M}\left(1 + \beta\alpha_m u_m(z_m)\right) - 1\right) \tag{18}$$

where $0 < \alpha_m < 1$ for all m, $\beta > -1$, and $\beta \neq 0$. For both (17) and (18), $u(\boldsymbol{z})$ takes values in the unit interval, provided each u_m maps to the unit interval. For attributes $\boldsymbol{z}_A$ observed on system A and $\boldsymbol{z}_B$ observed on system B, system A is preferred to B if $u(\boldsymbol{z}_A) > u(\boldsymbol{z}_B)$. If any of the attributes are stochastic, then system A is preferred to B if $E\left(u(\boldsymbol{z}_A)\right) > E\left(u(\boldsymbol{z}_B)\right)$, where the expectation is taken over the joint distribution of $\boldsymbol{z}$.

The additive utility function[7] in (17) is appropriate under the assumption of additive independence and requires $\sum_{m=1}^{M} \alpha_m = 1$. There are a number of approaches for determining whether additive independence is a valid assumption (Keeney & Raifa 1976, Delquié 1997) and for choosing the attributes such that additive independence is preserved (Edwards, et al 2007). Likewise, there are several approaches for choosing the weights, α_m, that reflect the value investigators place on each attribute (Keeney 1972, Ma et al. 1999, Wang & Parkan 2005, Wang & Luo 2010, and Goicoechea et al. 1982, Section 4.3.3). While the form of the multiplicative utility model (18) is more complex than (17), it does not require the specification of additional parameters beyond those required by the additive form, because the value of β can be obtained from the α_m. When additive independence is not satisfied, investigators should use the more appropriate multiplicative utility model. The value of β can be obtained by numerically solving the following equation using an iterative root-finding algorithm:

$$1 + \beta = \prod_{m=1}^{M} (1 + \beta\alpha_m) \tag{19}$$

If $\sum_{m=1}^{M} \alpha_m > 1$, the solution will lie in the open interval $(-1,0)$. If $\sum_{m=1}^{M} \alpha_m < 1$, the solution will be greater than 0. However, if $\sum_{m=1}^{M} \alpha_m = 1$, the additive model (17) should be used.

3.5.1.1 Determining the Single Attribute Utility Functions

We now demonstrate the utility analysis for the Bayes net combinations. To begin, we must identify the single attribute utility functions for investigation cost, sample size, and assay cost. While we refer to these single attribute utility functions as the "utility of investigation cost" or the "utility of the sample size," it is important to clarify that these single attribute utility functions actually represent the utility (or benefit) we ascribe *to a Bayes net combination* that has a specified value of the attribute. For example, the single attribute utility function for sample size provides the utility we ascribe *to the Bayes net combinations* that have a given sample size. We begin by presenting a generalized version of a single attribute exponential utility function. We will then apply it to our three attributes of interest.

For simplicity, we temporarily omit the subscript, m, that distinguishes one attribute and another. As before, let z denote the value of single attribute, where $z' \leq z \leq z''$. Here, z' could be the worst, least preferable value of z, and that z'' is the best, or most preferable. Alternatively, z'' could be the worst and z' the best. Either way works, so long as these endpoints represent the two preferential extremes over the domain of z. Let $z^* = (z - z') / (z'' - z')$ denote a linear normalization of z such that $0 \leq z^* \leq 1$. The single attribute exponential utility function is given by

$$\begin{aligned} u(z) &= u' + (u'' - u')\left(\frac{exp(\gamma\theta z^*) - 1}{exp(\gamma\theta) - 1}\right), & \theta \neq 0 \\ u(z) &= u' + (u'' - u')z^*, & \theta = 0 \end{aligned} \tag{20}$$

where $u' = u(z')$, $u'' = u(z'')$, $\gamma = \text{sign}(u'' - u')$, and θ is a shape parameter that governs the extent of the convexity or concavity of $u(z)$. The interval $[\min(u', u''), \max(u', u'')]$ is the range of $u(z)$. The first equation in (20) is a parametric curve that can be increasing, decreasing, convex, or concave. The second equation is linear, and it can also be increasing or decreasing.

