Bayesian Model Specification: Paying the Right Price for Model Uncertainty and Data Collection

David Draper (joint work with **Dimitris Fouskakis**, **Milovan Krnjajić** and **Ioannis Ntzoufras**)

Department of Applied Mathematics and Statistics

University of California, Santa Cruz

draper@ams.ucsc.edu

www.ams.ucsc.edu/∼draper

University of Florida

Tenth Annual Winter Workshop on

Bayesian Model Selection and Objective Methods

12 January 2008

What is a Bayesian Model?

Definition: A Bayesian model is a **mathematical framework** (embodying **assumptions** and **judgments**) for **quantifying uncertainty about unknown quantities** by relating them to **known quantities**.

Desirable for the **assumptions** and **judgments** in the model to arise as directly as possible from **contextual information** in the problem under study.

The most satisfying approach to **achieving this goal** appears to be that of de Finetti (1930): a **Bayesian model** is a joint predictive distribution

$$p(y) = p(y_1, \ldots, y_n) \tag{1}$$

for as-yet-unobserved **observables** $y = (y_1, \ldots, y_n)$.

Example 1: Data = **health outcomes** for all patients at one hospital with heart attack admission diagnosis.

Simplest possible: $y_i = 1$ if patient i **dies within 30 days of admission**, 0 otherwise.

Exchangeability

de Finetti (1930): **in absence of any other information**, my predictive uncertainty about y_i is **exchangeable**.

Representation theorem for binary data: if I'm willing to regard $(y_1, \ldots, y_n)$ as part of an **infinitely exchangeable sequence** (meaning that I judge all **finite subsets** exchangeable; this is like **thinking** of the y_i as having been **randomly sampled** from the **population** $(y_1, y_2, \ldots)$), then to be **coherent** my joint predictive distribution $p(y_1, \ldots, y_n)$ must have the simple **hierarchical** form

$$\begin{aligned} \theta & \sim p(\theta) \\ (y_i | \theta) & \stackrel{\text{IID}}{\sim} \text{Bernoulli}(\theta), \end{aligned} \tag{2}$$

where $\theta = P(y_i = 1) =$ **limiting value of mean of** y_i in infinite sequence.

Mathematically $p(\theta)$ is **mixing distribution** in

$$p(y_1, \ldots, y_n) = \int_0^1 \prod_{i=1}^n p(y_i | \theta) \, p(\theta) \, d\theta \ . \tag{3}$$

Model = Exchangeability + Prior

Statistically, $p(\theta)$ provides opportunity to quantify **prior information** about θ and combine with information in y.

Thus, in simplest situation, **Bayesian model specification** = choice of **scientifically appropriate prior distribution** $p(\theta)$.

Example 2 (elaborating Example 1): Now I want to predict real-valued **sickness-at-admission score** instead of mortality (still no **covariates**).

Uncertainty about y_i still **exchangeable**; de Finetti's (1937) **representation theorem** for real-valued data: if $(y_1, \ldots, y_n)$ part of **infinitely exchangeable sequence**, all **coherent** joint predictive distributions $p(y_1, \ldots, y_n)$ must have hierarchical form

$$\begin{aligned} F & \sim p(F) \\ (y_i|F) & \overset{\text{IID}}{\sim} F, \end{aligned} \tag{4}$$

where F = **limiting empirical cumulative distribution function** (CDF) of infinite sequence $(y_1, y_2, \ldots)$.

Bayesian Nonparametrics

Thus here Bayesian model specification = choosing **scientifically appropriate mixing (prior) distribution** $p(F)$ for F.

However, F is **infinite-dimensional parameter**; putting probability distribution on $\mathcal{D} = \{\text{all possible CDFs}\}$ is harder.

Specifying distributions on **function spaces** is task of **Bayesian nonparametric** (BNP) modeling (e.g., Dey et al. 1998).

Example 3 (elaborating Example 2): In practice, in addition to **outcomes** y_i, **covariates** x_{ij} will typically be available.

For instance (Hendriksen et al. 1984), 572 elderly people **randomized**, 287 to **control** (C) group (standard care) and 285 to **treatment** (T) group (standard care plus **in-home geriatric assessment** (IHGA): **preventive medicine** in which each person's medical/social needs assessed, acted upon individually).

One important **outcome** was **number of hospitalizations** (in two years): y_i^T, y_j^C = numbers of hospitalizations for **treatment** person i, **control** person j, respectively.

Conditional Exchangeability

Suppose **treatment/control** (T/C) status is **only available covariate**.

Unconditional judgment of exchangeability across all 572 outcomes **no longer automatically scientifically appropriate**.

Instead **design of experiment** compels (at least initially) judgment of conditional exchangeability **given T/C status** (e.g., de Finetti 1938, Draper et al. 1993), as in

$$\begin{array}{ccc} (F_T, F_C) & \sim & p(F_T, F_C) \\ (y_i^T | F_T, F_C) \overset{\text{IID}}{\sim} F_T & \Big| & (y_j^C | F_T, F_C) \overset{\text{IID}}{\sim} F_C \end{array} \qquad (5)$$

This framework, in which (a) **covariates** specify **conditional exchangeability judgments**, (b) de Finetti's **representation theorem** reduces model specification task to placing appropriate prior distributions on CDFs, covers much of field of **statistical inference/prediction**.

Data-Analytic Model Specification

Note that even in this **rather general nonparametric framework** it will be necessary to have a **good tool** for **discriminating between the quality of two models** (here: **unconditional** exchangeability ($F_T = F_C$; T has **same effect** as C) versus **conditional** exchangeability ($F_T \neq F_C$; T and C effects **differ**)).

Basic problem of Bayesian **model choice:** Given future observables $y = (y_1, \ldots, y_n)$, I'm **uncertain** about y (**first-order**), but I'm also **uncertain about how to specify my uncertainty** about y (**second-order**); I want to cope with **both of these kinds of uncertainty** in a **well-calibrated** manner.

Standard (**data-analytic**) approach to model specification involves initial choice, for **structure** of model, of **standard parametric family**, followed by **modification** of initial choice—once data begin to arrive—if data suggest **deficiencies** in original specification.

This approach (e.g., Draper 1995) is **incoherent** (unless I pay an **appropriate price** for **shopping around** for the model).

Cromwell's Rule

The **data-analytic** approach uses data both to specify **prior distribution on structure space** and to **update** using **data-determined prior** (result will typically be **uncalibrated** (too narrow) predictive distributions for future data).

Dilemma is example of **Cromwell's Rule** (if $p(\theta) = 0$ then $p(\theta|y) = 0$ for all y): initial model choice placed **0 prior probability** on **large regions of model space**; formally all such regions **must also have 0 posterior probability** even if data indicate **different prior on model space** would have been better.

Two possible solutions:

- **BNP** (which solves the problem by **"not putting zero probability on anything"**), and
- **3CV** (a modification of the usual **cross-validation** approach, which solves the problem by **paying an appropriate price for model exploration**).

Two Solutions: BNP and 3CV

- If use prior on F that places **non-zero probability on all Kullback-Leibler neighborhoods of all densities** (Walker et al. 2003; e.g., Pólya trees, Dirichlet process mixture priors, when chosen well), then BNP **directly avoids** Cromwell's Rule dilemma, at least for large n: as $n \rightarrow \infty$ posterior on F will **shrug off** any incorrect details of prior specification, will **fully adapt** to actual data-generating F ($\boxed{\textbf{NB}}$ this assumes **correct exchangeability judgments**).

- **Three-way cross-validation** (3CV; Draper and Krnjajić 2007): taking usual cross-validation idea one step further,

(1) Partition data at random into *three* (non-overlapping and exhaustive) subsets S_i.

(2) Fit tentative {likelihood + prior} to S_1. **Expand** initial model in all feasible ways suggested by data exploration using S_1. **Iterate** until you're happy.

3CV (continued)

(3) Use final model (fit to S_1) from (2) to create predictive distributions for all data points in S_2. Compare actual outcomes with these distributions, checking for **predictive calibration**. Go back to (2), change likelihood as necessary, **retune prior** as necessary, to get good calibration. **Iterate** until you're happy.

(4) Announce **final model** (fit to $S_1 \cup S_2$) from (3), and report **predictive calibration** of this model on data points in S_3 as indication of how well it would perform with new data.

With **large** n probably only need to do this **once**; with **small** and **moderate** n probably best to **repeat** (1–4) several times and **combine** results in some appropriate way (e.g., **model averaging**).

How large should the S_i be? (**Preliminary answer** below.)

Model Selection as Decision Problem

Given method like 3CV which permits **hunting around in model space** without forfeiting calibration, two kinds of model specification questions (in both **parametric** and **nonparametric** Bayesian modeling) arise:

(1) Is M_1 **better than** M_2? (this tells me **when it's OK to discard a model in my search**)

(2) Is M_1 **good enough**? (this tells me **when it's OK to stop searching**)

It would seem self-evident that **to specify a model you have to say to what purpose the model will be put**, for how else can you answer these two questions?

Specifying this purpose demands **decision-theoretic basis for model choice** (e.g., Draper 1996; Key et al. 1998).

To take **two examples**,

(Case 1) If you're going to choose which of several ways to behave in future, then the model has to be **good enough to reliably aid in choosing the best behavior**.

Choosing Utility Function

(Case 2) If you wish to make scientific summary of what's known, then—remembering that hallmark of good science is good prediction—the model has to be **good enough to make sufficiently accurate predictions of observable outcomes** (in which dimensions along which accuracy is to be monitored are driven by what's **scientifically relevant**).

