Statistics and Its Interface Volume 4 (2011) 27–36

Bayesian decision analysis for choosing between diagnostic/prognostic prediction procedures*

John Kornak† and Ying Lu

New diagnostic procedures and prognostic markers are contnually being developed for a wide range of medical complaints. Medical institutions are therefore regularly faced with the decision as to whether to replace an existing procedure with a new one. The decision to adopt a new method is primarily based on diagnostic/predictive accuracy and cost-effectiveness, but this trade-off is not usually considered in a formal decision-theoretic way. The decision process for diagnostic procedures is complicated by the fact that diagnostic decisions are typically based on thresholding one or more continuous variables. Therefore, a formal decision process should account for uncertainty in the optimal threshold value for each diagnostic procedure. We here propose a Bayesian decision approach based on maximizing expected utility (incorporating accuracy and costs) with respect to diagnostic procedure and threshold level simultaneously. The Bayesian decision approach is illustrated via an application comparing the utility of different bone mineral density (BMD) measurements for determining the need for preventative treatment of osteoporotic hip fracture in elderly patients.

Keywords and phrases: Bayesian decision analysis, Decision theory, Diagnostic methods.

1. INTRODUCTION

Diagnostic technologies evolve rapidly, forcing medical institutions, insurance companies, policy makers, and clinicians to make difficult decisions as to whether and how to incorporate the new diagnostic procedures. These new procedures can be used to determine patient disease status (diagnosis) or to predict adverse outcomes (prognosis). In either situation, treatment often follows a positive diagnosis in attempt to prevent, delay, or ameliorate more costly outcomes.

Often a selection of different diagnostic p rocedures is available, but when choosing which diagnostic procedure to employ, cost-benefit considerations are typically made outside a formal decision-theoretic framework (e.g. based on informal judgments of required sensitivity and specificity). An additional level of complexity in the decision process occurs because most diagnostic procedures do not directly produce a definitive diagnosis. Instead, some more or less arbitrary rule uses a pre-defined threshold value to convert an underlying continuous measurement into a categorical (usually binary) diagnostic decision. These threshold values, which are typically based on trading off sensitivity and specificity or other more *ad hoc* mechanisms, can vary for different target healthcare populations.

We propose here to incorporate threshold optimization directly into the decision process so that the decision space is extended to optimize over diagnostic procedure–threshold combinations. The integration of optimal threshold level estimation into the diagnostic/prognostic procedure decision process constitutes the primary methodological development of this paper.

Decisions pertaining to the choice of diagnostic procedure depend on the perspective of the decision-maker(s). From the institutional perspective that we focus on, a hospital department often must decide whether to adopt a new diagnostic procedure, continue with an existing one, or possibly to employ two or more methods side-by-side. In this paper we concentrate on the process of choosing an optimal diagnostic procedure in the case where the institution is required to make a decision between two (or more) diagnostic procedures for hospital/departmental-level implementation.

Bayesian decision analysis [28] has received considerable attention in medical statistics. An extensive literature addresses the application of Bayesian decision analysis in clinical trials where the decision to accept a new treatment over an old one depends directly on the cost-benefit trade-off [50, 17, 41, 42, 38–40, 53, 37, 51]. Other medically related areas where Bayesian decision analysis has been used include optimal sample size determination [29, 47, 4, 54, 16, 56], drug screening designs [48], bioequivalence trials [30], evidence-based medicine [3], clinical and public health research policy [45, 52] and choosing optimal experimental designs [35].

The diagnostic procedure decision problem has previously been considered from a Bayesian perspective, but only where no optimization is required for the procedures themselves, i.e. when the procedures provide direct diagnosis or when threshold levels have already been defined. Murray et al. [36] provide a Bayesian analysis approach for determining the utility of a diagnostic procedure based on the

*Supported by National Institutes of Health R01 EB0047079.
†Corresponding author.

ability to detect the presence of disease for a given prevalence: the difference in posterior probability of having disease given that diagnosis was positive rather than negative. Parmigiani [44] considers a multi-stage utility-based analysis of diagnostic decisions and subsequent treatment options, where the expected utility is maximized over all possible paths (diagnosis-treatments-outcome combinations) in the decision tree.

In addition, a series of related papers has appeared that examines the use of Bayesian decision analysis for variable selection in generalized linear models [9, 11–13]. Among these, the paper by Fouskakis and Draper [12] provides an MEU-based approach. Their idea is to explicitly include costs, benefits and predictive accuracy into their utility function when deciding which subset of variables to select for the purpose of health care evaluation. They construct proxy sets of future patients to measure predictive accuracy by repeatedly partitioning the data into modeling and validation subsets for cross-validation. For each partition, a logistic regression model is fit via maximum likelihood to the modeling dataset (for a particular subset of predictors) and this is evaluated against the validation set. The fitted posterior probabilities of the validation set are thresholded to mimic the discrete decision to perform or not perform a process audit, with the probability threshold chosen to maximize predictive accuracy. The expected utility is then maximized over all possible subsets of variables.

There is also a long history of applying frequentist methods to the choice of diagnostic procedures. These methods are primarily focused on Receiver Operator Characteristics (ROC) curves and area under the ROC curve (AUC) [57, 46] or non-inferiority methods based on differences between AUCs [18, 19]. Except for a classification tree algorithm by Li and Lu [27] that selects the optimal diagnostic procedure based on expected cost-effectiveness differences and patient characteristics, we are not aware of other work that selects the optimal diagnostic procedure based on accuracy and cost combined. However, Li and Lu did not search for optimal thresholding, but rather the optimal decisions based on given thresholds.

We are unaware of any publications that are more directly related to the present work, i.e. that apply Bayesian decision analysis to simultaneously choosing between diagnostic procedures and optimal thresholds. The present paper specifically considers comparisons between diagnostic procedures for which optimal thresholds should be determined.

The remainder of this paper has the following format. Section 2 describes the data structure for the diagnostic procedure decision process. Section 3 gives the methodology for using plug-in estimates (non-Bayesian) of model parameters to obtain the MEU of diagnostic procedures, when the procedures are based on thresholding continuous measures but with unknown optimal thresholds. Section 4 expands the MEU procedure in Section 3 into a Bayesian approach by integrating over parameter uncertainty in the posterior distribution. In Section 5 we illustrate the methodology via an example comparing diagnostic procedures for prognosing osteoporotic hip fracture. Finally, in Section 6 we offer some discussion and conclusions.

