Testing Hypotheses about Medical Test Accuracy: Considerations for Design and Inference

Citation	Branscum, A. J., Cheng, D., & Lee, J. J. (2015). Testing hypotheses about medical test accuracy: considerations for design and inference. Journal of Applied Statistics, 42(5), 1106-1119. doi:10.1080/02664763.2014.995608
DOI	10.1080/02664763.2014.995608
Publisher	Taylor & Francis
Version	Accepted Manuscript
Terms of Use	http://cdss.library.oregonstate.edu/sa-termsofuse

Testing Hypotheses about Medical Test Accuracy: Considerations for Design and Inference

Adam J. Branscum[1*], Dunlei Cheng[2*^◊], and J. Jack Lee[3]

[1] *Biostatistics Program, School of Biological and Population Health Sciences, Oregon State University, Corvallis, OR, 97331, USA.*

[2] *Division of Biostatistics, The University of Texas School of Public Health Dallas Regional Campus, Dallas, TX, 75390, USA.*

[3] *Department of Biostatistics, Division of Quantitative Sciences, The University of Texas MD Anderson Cancer Center, Houston, TX, 77030, USA.*

[*] These authors contributed equally to this work.

^ This author is currently an employee at American Thrombosis and Hemostasis Network.

[◊] Correspondence to: Dunlei Cheng, Associate Director of Health Services Research/Biostatistics, American Thrombosis and Hemostasis Network, 72 Treasure Lane, Riverwoods, IL 60015, Telephone: 800-360-2846 ext 114. Fax: 847-572-0967. E-mail: dcheng@athn.org.

Abstract

Developing new medical tests and identifying single biomarkers or panels of biomarkers with superior accuracy over existing classifiers promotes lifelong health of individuals and populations. Before a medical test can be routinely used in clinical practice, its accuracy within diseased and non-diseased populations must be rigorously evaluated. We introduce a method for sample size determination for studies designed to test hypotheses about medical test or biomarker sensitivity and specificity. We show how a sample size can be determined to guard against making type I and/or type II errors by calculating Bayes factors from multiple data sets simulated under null and/or alternative models. The approach can be implemented across a variety of study designs, including investigations into one test or two conditionally independent or dependent tests. We focus on a general setting that involves non-identifiable models for data when true disease status is unavailable due to the nonexistence of or undesirable side effects from a perfectly accurate (i.e., "gold standard") test; special cases of the general method apply to identifiable models with or without gold-standard data. Calculation of Bayes factors is performed by incorporating prior information for model parameters (e.g., sensitivity, specificity, and disease prevalence) and augmenting the observed test-outcome data with unobserved latent data on disease status to facilitate Gibbs sampling from posterior distributions. We illustrate our methods using a thorough simulation study and an application to toxoplasmosis.

Key Words: comparing diagnostic tests, power and sample size, sensitivity and specificity

1. Introduction

Sample size determination is a key component to designing a study of medical test accuracy. Throughout this paper, we use the terms “medical test” and “test” to broadly include any binary classifier for a well-defined condition (termed “disease,” with “non-diseased” used to indicate absence of the condition). Medical examples include antibody and antigen detection tests, and single biological markers or collections of biomarkers that are associated with a disease. A non-medical example is the examination of system parts, processes or products for quality control in manufacturing. For tests that produce binary outcomes, including dichotomization of continuous test data, accuracy is routinely characterized by the tests' sensitivity and specificity. Let T denote the outcome of a test (T=1 indicates testing positive and T=0 indicates testing negative), and let D denote disease status (D=1 indicates disease positive and D=0 indicates disease negative). The sensitivity (S) of a test is defined to be its accuracy among diseased individuals, namely $S=\Pr(T=1|D=1)$, while its specificity (C) is the probability of testing negative for non-diseased individuals, $C=\Pr(T=0|D=0)$.

When disease status is ascertainable from a flawless (i.e., “gold standard”) procedure, statistical methods for inference about sensitivity and specificity are relatively straightforward compared to when disease status is unknown. There are many popular study designs used to investigate the accuracy of a single test or to compare two or more medical tests in the absence of a gold standard (e.g., Hui and Walter, 1980; Joseph et al., 1995; Dendukuri and Joseph, 2001; Johnson et al., 2001; Georgiadis et al., 2003). Many of these models contain more parameters than can be uniquely estimated from the data and model alone, and constraints are needed to counteract this lack of identifiability. Commonly used constraints involve model contraction or model expansion (Gustafson, 2005). For instance, we might set some parameters equal to constants (e.g., set C=1 for a test known to have high specificity) or obtain data from a second population that has different disease prevalence (Hui and Walter, 1980). An alternative approach uses informative prior distributions in a Bayesian data analysis (e.g., place a prior on C that has most of its probability mass above 0.95, for a test thought to have high specificity).