A detailed look at Case 1 (Fouskakis and Draper, *JASA*, 2008; Fouskakis, Ntzoufras and Draper (FND), submitted, 2007a, 2007b). In the field of **quality of health care measurement**, patient **sickness at admission** is traditionally assessed by using **logistic regression** of **mortality within 30 days of admission** on a fairly large number of **sickness indicators** (on the order of **100**) to construct a sickness scale, employing standard **variable selection methods** (e.g., **backward selection** from a model with all predictors) to find an **"optimal"** subset of **10–20** indicators.

Such **"benefit-only"** methods ignore the considerable **differences** among the sickness indicators in **cost of data collection**, an issue that's **crucial** when admission sickness is used to drive programs (now implemented or

Choosing Utility Function (continued)

under consideration in several countries, including the U.S. and U.K.) that attempt to **identify substandard hospitals** by comparing **observed** and **expected mortality rates (given admission sickness)**.

When both **data-collection cost** and **accuracy of prediction** of 30-day mortality are considered, a large **variable-selection problem** arises in which **costly variables** that **do not predict well enough** should be **omitted** from the final scale.

We argue that there are **three main ways** to solve this problem:

(1) a **decision-theoretic cost-benefit approach** based on **maximizing expected utility** (Fouskakis and Draper, 2008),

(2) an **alternative cost-benefit approach** based on **posterior model odds** (FND, 2007a), and

(3) a **cost-restriction-benefit analysis** that **maximizes predictive accuracy** subject to a **bound on cost** (FND, 2007b).

The Data

$\boxed{\textbf{Data}}$ (Kahn et al., *JAMA*, 1990): $p = 83$ **sickness indicators** gathered on **representative sample** of $n = 2,532$ elderly American patients hospitalized in the period 1980–86 with **pneumonia**; original RAND **benefit-only scale** based on **subset** of 14 predictors:

Variable	Cost (U.S.$)	Correlation	Good?
Total APACHE II score (36-point scale)	3.33	0.39	
Age	0.50	0.17	$*$
Systolic blood pressure score (2-point scale)	0.17	0.29	$**$
Chest X-ray congestive heart failure score (3-point scale)	0.83	0.10	
Blood urea nitrogen	0.50	0.32	$**$
APACHE II coma score (3-point scale)	0.83	0.35	$**$
Serum albumin (3-point scale)	0.50	0.20	$*$
Shortness of breath (yes, no)	0.33	0.13	$**$
Respiratory distress (yes, no)	0.33	0.18	$*$
Septic complications (yes, no)	1.00	0.06	
Prior respiratory failure (yes, no)	0.67	0.08	
Recently hospitalized (yes, no)	0.67	0.14	
Ambulatory score (3-point scale)	0.83	0.22	
Temperature	0.17	-0.16	$*$

Decision-Theoretic Cost-Benefit Approach

$\boxed{\textbf{Approach (1)}}$ (decision-theoretic cost-benefit). **Problem formulation:** Suppose (a) the 30–day **mortality outcome** y_i and data on p **sickness indicators** $(x_{i1}, \dots, X_{ip})$ have been collected on n individuals sampled exchangeably from a **population** $\mathcal{P}$ of patients with a given disease, and (b) the goal is to **predict** the death outcome for n^* **new patients** who will in the future be sampled exchangeably from $\mathcal{P}$, (c) on the basis of some or all of the predictors $X_{\cdot j}$, when (d) the **marginal costs of data collection** per patient $c_1, \dots, c_p$ for the $X_{\cdot j}$ **vary considerably**.

What is the **best subset** of the $X_{\cdot j}$ to choose, if a **fixed amount of money** is available for this task and you're **rewarded** based on the **quality** of your predictions?

Since data on **future patients** are **not available**, we use a **cross-validation** approach in which (i) a random subset of n_M observations is drawn for creation of the mortality predictions (the **modeling** subsample) and (ii) the quality of those predictions is assessed on the remaining $n_V = (n - n_M)$ observations (the **validation** subsample, which serves as a proxy for future patients).

Utility Elicitation

In our approach **utility** is quantified in **monetary terms**, so that the **data collection** part of the **utility function** is simply the **negative** of the **total amount of money** required to gather data on the specified predictor subset.

Letting $I_j = 1$ if $X_{\cdot j}$ is included in a given model (and 0 otherwise), the **data-collection utility** associated with subset $I = (I_1, \ldots, I_p)$ for patients in the **validation subsample** is

$$U_D(I) = -n_V \sum_{j=1}^{p} c_j I_j, \tag{6}$$

where c_j is the **marginal cost per patient of data abstraction** for variable j (the second column in the table above gave examples of these marginal costs).

To measure the **accuracy** of a model's predictions, a metric is needed that quantifies the **discrepancy** between the actual and predicted values, and in this problem **the metric must come out in monetary terms** on a scale comparable to that employed with the data-collection utility.

Utility Elicitation (continued)

In the setting of this problem the outcomes Y_i are **binary death indicators** and the **predicted values** $\hat{p}_i$, based on statistical modeling, take the form of **estimated death probabilities**.

We use an approach to the comparison of **actual** and **predicted** values that involves **dichotomizing** the $\hat{p}_i$ with respect to a **cutoff**, to mimic the decision-making reality that **actions** taken on the basis of observed-versus-expected quality assessment will have an **all-or-nothing character** at the hospital level (for example, regulators must decide either to subject or not subject a given hospital to a more detailed, more expensive quality audit based on **process criteria**).

In the first step of our approach, given a particular **predictor subset** I, we fit a **logistic regression model** to the **modeling** subsample M and apply this model to **validation** subsample V to create predicted death probabilities $\hat{p}_i^I$.

In more detail, letting $Y_i = 1$ if patient i dies and 0 otherwise, and taking $X_{i1}, \ldots, X_{ik}$ to be the k **sickness predictors** for this patient under model I, the usual **sampling model** which underlies logistic regression in this case is

Utility Elicitation (continued)

$$\begin{aligned} (Y_i \,|\, p_i^I) &\stackrel{\text{indep}}{\sim} \text{Bernoulli}(p_i^I), \\ \log(\tfrac{p_i^I}{1-p_i^I}) &= \beta_0 + \beta_1 X_{i1} + \ldots + \beta_k X_{ik}. \end{aligned} \qquad (7)$$

We use **maximum likelihood** to fit this model (as a computationally efficient approximation to Bayesian fitting with relatively diffuse priors), obtaining a vector $\hat{\beta}$ of estimated logistic regression coefficients, from which the **predicted death probabilities** for the patients in subsample V are as usual given by

$$\hat{p}_i^I = \left[1 + \exp\left(- \sum_{j=0}^{k} \hat{\beta}_j X_{ij} \right) \right]^{-1}, \qquad (8)$$

where $X_{i0} = 1$ ($\hat{p}_i^I$ may be thought of as the **sickness score** for patient i under model I).

In the second step of our approach we **classify** patient i in the validation subsample as **predicted dead or alive** according to whether $\hat{p}_i^I$ exceeds or falls short of a **cutoff** p^*, which is chosen — by searching on a discrete grid from 0.01 to 0.99 by steps of 0.01 — to **maximize the predictive accuracy** of model I.

Utility Elicitation (continued)

We then cross-tabulate actual versus predicted death status in a 2×2 **contingency table**, **rewarding** and **penalizing** model I according to the numbers of patients in the **validation sample** which fall into the cells of the right-hand part of the following table.

		Rewards and Penalties		Counts	
		Predicted		Predicted	
		Died	Lived	Died	Lived
Actual	Died	C_{11}	C_{12}	n_{11}	n_{12}
	Lived	C_{21}	C_{22}	n_{21}	n_{22}

The left-hand part of this table records the **rewards and penalties** in US$.

The **predictive utility** of model I is then

$$U_P(I) = \sum_{l=1}^{2} \sum_{m=1}^{2} C_{lm} \, n_{lm} \, . \tag{9}$$

To **elicit** the **utility values** C_{lm} we reason as follows.

Utility Elicitation (continued)

(1) Clearly C_{11} (the **reward** for correctly predicting death at 30 days) and C_{22} (the **reward** for correctly predicting living at 30 days) should be **positive**, and C_{12} (the **penalty** for a false prediction of living) and C_{21} (the **penalty** for a false prediction of death) should be **negative**.

(2) Since it is **easier** to correctly predict that a person lives than dies with these data (the overall pneumonia 30–day death rate in the RAND sample was 16%, so a prediction that every patient lives would be right about **84%** of the time), it is natural to specify that $C_{11} > C_{22}$.

(3) Since it is arguably **worse** to label a "bad" hospital as "good" than the other way around, one should take $|C_{12}| > |C_{21}|$, and furthermore it is natural that the magnitudes of the **penalties** should exceed those of the **rewards**.

(4) We completed the utility specification by **eliciting** information from **health experts** in the U.S. and U.K, first to **anchor** C_{21} to the cost of subjecting a "good" hospital to an unnecessary process audit and then to obtain **ratios** relating the other C_{lm} to C_{21}.

Utility Elicitation (continued)

Since the **utility structure** we use is based on the idea that hospitals have to be treated in an **all-or-nothing** way in acting on the basis of their apparent quality, the approach taken was (i) to quantify the **monetary loss** L of incorrectly subjecting a "good" hospital to a detailed but unnecessary process audit and then (ii) to **translate** this from the hospital to the patient level.