2. DATA STRUCTURE

We consider the hospital level decision problem to determine which diagnostic procedure(s) should be implemented for a particular medical problem. The decision is based on datasets that include measurements from two or more diagnostic procedures for the same medical problem and the same patient population, along with a gold standard diagnosis. The gold standard, could come from pathology or be determined by clinical outcomes (as with the osteoporosis example described in Section 5). It is important for optimal decision making that this dataset be representative of the population to be referred for diagnosis, unless the differences in population composition can be properly compensated for.

Setting up the notation, for patient $i = 1, \ldots, I$ and diagnostic procedure $j = 1, \ldots, J$ we define binary variables: disease state y_i (with $y_i = 1$ indicating that the patient has the disease and 0 if not); and diagnosis d_{ij} (with $d_{ij} = 1$ indicating a positive diagnosis and 0 negative).

We are specifically concerned with diagnostic procedures based on an underlying continuous variable x_{ij}. A continuous diagnostic variable x_{ij} is typically dichotomized at some (diagnostic procedure-specific) threshold a_j to produce a diagnosis. In anticipation of our example of Section 5, we develop the model for the case when a low value implies a positive diagnosis, i.e. $x_{ij} \leq a_j \Leftrightarrow d_{ij} = 1$. The changes required for the reverse situation of a high value implying positive diagnosis are straightforward.

For subject i and diagnostic procedure j we therefore have $d_{ij} = I(x_{ij} \leq a_j)$, where I is the indicator function. An optimal decision process must find an optimal procedure-threshold combination among the set of diagnostic procedures and their possible associated thresholds.

3. MAXIMUM EXPECTED UTILITY (MEU)

We consider the contributing factors towards utility in the diagnostic procedure decision problem to be: a) cost of the jth diagnostic procedure, c_j^D; b) cost of preventative treatment, c^T; and c) cost of disease onset and progression, c^P, where the "cost" of disease progression includes both money and quality of life. In addition, we assume that the preventative treatment has a constant efficacy rate, Λ across subjects and that conditional on knowing Λ, preventative treatment acts independently across subjects[1].

We proceed by defining the utility of the possible outcomes for each of the diagnostic procedures: false negative

[1]This assumption could be relaxed by modeling Λ as a function of covariates, possibly including the (continuous) diagnostic variable of interest. Implementation would require data or/and prior knowledge relating covariates to treatment outcomes.

(FN): $u_j^{FN} = -(c_j^D + c^P)$; true negative (TN): $u_j^{TN} = -c_j^D$; false positive (FP): $u_j^{FP} = -(c_j^D + c^T)$; true positive (TP):

$$u_j^{TP} = \begin{cases} -(c_j^D + c^T) & \text{if treatment succeeds} \\ -(c_j^D + c^T + c^P) & \text{if treatment fails.} \end{cases}$$

Taking the expectation of the utility with respect to the possible outcomes leads to

$$\begin{aligned} E(u_j) = & E(u_j^{TP})p(y_i = 1, d_{ij} = 1) + u_j^{TN}p(y_i = 0, d_{ij} = 0) \\ & + u_j^{FP}p(y_i = 0, d_{ij} = 1) + u_j^{FN}p(y_i = 1, d_{ij} = 0) \\ = & -\{c_j^D + c^T p(d_{ij} = 1) \\ & + c^P[(1 - \Lambda)p(y_i = 1, d_{ij} = 1) + p(y_i = 1, d_{ij} = 0)]\} \end{aligned} \tag{1}$$

Note that if the costs were to vary across individuals we could substitute "expected costs" for actual costs in Equation 1 provided the costs could be considered independent of diagnostic variables and disease status. However, for utility functions that are nonlinear in the costs, the expected utilities for each possible outcome would need to be integrated over the joint distribution of the costs. Similarly, we can substitute "expected efficacy" for actual efficacy in Equation 1 provided that we are willing to make the assumption that the efficacy is independent of diagnostic variables, disease status and costs.

When the diagnostic procedure depends on an underlying continuous diagnostic variable, the expected utility depends on the associated threshold a_j. For the case when a low value of the continuous variable leads to a positive diagnosis then Equation 1 becomes:

$$\begin{aligned} E(u_j|a_j) = & -\{c_j^D + c^T p(x_{ij} \le a_j) \\ & + c^P[(1 - \Lambda)p(y_i = 1, x_{ij} \le a_j) \\ & + p(y_i = 1, x_{ij} > a_j)]\} \end{aligned} \tag{2}$$

As previously stated, a major component of the problem when continuous diagnostic variables are used, is to obtain optimal threshold values for each procedure (a_j^*). Our approach is to optimize threshold values as part of the MEU procedure, and we implement this as a 2-step process:

1. for each diagnostic procedure, maximize the expected utility $E(u_j|a_j^*) = \max_{a_j} E(u_j|a_j)$
2. select the diagnostic procedure $j \in 1, \ldots, J$ with the highest expected utility at a_j^*

3.1 Condition on y_i or $x_{ij} \le a_j$?

Given Equation 2, we can expand $p(y_i = k, x_{ij} \le a_j)$, $k = 0, 1$, by conditioning on either the event $y_i = k$ to give $p(x_{ij} \le a_j|y_i = 1)p(y_i = 1)$ or the event $x_{ij} \le a_j$ to give $\int_{-\infty}^{a_j} p(y_i = 1|x_{ij} = x)\mathrm{d}p(x_{ij} \le x)$. The model for conditioning on $y_i = 1$ is typically easier to define when the value of the continuous variable depends directly on whether the subject has the disease or not. That is, the distribution for the continuous variable differs depending on whether or not the subject has the disease. An example would be diagnosing influenza based on body temperature; when you develop the flu you 'move' to a different distribution of body temperature. By contrast, conditioning on $x_{ij} \le a_{ij}$ is more intuitive when the definition of the disease depends directly on the magnitude of the continuous variable. For example, hypertension is typically defined in terms of whether a patient has high blood pressure, and is a mediator/marker for cardiovascular disease and stroke. It therefore seems appropriate to define a model for the marginal distribution of the population as a whole.

We hereafter focus on conditioning on y_i. Primarily because in practice we found that conditioning on $x_{ij} \le a_{ij}$ (using logistic regression) has specific disadvantages. In particular, the difference between expected utility at the lowest (no positive diagnoses) and highest (no negative diagnoses) thresholds does not generally equal the difference in costs of the diagnostic procedures as expected. The discrepancy occurs because when conditioning on $x_{ij} \le a_{ij}$ the posterior distribution of disease prevalence cannot be constrained to be the same for different diagnostic procedure models.