We focus on Bayesian models with informative priors for studies designed to test hypotheses about the accuracy of one test or to compare the accuracies of two tests using paired data (i.e., both tests are applied to each sampled individual). For example, a study might aim to determine if the sensitivity and specificity of a newly developed test exceed those for a standard test. We develop a simulation-based procedure to assist in sample size determination for both types of study designs. In our approach, hypothesis testing using Bayes factors proceeds by simulating multiple data sets assuming either a null model is true (to guard against making a type I error) or an alternative model is true (to guard against making a type II error). For instance, our approach can be used to find a sample size that yields an appropriately high probability of producing null-supporting Bayes factors across the conditions that exist in the population when the null model is true. We also demonstrate methods for addressing both types of errors by simulating data and calculating corresponding Bayes factors under both the null and alternative models.

A complication arises in the process of making a statistical inference for the sensitivities and specificities of two tests with paired data due to the potential for correlation among test outcomes from the same individual. Correlated test results might occur, for instance, when both tests target similar biological mechanisms (e.g., comparison of antibody-antibody or antigen-antigen detection tests). There are four possibilities to consider: correlation among test results for diseased but not non-diseased individuals, correlation among test results for non-diseased but not diseased individuals, correlation among test results for both populations of individuals, and uncorrelated test results for both populations. We use Bayes factors for the inferential task of determining whether estimates and comparisons of sensitivities and specificities of two tests should be adjusted for correlation or not. This provides an attractive alternative to current practices, which include use of a model selection statistic such as the Deviance Information Criterion (e.g., Pan-ngum et al., 2013), model averaging (e.g., Black and Craig, 2002), or relying solely on a sensitivity analysis to compare how estimates change under different assumptions about correlation (e.g., Branscum et al., 2005).

We present a method for finding a sample size such that the Bayes factor from the future study will be highly likely to support an assumed correct model. Although we focus on studies about medical test accuracy, the ideas and methods can be used in a variety of other contexts and disciplines. We focus on a general scenario involving test-accuracy models that lack identifiability due to over parameterization (e.g., Joseph et al., 1995; Johnson et al., 2001; Jones et al., 2010); simpler settings and models are special cases. In our proposed sample size analysis, non-identifiability is offset by using informative prior distributions on some model parameters. This concentrates posterior sampling to regions of high prior probability, and hence the chosen prior will influence posterior inference even when the sample size is large.

This paper presents new approaches to aspects of study design and statistical inference for medical test accuracy. The design consideration we address is the development of a novel simulation-based method for sample size determination for studies that investigate hypotheses about medical test accuracy, while the important inferential task we address involves testing competing models that make different assumptions about correlation among test outcomes with paired data. Although the study designs we consider in this paper are used to investigate test accuracy (i.e., sensitivity and specificity), it is possible to alternatively use them to evaluate diagnostic accuracy (i.e., positive and negative predictive value).

2. Models and Methods

2.1. One test

A study will be conducted to decide between competing hypotheses about the sensitivity and specificity of a single test. For instance, the null might hypothesize low *S* and high *C*, while the alternative hypothesizes high *S* and high *C*. The study will enroll a random sample of *n* individuals from a large population that has unknown disease prevalence π. Let n_1 and n_2 denote the number out of the *n* individuals who test positive or negative, respectively. When a gold standard test is unavailable, we will model the future data as $n_1|(\pi, S, C)$ ~ binomial(n, p), where

$$\begin{aligned} p = \Pr(T=1) &= \Pr(T=1 \mid D=1)\Pr(D=1) + \Pr(T=1 \mid D=0)\Pr(D=0) \\ &= S\pi + (1-C)(1-\pi). \end{aligned}$$

We will assign independent beta prior distributions to π, *S*, and *C*, namely π ~ beta(a_π, b_π), *S* ~ beta(a_s, b_s), and *C* ~ beta(a_c, b_c), which is very common in practice (e.g., Joseph et al. 1995; Dendukuri and Joseph, 2001; Georgiadis et al. 2003; Branscum et al. 2005; among many). The model lacks identifiability because there are 3 unknown parameters (π, *S*, and *C*) and only one degree of freedom (n_1). Therefore, we require informative prior distributions for at least two model parameters (often for the prevalence and either *S* or *C*).