The utility function is increasing in z if $u' < u''$ and decreasing if $u' > u''$. The three parameters, u', u'', and θ, must be elicited from the investigator. The extreme values of $u(z)$, u' and u'', are relatively straightforward to estimate as they indicate the minimal or maximal values of the utility function. Typically, they will be 0 and 1, or 1 and 0. For $\theta > 0$, $u(z)$ is convex, and the larger the value

[7] The additive utility model is a special case of the multiplicative model. Consequently, (18) reduces to (17) when $\sum_{m=1}^{M} \alpha_m = 1$.

of θ, the more convex the function becomes. Likewise, $\theta < 0$ results in $u(z)$ being concave. For $\theta \approx 0$, $u(z)$ is nearly linear. Not surprisingly, it is straightforward to show using l'Hôpital's rule that the convex or concave exponential utility function converges to the linear function as θ becomes small:

$$\lim_{\theta \to 0} \left\{ u' + (u'' - u') \left(\frac{exp(\gamma\theta z^*) - 1}{exp(\gamma\theta) - 1} \right) \right\} = u' + (u'' - u')z^*. \tag{21}$$

The parameter θ is arguably the most difficult to determine. It is chosen to reflect the investigator's preference in how the utility changes as z changes throughout its domain. Once u' and u'' are specified, a single point $(z_0,\ u(z_0))$ can be used to solve for the value of θ. A technique motivated by utility theory that is often employed (Keeney & Raiffa, 1976) to elicit the value of θ involves an investigator determining a *certainty equivalent* for the attribute. To illustrate the concept of a certainty equivalent, consider an example where you are offered 1) either a guaranteed amount of money, $\$g$, where $0 \leq g \leq 100$, or 2) a lottery with a 50% chance of winning \$100. How large or small would g have to be for you to be *just* willing to risk losing $\$g$ for a 50% chance to win \$100? Alternatively, how much would you be willing to pay in order to play the lottery?

Speaking now more generally, an estimate of the certainty equivalent, g_z, can be found by identifying the value of g_z such that an investigator is indifferent between having g_z for certain versus a 50% / 50% lottery of having z' or z''. Once the certainty equivalent is specified, the value of θ can be obtained as follows. To simplify the following algorithm, denote the midpoints of the domain and range of the utility function by $\tilde{z} = \frac{z' + z''}{2}$ and $\tilde{u} = \frac{u' + u''}{2}$, respectively.

1. If $g_z = \tilde{z}$, the utility function is linear, and $\theta = 0$.
2. If $g_z > \tilde{z}$ and $\gamma = 1$, or if $g_z < \tilde{z}$ and $\gamma = -1$, the solution for θ will be positive. Set the first equation in (20) equal to $\tilde{u}$ and solve for θ numerically over the range of $(0,\ \theta^*)$, for suitably large $\theta^* > 0$.
3. If $g_z < \tilde{z}$ and $\gamma = 1$, or if $g_z > \tilde{z}$ and $\gamma = -1$, the solution for θ will be negative. Set the first equation in (20) equal to $\tilde{u}$ and solve for θ numerically over the range of $(-\theta^*,\ 0)$, for suitably large $\theta^* > 0$.

While the simplicity of the exponential utility function is appealing, it also has its limitations. Other functional forms for single attribute utility functions are possible, and there are more comprehensive methods for eliciting utility functions. For example, multiple points in the utility space may be elicited from a decision maker and the curve that best fits the points may be used as the utility function. See Keeney & Raiffa (1976) and Berger (1985) for more discussion and examples.

We will construct a single attribute utility function for each of the three attributes of interest, investigation cost, sample size, and assay cost, which we will distinguish using the subscripts $1, 2,$ and 3, respectively. For example, we write the utility function for investigation cost as $u_1(z_1)$. The parameter values of the utility functions reflect our preferences—but we do not assert that these values are best. Furthermore, the preferences of other decision makers would likely lead to different parameter choices. The parameter values we chose for the three single attribute utility functions are shown Table 7, with corresponding graphs shown Figure 5.

For each attribute, we set z' equal to the smallest value of the attribute observed in the original test dataset, and z'' equal to the largest value of the attribute in the original test dataset. For all three attributes, we chose decreasing utility functions (evidenced by $u' > u''$), because low investigation cost, small sample size, and low assay cost correspond to high utility. The admittedly unusual values of u_1' and u_1'' where chosen to ensure that the range of the expected utility of investigation cost, $E(u_1(z_1))$, was [0,1]. This is discussed in more detail in Section 3.5.1.3. Using the lottery method discussed above, we chose values of the certainty equivalents for each attribute that seemed reasonable and solved for the corresponding value of θ.