Rough **correspondence** may be made between left-hand part of contingency table above at **patient level** and **hospital-level** table with rows representing **truth** ("bad" in row 1, "good" in row 2) and columns representing **decision taken** ("process audit" in column 1, "no process audit" in column 2).

Unnecessary process audits then correspond to cell $(2, 1)$ in these tables (hospitals where a process audit is **not needed** will typically have an **excess** of patients who are predicted to die but actually live).

Discussions with health experts in the U.S. and U.K. suggested that **detailed process audits** cost on the order of $L =$**$5,000** per hospital (in late 1980s U.S. dollars), and RAND data indicated that the mean number of pneumonia patients per hospital per year in the U.S. at the time of the RAND quality of care study was **71.8**.

Utility Elicitation (continued)

This **fixed** C_{21} at approximately $\frac{-\$5{,}000}{71.8} = -\$69.6$.

Our **health experts** judged that C_{12} should be the **largest** in absolute value of the C_{lm}, and averaging across the expert opinions, expressed as orders of magnitude base 2, the elicitation results were $\left| \frac{C_{12}}{C_{21}} \right| = 2$, $\left| \frac{C_{11}}{C_{21}} \right| = \frac{1}{2}$, and $\left| \frac{C_{22}}{C_{21}} \right| = \frac{1}{8}$, finally yielding $(C_{11}, C_{12}, C_{21}, C_{22}) =$ **\$(34.8, –139.2, –69.6, 8.7)**.

The results in Fouskakis and Draper (2008) use these values; Draper and Fouskakis (2000) present a **sensitivity analysis** on the choice of the C_{lm} which demonstrates **broad stability** of the findings when the utility values mentioned above are **perturbed** in reasonable ways.

With the C_{lm} in hand, the **overall expected utility function** to be maximized over I is then simply

$$E\left[U(I)\right] = E\left[U_D(I) + U_P(I)\right], \tag{10}$$

where this expectation is over **all possible cross-validation splits** of the data.

Results

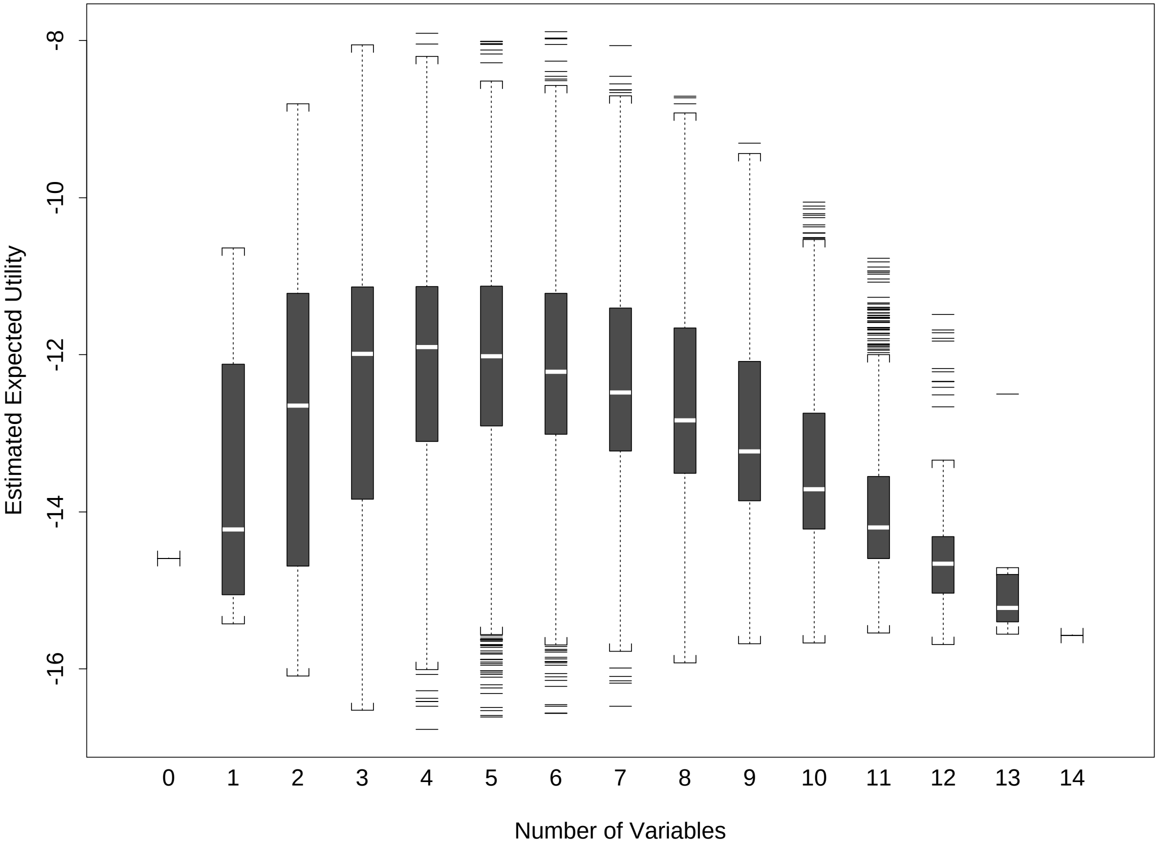
The number of possible cross-validation splits is **far too large** to evaluate the expectation in (10) directly; in practice we therefore use **Monte Carlo methods** to evaluate it, **averaging** over N random modeling and validation **splits**.

$\boxed{\textbf{Results.}}$ We explored this approach in **two settings**:

- a **Small World** created by focusing only on the $p = 14$ variables in the **original RAND scale** ($2^{14} = 16,384$ is a **small enough number of possible models** to do **brute-force enumeration** of the estimated expected utility of all models), and

- the **Big World** defined by all $p = 83$ available predictors ($2^{83} \doteq 10^{25}$ is **far too large** for brute-force enumeration; we compared a variety of **stochastic optimization methods** — including **simulated annealing**, **genetic algorithms**, and **tabu search** — on their ability to find **good variable subsets**).

Results: Small World



The **20 best models** included the **same three variables** 18 or more times out of 20, and never included six other variables; the **five best models** were minor variations on each other, and included **4–6 variables** (last column in table on page 14).

Approach (2)

The best models **save almost $8 per patient** over the full 14-variable model; this would amount to **significant savings** if the observed-versus-expected assessment method were **applied widely**.

Approach (2) (alternative cost-benefit) **Maximizing expected utility**, as in Approach (1) above, is a natural Bayesian way forward in this problem, but (a) the elicitation process was **complicated** and **difficult** and (b) the **utility structure** we examine is only one of a number of plausible alternatives, with utility framed from **only one point of view**; the broader question for a decision-theoretic approach is **whose utility should drive the problem formulation**.

It is well known (e.g., Arrow, 1963; Weerahandi and Zidek, 1981) that **Bayesian decision theory** can be **problematic** when used **normatively** for **group decision-making**, because of **conflicts** in **preferences** among members of the group; in the context of the problem addressed here, it can be **difficult** to identify a **utility structure acceptable to all stakeholders** (including patients, doctors, hospitals, citizen watchdog groups, and state and federal regulatory agencies) in the quality-of-care-assessment process.

Approach (2) (continued)

As an **alternative**, in Approach (2) we propose a **prior distribution** that accounts for the **cost** of each variable and results in a set of **posterior model probabilities** which correspond to a **generalized cost-adjusted version of the Bayesian information criterion** (BIC).

This provides a **principled approach** to performing a **cost-benefit trade-off** that **avoids ambiguities** in identification of an **appropriate utility structure**.

Details. Bayesian **parametric model comparison** and **variable selection** are based on specifying a model m, its likelihood $f(\boldsymbol{y}|\boldsymbol{\theta}_m, m)$, the prior distribution of model parameters $f(\boldsymbol{\theta}_m|m)$ and the corresponding prior model weight (or probability) $f(m)$, where $\boldsymbol{\theta}_m$ is a parameter vector under model m and $\boldsymbol{y}$ is the data vector.

Parametric inference is based on the posterior distribution $f(\boldsymbol{\theta}_m|\boldsymbol{y}, m)$, and quantifying **model uncertainty** by estimating the posterior model probability $f(m|\boldsymbol{y})$ is also an important issue.

Parametric Model Comparison

Hence, when we consider a set of competing models $\mathcal{M} = \{m_1, m_2, \cdots, m_{|\mathcal{M}|}\}$, we focus on the **posterior probability** of model $m \in \mathcal{M}$, defined as

$$\begin{aligned} f(m|\boldsymbol{y}) &= \frac{f(\boldsymbol{y}|m)f(m)}{\sum_{m_l \in \mathcal{M}} f(\boldsymbol{y}|m_l)f(m_l)} = \left(\sum_{m_l \in \mathcal{M}} PO_{m_l,m} \right)^{-1} \qquad (11) \\ &= \left[\sum_{m_l \in \mathcal{M}} B_{m_l,m} \frac{f(m_l)}{f(m)} \right]^{-1}, \end{aligned}$$

where $PO_{m_i,m_j} = \frac{f(m_i|\boldsymbol{y})}{f(m_j|\boldsymbol{y})}$ is the **posterior model odds** and B_{m_i,m_j} is the **Bayes factor** for comparing models m_i and m_j.

When we limit ourselves in the comparison of only **two models** we typically focus on PO_{m_i,m_j} and B_{m_i,m_j}, which have the desirable property of **insensitivity** to the selection of the model space $\mathcal{M}$.