The model yields unrecognizable full conditional distributions in a Gibbs sampler. Therefore, we incorporate latent data in such a way that a Gibbs sampler involves simulation from only standard distributions. To this end, let Z_1 and Z_2 denote the latent number of individuals who are diseased out of the n_1 and n_2 individuals, respectively. These variables are also chosen because, together with the observed data n_1 and n_2, they constitute the data that would have been available under the ideal scenario in which disease status is known, as depicted below:

		D		
		1	0	
T	1	Z_1	n_1-Z_1	n_1
	0	Z_2	n_2-Z_2	n_2

The distribution of the augmented data (Z_1, Z_2, n_1-Z$_1$, n_2-Z$_2$) is multinomial with parameter vector (Pr(D = 1, T = 1), Pr(D = 1, T = 0), Pr(D = 0, T = 1), Pr(D = 0, T = 0)). Hence, the augmented data likelihood is proportional to $(\pi S)^{Z_1}\{\pi(1-S)\}^{Z_2}\{(1-\pi)(1-C)\}^{n_1-Z_1}\{(1-\pi)C\}^{n_2-Z_2}$, with corresponding augmented data posterior proportional to

$$\pi^{Z_1+Z_2+a_\pi-1}(1-\pi)^{n_1+n_2-Z_1-Z_2+b_\pi-1}S^{Z_1+a_s-1}(1-S)^{Z_2+b_s-1}C^{n_2-Z_2+a_c-1}(1-C)^{n_1-Z_1+b_c-1}.$$

It immediately follows that π, S, and C are sampled from beta distributions in a Gibbs sampler that simulates from the augmented data posterior (e.g., $\pi|(n_1,n_2,Z_1,Z_2)$~beta($Z_1+Z_2+a_\pi$, n_1+n_2-Z_1-Z_2+b_π)). The latent data are also easy to update since $Z_1|(n_1, \pi, S, C)$~binomial(n_1, $\frac{\pi S}{\pi S+(1-\pi)(1-C)}$) and $Z_2|(n_2, \pi, S, C)$ ~binomial(n_2, $\frac{\pi(1-S)}{\pi(1-S)+(1-\pi)C}$).

2.2. Comparing two tests

Consider a comparative study of test accuracy that will use the same sampling scheme described in Section 2.1, but where two tests will be applied to each sampled individual. The goal of the future study is to decide between two hypotheses that make different conjectures about S_1, S_2, C_1, and C_2, the sensitivities and specificities of the two tests. A common application involves testing whether the sensitivity and specificity of a newly developed test exceed those for a standard test.

The data from each of n randomly sampled individuals are pairs (T_1, T_2) of binary test outcomes. We assume that the tests are independent, conditional on disease status. That is, we assume $\Pr(T_1=i, T_2=j \mid D=k)=\Pr(T_1=i \mid D=k)\Pr(T_2=j \mid D=k)$, for i, j, k = 0, 1. Biologically, this assumption is often supported when the tests detect different mechanisms (e.g., an organism detection test and an antibody detection test). Then, the probabilities of the four possible combinations of paired outcomes are:

$$p_{11} = \Pr(T_1 = 1, T_2 = 1) = \pi S_1 S_2 + (1 - \pi)\,(1 - C_1)(1 - C_2),$$

$$p_{10} = \Pr(T_1 = 1, T_2 = 0) = \pi S_1(1 - S_2) + (1 - \pi)\,(1 - C_1)C_2,$$

$$p_{01} = \Pr(T_1 = 0, T_2 = 1) = \pi\,(1 - S_1)S_2 + (1 - \pi)\,C_1(1 - C_2),$$

$$p_{00} = \Pr(T_1 = 0, T_2 = 0) = \pi\,(1 - S_1)(1 - S_2) + (1 - \pi)\,C_1C_2.$$