Table 7: Parameter values for single attribute utility functions

Attribute	Subscript	z'	z''	u'	u''	g	θ
Investigation cost	1	1	83.5	2.503	-2.916	25	1.906
Sample size	2	0.01	1.41	1	0	1	-1.884
Assay cost	3	100	720	1	0	441	0

We elected to use a convex nonlinear utility function for investigation cost to reflect the fact that an increase in investigation cost from 1 to 20 is less preferred than an increase from 60 to 80. This emphasizes the value of having a low investigation cost, which results when the correct, culpable, institution is investigated early on. In other words, we would be willing to risk incurring an additional investigation cost of $(83.5 - 25 = 58.5 \text{ institutions})$ in order to have a 50% chance of correctly picking the institution at the beginning of the investigation and thus have an investigation cost of one institution. This type of risk preference is an example of what is commonly known as *risk seeking*, or risk prone (Keeney & Raiffa, 1976).

We used a concave nonlinear utility function for sample size to reflect the fact that a change from, say, 0.1 mg to 0.5 mg would be preferred over a change from 1 mg to 1.5 mg, because the sample is a finite resource. In other words, when the sample is nearly consumed by various test assays, it becomes increasingly valuable. Thus, Bayes net combinations that consume more of the sample have less utility than those which consume less, and the rate of decrease in utility for combinations with higher consumption is more pronounced than for combinations with lower consumption. This concave utility function reflects a *risk averse* preference (Keeney & Raifa, 1976).

We assumed a *risk neutral*, linear utility function for assay cost, which seemed reasonable because the assay cost ranges over a relatively small interval (from \$100 to \$720).

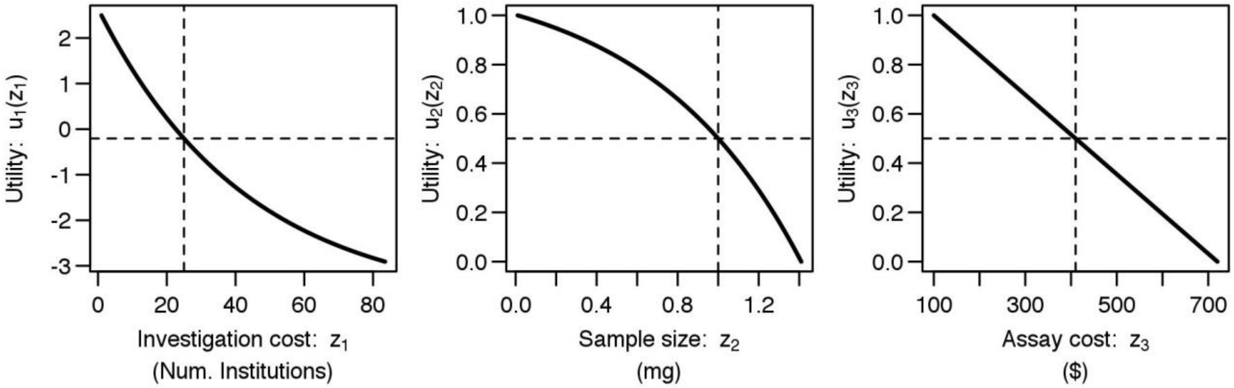

Figure 5: Single attribute utility functions for investigation cost, sample size, and assay cost. Dotted lines indicate the certainty equivalents and their corresponding utility of 0.5.

3.5.1.2 Determining the Weights

Having constructed the utility single attribute utility functions, we can now determine the relative value weights, α_1, α_2, and α_3, for risk, sample size, and cost, respectively. We use the methodology discussed by Keeney (1974).

We begin by considering five extreme, hypothetical outcomes that could arise from various Bayes net combinations. These five cases, labeled M', L', C', S', and A', are presented in Table 8. Note these five cases are not the same as the 15 combinations, c1-c15, discussed earlier. Specifically, M' and L' represent the two hypothetical Bayes net combinations that would take on the most desirable or least desirable values of each attribute.[8] This is important because it ensures that the utility of the most desirable hypothetical combination, M', is 1, i.e. $u(z_1 = 1, z_2 = 0.01, z_3 = 100) = 1$, and the utility of the least desirable hypothetical combination, L', is 0, i.e. $u(z_1 = 83.5, z_2 = 1.41, z_3 = 720) = 0$.