By definition the **Bayes factor** is the ratio of the **posterior model odds** over the **prior model odds**; thus **large values** of B_{m_i,m_j} (usually greater than **12**, say) indicate **strong posterior support** of model m_i against model m_j.

Variable Selection in Logistic Regression

The **posterior model probabilities** and **integrated likelihoods** $f(\boldsymbol{y}|m_i)$ in (11) are **rarely analytically tractable**; we use a combination of **Laplace approximations** and **Markov Chain Monte Carlo** (MCMC) methodology to approximate posterior odds and Bayes factors.

In the sickness-at-admission problem at issue here, we use a simple **logistic regression** model with response $Y_i = 1$ if patient i dies and 0 otherwise.

We further denote by X_{ij} the **sickness predictor variable** j for patient i and by γ_j an **indicator**, often used in Bayesian variable selection problems, taking the value 1 if variable j is included in the model and 0 otherwise; thus in this case $\mathcal{M} = \{0,1\}^p$, where p is the total number of variables.

In order to map the set of **binary model indicators** $\boldsymbol{\gamma}$ onto a model m we can use a **representation** of the form $m(\boldsymbol{\gamma}) = \sum_{i=1}^p 2^{i-1}\gamma_i$.

Hence the **model formulation** can be summarized as

$$\begin{aligned} (Y_i \mid \boldsymbol{\gamma}) &\overset{\text{indep}}{\sim} \text{Bernoulli}[p_i(\boldsymbol{\gamma})], \\ \eta_i(\boldsymbol{\gamma}) = \log\left[\frac{p_i(\boldsymbol{\gamma})}{1-p_i(\boldsymbol{\gamma})}\right] &= \sum_{j=0}^{p} \beta_j \gamma_j X_{ij}, \end{aligned} \qquad (12)$$

Prior on Model Parameters

$$\boldsymbol{\eta}(\boldsymbol{\gamma}) = \boldsymbol{X}\,\mathrm{diag}(\boldsymbol{\gamma})\,\boldsymbol{\beta} = \boldsymbol{X}_{\boldsymbol{\gamma}}\,\boldsymbol{\beta}_{\boldsymbol{\gamma}},$$

defining $X_{i0} = 1$ for all $i = 1, \ldots, n$ and $\gamma_0 = 1$ with **prior probability one** since here the intercept is always included in all models.

Here $p_i(\boldsymbol{\gamma})$ is the **death probability** (which may be thought of as the **sickness score**) for patient i under model $\boldsymbol{\gamma}$, $\boldsymbol{\eta}(\boldsymbol{\gamma}) = [\eta_1(\boldsymbol{\gamma}), \ldots, \eta_n(\boldsymbol{\gamma})]^T$, $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \ldots, \gamma_p)^T$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$, and $\boldsymbol{X} = (X_{ij}, i = 1, \ldots, n; j = 0, 1, \ldots, p)$; the vector $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ stands for the subvector of $\boldsymbol{\beta}$ which is included in the model specified by $\boldsymbol{\gamma}$, i.e., $\boldsymbol{\beta}_{\boldsymbol{\gamma}} = (\beta_i : \gamma_i = 1, i = 0, 1, \ldots, p)$, and is equivalent to the $\boldsymbol{\theta}_m$ vector defined above; similarly $\boldsymbol{X}_{\boldsymbol{\gamma}}$ is the submatrix of $\boldsymbol{X}$ with columns corresponding to variables included in the model specified by $\boldsymbol{\gamma}$.

Prior on model parameters. We proceed in **two steps**:

(1) First we build a **prior** on $\boldsymbol{\beta}$ that is a modified version of the **unit information prior** for this problem (to avoid **Lindley's paradox**); then

(2) We **adjust** this prior for **differences** in **marginal costs** of variables.

Sensitivity to Prior Variance

$\boxed{\textbf{Step (1).}}$ One important problem in **Bayesian model evaluation** using **posterior model probabilities** is their **sensitivity** to the **prior variance** of the model parameters: large variance of the $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ (used to represent prior ignorance) will **increase** the posterior probabilities of the **simpler** models considered in the model space $\mathcal{M}$ (Lindley's paradox).

We address this issue by using ideas proposed by Ntzoufras *et al.* (2003): we use a **prior distribution** of the form

$$f(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\boldsymbol{\gamma}) = N(\boldsymbol{\mu}_{\boldsymbol{\gamma}}, \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}) \tag{13}$$

with **prior covariance matrix** given by $\boldsymbol{\Sigma}_{\boldsymbol{\gamma}} = n\left[\mathcal{I}(\boldsymbol{\beta}_{\boldsymbol{\gamma}})\right]^{-1}$, where n is the total sample size and $\mathcal{I}(\boldsymbol{\beta}_{\boldsymbol{\gamma}})$ is the information matrix

$$\mathcal{I}(\boldsymbol{\beta}_{\boldsymbol{\gamma}}) = \boldsymbol{X}_{\boldsymbol{\gamma}}^T \boldsymbol{W}_{\boldsymbol{\gamma}} \boldsymbol{X}_{\boldsymbol{\gamma}};$$

here $\boldsymbol{W}_{\boldsymbol{\gamma}}$ is a diagonal matrix which in the Bernoulli case takes the form

$$\boldsymbol{W}_{\boldsymbol{\gamma}} = \text{diag}\left\{p_i(\boldsymbol{\gamma})[1 - p_i(\boldsymbol{\gamma})]\right\}.$$

Unit Information Prior

This is the **unit information prior** of Kass and Wasserman (1996), which corresponds to adding **one data point** to the data.

Here we use this prior as a **base**, but we specify $p_i(\boldsymbol{\gamma})$ in the information matrix according to our prior information; in this manner we **avoid (even minimal) reuse** of the data in the prior.

When little prior information is available, a reasonable prior mean for $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ is $\boldsymbol{\mu}_{\boldsymbol{\gamma}} = \mathbf{0}$.

This corresponds to a prior mean on the log-odds scale of zero, from which a sensible prior estimate for all model probabilities is $p_i(\boldsymbol{\gamma}) = 1/2$; with this choice (13) becomes

$$f(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\boldsymbol{\gamma}) = N\left[\mathbf{0}, 4n\left(\boldsymbol{X}_{\boldsymbol{\gamma}}^T \boldsymbol{X}_{\boldsymbol{\gamma}}\right)^{-1}\right]. \qquad (*) \tag{14}$$

This prior distribution can also be motivated by combining the idea of **imaginary data** with the **power prior** approach of Chen *et al.* (2000); it turns out that (14) introduces additional information to the posterior equivalent to adding **one data point** to the likelihood and therefore we support **a priori** the simplest model with a weight of one data point.

Laplace Approximation

$\boxed{\textbf{Step (2).}}$ To introduce **costs** we again proceed in **two sub-steps**:

(2a) First we specify a **Laplace approximation** (and the **BIC approximation** that corresponds to it) for the **posterior model odds** in our problem, using the **prior** in Step (1), and

(2b) Then we see how to **adjust** the approximations in Step (2a) to account for **cost differences** among the variables.

$\boxed{\textbf{Step (2a).}}$ We denote by $PO_{k\ell}$ the **posterior odds** of model $\boldsymbol{\gamma}^{(k)}$ versus model $\boldsymbol{\gamma}^{(\ell)}$; then we have

$$-2\log PO_{k\ell} = -2\left[\log f(\boldsymbol{\gamma}^{(k)}|\boldsymbol{y}) - \log f(\boldsymbol{\gamma}^{(\ell)}|\boldsymbol{y})\right]. \tag{15}$$

Following the approach of Raftery (1996), we can approximate the posterior distribution of a model $\boldsymbol{\gamma}$ using the following **Laplace approximation**:

$$\begin{aligned} -2\log f(\gamma|\boldsymbol{y}) \quad = \quad & -2\log f(\boldsymbol{y}|\tilde{\boldsymbol{\beta}}_{\boldsymbol{\gamma}},\gamma) - 2\log f(\tilde{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}|\gamma) - d_{\boldsymbol{\gamma}}\log(2\pi) \\ & -\log|\boldsymbol{\Psi}_{\boldsymbol{\gamma}}| - 2\log f(\gamma) + O(n^{-1}), \end{aligned} \tag{16}$$

Details

where $\tilde{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}$ is the posterior mode of $f(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\boldsymbol{y},\gamma)$, $d\boldsymbol{\gamma} = \sum_{j=0}^{p} \gamma_j$ is the dimension of the model $\boldsymbol{\gamma}$, and $\boldsymbol{\Psi}_{\boldsymbol{\gamma}}$ is minus the inverse of the Hessian matrix of $h(\boldsymbol{\beta}_{\boldsymbol{\gamma}}) = \log f(\boldsymbol{y}|\boldsymbol{\beta}_{\boldsymbol{\gamma}},\gamma) + \log f(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\gamma)$ evaluated at the posterior mode $\tilde{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}$.