3.2 Conditioning on y_i

When expanding by conditioning on the event $y_i = k$, Equation 2 becomes

$$E(u_j|a_j) = -\left\{\begin{array}{r} c_j^D + c^T[p(y_i = 0)F^{j0}(a_j) \\ + p(y_i = 1)F^{j1}(a_j)] \\ + c^P p(y_i = 1)(1 - \Lambda F^{j1}(a_j)) \end{array}\right\}, \tag{3}$$

where $F^{jk}(x)$ is the CDF of the continuous variable x for diagnostic procedure j conditional on disease state k.

Theorem 3.1. *Let each $F^{jk}(x)$ be a differentiable CDF with associated pdf $f^{jk}(x)$. Furthermore, let each $G^j(x) = \frac{f^{j1}(x)}{f^{j0}(x)}$ be a strictly decreasing continuous function of x with $\lim_{x\to-\infty} G^j(x) > \frac{c^T p(y_i=0)}{p(y_i=1)(c^P\Lambda - c^T)}$, $\lim_{x\to\infty} G^j(x) < \frac{c^T p(y_i=0)}{p(y_i=1)(c^P\Lambda - c^T)}$, and $c^P\Lambda - c^P > 0$. Then $\max_{a_j} E(u_j|a_j)$ exists and occurs at $a_j^* = (G^j)^{-1}(\frac{c^T p(y_i=0)}{p(y_i=1)(c^P\Lambda - c^T)})$.*

Proof in Appendix A.

To obtain the optimal diagnostic procedure and threshold choice we finally compare the expected utility for each a_j^* to determine the diagnostic procedure with maximum expected utility (MEU).

4. INCORPORATING PARAMETER UNCERTAINTY

The methodology developed thus far assumes that all model parameters are known *a priori*. We can estimate and "plug-in" any parameters, but the plug-in approach ignores

parameter uncertainty when estimating utility. That is, in general $\max_{a_j} E_{x,\boldsymbol{\xi}}(u_j|a_j) \neq \max_{a_j} E_x(u_j|a_j, \hat{\boldsymbol{\xi}})$, where, $\boldsymbol{\xi}$ is a vector of the unknown parameters and $\hat{\boldsymbol{\xi}}$ is some estimate of these parameters. We here develop a fully Bayesian approach based on Markov chain Monte Carlo (MCMC) sampling that incorporates parameter uncertainty when calculating MEU. The approach takes the following steps:

1. Simulate from the posterior distribution of $\boldsymbol{\xi}$ using MCMC.
2. For an appropriately finely sampled set of values for a, use the generated MCMC sample to estimate $E_{x,\boldsymbol{\xi}}(u_j|a)$ at all values (for each diagnostic procedure).
3. Determine a_j^* and $E_{x,\boldsymbol{\xi}}(u_j|a_j^*)$ for each diagnostic procedure by selecting the a that leads to the largest $E_{x,\boldsymbol{\xi}}(u_j|a)$.
4. Determine the procedure j that corresponds to $\max_j E_{x,\boldsymbol{\xi}}(u_j|a_j^*)$ – the MEU across all diagnostic procedures.

Here, $E_{x,\boldsymbol{\xi}}(u_j|a)$ is calculated by approximating the integral $\int p(\boldsymbol{\xi}|\mathbf{x},\mathbf{y})E_x(u_j|a_j,\boldsymbol{\xi})\,\mathrm{d}\boldsymbol{\xi}$ using MCMC samples; the vector $\mathbf{x}$ is the complete set of continuous diagnostic variables across all procedures, i.e. $\{x_{ij} : i = 1\ldots I, j = 1\ldots J\}$, and $\mathbf{y} = \{y_i : i = 1\ldots I\}$ is the set of gold standard diagnoses.

The expected utility is integrated over the posterior distribution with $f^{jk}(x)$ and $p(y_i = k)$ considered as conditional on their parameters $\boldsymbol{\gamma}_{jk}$ and δ respectively. When incorporating parameter uncertainty Equation 3 expands to

$$E_{\{x,\boldsymbol{\gamma}_{jk},\delta\}}(u_j|a_j) = - \begin{bmatrix} c_j^D + c^T \sum_{k=0}^{1} \int_{\boldsymbol{\gamma}_{jk}} \int_{\delta} \int_{-\infty}^{a_j} f^{jk}(z|\boldsymbol{\gamma}_{jk})[k\delta \\ + (1-k)(1-\delta)]\pi(\boldsymbol{\gamma}_{jk}|\mathbf{x},\mathbf{y})\pi(\delta|\mathbf{x},\mathbf{y})\,\mathrm{d}z\,\mathrm{d}\delta\,\mathrm{d}\boldsymbol{\gamma}_{jk} \\ + c^P \left\{ \begin{matrix} \int_{\delta} \delta\pi(\delta|\mathbf{x},\mathbf{y}) \int_{\boldsymbol{\gamma}_{j1}} \int_{-\infty}^{a_j} \left[1 - \Lambda f^{j1}(z|\boldsymbol{\gamma}_{j1})\right] \\ \pi(\boldsymbol{\gamma}_{j1}|\mathbf{x},\mathbf{y})\,\mathrm{d}z\,\mathrm{d}\boldsymbol{\gamma}_{j1}\,\mathrm{d}\delta \end{matrix} \right\} \end{bmatrix}$$

In the above expression we have implicitly assumed that $\boldsymbol{\gamma} = \{\boldsymbol{\gamma}_{jk} : j = 1\ldots J, k = 0, 1\}$ and δ are independent of each other.

The integrals with respect to $\boldsymbol{\gamma}$ terms and δ are estimated by averaging the expectation over an MCMC sample of the posterior distribution, i.e.:

$$(4)\quad E_{\{x,\boldsymbol{\gamma}_{jk},\delta\}}(u_j|a_j) \approx - \frac{1}{N} \sum_{s \in S} \begin{bmatrix} c_j^D + c^T \sum_{k=0}^{1} \int_{-\infty}^{a_j} f^{jk}(z|\boldsymbol{\gamma}_{jk}^s)[k\delta^s \\ + (1-k)(1-\delta^s)]\pi(\boldsymbol{\gamma}_{jk}^s|\mathbf{x},\mathbf{y})\pi(\delta^s|\mathbf{x},\mathbf{y})\,\mathrm{d}z \\ + c^P \left\{ \begin{matrix} \delta^s \left\{ \pi(\delta^s|\mathbf{x},\mathbf{y}) \int_{-\infty}^{a_j} [1 - \Lambda f^{j1}(z|\boldsymbol{\gamma}_{j1}^s)] \right. \\ \left. \pi(\boldsymbol{\gamma}_{j1}^s|\mathbf{x},\mathbf{y})\,\mathrm{d}z \right\} \end{matrix} \right\} \end{bmatrix}$$

where s denotes a single realization of the parameter set $(\boldsymbol{\gamma}, \delta)$ from the complete set S of N MCMC sample realizations.