To determine α_1, we must determine the probability, p_1, such that we are indifferent between 1) having C' for certain versus 2) a gamble of having M' with probability p_1, or having L' with probability $1 - p_1$. Another way to express it: Given we have C' for certain, we must determine the smallest probability p_1 such that we would accept a $p_1 \times 100\%$ chance of exchanging C' for M' (and gain the lowest possible sample size and cost) versus the $(1 - p_1) \times 100\%$ chance of exchanging C' for L' (and raise the investigation cost to its worst possible value). The value of α_1 is then set equal to p_1. Because the investigation cost is our principal measure of the performance of the Bayes net in predicting the culpable institution, it is far more important than sample size or assay cost, and consequently we chose $\alpha_1 = p_1 = 0.90$.

Table 8: Hypothetical cases for assessing the relative value weights

Combination description		Label	Risk (Number of Institutions)	Sample Size (mg)	Cost ($)
Most extreme cases	Most desirable	M'	1	0.01	100
	Least desirable	L'	83.5	1.41	720
Hypothetical cases where one attribute is the best and the others are the worst	Invest. cost best	C'	1	1.41	720
	Sample size best	S'	83.5	0.01	720
	Assay cost best	A'	83.5	1.41	100

Employing a similar procedure for α_2, suppose we have S' for certain. Because sample size is important (but not as much as investigation cost), we would be willing to accept as small as a 20% chance of obtaining the lowest possible investigation cost and assay cost (i.e. exchanging S' for M') versus an 80% chance of raising the sample size to its worst possible value, 1.41 mg (i.e. exchanging S' for L'). Thus, $\alpha_2 = 0.20$. Finally, suppose we have A' for certain. Because assay cost is the least important attribute, we would be willing to accept as little as a 5% chance of obtaining the lowest possible risk and sample size (i.e. exchanging A' for M') versus an 95% chance of raising the assay cost to its worst possible value, $720 (i.e. exchanging A' for L'). Thus, $\alpha_3 = 0.05$.

In this instance, because $\alpha_1 + \alpha_2 + \alpha_3 = 0.90 + 0.20 + 0.05 = 1.15 \neq 1$, the multiplicative form of the utility function, (18), is appropriate. Using these values of α_1, α_2, and α_3 defined previously, solving (19) gives $\beta \approx -0.6547$. Having now determined each of the components of the multiplicative utility function (u_1, u_2, u_3, α_1, α_2, α_3, and β) we can calculate the utility for each Bayes net combination.

[8] The best and worst values of investigation cost, sample size, and assay cost in Table 8 correspond to rounded values of either z' or z'' from Table 7.

3.5.1.3 The completed multi attribute utility function

At this point, it is useful write the complete multiplicative utility function and its expected value. Beginning with the single attribute utility functions for sample size and assay cost, inserting the values of Table 7 into (20) and gives:

$$u_2(z_2^c) = 1 - \left(\exp\left[1.88\left(\frac{z_2^c - 0.01}{1.41 - 0.01}\right)\right] - 1\right)/(\exp[1.88] - 1) \tag{22}$$

$$u_3(z_3^c) = 1 - (z_3^c - 100)/(720 - 100) \tag{23}$$

where z_2^c denotes the sample size and z_3^c denotes the assay cost for Bayes net combination c. The utility function for investigation cost is a function of the true institution, y and the assay data, $\boldsymbol{x}$:

$$u_1\big(y, \boldsymbol{r}_c(\boldsymbol{x})\big) = 2.503 - 5.416\left(\exp\left[-1.91\left(\frac{r_c^{[y]}(\boldsymbol{x}) - 1}{83.5 - 1}\right)\right] - 1\right)/(\exp[-1.91] - 1) \tag{24}$$

where $\boldsymbol{r}_c(\boldsymbol{x})$, the ranked posterior probabilities for combination c, is defined in Section 3.4.2, and $r_c^{[y]}(\boldsymbol{x})$ is defined in (11). The multiplicative utility function (18) of the three attributes can be expressed as follows:

$$\begin{aligned} &u(y, \boldsymbol{r}_c(\boldsymbol{x}), z_2^c, z_3^c) = \\ &\quad \alpha_1 u_1\big(y, \boldsymbol{r}_c(\boldsymbol{x})\big) + \alpha_2 u_2(z_2^c) + \alpha_3 u_3(z_3^c) \\ &\quad + \beta\left(\alpha_1\alpha_2 u_1\big(y, \boldsymbol{r}_c(\boldsymbol{x})\big)u_2(z_2^c) + \alpha_1\alpha_3 u_1\big(y, \boldsymbol{r}_c(\boldsymbol{x})\big)u_3(z_3^c) + \alpha_2\alpha_3 u_2(z_2^c)u_3(z_3^c)\right) \\ &\quad + \beta^2\alpha_1\alpha_2\alpha_3 u_1\big(y, \boldsymbol{r}_c(\boldsymbol{x})\big)u_2(z_2^c)u_3(z_3^c) \end{aligned} \tag{25}$$

where $\alpha_1 = 0.90,\ \alpha_2 = 0.20,\ \alpha_3 = 0.05,$ and $\beta = -0.6547$, as discussed previously.

The expected utility is given by integrating (25) with respect to the joint probability distribution of $y, \boldsymbol{x}, z_2,$ and z_3. The expected utility, $E[u(y, \boldsymbol{r}_c(\boldsymbol{x}), z_2, z_3)]$, is analogous to the Bayes risk (15), except the utility function is used instead of the loss function. Because the sample size and assay cost are not stochastic, their distributions are degenerate. Thus, the overall joint distribution reduces simply to $F(\boldsymbol{x}, y) = F(\boldsymbol{x}|y)F(y)$. Because the expected value of a linear combination is the linear combination of the expected values, we have

$$\begin{aligned} &E[u(y, \boldsymbol{r}_c(\boldsymbol{x}), z_2^c, z_3^c)] = \\ &\quad \alpha_1 E\big[u_1\big(y, \boldsymbol{r}_c(\boldsymbol{x})\big)\big] + \alpha_2 u_2(z_2^c) + \alpha_3 u_3(z_3^c) \\ &\quad + \beta\left(\alpha_1\alpha_2 u_2(z_2^c)E[u_1\big(y, \boldsymbol{r}_c(\boldsymbol{x})\big)] + \alpha_1\alpha_3 u_3(z_3^c)E[u_1\big(y, \boldsymbol{r}_c(\boldsymbol{x})\big)] + \alpha_2\alpha_3 u_2(z_2^c)u_3(z_3^c)\right) \\ &\quad + \beta^2\alpha_1\alpha_2\alpha_3 u_2(z_2^c)u_3(z_3^c)E[u_1\big(y, \boldsymbol{r}_c(\boldsymbol{x})\big)] \end{aligned} \tag{26}$$

where the expectations are with respect to $F(\boldsymbol{x}, y)$. We calculate $E\big[u_1\big(y, \boldsymbol{r}_c(\boldsymbol{x})\big)\big]$ using (16), except the loss function L is replaced by u_1 as defined in (24).

By design, the range of u_2 and u_3 is the unit interval, [0,1], i.e., the least preferable values of the attribute are assigned a single attribute utility score of 0, while the most preferable are assigned a score of 1. However, based on the data in the test datasets, we chose the range of u_1 to be $[-2.916,\ 2.503]$ so that $E\big[u_1\big(y, \boldsymbol{r}_c(\boldsymbol{x})\big)\big]$ would range from 0 to 1, thereby making it comparable to the ranges of u_2 and u_3. This is essential in order for the α parameters to weight the attributes in the way that reflects our preferences. Had we chosen $u_1' = 1$ and $u_1'' = 0,\ E\big[u_1\big(y, \boldsymbol{r}_c(\boldsymbol{x})\big)\big]$ would have ranged from 0.538 to 0.723, effectively diminishing the influence of investigation cost in discriminating between the Bayes net combinations—precisely the opposite of what we intended to achieve by assigning $\alpha_1 = 0.90$. The need to set $z_1' =\ 2.503$ and $z_1'' = -2.916$ occurred because our three attributes are a mix of stochastic (z_1) and non-stochastic (z_2 and z_3) attributes.

As we did with the EIC in Figure 3, we calculated the expected utility (26) for each of the 1000 bootstrap samples in order to reflect the variability inherent in the data. The results are shown in Figure 6.