Under the **model formulation** given by equation (12) and the **prior distribution** (14) we have that

$$\begin{aligned}\boldsymbol{\Psi}_{\boldsymbol{\gamma}} &= \left[- \left. \frac{\partial^2 \log f(\boldsymbol{y}|\boldsymbol{\beta}_{\boldsymbol{\gamma}},\gamma)}{\partial \boldsymbol{\beta}_{\boldsymbol{\gamma}}^2} \right|_{\boldsymbol{\beta}_{\boldsymbol{\gamma}}=\tilde{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}} - \left. \frac{\partial^2 \log f(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\gamma)}{\partial \boldsymbol{\beta}_{\boldsymbol{\gamma}}^2} \right|_{\boldsymbol{\beta}_{\boldsymbol{\gamma}}=\tilde{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}} \right]^{-1} \\ &= \left(\boldsymbol{X}_{\boldsymbol{\gamma}}^T \,\mathrm{diag} \left\{ \frac{\exp\left(\boldsymbol{X}_{\boldsymbol{\gamma},i}\, \tilde{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}\right)}{\left[1+\exp\left(\boldsymbol{X}_{\boldsymbol{\gamma},i}\, \tilde{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}\right)\right]^2} + \frac{1}{4n} \right\} \boldsymbol{X}_{\boldsymbol{\gamma}} \right)^{-1}, \qquad (17)\end{aligned}$$

where $\boldsymbol{X}_{\boldsymbol{\gamma},i}$ is row i of the matrix $\boldsymbol{X}_{\boldsymbol{\gamma}}$ for $i = 1, \ldots, n$.

By substituting the **prior** (14) in expression (16) we get

$$-2\log f(\gamma|\boldsymbol{y}) = -2\log f(\boldsymbol{y}|\tilde{\boldsymbol{\beta}}_{\boldsymbol{\gamma}},\gamma) + \phi(\gamma) - 2\log f(\gamma) + O(n^{-1}), \qquad (18)$$

Penalized Log Likelihood Ratio

$$\text{where } \phi(\boldsymbol{\gamma}) = \frac{1}{4n}\tilde{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}^{T}\boldsymbol{X}_{\boldsymbol{\gamma}}^{T}\boldsymbol{X}_{\boldsymbol{\gamma}}\tilde{\boldsymbol{\beta}}_{\boldsymbol{\gamma}} + d_{\boldsymbol{\gamma}}\log(4n) + \log\frac{|\boldsymbol{\Psi}_{\boldsymbol{\gamma}}^{-1}|}{|\boldsymbol{X}_{\boldsymbol{\gamma}}^{T}\boldsymbol{X}_{\boldsymbol{\gamma}}|}. \tag{19}$$

From the above expression it's clear that the logarithm of a posterior model probability can be regarded as a **penalized log-likelihood** evaluated at the posterior mode of the model, in which the term $\phi(\boldsymbol{\gamma}) - 2\log f(\boldsymbol{\gamma})$ can be interpreted as the **penalty** imposed upon the log-likelihood.

In **pairwise model comparisons**, we can directly use the **posterior model odds** (15), which can now be written as

$$\begin{aligned} -2\log PO_{k\ell} \quad = \quad & -2\log\left\{\frac{f(\boldsymbol{y}|\tilde{\boldsymbol{\beta}}_{\boldsymbol{\gamma}^{(k)}}, \boldsymbol{\gamma}^{(k)})}{f(\boldsymbol{y}|\tilde{\boldsymbol{\beta}}_{\boldsymbol{\gamma}^{(\ell)}}, \boldsymbol{\gamma}^{(\ell)})}\right\} + \phi\left(\boldsymbol{\gamma}^{(k)}\right) - \phi\left(\boldsymbol{\gamma}^{(\ell)}\right) \\ & -2\log\frac{f(\boldsymbol{\gamma}^{(k)})}{f(\boldsymbol{\gamma}^{(\ell)})} + O(n^{-1}). \end{aligned} \tag{20}$$

Therefore, the comparison of the two models is based on a **penalized log-likelihood ratio**, where the penalty is now given by

$$\psi(\boldsymbol{\gamma}^{(k)}, \boldsymbol{\gamma}^{(\ell)}) = \phi(\boldsymbol{\gamma}^{(k)}) - \phi(\boldsymbol{\gamma}^{(\ell)}) - 2\log\frac{f(\boldsymbol{\gamma}^{(k)})}{f(\boldsymbol{\gamma}^{(\ell)})}.$$

Decomposing the Penalty Term

Each **penalty term** is divided into **two parts**: $\phi(\boldsymbol{\gamma})$ and $-2\log f(\boldsymbol{\gamma})$.

The first term, $\phi(\boldsymbol{\gamma})$, has its source in the **marginal likelihood** $f(\boldsymbol{y}|\boldsymbol{\gamma})$ of model $\boldsymbol{\gamma}$ and can be thought of as a measure of **discrepancy** between the **data** and the **prior information** for the model parameters; the second part comes from the **prior model probabilities** $f(\boldsymbol{\gamma})$.

Indifference on the space of all models, usually expressed by the **uniform distribution** (i.e., $f(\boldsymbol{\gamma}) \propto 1$), eliminates the second term from the model comparison procedure, since the penalty term in (20) will then be based only on the difference of the first penalty terms $\phi(\boldsymbol{\gamma}^{(k)}) - \phi(\boldsymbol{\gamma}^{(\ell)})$.

For this reason the penalty term $\phi(\boldsymbol{\gamma})$ is the **imposed penalty** which appears in the penalized log-likelihood expression of the **Bayes factor** $BF_{k\ell}$ with a uniform prior on model space.

A **simpler** but **less accurate** approximation of $\log PO_{k\ell}$ can be obtained following the arguments of Schwartz (1978):

BIC Approximation

$$
\begin{aligned}
-2\log PO_{k\ell} &= -2\log\left[\frac{f(\boldsymbol{y}|\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}^{(k)}},\boldsymbol{\gamma}^{(k)})}{f(\boldsymbol{y}|\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}^{(\ell)}},\boldsymbol{\gamma}^{(\ell)})}\right] + \left(d_{\boldsymbol{\gamma}^{(k)}} - d_{\boldsymbol{\gamma}^{(\ell)}}\right)\log n \\
&\quad -2\log\frac{f(\boldsymbol{\gamma}^{(k)})}{f(\boldsymbol{\gamma}^{(\ell)})} + O(1) \qquad (21) \\
&= BIC_{k\ell} - 2\log\frac{f(\boldsymbol{\gamma}^{(k)})}{f(\boldsymbol{\gamma}^{(\ell)})} + O(1),
\end{aligned}
$$

where $BIC_{k\ell}$ is the **Bayesian Information Criterion** for choosing between models $\boldsymbol{\gamma}^{(k)}$ and $\boldsymbol{\gamma}^{(\ell)}$ and $\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}$ is vector of maximum likelihood estimates of $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$.

Since $BIC_{k\ell}$ is an $O(1)$ approximation, it might **diverge** from the exact value of the logarithm of the Bayes factor even for large samples; even so, it has often been shown to provide a **reasonable measure of evidence** (for finite n) and its straightforward calculation has encouraged its **widespread use** in practice.

Step (2b). From the above argument and equations (18) and (20), it's clear that an **additional penalty** can be directly imposed on the posterior model probabilities and odds via the **prior model probabilities** $f(\boldsymbol{\gamma})$.

Cost Adjustment

Therefore we may use **prior model probabilities** to **induce prior preferences** for specific variables depending on their **costs**.

For this reason we propose to use **prior model probabilities** of the form

$$(*) \qquad f(\gamma_j) \propto \exp\left[-\frac{\gamma_j}{2}\left(\frac{c_j - c_0}{c_0}\right)\log n\right] \text{ for } j = 1, \ldots, p, \tag{22}$$

where c_j is the **marginal cost per observation** for variable X_j and (as will be seen below) the desire for our approach to yield a **cost-adjusted generalization of BIC** compels the definition $c_0 = \min\{c_j, j = 1, \ldots, p\}$.

We further assume that the **constant term** is included in all models by specifying $f(\gamma_0 = 1) = 1$, resulting in

$$-2\log f(\boldsymbol{\gamma}) = \sum_{j=1}^{p} \gamma_j \frac{c_j}{c_0} \log n - d_{\boldsymbol{\gamma}} \log n + 2\sum_{j=1}^{p} \log\left[1 + n^{-\frac{1}{2}\left(1 - \frac{c_j}{c_0}\right)}\right]. \tag{23}$$

If all variables have the **same cost** or we're indifferent concerning the cost then we can set $c_j = c_0$ for $j = 1, \ldots, p$, which reduces to the **uniform prior** on model space ($f(\boldsymbol{\gamma}) \propto 1$) and posterior odds equal to the **usual Bayes factor**.

Cost Adjustment (continued)

When comparing two models $\boldsymbol{\gamma}^{(k)}$ and $\boldsymbol{\gamma}^{(\ell)}$, the **additional penalty** imposed on the log-likelihood ratio due to the **cost-adjusted prior model probabilities** is given by

$$\begin{aligned} -2\log\left[\frac{f(\boldsymbol{\gamma}^{(k)})}{f(\boldsymbol{\gamma}^{(\ell)})}\right] &= \sum_{j=1}^{p}\left(\gamma_j^{(k)}-\gamma_j^{(\ell)}\right)\frac{c_j}{c_0}\log n-\left(d_{\boldsymbol{\gamma}^{(k)}}-d_{\boldsymbol{\gamma}^{(\ell)}}\right)\log n \\ &= \left[\frac{C_{\boldsymbol{\gamma}^{(k)}}-C_{\boldsymbol{\gamma}^{(\ell)}}}{c_0}-\left(d_{\boldsymbol{\gamma}^{(k)}}-d_{\boldsymbol{\gamma}^{(\ell)}}\right)\right]\log n, \qquad (24) \end{aligned}$$

where $C_{\boldsymbol{\gamma}}=\sum_{j=1}^{p}\gamma_j c_j$ is the **total cost** of model $\boldsymbol{\gamma}$; thus two models of the **same dimension and cost** will have the **same prior weight**.