Let n_{ij} denote the number of subjects for whom $T_1=i$ and $T_2=j$, for i, j=0,1. These observed counts are distributed as (n_{11}, n_{10}, n_{01}, n_{00}) ~ multinomial(n, (p_{11}, p_{10}, p_{01}, p_{00})). We again introduce latent data to ease the implementation of Gibbs sampling from the augmented data posterior distribution. Let Z_{ij} denote the unobserved number of individuals who are diseased out of n_{ij}, for i, j=0,1. Then, the augmented data likelihood is the multinomial mass function for (Z_{11}, Z_{10}, Z_{01}, Z_{00}, n_{11}-Z_{11}, n_{10}-Z_{10}, n_{01}-Z_{01}, n_{00}-Z_{00}). Since the tests are conditionally independent, the element of the multinomial probability vector associated with Z_{11} is $\Pr(T_1=1,T_2=1,D=1)=\Pr(T_1=1|D=1)\Pr(T_2=1|D=1)\Pr(D=1)=S_1S_2\pi$. The other multinomial probabilities are derived similarly to obtain an augmented data likelihood function proportional to

$$(\pi S_1 S_2)^{Z_{11}}\{\pi S_1(1-S_2)\}^{Z_{10}}\{\pi(1-S_1)S_2\}^{Z_{01}}\{\pi(1-S_1)(1-S_2)\}^{Z_{00}}\times$$

$$\{(1-\pi)(1-C_1)(1-C_2)\}^{n_{11}-Z_{11}}\{(1-\pi)(1-C_1)C_2\}^{n_{10}-Z_{10}}\{(1-\pi)C_1(1-C_2)\}^{n_{01}-Z_{01}}\{(1-\pi)C_1C_2\}^{n_{00}-Z_{00}}.$$

With independent beta priors on π, S_1, S_2, C_1, and C_2, a Gibbs sampler is straightforward to implement because it contains only beta and binomial full conditional distributions (Table 1).

It is important to note that we again require informative prior distributions because, since n is fixed, there are only 3 degrees of freedom (n_{11}, n_{10}, and n_{01}) to estimate 5 parameters (prevalence, two sensitivities, and two specificities). Therefore, in practice, this design is commonly used to compare a new test to a standard test when the disease prevalence is approximately known. Then, informative priors are placed on π and the sensitivity and/or specificity of the standard test.

2.3. Conditional dependence

In order to apply the model in Section 2.2, we must rule out conditional dependence between the two tests. Our approach uses a Bayes factor to compare the conditional independence model to a conditional dependence model. We proceed by using the conditional dependence model developed by Georgiadis et al. (2003) (see also Dendukuri and Joseph, 2001). Briefly, the model parameters are the prevalence and sensitivities and specificities of the two tests, plus two conditional covariance parameters. The covariance between test outcomes for diseased individuals is defined as $\text{Cov}(T_1,T_2|D=1)=\text{E}(T_1T_2|D=1)-\text{E}(T_1|D=1)\text{E}(T_2|D=1)=S_{11}-S_1S_2$, where $S_{11}=\Pr(T_1=1, T_2=1 \,|D=1)$. For non-diseased individuals, the covariance is defined to be $C_{00}-C_1C_2$, where $C_{00}=\Pr(T_1=0, T_2=0 \mid D=0)$. Data augmentation by the same Z_{ij}'s in Section 2.2 leads to a simple Gibbs sampler that contains only beta and binomial distributions (Georgiadis et al. 2003). Informative prior distributions are required (e.g., on the prevalence, and the sensitivity and specificity of the standard test).

2.4. Bayes factor

We use Bayes factors for selecting between competing models (Kass and Raftery, 1995). A sample size can be chosen so that the Bayes factor from a future study is unlikely to support an incorrect alternative model (low probability of making a type I error) and/or highly likely to support a correct alternative model (high power). To accomplish these goals, the sampling distribution of a Bayes factor is simulated under either a null or alternative model, or under both models.

For concreteness, the remainder of this subsection presents details for addressing both type I and type II errors; obvious omissions and alterations of the procedure are needed when the study goals dictate focusing on only one type of error. Multiple data sets are simulated under H_0 and H_1 to approximate the distributions of Bayes factors under both models. Specifically, data y_0 and y_1 are generated under models H_0 and H_1, respectively, and Bayes factors are given by

$$BF_{01 \cdot y_j} = \frac{p(y_j \mid H_0)}{p(y_j \mid H_1)} = \frac{\int p(y_j \mid \theta_0, H_0) p_0(\theta_0) d\theta_0}{\int p(y_j \mid \theta_1, H_1) p_1(\theta_1) d\theta_1},$$

where $p_k(\theta_k)$ and $p(y_j \mid \theta_k, H_k)$ denote the prior and likelihood function, respectively, for parameter vector θ_k under model H_k, for j, k = 0,1. High values of BF_{01} (e.g., >10) support model H_0, while low values (e.g., < 0.10) support H_1; the decision between H_0 and H_1 is inconclusive when BF_{01} is near 1, in which case the more parsimonious model is often selected. To improve computational stability, we work with log-transformed Bayes factors (lnBF_{01}).