3.5.2 Comparing the Bayes Net Combinations

The expected multi attribute utility function (26) is a metric we may use to compare the combinations on the Pareto frontier (c3, c10, c11, c12, c13, and c15). Referring to Figure 6, we can clearly rule out c10, c12, c13, and c15. While combinations c3 and c11 have nearly overlapping utility distributions, c3 has the highest overall expected utility, reflecting our preferences for the attributes via the weights $(\alpha_1, \alpha_2, \alpha_3)$ and the shapes of the single utility functions (u_1, u_2, u_3).

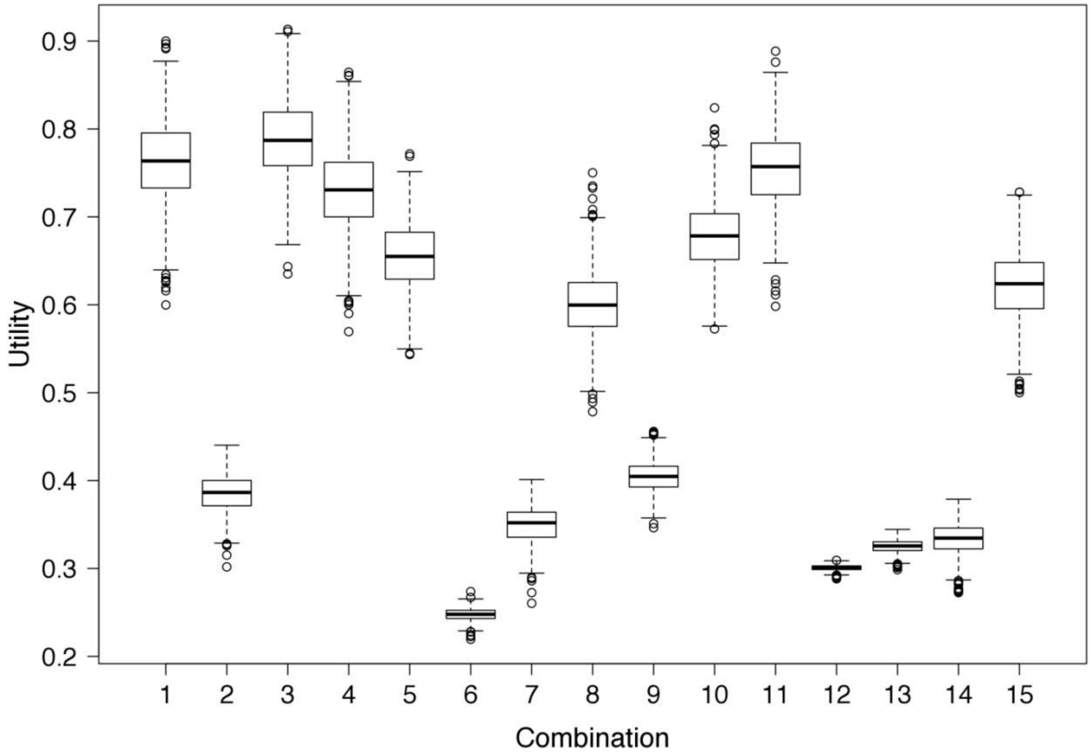

Figure 6: Boxplots of the expected multiplicative utility (26) of 1000 bootstrapped samples for each test combination

To further investigate the influence of the four assays on the Bayes net, note that c1, c3, c4, and c11 all have statistically similar utility. Referring back to Table 1, we see that c1 contains all the assays, c3 excludes SIMS, c4 excludes MALDI, and c11 excludes both SIMS and MALDI. This suggests that SIMS and MALDI are the least informative of the tests, while ESI-MS and IRMS are the most valuable. For the combinations where ESI-MS or IRMS were not performed (this includes all combinations except c1, c3, c4, and c11), the utility was considerably lower. This suggests the C/N isotope ratio and the presence/absence of Agar result in the greatest utility, while tests for heme and metals (MALDI and SIMS) are the least useful. However, it is likely that the primary reason why SIMS and MALDI appear to be less useful arises from the fact that the MALDI test for heme is indicative of only one culture medium, blood agar (BA), and the SIMS test for metals is only indicative of two culture media, glucose agar (GA) and glucose broth (GB) (refer to Table 2). It may be that these tests perform quite well in identifying these particular culture media—but unless these media have a very high prevalence within the literature, the performance of SIMS and MALDI will not easily be reflected in the aggregate measures of EIC and utility that we used, which also include the performance of the other eight culture media. Table 4 indicates that BA, GA, and GB are rather rare, and so their influence on the EIC and utility calculations would be minimal. Likewise, the fidelity metrics were aggregated across all culture media. Consequently, excluding SIMS, MALDI, or both from the Bayes net did not severely impact fidelity of the Bayes net, which likely explains why c1, c3, c4, and c11 perform well in Figure 2.