5. OSTEOPOROTIC HIP FRACTURE EXAMPLE

Osteoporosis is a major public health problem estimated to have cost on the order of $19 billion in the USA during 2005. The worst outcome of osteoporosis, hip fracture, is extremely painful, debilitating, and in 10% of cases leads to death. The World Health Organization (WHO) defines osteoporosis as having bone mineral density (BMD) or bone mineral content (BMC) below a "T-score" of -2.5, where the T-score is defined as an individual's BMD or BMC normalized to that of a young adult reference range from a historical/population dataset [23]. The WHO definition does not consider the optimality of the BMD/BMC threshold in terms of utility with respect to potential treatment. In addition, BMD/BMC estimation, diagnostic accuracy, and test cost all vary by skeletal site; the WHO definition fails to specify which skeletal sites to use when measuring BMD or BMC for the diagnosis of osteoporosis [24].

We provide an example of the Bayesian diagnostic procedure decision process applied to the prognosis of oseoporotic hip fracture in The Study of Osteoporotic Fractures (SOF) [7, 8]. In the SOF, 7071 randomly selected post-menopausal Caucasian women had distal forearm BMD measured by single X-ray absorptiometry (SXA) and femoral neck BMD measured by dual x-ray absorptiometry (DXA).

Validity of the SOF population to study osteoporotic hip fracture risk is well established. The population in the study is reasonably representative of the untreated Caucasian post-menopausal women aged 65 and older in the US that would currently be considered for osteoporosis screening via BMD measurement. We use the subject outcome of 5-year post-examination hip fracture as the objective standard against which we compare the BMD measurement procedures.

We wish to emphasize here that in the interest of providing a clear illustration of the methodology we have made simplifying assumptions with respect to this study and do not go into detail as to how costs/utilities were evaluated. Therefore, this is not meant as an authoritative analysis of this dataset, and in no way do we mean to encourage changes in medical practice based on the results.

Figure 1 displays summary BMD histograms classified by measurement location and 5 year hip fracture status ($y = 0$ no fracture, $y = 1$ fracture). The primary observation to note is that there appears to be poor separation between the fracture and non-fracture individuals with respect to BMD. The separation appears particularly limited for distal forearm BMD. The femoral neck measurements do show a clear distributional shift towards lower BMD for the fracture cases. However, there is very large overlap between the fracture and non-fracture individuals.

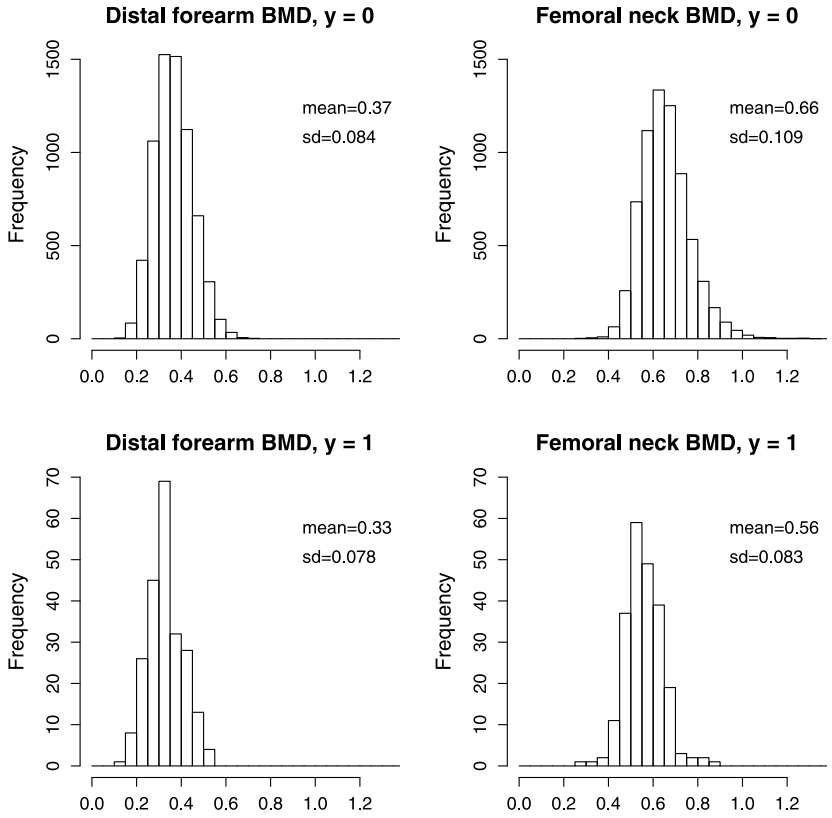

Figure 1. Histograms and associated sample mean and standard deviation estimates illustrating distributions of BMD measurements for distal forearm BMD (Left) and femoral neck BMD (Right) for each of the non-fracture ($y = 0$, Top) and fracture ($y = 1$, Bottom) outcome states.

5.1 Determining costs/utilities

Because the determination of component costs is not the focus of this paper, we only give a brief overview of how costs/utilities were determined. The costs we use for the osteoporosis diagnostic procedures are based on medicare reimbursement values for Current Procedural Terminology (CPT) 7605 DXA measurements: distal forearm BMD measurement costs $c_1^D = \$42$ and femoral neck BMD measurement costs $c_2^D = \$139$. The cost of preventative treatment we use is $c^T = \$12,000$ (based on the product of estimates of annual medication costs [10] and life expectancy for this group of post-menopausal women [1]); and we use an estimated preventative treatment efficacy rate of $\Lambda = 0.55$ [34, 5, 25]. Assessment of the expected cost of disease progression (i.e., hip fracture) is more complex, requiring consideration of medical costs [33, 14], loss of life [34, 26, 32] and loss of quality of life [33, 49, 31, 6, 20, 22, 21] among hip fracture patients. The total expected cost of disease progression we use is $c^P = \$234,000$.