In the simpler case where we compare **two nested models** that differ only on the status of variable j, the prior model ratio simplifies to

$$-2\log\left[\frac{f(\gamma_j=1,\boldsymbol{\gamma}_{\setminus j})}{f(\gamma_j=0,\boldsymbol{\gamma}_{\setminus j})}\right]=\left(\frac{c_j}{c_0}-1\right)\log n, \qquad (25)$$

where $\boldsymbol{\gamma}_{\setminus j}$ is the vector of $\boldsymbol{\gamma}$ **excluding** element γ_j.

Cost-Adjusted Laplace Approximation

The above expression can be viewed as a **prior penalty** for including the variable j in the model, while the term $\left(\frac{c_j}{c_0} - 1\right)$ can be interpreted as the **proportional additional penalty** imposed upon $(-2\log BF)$ if the variable X_j is included in the model due to its **increased cost**.

Using the **prior model odds** (24) in the **approximate posterior model odds** (20) we obtain

$$-2\log PO_{k\ell} = -2\log\left[\frac{f(\boldsymbol{y}|\tilde{\boldsymbol{\beta}}_{\boldsymbol{\gamma}^{(k)}}, \boldsymbol{\gamma}^{(k)})}{f(\boldsymbol{y}|\tilde{\boldsymbol{\beta}}_{\boldsymbol{\gamma}^{(\ell)}}, \boldsymbol{\gamma}^{(\ell)})}\right] + \psi(\boldsymbol{\gamma}^{(k)}, \boldsymbol{\gamma}^{(\ell)}) + O(n^{-1}), \qquad (26)$$

where the **penalty term** is given by

$$\begin{aligned}\psi(\boldsymbol{\gamma}^{(k)}, \boldsymbol{\gamma}^{(\ell)}) \quad &= \quad \frac{1}{4n}\left(\tilde{\beta}^T_{\boldsymbol{\gamma}^{(k)}}\boldsymbol{X}^T_{\boldsymbol{\gamma}^{(k)}}\boldsymbol{X}_{\boldsymbol{\gamma}^{(k)}}\tilde{\beta}_{\boldsymbol{\gamma}^{(k)}} - \tilde{\beta}^T_{\boldsymbol{\gamma}^{(\ell)}}\boldsymbol{X}^T_{\boldsymbol{\gamma}^{(\ell)}}\boldsymbol{X}_{\boldsymbol{\gamma}^{(\ell)}}\tilde{\beta}_{\boldsymbol{\gamma}^{(\ell)}}\right) \\ &\quad + \left(d_{\boldsymbol{\gamma}^{(k)}} - d_{\boldsymbol{\gamma}^{(\ell)}}\right)\log(4) + \log\frac{|\boldsymbol{\Psi}^{-1}_{\boldsymbol{\gamma}^{(k)}}|}{|\boldsymbol{X}^T_{\boldsymbol{\gamma}^{(k)}}\boldsymbol{X}_{\boldsymbol{\gamma}^{(k)}}|} \qquad (27) \\ &\quad - \log\frac{|\boldsymbol{\Psi}^{-1}_{\boldsymbol{\gamma}^{(\ell)}}|}{|\boldsymbol{X}^T_{\boldsymbol{\gamma}^{(\ell)}}\boldsymbol{X}_{\boldsymbol{\gamma}^{(\ell)}}|} + \frac{C_{\boldsymbol{\gamma}^{(k)}} - C_{\boldsymbol{\gamma}^{(\ell)}}}{c_0}\log n.\end{aligned}$$

Cost-Adjusted BIC

Finally we consider the **BIC-based approximation** (21) to the logarithm of the posterior model odds with the prior model odds (24), yielding (*)

$$-2\log PO_{k\ell} = -2\log\left[\frac{f(\boldsymbol{y}|\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}^{(k)}}, \boldsymbol{\gamma}^{(k)})}{f(\boldsymbol{y}|\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}^{(\ell)}}, \boldsymbol{\gamma}^{(\ell)})}\right] + \frac{C_{\boldsymbol{\gamma}^{(k)}} - C_{\boldsymbol{\gamma}^{(\ell)}}}{c_0}\log n + O(1). \quad (28)$$

The **penalty term** $d_{\boldsymbol{\gamma}}\log n$ of model $\boldsymbol{\gamma}$ used in (21) has been replaced in the above expression by the **cost-dependent penalty** $c_0^{-1}C_{\boldsymbol{\gamma}}\log n$; **ignoring costs** is equivalent to taking $c_j = c_0$ for all j, yielding $c_0^{-1}C_{\boldsymbol{\gamma}} = d_{\boldsymbol{\gamma}}$, the **original BIC expression**.

Therefore, we may interpret the quantity $\log n$ as the **imposed penalty** for each variable included in the **model** $\boldsymbol{\gamma}$ **when no costs are considered** (or when costs are equal).

Moreover, this baseline penalty term is inflated **proportionally** to the cost ratio $\frac{c_j}{c_0}$ for each variable X_j; for example, if the cost of a variable X_j is **twice** the minimum cost ($c_j = 2\,c_0$) then the imposed penalty is equivalent to **adding two variables with the minimum cost**.

MCMC Implementation

For this reason, (28) can be considered as a **cost-adjusted generalization of BIC** when prior model probabilities of type (22) are adopted.

MCMC implementation. As noted earlier, in our quality of care study with $p = \mathbf{83}$ predictors there are on the order of $\mathbf{10^{25}}$ possible models.

In such situations, **sampling algorithms** will not be able to estimate posterior model probabilities with **high accuracy** in a reasonable amount of CPU time due to the **large model space**.

For this reason, we implemented the following **two-step method**:

(1) First we use a **model search tool** to identify variables with **high marginal posterior inclusion probabilities** $f(\gamma_j|\boldsymbol{y})$, and we create a **reduced model space** consisting only of those variables whose marginal probabilities are above a **threshold value**.

According to Barbieri and Berger (2004) this method of selecting variables based on their **marginal probabilities** may lead to the identification of models with **better predictive abilities** than approaches based on maximizing posterior model probabilities.

MCMC Implementation (continued)

Although Barbieri and Berger proposed **0.5** as a threshold value for $f(\gamma_j = 1|\boldsymbol{y})$, we used the lower value of **0.3**, since our aim was only to **identify and eliminate** variables not contributing to models with high posterior probabilities.

(2) Then we use a **model search tool** in the **reduced model space** to estimate **posterior model probabilities** (and the corresponding odds).

To ensure **stability** of our findings we explored the use of **two model search tools** in step (1):

- A **reversible-jump MCMC algorithm** (RJMCMC), as implemented for variable selection in generalized linear models by Dellaportas *et al.* (2002) and Ntzoufras *et al.*(2003); and
- the **MCMC model composition** (MC^3) algorithm (Madigan and York, 1995).

More specifically, we implemented **reversible-jump moves within Gibbs** for the model indicators γ_j, by proposing the new model to differ from the current one in each step by a **single term** j with probability one.

MCMC Implementation (continued)

The **algorithm** can be summarized as follows:

(1) For $j = 1, \ldots, p$, use **RJMCMC** to compare the current model $\boldsymbol{\gamma}$ with the proposed one $\boldsymbol{\gamma}'$ with components $\gamma'_j = 1 - \gamma_j$ and $\gamma'_k = \gamma_k$ for $k \neq j$ with probability one; the **updating sequence** of γ_j is randomly determined in each step.

(2) For $j = 0, \ldots, p$, if $\gamma_j = 1$ then **generate** model parameters β_j from the corresponding posterior distribution $f(\beta_j | \boldsymbol{\beta}_{\setminus j}, \boldsymbol{\gamma}, \boldsymbol{y})$, otherwise set $\beta_j = 0$.

In our context the $\boldsymbol{MC^3}$ **algorithm** may be summarized by the following steps:

(1) For $j = 1, \ldots, p$, propose a **move** from the current model $\boldsymbol{\gamma}$ to a new one $\boldsymbol{\gamma}'$ with components $\gamma'_j = 1 - \gamma_j$ and $\gamma'_k = \gamma_k$ for $k \neq j$ with probability one; the **updating sequence** of γ_j is randomly determined in each step.

(2) **Accept** the proposed model $\boldsymbol{\gamma}'$ with probability

$$\alpha = \min\left[1, \frac{f(\boldsymbol{\gamma}'|\boldsymbol{y})}{f(\boldsymbol{\gamma}|\boldsymbol{y})}\right] = \min\left(1, PO_{\boldsymbol{\gamma},\boldsymbol{\gamma}'}\right).$$

MCMC Implementation (continued)

Since the **posterior model odds** $PO_{\boldsymbol{\gamma},\boldsymbol{\gamma}'}$ used in MC^3 are **not analytically available** here, we also explored **two methods** for calculating them — approximating the acceptance probabilities with **cost-adjusted Laplace** (equation 26) and **cost-adjusted BIC** (equation 28) — and in addition we further explored one additional form of **sensitivity analysis**: initializing the MCMC runs at the **null model** (with no predictors) and the **full model** (with all predictors).

All of this was done both for the **benefit-only analysis** (specified by **setting all variable costs equal**) and the **cost-benefit approach**.