4.0 Conclusion

We demonstrated an approach for assessing the quality of a signature system designed to predict the culture medium used to grow *B. anthracis*. The system was comprised of four chemical assays designed to 1) identify the presence of heme, 2) identify the presence of agar, 3) categorize the culture media food content via analysis of the isotopic ratios of C and N, and 4) identify the presence of Cu^{2+} or Zn^{2+}. The analytical results from any combination of these four assays can be used in a Bayesian network to predict the probabilities that the spores were grown using one of eleven culture media. We evaluated combinations of the signature system by removing one or more of the assays from the Bayes net. We measured and compared the quality of the various Bayes nets in terms of fidelity, cost, risk, and utility, a method we refer to as Signature Quality Metrics (SQM).

Because the primary output of the Bayes net was a vector of posterior probability estimates of the culture medium, we measured fidelity using two proper scoring rules, the logarithmic score and the Brier score. We also considered respective decompositions of those scores, the information and sharpness, which measure the strength of the probability predictions of the Bayes net. We obtained estimates of the cost for commercial laboratories to conduct each of the four assays under consideration. Combining the results of the Bayes net with information obtained from the scientific literature, we developed a risk model to calculate the expected number institutions that would have to be investigated before discovering the culpable party. Another important attribute in this problem was the sample size, i.e., the amount of spores consumed by a particular combination of assays.

We illustrated a multiattribute assessment of the fifteen Bayes net combinations with a utility analysis that reflected the value we placed on each of the attributes of concern: investigation cost, sample size, and assay cost. We accounted for variability in the estimates of fidelity, risk, and utility by calculating these quantities repeatedly on bootstrapped samples of the test data. We demonstrated that the assays which detect heme and metals (Cu^{2+}, Zn^{2+}) were the least useful. That is, removing one or both of these assays from the Bayes net did not appreciably diminish the quality of the Bayes net in terms of fidelity, risk, nor utility. However, both these assays are each only indicative of a small number of culture media (blood agar for the heme assay, glucose agar or glucose broth for the metals assay). Furthermore, based on the literature sample, blood agar, glucose agar, and glucose broth are not likely to be used often. Thus, it is likely that while assays for heme and metals may be quite proficient in correctly predicting blood agar or glucose agar and glucose broth, their overall utility is diminished because they are of no value in predicting the remaining culture media which are more prevalent. The question of whether the heme and metals assays are of merit could be addressed by dividing the culture media into two groups and repeating a similar SQM analysis for 1) blood agar, glucose agar and glucose broth, and 2) for the remaining eight culture media.

As is the case for most mathematical models, our analysis required various assumptions—especially for the calculation of risk and utility. One of the most important assumptions we employed for the risk and utility analysis was that the natural prevalence of the culture media was well-estimated by the extent to which they were mentioned in a sample of the scientific literature. We also assumed that the *a priori* likelihoods of institutional culpability (where culpability occurs if investigation of an institution eventually leads to correct identification of the culprit) were well-estimated by the amount of relevant literature produced by the institution. However, there are likely numerous other factors beyond the extent of their publications that may influence the *a priori* probability that an institution is culpable. We also assumed that each institution, regardless of size or location, was equally costly to investigate and that if a culpable institution were investigated, the perpetrator would be identified.

The data set used to develop (i.e. train) the Bayes net was gathered from a variety of sources in an opportunistic fashion. This somewhat limits our ability to trust the data because it was not obtained from a single study whose purpose was to demonstrate the feasibility of the Bayes net. Likewise, since the data used to evaluate the Bayes net combinations were derived from the training data, the performance of the Bayes net combinations is likely overstated. However, the drawbacks of these limitations and assumptions are somewhat mitigated because the principal objective of our work was to make relative comparisons between the fifteen Bayes net combinations, as opposed to making absolute assessments of signature quality. Because each of the Bayes net combinations are likely to be influenced by these limitations in a similar way, relative comparisons may still be meaningful.