5.2 Model specification and implementation

We model the joint distribution of the continuous diagnostic variables (distal forearm and femoral neck BMD) conditional on each disease state as bivariate lognormal. For each disease state k (hip fracture versus no hip fracture), the vector of BMD measurements for each subject, $\boldsymbol{x}_i$, is distributed as

$$\log(\boldsymbol{x}_i)|y_i = k \sim MVN(\boldsymbol{\mu}_k, \Sigma_k),$$

where $\boldsymbol{\mu}_k$ is the mean vector of the natural logarithm of the diagnostic variables and Σ_k is the corresponding covariance matrix. The log-normal distribution appeared to provide a reasonable fit to the data (based on diagnostic plots – not shown).

MCMC sampling for this application was coded in WinBUGS (http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml) and was called from the R (http://www.r-project.org/) library R2WinBUGS (http://cran.r-project.org/web/packages/R2WinBUGS/index.html). Weak prior distributions are placed on $\boldsymbol{\mu}_k$ and Σ_k, consisting of a bivariate normal prior for $\boldsymbol{\mu}_k$ ($MVN(\mathbf{0}, 10^3\mathrm{I}_2)$) and a weak Wishart prior for Σ_k^{-1} ($W_2(2, 10^4\mathrm{I}_2)$) [2]. There are implied constraints when using Wishart priors for Σ_k^{-1} [15, p.p. 284–287]. However, for this study the posterior estimates of the SDs and correlations for the μ_k's are very close to the sample estimates, and therefore, the prior does

Table 1. Comparison of MEU between prognostic procedures for hip fracture. The plug-in estimates were evaluated using Theorem 3.1 for which all assumptions were met if the MLE estimates were considered as true

	Plug-in	*Fully Bayes*
distal a_j^*	0.189	0.191
neck a_j^*	0.498	0.497
distal MEU	−7,513	−7,507
neck MEU	−7,335	−7,329
optimal method	neck	neck

not perceivably constrain the posterior results. If a prior that does not control the precision of all elements of Σ with a single parameter is desired, then a scaled inverse Wishart could be used instead [43]; [15, p.p. 286–287]. The prior distribution for disease prevalence is a weak beta distribution ($Be(10^{-5}, 10^{-5})$), approximating the improper and non-informative $Be(0, 0)$ prior [58]. The WinBUGS code for this model is given in Appendix B.

5.3 Results

Figure 2 (a) plots expected utility against BMD threshold for distal forearm and femoral neck BMD. The plot displays both fully Bayesian and plug-in (maximum likelihood) estimation of utility curves evaluated by a grid search over the range $[0, 1.5]$ with threshold increments of 0.001 between evaluations (1.5 is well beyond the range of any BMD measurements in the data). The fully Bayesian and plug-in methods lead to similar utility curves, though there are slight differences visible in the enlarged region of Figure 2 (b). Regardless of whether plug-in or fully Bayesian MEU is used, the decision is to use femoral neck BMD as the optimum prognostic procedure for hip fracture. This is reflected in the quantitative results in Table 1, that also provides values for optimal threshold.

The fact that low threshold values have similarly high utility in Figure 2 (a) is a consequence of the treatment being expensive combined with the relatively low prevalence of fracture in the randomly sampled elderly population ($\approx$ 3%). Overall, it is cost-effective to accept that most patients who will experience fractures will go untreated rather than risk a large number of false positives that will be treated unnecessarily. However, there is some gain in MEU to be obtained by using femoral neck BMD at its optimal threshold value rather than not treating anyone. This value is in the lower range of BMD values as can be seen by relating the optimal threshold in Table 1 back to the histograms of Figure 1.

The fully Bayesian and plug-in curves can be similar for three reasons: 1) weak prior information, 2) strong data information and 3) near symmetry of the fracture and non-fracture BMD data subsets. For this dataset, the large sample and weak priors combination led to very precise posteriors for the parameters of the fracture and non-fracture BMD distributions (with posterior means similar to the ML estimates). To examine the extent to which the size of the dataset contributed to the similarity between plug-in and fully Bayesian MEU, we repeated the analysis based on a randomly selected sub-sample of 100 individuals (which contained only 4 fracture patients). Figure 2 (c) of the ensuing expected utility plot shows increased difference between plug-in and fully Bayes expected utility curves (and the associated MEUs).

5.4 Strong prior information

Thus far we have only considered weak, uninformative prior information when comparing the plug-in and fully Bayesian approach. At least for this dataset, the differences between the two approaches have proved relatively minor — not affecting the overall decision much. However, increased prior information can lead to larger differences between the plug-in and fully Bayes approaches. The most obvious prior information that could be used here would come from knowledge about disease prevalence in the target population, which we incorporate through the prior distribution on δ.

For illustration purposes, we employ a very tight prior ($Be(5 \times 10^5, 95 \times 10^5)$) for δ centered at 0.05 (contrasting with the disease prevalence in the data of approximately 0.03). This is an unrealistically tight prior that we have adopted in order to force the posterior estimate of δ to be close to 0.05 in contradiction of the strong information in the data.

Figure 2 (d) shows a plot of utility against threshold for this strong prior on δ. The increased posterior expectation of prevalence level induced by the strong prior leans the optimal decision towards treating more subjects based on femoral neck BMD; the importance of threshold choice is more obvious in this plot than in those with weak prior information because there is more of a balance between expected treatment and fracture costs induced by the increased posterior prevalence rate.

6. DISCUSSION

6.1 Alternate decision perspectives and modeling extensions

In this paper we have presented a model of institutional/departmental decision-making for choosing between diagnostic procedures. The perspective would be different for other decision makers necessitating modifications to costs/utility and potentially the utility model structure.

The decision from the clinician's perspective requires that individuals be assigned to particular diagnostic procedures based on patient circumstances, e.g. based on the patient's subgroup classification or patient-level covariates. However, there are logistical, legal and ethical issues that would have to be overcome before clinicians would be motivated to consider assigning personalized diagnostic procedures based on