We use the distribution of lnBF_{01} under two competing models to determine an appropriate sample size as follows. First, for a fixed sample size, simulate multiple data sets under model H_0 and use the data sets to determine a value ω such that $\Pr(\ln BF_{01 \cdot y_0} < \omega) = \alpha$, for a pre-specified (small) value of α. Then, use the same sample size to simulate data under H_1 and approximate the following measure of “power”: $\Pr(\ln BF_{01\cdot y_1} < \omega)$. The current sample size is selected for use in the future study if the power is acceptably high. The future study will use ω as the decision threshold; we will select H_0 if the log Bayes factor from the future study is $> \omega$, and select H_1 otherwise. Note that Bayes factors are calculated by integrating the likelihood against the prior, which appropriately propagates uncertainty about parameter values under the null and alternative models instead of treating parameters as known constants for power analysis.

2.5. Computing lnBF_{01}

For the models we consider, the analytic forms of the marginal densities in the numerator and denominator of BF_{01} are unattainable because they depend on intractable high-dimensional integrals. Diverse approaches have been developed to approximate marginal likelihoods (e.g., Newton and Raftery, 1994; Green, 1995; Chib, 1995; Raftery, 1996; Chib and Jeliazkov, 2001; Neal, 2001; Friel and Pettitt, 2008). We used Chib’s 1995 method because a Gibbs sampler contingent upon latent data for disease status is easy to implement for the models we consider.

In general, a log marginal likelihood function is given by

$$\ln p(y) = \ln p(y|\theta) + \ln p(\theta) - \ln p(\theta|y),$$

where the parameter vector θ has length s. Although this equation holds for any value of θ, efficiency considerations led Chib (1995) to suggest using a value $\theta^*=(\theta_1^*,\ldots,\theta_s^*)$ that has relatively high posterior ordinate. The terms ln $p(y|\theta^*)$ and ln $p(\theta^*)$ are straightforward to compute. The following procedure was used to approximate ln $p(\theta^*|y)$.

Start with the decomposition:

$$\ln p\,(\theta^* \mid y) = \ln p(\theta_1^* \mid y) + \ln p(\theta_2^* \mid \theta_1^*, y) + \ln p(\theta_3^* \mid \theta_2^*, \theta_1^*, y) + \cdots + \ln p(\theta_s^* \mid \theta_{s-1}^*, \ldots, \theta_1^*, y). \quad (1)$$

Our applications involve latent data, so we used the technique of Rao-Blackwellization (Gelfand and Smith, 1990; Chib, 1995; Robert and Casella, 2010) to approximate $p(\theta^* \mid y)$ by averaging over simulated realizations of the latent data.

Each conditional distribution on the right hand side of (1) gets approximated in turn by using a sequence of different Gibbs samplers. For the general setting, denote the latent data by $Z=(Z_1,\ldots,Z_t)$ and let $Z^{(u,g)}=(Z_1^{(u,g)},\ldots,Z_t^{(u,g)})$ be the (post-convergence) simulated values from the uth iteration of the gth Gibbs sampler, for $g = 1,\ldots,s$. The first Gibbs sampler is used to simulate from the augmented data posterior to calculate $p(\theta_1^* \mid y)=\int p(\theta_1^* \mid y,Z)\,p(Z \mid y)dZ \approx U_1^{-1}\sum_{u=1}^{U_1} p(\theta_1^* \mid y,Z^{(u,1)})$. In a second Gibbs sampler, θ_1 is set equal to θ_1^* and we generate $Z^{(u,2)}\sim p(Z|y,\theta_1^*)$ for $u=1,\ldots,U_2$; these values are then used to calculate $p(\theta_2^* \mid \theta_1^*,y)\approx U_2^{-1}\sum_{u=1}^{U_2} p(\theta_2^* \mid \theta_1^*,y,Z^{(u,2)})$. This process is continued for a total of s Gibbs samplers to estimate ln $p(\theta^*|y)$ by

$$\ln \hat{p}\,(\theta^* \mid y) =$$

$$\ln[U_1^{-1}\sum_{u=1}^{U_1} p(\theta_1^* \mid y,Z^{(u,1)})]+\ln[U_2^{-1}\sum_{u=1}^{U_2} p(\theta_2^* \mid \theta_1^*,y,Z^{(u,2)})]+\cdots+\ln[U_s^{-1}\sum_{u=1}^{U_s} p(\theta_s^* \mid \theta_{s-1}^*,\ldots,\theta_1^*,y,Z^{(u,s)})].$$

The number of draws in each Gibbs sampler can be different or the same (i.e., $U_1 = \cdots = U_s$).