The arguably subjective choices for the components of the utility analysis (i.e., the forms of the single attribute utility functions and the attribute weights) may also be a concern. However, a sensitivity analysis to explore the impact of those parameters could be conducted to determine the extent to which those parameters influence the comparisons.

5.0 References

Atlas, R.M. (2010) *Handbook of Microbiological Media*, 4th edition, American Society for Microbiology Press, Washington, DC and the CRC Press, Boca Raton, FL.

Berger, J.O. (1985). *Statistical Decision Theory and Bayesian Analysis*, 2nd edition, Springer.

Bickel, J.E. (2007). Some Comparisons among quadratic, spherical, and logarithmic scoring rules. *Decision Analysis*. 4:49-65.

Bickel, J.E. (2010). Scoring rules and decision analysis education. *Decision Analysis* 7(4): 346-357.

Cliff, J.B., K.H. Jarman, N.B. Valentine, S.L. Golledge, D.J. Gaspar, D.S. Wunschel, and K.L. Wahl (2005). Differentiation of Spores of *Bacillus subtilis* Grown in Different Media by Elemental Characterization Using Time-of-Flight Secondary Ion Mass Spectrometry. *Applied and Environmental Microbiology*. 71(11):6524-6530.

Delquié, P. and M. Luo (1997). A simple trade-off condition for additive multiattribute utility. *Journal of Multi-Criteria Decision Analysis*. 6(5):248-252.

Edwards W, Miles RF Jr., von Winterfeldt D, eds. (2007). *Advances in Decision Analysis: From Foundations to Applications*, Cambridge University Press.

Hand, D. J. (1997). *Construction and Assessment of Classification Rules.* John Wiley & Sons.

Henderson, J.M. and R.E. Quandt (1980). *Microeconomic Theory: A Mathematical Approach*. McGraw-Hill Book Company. New York, New York.

Jarman, K. H., H. W. Kreuzer-Martin, et al. (2008). Bayesian-integrated microbial forensics. *Applied Environmental Microbiology* 74(11): 3573-82.

Keeney, R.L. (1974) Multiplicative utility functions. *Operations Research*. 22(1):22-34.

Keeney R.L. and Raiffa H. (1976) *Decisions with Multiple Objectives: Preferences and Value Tradeoffs.* John Wiley & Sons, Inc.

Kohavi, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence (IJCAI),* pages 1137–1143, San Mateo, CA, 1995. Morgan Kaufmann.

Kreuzer-Martin, H. W. and K. H. Jarman (2007). Stable isotope ratios and forensic analysis of microorganisms. *Applied Environmental Microbiology* 73(12): 3896-908.

Ma J, Fan Z, Huang L. (1999). A subjective and objective integrated approach to determine attribute weights. *European Journal of Operational Research*, 112:397—404.

Murphy, A.H. (1973). A new vector partition of the probability score. *Journal of Applied Meteorology*, 12:595-600.

Rousseeuw, P.J., I. Ruts, and J.W. Tukey (1999). The bagplot: a bivariate boxplot. *The American Statistician* 53(4): 382-387.

Toda , M. (1963). Measurement of subjective probability distributions. Report ESD-TDR-63-407, Decision Sciences Laboratory, Electronic Systems Division, Air Force Systems Command, United States Air Force, L.G. Hanscom Field, Bedford, MA.

Wang YM, Parkan C. (2005). Multiple attribute decision making based on fuzzy preference information on alternatives: Ranking and weighting. *Fuzzy Sets and Systems*, 153:331-346.

Wang YM, Luo Y. (2010). Integration of correlations with standard deviations for determining attribute weights in multiple attribute decision making. *Mathematical and Computer Modeling*, 51:1-12.

Winkler, R., J. Muñoz, J. Cervera, J. Bernardo, G. Blattenberger, J. Kadane, D. Lindley, A. Murphy, R. Oliver, and D. Ríos-Insua, (1996). Scoring rules and the evaluation of probabilities. *TEST: An Official Journal of the Spanish Society of Statistics and Operations Research*, 5(1):1-60.

Webb-Robertson, BJM, C Corley, LA McCue, K Wahl and H Kreuzer. (2012). Fusion of laboratory and textual data for investigative bioforensics. *Forensic Sciences International*. In press.