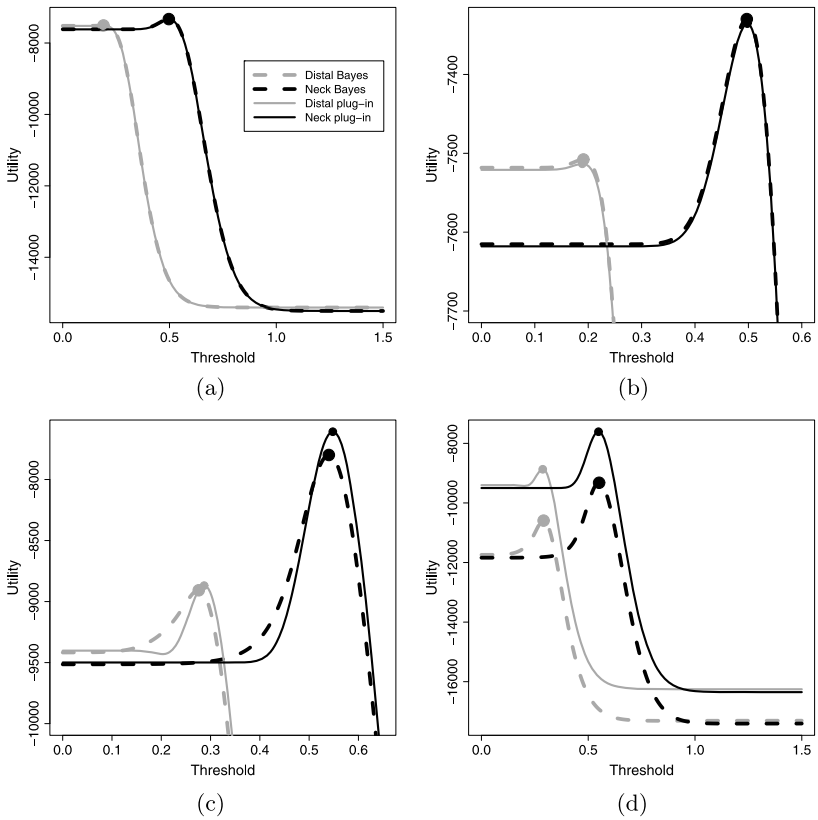

Figure 2. Plots of expected utility against threshold (a) for the model conditioning on y_i: dots correspond to the MEU for each prognostic procedure for hip fracture. Panel (a) shows the expected utility plot for the full sample and panel (b) shows the same plot zooming in on the peaks; panel (c) shows the same (zoomed) plot generated from a random sub-sample of 100 individuals (note that the range on the utility axis is changed) and panel (d) shows an expected utility plot based on a strong prior for δ with mean of 0.05 (c.f. fracture prevalence $\approx$ 0.03) using the full dataset. The dots at the top of the curves give the locations of the MEU estimates. The plug-in MEU estimates were evaluated using Theorem 3.1 for which all assumptions were met once the MLE estimates were considered as true.

patient-level covariates. The hospital-level decision could also be affected by other (measured) known covariates in which case they might be integrated marginally into the model. The perspective of the insurance company may vary from global-level decisions applied to a complete insured group to whether or not to provide coverage of a specific diagnostic procedure for an individual. We do not directly consider additional covariates here, but provide some discussion in Section 6.

Further potential extensions to the model include consideration of start-up costs for switching to a new diagnostic procedure (potentially allowing for risk aversion) and allowing for the possibility of running diagnostic procedures side-by-side which relates back to the clinician's perspective.

6.2 Comparison of MEU for a range of disease progression costs

Determining the costs to be used in the Bayesian decision analysis/MEU procedure is difficult and subject to criticism on ethical grounds. The main concern is that defining the utility of disease progression requires assigning relative values to loss of life and quality of life. Defining a relative value on life (even in non-monetary terms) has ethical implications that need to be considered carefully. The loss of life and quality of life needs to be converted to a scale that can be compared with the other costs (or vice-versa). A graphical approach plotting MEU against a range of costs for disease progression could be used to aid decision making when the decision maker is unable or unwilling to assign a specific cost to loss of life or quality of life. This approach allows the decision-maker to determine the range of loss of life/quality of life for which each diagnostic procedure is optimal. The graphical approach is similar to that of the Cost-Effectiveness Acceptability Curve (CEAC) for clinical trials proposed by van Hout et al. [55]. The CEAC plots the probability that treatment 1 is more cost-effective than treatment 2 against a quantity K that describes the relative willingness to pay for 1 unit of treatment effectiveness. O'Hagan and Stevens [40] extend this idea to plot the mean

incremental net benefit (INB) – the improvement in cost-effectiveness of treatment 1 over treatment 2 – against K.

6.3 Computational overhead

Computation was very quick for the models considered in this paper. 20,000 MCMC iterations took 1 minute on a Mac OS X laptop with a 2GHz Intel Core 2 Duo and 4 GB 667 MHz DDR2 SDRAM. We used 10,000 burn-in samples and 10,000 samples for evaluation. Good convergence (and mixing) of MCMC output was achieved within a few hundred iterations (based on visual diagnosis of MCMC output for model parameters) and so the burn-in period was perhaps conservative. The estimated Monte Carlo error for all parameters was consistently less than 1% of the associated sample standard deviation for all parameters of the model.

6.4 Conclusions

The work presented here provides a Bayesian utility framework for choosing between diagnostic procedures. We have shown that a Bayesian utility based approach is feasible for choosing between diagnostic procedures that are derived from threshold values for continuous diagnostic variables. The fully Bayesian decision is different from an 'estimate and plug-in' approach in that the fully Bayesian approach appropriately incorporates uncertainty in parameter values and can incorporate other prior information. As illustrated in the osteoporosis hip fracture example, the fully Bayesian decision approach provides maximum benefit over 'plug-in' when there is a) considerable uncertainty in parameter estimates — small n, and b) strong prior information.

APPENDIX A. PROOF OF THEOREM 3.1

Proof.

$$\frac{\mathrm{d}E(u_j|a_j)}{\mathrm{d}a_j} = -c^T\left[p(y_i=0)f^{j0}(a_j)+p(y_i=1)f^{j1}(a_j)\right] + c^P p(y_i=1)\Lambda f^{j1}(a_j) = 0$$

$$\Rightarrow f^{j0}(a_j)\left\{G^j(a_j)p(y_i=1)\left(c^P\Lambda - c^T\right) - c^T p(y_i=0)\right\} = 0. \quad (5)$$

Because $G^j(x)$ is a strictly decreasing continuous function, it is one-to-one and hence invertible. Therefore, there is at most one a_j that satisfies Equation 5. Furthermore, because $(c^P\Lambda - c^T) > 0$ and $\lim_{x\to\infty} G^j(x) < \frac{c^T p(y_i=0)}{p(y_i=1)(c^P\Lambda - c^T)}$, $\frac{\mathrm{d}E(u_j|a_j)}{\mathrm{d}a_j} < 0$ for $a_j > a_j^*$. Similarly, $\lim_{x\to-\infty} G^j(x) > \frac{c^T p(y_i=0)}{p(y_i=1)(c^P\Lambda - c^T)}$ implies $\frac{\mathrm{d}E(u_j|a_j)}{\mathrm{d}a_j} > 0$ for $a_j < a_j^*$. Thus, $E(u_j|a_j^*) = \max_{a_j} E(u_j|a_j)$. □