For the one test setting, we have $\theta=(\pi,S,C)$, $y=(n_1,n_2)$, $Z=(Z_1,Z_2)$, and

$$\ln \hat{p}(\theta^* \mid n_1,n_2)=\ln[U_1^{-1}\sum_{u=1}^{U_1} p(\pi^* \mid n_1,n_2,Z_1^{(u,1)},Z_2^{(u,1)})]+\ln[U_2^{-1}\sum_{u=1}^{U_2} p(S^* \mid \pi^*,n_1,n_2,Z_1^{(u,2)},Z_2^{(u,2)})]\ +$$

$$\ln[U_3^{-1}\sum_{u=1}^{U_3} p(C^* \mid S^*,\pi^*,n_1,n_2,Z_1^{(u,3)},Z_2^{(u,3)})]. \qquad (2)$$

For every draw ($Z_1^{(u,1)}$, $Z_2^{(u,1)}$), we set $\pi^{(u,1)}$ equal to the mean of the beta full conditional for π, i.e., $\pi^{(u,1)} = E(\pi \mid n_1, n_2, Z_1^{(u,1)}, Z_2^{(u,1)}) = \dfrac{Z_1^{(u,1)} + Z_2^{(u,1)} + a_\pi}{n_1 + n_2 + a_\pi + b_\pi}$. The value of π^* that we used was $\pi^* = U_1^{-1}\sum_{u=1}^{U_1}\pi^{(u,1)}$. The first term on the right side of (2) is then calculated as $\ln[U_1^{-1}\sum_{u=1}^{U_1}\beta(\pi^* \mid Z_1^{(u,1)} + Z_2^{(u,1)} + a_\pi, n_1 + n_2 - Z_1^{(u,1)} - Z_2^{(u,1)} + b_\pi)]$, where $\beta(x|a,b)$ denotes the beta(a,b) density evaluated at x.

The second Gibbs sampler is run with π fixed at π^* throughout. Calculation of the second term on the right side of (2) is also easy because $S \mid (\pi^*, n_1, n_2, Z_1^{(u,2)}, Z_2^{(u,2)}) \sim beta(Z_1^{(u,2)} + a_s, Z_2^{(u,2)} + b_s)$. We used $S^* = U_2^{-1}\sum_{u=1}^{U_2} S^{(u,2)}$, where $S^{(u,2)} = \dfrac{Z_1^{(u,2)} + a_S}{Z_1^{(u,2)} + Z_2^{(u,2)} + a_S + b_S}$. For the last Gibbs sampler, only the specificity parameter is updated (along with the latent true positives and false negatives). The fixed values π^* and S^* from the two previous Gibbs runs are used here. We have $C \mid (S^*, \pi^*, n_1, n_2, Z_1^{(u,3)}, Z_2^{(u,3)}) \sim beta(n_2 - Z_2^{(u,3)} + a_c, n_1 - Z_1^{(u,3)} + b_c)$ and we set $C^* = U_3^{-1}\sum_{u=1}^{U_3} C^{(u,3)}$, where $C^{(u,3)} = \dfrac{n_2 - Z_2^{(u,3)} + a_C}{n_1 + n_2 - Z_1^{(u,3)} - Z_2^{(u,3)} + a_C + b_C}$.

Similar methods were used to compute $\ln BF_{01}$ in the case of two conditionally independent tests, which involves 5 Gibbs samplers and latent data (Z_{11}, Z_{10}, Z_{01}, Z_{00}); we averaged over the Rao-Blackwellized values in Table 1 to determine θ^*. Similar methods were also used for the case of two conditionally dependent tests.

3. Illustrations

Our procedure requires prior distributions on prevalence and test accuracy parameters when fitting models to simulated data sets. We also require distributions for prevalence and test accuracy parameters for use in simulating data sets under null and alternative models. In all our examples, the same distributions were used for priors and data simulation. The results in this section are based on 1000 simulated data sets under H_0 and H_1 at each sample size considered. For every Gibbs sampler, 5000 iterates were obtained with the first 1000 values discarded as burn in.