In moving from the **full** to the **reduced** model space to implement step (1) of our two-step method, for both the benefit-only and cost-benefit analyses we found a **striking level of agreement** — across (a) the two model search tools, (b) the two methods to approximate the acceptance probabilities in MC^3, and (c) the two choices for initializing the MCMC runs — in the **subset of variables** defining the reduced model space; this made it **unnecessary** to perform a similar sensitivity analysis in step (2).

Results

Results are therefore presented below **only for RJMCMC** (starting from the full model).

Convergence of the RJMCMC algorithm was checked using **ergodic mean plots** of the **marginal inclusion probabilities** for the full model space and the **posterior model probabilities** for the reduced space.

In what follows we refer to the **cost-benefit results** as "**RJMCMC**," but we could equally well have used the term "$\boldsymbol{MC^3}$ **with cost-adjusted BIC**" (or just "**cost-adjusted BIC**" for short), because the results from the two methods were in such **close agreement**.

$\boxed{\textbf{Results.}}$ The table below presents the **marginal posterior probabilities** of the variables that exceeded the threshold value of 0.30, in each of the **benefit-only** and **cost-benefit** analyses, together with their data collection costs (in minutes of abstraction time rather than US$), in the **Big World** of all 83 predictors.

In both the **benefit-only** and **cost-benefit** situations our methods reduced the initial list of $p = 83$ available candidates down to **13** predictors.

Results (continued)

			Marginal Posterior Probabilities	
	Variable		Analysis	
Index	Name	Cost	Benefit-Only	Cost-Benefit
1	SBP Score	0.50	0.99	0.99
2	Age	0.50	0.99	0.99
3	Blood Urea Nitrogen	1.50	1.00	0.99
4	Apache II Coma Score	2.50	1.00	
5	Shortness of Breath Day 1?	1.00	0.97	0.79
8	Septic Complications?	3.00	0.88	
12	Initial Temperature	0.50	0.98	0.96
13	Heart Rate Day 1	0.50		0.34
14	Chest Pain Day 1?	0.50		0.39
15	Cardiomegaly Score	1.50	0.71	
27	Hematologic History Score	1.50	0.45	
37	Apache Respiratory Rate Score	1.00	0.95	0.32
46	Admission SBP	0.50	0.68	0.90
49	Respiratory Rate Day 1	0.50		0.81
51	Confusion Day 1?	0.50		0.95
70	Apache pH Score	1.00	0.98	0.98
73	Morbid + Comorbid Score	7.50	0.96	
78	Musculoskeletal Score	1.00		0.54

Note that the **most expensive** variables with high marginal posterior probabilities in the **benefit-only** analysis were **absent** from the set of promising variables in the **cost-benefit** analysis (e.g., Apache II Coma Score).

Results (continued)

Common variables in both analyses: $X_1+X_2+X_3+X_5+X_{12}+X_{70}$

Benefit-Only Analysis

	Common Variables	Additional	Model	Posterior	
k	Within Each Analysis	Variables	Cost	Probabilities	PO_{1k}
1	$X_4+X_{15}+X_{37}+X_{73}$	$+X_8+X_{27}+X_{46}$	22.5	0.3066	1.00
2		$+X_8+X_{27}$	22.0	0.1969	1.56
3		$+X_8$	20.5	0.1833	1.67
4		$+X_{27}+X_{46}$	19.5	0.0763	4.02
5			17.5	0.0383	8.00

Cost-Benefit Analysis

	Common Variables	Additional	Model	Posterior	
k	Within Each Analysis	Variables	Cost	Probabilities	PO_{1k}
1	$X_{46}+X_{51}$	$+X_{49}+X_{78}$	7.5	0.1460	1.00
2		$+X_{14}+X_{49}+X_{78}$	7.5	0.1168	1.27
3		$+X_{13}+X_{49}+X_{78}$	7.5	0.0866	1.69
4		$+X_{13}+X_{14}+X_{49}+X_{78}$	8.0	0.0665	2.20
5		$+X_{14}+X_{49}$	7.0	0.0461	3.17
6		$+X_{49}$	6.5	0.0409	3.57
7		$+X_{37}+X_{78}$	7.5	0.0382	3.82
8		$+X_{13}+X_{14}+X_{49}$	7.5	0.0369	3.96
9		$+X_{13}$	6.5	0.0344	4.25

Results (continued)

	Analysis		Percentage
	Benefit-Only	Cost-Benefit	Difference
Minimum Deviance	1553.2	1635.8	+5.3
Median Deviance	1564.5	1644.8	+5.1
Cost	22.5	7.5	–66.7
Dimension	13	10	–23.1

The table above presents a comparison of **measures of fit, cost and dimensionality** between the best models in the reduced model space of the benefit-only and cost-benefit analyses (percentage difference is in relation to benefit-only).

- The **deviance statistic** for the benefit-only RAND model summarized in Table 1 turned out to be **1587.3** (achieved with **14** predictors), **substantially worse** than the median deviance (**1564.5**, achieved with **13** predictors) of the best model visited by the **benefit-only** approach we investigate; in other words, in this case study, **frequentist backward selection** from the model with all predictors (the RAND approach) was **substantially out-performed** by Bayesian RJMCMC.

Results (continued)

	Analysis		Percentage
	Benefit-Only	Cost-Benefit	Difference
Minimum Deviance	1553.2	1635.8	+5.3
Median Deviance	1564.5	1644.8	+5.1
Cost	22.5	7.5	–66.7
Dimension	13	10	–23.1

- The minimum and median values of the posterior distribution of the **deviance** statistic for the benefit-only analysis were **lower** by a **relatively modest 5.3% and 5.1%** compared to the corresponding values of the cost-benefit analysis, but the **cost** of the best model in the cost-benefit analysis was almost **67% lower** than that for the benefit-only analysis; similarly, the dimensionality of the best model in the cost-benefit analysis was about **23% lower** than that for the benefit-only analysis.

These values indicate that the **loss of predictive accuracy** with the **cost-benefit analysis** is **small** compared to the **substantial gains** achieved in **cost** and **reduced model complexity**.

Utility Versus Cost-Adjusted BIC

Variable			Method		
			Utility	RJMCMC	
		Cost			Posterior
Index	Name	(Minutes)	Good?	Good?	Probability
1	Systolic Blood Pressure Score (2-point scale)	0.5	**	**	0.99
2	Age	0.5	*	**	0.99
3	Blood Urea Nitrogen	1.5	**	**	1.00
4	APACHE II Coma Score (3-point scale)	2.5	**	**	1.00
5	Shortness of Breath Day 1 (yes, no)	1.0	**	**	0.99
6	Serum Albumin (3-point scale)	1.5	*	**	0.55
7	Respiratory Distress (yes, no)	1.0	*	**	0.92
8	Septic Complications (yes, no)	3.0			0.00
9	Prior Respiratory Failure (yes, no)	2.0			0.00
10	Recently Hospitalized (yes, no)	2.0			0.00
12	Initial Temperature	0.5	*	**	0.95
17	Chest X-ray Congestive Heart Failure Score (3-point scale)	2.5			0.00
18	Ambulatory Score (3-point scale)	2.5			0.00
48	Total APACHE II Score (36-point scale)	10.0			0.00

It's clear that the **utility** and **cost-adjusted BIC** approaches have reached **nearly identical conclusions** in the **Small World** of $p = 14$ predictors.

Utility Versus Cost-Adjusted BIC (continued)

With $p = 83$ the **agreement** between the two methods is also **strong** (although not as strong as with $p = 14$): using a **star system** for variable importance given in FND (2007a), **60** variables were **ignored** by both methods, **8** variables had **identical** star patterns, **3** variables were chosen as **important by both methods** but with different star patterns, **10** variables were marked as important by the utility approach and not by RJMCMC, and **2** variables were singled out by RJMCMC and not by utility: thus the two methods **substantially** agreed on the importance of **71 (86%) of the 83 variables**.

				Median	
p	Method	Model	Cost	Deviance	LS_{CV}
14	RJMCMC	$X_1 + X_2 + X_3 + X_4 + X_5 + X_6 + X_7 + X_{12}$	9.0	1654	-0.329
		$X_1 + X_2 + X_3 + X_4 + X_5 + X_7 + X_{12}$	7.5	1676	-0.333
	Utility	$X_1 + X_3 + X_4 + X_5$	5.5	1726	-0.342
83	RJMCMC	$X_1 + X_2 + X_3 + X_5 + X_{12}$ $+X_{46} + X_{49} + X_{51} + X_{70} + X_{78}$	7.5	1645	-0.327
	Utility	$X_1 + X_3 + X_4 + X_{12}$ $+X_{46} + X_{49} + X_{57}$	6.5	1693	-0.336

To the extent that the two methods **differ**, the **utility** method favors models that **cost somewhat less** but also **predict somewhat less well**.

Utility Versus Cost-Adjusted BIC (continued)

The fact that the **two methods** may yield **somewhat different results** in **high-dimensional problems** does not mean that either is **wrong**; they are both **valid solutions** to **similar but not identical problems**.

Both methods lead to **noticeably better models** (in a **cost-benefit** sense) than frequentist or Bayesian **benefit-only** approaches, when — as is often the case — **cost** is an issue that must be included in the **problem formulation** to arrive at a **policy-relevant solution**.

Summary. In **comparing two or more models**, to say whether one is **better** than another I have to face the question: **better for what purpose**?

This makes **model specification** a **decision problem**: I need to either

(a) elicit a **utility structure** that's specific to the **goals** of the current study and **maximize expected utility** to find the best models, or

(b) (if (a) is too hard, e.g., because the problem has a **group decision** character) I can look for a **principled alternative** (like the **cost-adjusted Laplace and BIC methods** described here) that **approximates** the utility approach while **avoiding ambiguities in utility specification**.