APPENDIX B. WINBUGS CODE

```
model {
for (i in 1:N) {
  # Model specification
  # y[i] is disease status
  y[i]~dbern(delta)
  # change outcome to 1 and 2 for matrix
  # indexing rather than 0 and 1
  ix[i] <- y[i] + 1
  # joint distal and neck BMD conditional
  # distributions, lgx is log(BMD), mu/tau are
  # prior mean/var vectors of distal and neck BMD
  lgx[i,1:2]~dmnorm(mu[ix[i],1:2],tau[ix[i],1:2,1:2])
  }
  smallnumber <- 1.0E-5
  # theta is marginal probability of disease
  # in study population
  theta~dbeta(smallnumber,smallnumber)
  for(j in 1:2) {
    # hyper-parameters of Mean/Precision from R
    mu[j,1:2]~dmnorm(Mean[],Prec[,])
    # hyper-parameters of Omega/degFdm from R
    tau[j,1:2,1:2]~dwish(Omega[,],degFdm)
  }
}
```


ACKNOWLEDGEMENTS

Thanks to Caixia Li for the research into costs incorporated into the model and to Bill Chu for comments, review and editing of this manuscript.

Received 18 June 2010

REFERENCES

[1] Arias, E. (2007). United States life tables, 2004. *Natl Vital Stat Rep* **56** 1–40.

[2] Arminger, G. and Muthén, B. O. (1998). A Bayesian approach to nonlinear latent variable models using the Gibbs sampler and the Metropolis-Hastings algorithm. *Psychometrika* **63** 271–300.

[3] Ashby, D. and Smith, A. F. M. (2000). Evidence-based medicine as Bayesian decision-making. *Statistics in medicine* **19**.

[4] Bernado, J. M. (1997). Statistical Inference as a Decision Problem: The Choice of Sample Size. *Statistician* **46** 151–153.

[5] Black, D. M., Thompson, D. E., Bauer, D. C., Ensrud, K., Musliner, T., Hochberg, M. C., Nevitt, M. C., Suryawanshi, S. and Cummings, S. R. (2000). Fracture Risk Reduction with Alendronate in Women with Osteoporosis: The Fracture Intervention Trial.

[6] Burström, K., Johannesson, M. and Diderichsen, F. (2001). Swedish population health-related quality of life results using the EQ-5D. *Quality of Life Research* **10** 621–635.

[7] Cummings, S. R., Black, D. M., Nevitt, M. C., Browner, W., Cauley, J., Ensrud, K., Genant, H. K., Palermo, L., Scott, J. and Vogt, T. M. (1993). Bone density at various sites for prediction of hip fractures. *Lancet (British edition)* **341** 72–75.

[8] Cummings, S. R., Nevitt, M. C., Browner, W. S., Stone, K., Fox, K. M., Ensrud, K. E., Cauley, J., Black, D. and Vogt, T. M. (1995). Risk Factors for Hip Fracture in White Women.

[9] Draper, D. and Fouskakis, D. (2000). A Case Study of Stochastic Optimization in Health Policy: Problem Formulation and Preliminary Results. *Journal of Global Optimization* **18** 399–416.

[10] Fleurence, R. L., Iglesias, C. P. and Johnson, M. J. (2007). The cost effectiveness of biophosphonates for the prevention and treatment of osteoporosis: a structured review of the literature. *Pharmacoeconomics* **25** 913–933.

[11] Fouskakis, D. and Draper, D. (2002). Stochastic Optimization: a Review. *International Statistical Review* **70** 315–349.

[12] Fouskakis, D. and Draper, D. (2008). Comparing Stochastic Optimization Methods for Variable Selection in Binary Outcome Prediction, With Application to Health Policy. *Journal of the American Statistical Association* **103** 1367–1381.

[13] Fouskakis, D., Ntzoufras, I. and Draper, D. (2009). Bayesian variable selection using cost-adjusted BIC, with application to cost-effective measurement of quality of health care. *Annals of Applied Statistics*. **3** 663–690.

[14] Gabriel, S., Gabriel, S., Tosteson, A., Leibson, C., Crowson, C., Pond, G., Hammond, C. and Melton, L. III (2002). Direct Medical Costs Attributable to Osteoporotic Fractures. *Osteoporosis International* **13** 323–330.

[15] Gelman, A. and Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press New York.

[16] Halpern, J., Brown, B. W. Jr and Hornberger, J. (2001). The Sample Size for a Clinical Trial: A Bayesian-Decision Theoretic Approach. *Statist. Med.* **20** 841–858.

[17] Heitjan, D. F., Moskowitz, A. J. and Whang, W. (1999). Bayesian Estimation of Cost-Effectiveness Ratios from Clinical Trials. *Health Economics* **8** 191–201.

[18] Jin, H. and Lu, Y. (2008). A Procedure for Determining Whether a Simple Combination of Diagnostic Tests May Be Noninferior to the Theoretical Optimum Combination. *Medical Decision Making* **28** 909.

[19] Jin, H. and Lu, Y. (2009). Permutation test for non-inferiority of the linear to the optimal combination of multiple tests. *Statistics and Probability Letters* **79** 664–669.

[20] Kanis, J., Johnell, O., Oden, A., Borgstrom, F., Zethraeus, N., Laet, C. D. and Jonsson, B. (2004). The risk and burden of vertebral fractures in Sweden. *Osteoporosis International* **15** 20–26.

[21] Kanis, J. A., Johnell, O., Oden, A., De Laet, C., Oglesby, A. and Jönsson, B. (2002). Intervention thresholds for osteoporosis. *Bone* **31** 26–31.

[22] Kanis, J. and Jonsson, B. (2002). Economic Evaluation of Interventions for Osteoporosis. *Osteoporosis International* **13** 765–767.

[23] Kanis, J., Melton, L., Christiansen, C., Johnston, C. and Khaltaev, N. (1994). The diagnosis of osteoporosis. *J Bone Miner Res* **9** 1137–1141.

[24] Kanis, J. and the WHO Study Group (1994). Assessment of fracture risk and its application to screening for postmenopausal osteoporosis: Synopsis of a WHO report. *Osteoporosis International* **4** 368–381.

[25] Karpf, D., Shapiro, D., Seeman, E., Ensrud, K., Johnston, C., Adami, S., Harris, S., Santora, A., Hirsch, L., Oppenheimer, L. et al. (1997). Prevention of nonvertebral fractures by alendronate. A meta-analysis. Alendronate Osteoporosis Treatment Study Groups. *JAMA* **277** 1159–1164.

[26] Leibson, C. L., Tosteson, A. N. A., Gabriel, S. E., Ransom, J. E. and Melton, L. J. (2002). Mortality, Disability, and Nursing Home Use for Persons with and without Hip Fracture: A Population-Based Study. *Geriatrics* **50** 1644–1650.

[27] Li, C. and Lu, Y. (2008). Tree-Structured Analysis for Determining Optimal Diagnostic Tests for Patients. Joint Statistical Meetings, Denver, CO, USA.

[28] Lindley, D. V. (1985). *Making Decisions*, second ed. Wiley. MR0892099

[29] Lindley, D. V. (1997). The Choice of Sample Size. *Statistician* **46** 129–138.

[30] Lindley, D. V. (1998). Decision analysis and bioequivalence trials. *Statistical Science* 136–141.

[31] Macran, S., Weatherly, H. and Kind, P. (2003). Measuring Population Health: A Comparison Of Three Generic Health Status Measures. *Medical Care* **41** 218.

[32] Magaziner, J., Simonsick, E. M., Kashner, T. M., Hebel, J. R. and Kenzora, J. E. (1989). Survival experience of aged hip fracture patients.

[33] Meadows, E. S., Klein, R., Rousculp, M. D., Smolen, L., Ohsfeldt, R. L. and Johnston, J. A. (2007). Cost-effectiveness of preventative therapies for postmenopausal women with osteopenia. *BMC Women's Health* **7** 6.

[34] Mobley, L. R., Hoerger, T. J., Wittenborn, J. S., Galuska, D. A. and Rao, J. K. (2006). Cost-Effectiveness of Osteoporosis Screening and Treatment with Hormone Replacement Therapy, Raloxifene, or Alendronate. *Medical Decision Making* **26** 194.

[35] Muller, P. (1999). Simulation based optimal design. MR1723509

[36] Murray, R., McKillop, J., Bessent, R., Hutton, I., Lorimer, A. and Lawrie, T. (1981). Bayesian analysis of stress thallium-201 scintigraphy. *European Journal of Nuclear Medicine and Molecular Imaging* **6** 201–204.

[37] O'Hagan, A. and Forster, J. (2004). *Bayesian Inference*, second ed. *Kendall's Advanced Theory of Statistics* **2B**. Wiley.

[38] O'Hagan, A. and Stevens, J. W. (2001). A framework for cost-effectiveness analysis from clinical trial data. *Health Economics* **10** 302–315.

[39] O'Hagan, A. and Stevens, J. W. (2002). Bayesian methods for design and analysis of cost-effectiveness trials in the evaluation of health care technologies. *Statistical Methods in Medical Research* **11** 469–490.

[40] O'Hagan, A. and Stevens, J. W. (2002). The probability of cost-effectiveness. *BMC Medical Research Methodology* **2** 5.

[41] O'Hagan, A., Stevens, J. W. and Montmartin, J. (2000). Inference for the Cost-Effectiveness Acceptibility Curve and Cost-Effectiveness Ratio. *Pharmacoeconomics* **17** 339–349.

[42] O'Hagan, A., Stevens, J. W. and Montmartin, J. (2001). Bayesian cost effectiveness analysis from clinical trial data. *Statist. Med.* **20** 733–753.

[43] OMalley, A. and Zaslavsky, A. (2005). Cluster-level covariance analysis for survey data with structured nonresponse. Technical report, Department of Health Care Policy, Harvard Medical School.

[44] Parmigiani, G. (2004). Uncertainty and the value of diagnostic information, with application to axillary lymph node dissection in breast cancer. *Statistics in medicine* **23** 843–855.

[45] Parmigiani, G., Ancukiewicz, M. and Matchar, D. (1996). Decision models in clinical recommendations development: The stroke prevention policy model. *Bayesian Biostatistics* 207–233.

[46] Pepe, M. S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press. MR2260483

[47] Pham-Gia, T. (1997). On Bayesian Analysis, Bayesian Decision Theory and the Sample Size problem. *Statistician* **46** 139–144.

[48] Rossell, D., Müller, P. and Rosner, G. L. (2007). Screening Designs for Drug Development. *Biostatistics* **8** 595–608.

[49] Schousboe, J. T., Nyman, J. A., Kane, R. L. and Ensrud, K. E. (2005). Cost-Effectiveness of Alendronate Therapy for Osteopenic Postmenopausal Women. *Annals of Internal Medicine* **142** 734–741.

[50] Simon, R. (1999). Bayesian design and Analysis of Active Control Clinical Trials. *Biometrika* **55** 484–487.

[51] Stangl, D. K. (1995). Prediction and Decision Making Using Bayesian Hierarchical Models Statistics in Medicine. *Statistics in Medicine*.

[52] Stangl, D. K. (2005). Bridging the gap between statistical analysis and decision making in public health research. *Statistics in medicine* **24**. MR2134520
[53] Stevens, J. W. and O'Hagan, A. (2002). Incorporation of genuine prior information in cost-effectiveness analysis. *International Journal of Technology Assessment in Health Care* **18** 782–790.
[54] Tan, S. B. and Smith, A. F. M. (1998). Exploratory Thoughts on Clinical Trials with Utilities. *Statist. Med.* **17** 2771–2791.
[55] van Hout, B., Al, M. J., Gordon, G. S. and Rutten, F. (1994). Costs, effects and C/E-ratios alongside a clinical trial. *Health Econ* **3** 309–19.
[56] Walker, S. G. (2003). How Many Samples? A Bayesian Nonparametric Approach. *Statistician* **52** 475–482. MR2011142
[57] Zhou, X. H., Obuchowski, N. A. and McClish, D. K. (2002). *Statistical Methods in Diagnostic Medicine*. Wiley-Interscience New York. MR1915698
[58] Zhu, M. and Lu, A. Y. (2004). The counter-intuitive non-informative prior for the Bernoulli family. *Journal of Statistics Education* **12**.

John Kornak
University of California, San Francisco
Department of Radiology and Biomedical Imaging and
Department of Epidemiology and Biostatistics
185 Berry St, Ste. 350
San Francisco, CA 94107
USA
E-mail address: john.kornak@ucsf.edu

Ying Lu
Palo Alto VA Health Care System and
Department of Health Research and Policy
Stanford University
259 Campus Drive
HRP/Redwood Building T152
Stanford, CA 94305
USA
E-mail address: ylu1@stanford.edu