Utility Specification

Toward the beginning I noted that in **specifying the utility structure for model choice** there were **two cases** to distinguish:

(Case 1) If you're going to choose which of several ways to behave in the future, then the model has to be **good enough to reliably aid in choosing the best behavior** (the case study above on **variable selection in generalized linear models under cost constraints** was an example of this).

(Case 2) If you wish to make scientific summary of what's known, then—remembering that hallmark of good science is good prediction—the model has to be **good enough to make sufficiently accurate predictions of observable outcomes** (in which dimensions along which accuracy is to be monitored are driven by what's **scientifically relevant**).

How can a **utility function** driven by predictive accuracy be specified in a **reasonably general way** to answer **model specification question (1)** mentioned on page 11? (Is M_1 **better than** M_2?)

Need **scoring rule** that measures **discrepancy** between observation y^* and predictive distribution $p(\cdot|y, M_i)$ for y^* under model M_i given data y.

Log Score as Utility

As noted (e.g.) by Good (1950) and O'Hagan and Forster (2004), **the optimal (impartial, symmetric, proper)** scoring rules are linear functions of

$$\boxed{\log p(y^* | y)}.$$

On **calibration** grounds it would $\boxed{\textbf{seem}}$ to be a mistake to **use data twice** in measuring this sort of thing (once to make predictions, again with same data to see how good they are; but see below).

Out-of-sample predictive validation (e.g., Geisser and Eddy 1979, Gelfand et al. 1992) addresses this apparent concern directly: e.g., successively remove each observation y_j one at a time, construct predictive distribution for y_j based on y_{-j} (data vector with y_j removed), see where y_j falls in this distribution.

This motivates **cross-validated** version of **log scoring rule** (e.g., Gelfand and Dey 1994; Bernardo and Smith 1994): with n data values y_j, when choosing among k models $M_i, i \in I$, find that model M_i which maximizes

$$LS_{CV}(M_i | y) = \frac{1}{n} \sum_{j=1}^{n} \log p(y_j | M_i, y_{-j}). \tag{29}$$

Full-Sample Log Score

It has been argued that this can be given direct **decision-theoretic justification**: with utility function for model i

$$U(M_i|y) = \log p(y^*|M_i, y), \tag{30}$$

where y^* is **future data value**, expectation in MEU is over **uncertainty about** y^*; Gelfand et al. (1992) and Bernardo and Smith (1994) claim that this expectation can be accurately **estimated** (assuming exchangeability) by LS_{CV}.

Draper and Krnjajić (2008) **show** that

(a) another **popular model choice method**, the **Deviance Information Criterion** (DIC; Spiegelhalter et al., 2002), can be viewed as a computationally-fast **approximation** to LS_{CV}, but DIC can be **unstable** with respect to **model parameterization**;

(b) an alternative log-scoring criterion, the **full-sample log score** LS_{FS}, specified (e.g.) in the one-sample situation by computing a **single predictive distribution** $p^*(\cdot|y, M_i)$ for a future data value with each model M_i under consideration, based on the **entire data set** y (without omitting any observations), and then defining (cf. Laud and Ibrahim 1995)

LS_{FS}

$$LS_{FS}(M_i|y) = \frac{1}{n}\sum_{j=1}^{n} \log p^*(y_j|y, M_i), \tag{31}$$

has the following **desirable properties**:

(i) The **naive** approach to calculating LS_{CV}, when MCMC is needed to compute the predictive distributions, requires n MCMC runs, **one for each omitted observation**; by contrast LS_{FS} needs only a **single** MCMC run, making its computational speed (a) n **times faster** than naive implementations of LS_{CV} and (b) **equivalent** to that of DIC.

(ii) The **log score approach** works equally well with **parametric** and **nonparametric** Bayesian models; DIC is **only defined for parametric models**.

(iii) LS_{FS} has both (a) **superior small-sample behavior** in relation to LS_{CV} and DIC and (b) **superior asymptotic behavior** in relation to LS_{CV} in **correctly discriminating** between two or more models.

Conclusions

- Standard (**data-analytic**) (DA) approach to **model specification** **“shops around** for the **‘right’** model,” thereby often yielding **poorly calibrated (too narrow) predictive intervals** (a symptom of **incoherence**).

- I’m aware of **two principled solutions** to this problem:

– {**Exchangeability judgments** plus **Bayesian nonparametric (BNP)** modeling} (this solves the problem by **avoiding (some of) the shopping**: with BNP (and enough data), **no need to use DA** to specify many modeling details (**error distributions, response surfaces**); but will often still need to **compare** models with **different sets of exchangeability judgments**); and

– **3-way out-of-sample predictive cross-validation (3CV)**, a modification of DA in which the data are **partitioned** into **3** (rather than the usual 2) subsets S_1, S_2, S_3; a **DA search** is undertaken iteratively, **modeling** with S_1 and **predictively validating** with S_2; and S_3 is **not used in quoting final uncertainty assessments**, but is instead used to **evaluate predictive calibration of the entire modeling process** (this solves the problem by paying the **“right” price** for shopping around).

Conclusions (continued)

- **Two basic kinds of model choices** need to be made in **both BNP and 3CV**: $\boxed{Q_1}$ Is M_1 **better than** M_2? $\boxed{Q_2}$ Is M_1 **good enough**?

- ($\boxed{Q_1}$ and $\boxed{Q_2}$) Model choice is really a **decision problem** and should be approached via **MEU**, with a utility structure that's **sensitive to the real-world context**.

- ($\boxed{Q_1}$) When the context explicitly involves **choosing between two or more actions** and a **utility structure** that **satisfies all relevant agents** is **difficult to find**, a **principled alternative** — such as the **cost-adjusted Laplace and BIC methods** presented here — may be available.

- ($\boxed{Q_1}$ and $\boxed{Q_2}$) When the goal is to make an **accurate scientific summary** of what's known about something, the **predictive log score** has a **sound generic utility basis** and can yield **stable** and **accurate** model specification decisions.

- ($\boxed{Q_1}$) DIC can be thought of as a fast approximation to the **leave-one-out predictive log score** (LS_{CV}), but DIC can behave **unstably** as a function of **parameterization**.

Conclusions (continued)

- ($\boxed{Q_1}$) The **full-sample log score** (LS_{FS}) is n times **faster** than naive implementations of LS_{CV}, has better **small-sample model discrimination accuracy** than either LS_{CV} or DIC, and **has better asymptotic behavior** than LS_{CV}.

  See Draper and Krnjajić (2008) for **details** on these points:

- ($\boxed{Q_1}$) **Generic Bayes factors** are **highly unstable** when context suggests **diffuse prior information**; many methods for fixing this have been proposed, most of which seem to require an **appeal to ad-hockery** which is **absent** from the LS_{FS} approach.

- ($\boxed{Q_2}$) The basic Gelman et al. (1996) method of **posterior predictive model-checking** is **badly calibrated**: when it gives you a tail area of, e.g., **0.4**, the calibrated equivalent may well be **0.04**.

- ($\boxed{Q_2}$) We have modified an **approach** suggested by Robins et al. (2000) to help answer the question "Could the data have arisen from M_1?" in a **well-calibrated** way.

Paying the Right Price For Model Uncertainty

- People often talk about **BNP modeling** as providing **"insurance"** against **mis-specified parametric models**:

(1) You can **simulate** from a known **("true") parametric model** M_1 and fit M_1 and BNP to the simulated data sets; both will be **valid** (both will **reconstruct** the **right answer averaging across simulation replications**) but the BNP **uncertainty bands** will typically be **wider**.

(2) You can also simulate from a **different parametric model** M_2 and fit M_1 and BNP to the simulated data sets; often now **only BNP will be valid**.

People refer to the **wider uncertainty bands** for BNP in (1) as the **"insurance premium"** you have to pay with BNP **to get the extra validity** of BNP in (2).

But this is **not a fair comparison**: the simulation results in (1) and (2) were all **conditional on a known "true" model**, and don't immediately apply to a **real-world setting** in which **you don't know what the "true" model is**; when you **pay an appropriate price** for shopping around for the **"right" parametric model** (as in 3CV), the **discrepancy** between the parametric and BNP uncertainty bands **vanishes**.

Paying the Right Price For Model Uncertainty

In my view, this is a good way to **quantify** the **price of model uncertainty**: give a **BNP model** — centered at an ***a priori* plausible parametric model not arrived at with pre-modeling** — all n of the data values, and find out how many data points $n_{P,DA} < n$ are needed by the **best parametric model discovered with a DA search** to achieve the **same inferential accuracy** as the BNP model; the difference $(n - n_{P,DA})$ is how much data should be **reserved** in 3CV subset S_3.

- In preliminary results (with random-effects models in T versus C randomized trials), the **right amount of data** to allocate to subset S_3 to make this happen with moderate sample sizes is about **25%**, leading to a **recommended allocation of data** across (S_1, S_2, S_3) in the vicinity of **(50%, 25%, 25%)**.

In other words, with $n =$ **1,000** I should be prepared to **pay about 250 observations worth of information** in **quoting my final uncertainty assessments** (i.e., make these uncertainty assessments **about** $\sqrt{\frac{n}{0.75 n}} \doteq 15\%$ **wider** than those based on the full data set), to **account in a well-calibrated manner** for my **search for a good model**.