3.1. One test

Consider a population with disease prevalence believed to be about 0.75. Suppose we are interested in comparing the following null model that specifies low test accuracy to an alternative model that specifies moderately high test accuracy:

H_0: $\pi \sim$ beta(37.5,12.5)	H_1: $\pi \sim$ beta(37.5,12.5)
$S \sim$ beta(25.5,24.5)	$S \sim$ beta(42.5,7.5)
$C \sim$ beta(25,25)	$C \sim$ beta(32.5,17.5)

Under H_0, the sensitivity and specificity are expected to be about 50%, while under H_1 they are expected to be 85% and 65%, respectively. Observe that $a+b$ =50 in all of the above beta(a,b) distributions. Figure 1 presents distributions of the log Bayes factor under H_0 and H_1 for sample sizes of 25, 50, and 75. As expected, there is greater separation between the null and alternative distributions of lnBF_{01} as the sample size increases. The power increases from 68% to 79% to 88% as n increases from 25 to 50 to 75. The corresponding cut-offs (ω) determined by the fifth percentile of the distribution for $\ln BF_{01 \cdot y_0}$ are -1.05, -1.11, and -0.65. By lowering α to 0.025, we obtained lower power values of 55%, 68%, and 84%, while increasing α to 0.10 gave power of 79%, 89%, and 95% for n = 25, 50, and 75, respectively.

Figure 1 shows that the variability of log Bayes factors in this example increases as the sample size increases, which occurs mainly because the variance of the marginal distribution increases with sample size when the data are inconsistent with the model. For example, compared to 4.9 and 3.5 at n=25, the variances of log-transformed $p(y_0|H_1)$ and $p(y_1|H_0)$ (where the data and model are mismatched) at n=75 are nearly doubled to 9.6 and 6.7, respectively. On the other hand, the corresponding variances for $p(y_0|H_0)$ and $p(y_1|H_1)$ were less than 0.05 for the three sample sizes considered.

Recall that in the one test setting, we model the probability of testing positive by $p=S\pi+(1-C)(1-\pi)$. Bayes factors in this context are highly dependent upon the difference in magnitude between the values of p under the null and alternative models. This is noteworthy because different combinations of π, S, and C can lead to similar values of p. Thus, data can be similar even when parameter values are different under competing models. As a simple example, suppose the null hypothesis is H_0: (π=0.50, S=0.10, C=0.90) and consider the alternatives H_1: (π=0.50, S=0.10, C=0.901) and H_2: (π=0.50, S=0.20, C=0.999). In all 3 cases, $p\approx0.10$. This will lead to a Bayes factor for comparing (the very similar) H_0 to H_1 that is similar to the Bayes factor for comparing (the much more different) H_0 to H_2. Thus, large differences between the null and alternative values of parameters do not necessarily translate into higher power compared to the power for similar null and alternative models. We therefore suggest monitoring the simulated mean of p. The same issue can occur in two test settings.

Using the previous specification of H_0, we studied changes in power for alternative models that conjecture different test sensitivity. The one-test study design is commonly used to determine whether the sensitivity or specificity (but not both) of a new test exceed the known value of a standard test. Since, in this example, the prevalence is high (approximately 75%), this setting would typically be used to compare sensitivities of a new and standard test. We therefore consider power calculations under five scenarios that posit increasing test sensitivity while leaving the specificity unaltered. In particular, five specifications for H_1 were considered, including the setting from above (details in the top half of Table 2). For cases 1-5, mean sensitivities are hypothesized under H_1 to be 70%, 75%, 80%, 85%, and 94%. We modeled the prevalence as π ~ beta(37.5,12.5) for each H_1. The results for cases 3-5 demonstrate that a test with higher sensitivity requires a much smaller sample size to achieve similar power as a test with lower sensitivity in this example. This is expected here because the prevalence is high ($\pi \approx 0.75$). For instance, a sample size of 50 in case 4 yields approximately 80% power, while in case 2 the power was only 74% even when the sample size was 5 times higher at n=250 (Table 3).

We also investigated the power obtained when using fixed values of some parameters instead of priors, since it is common to use fixed inputs when performing power calculations. The same priors on π and S as in case 4 were used, but with test specificity under H_0 and H_1 fixed at 50% and 65% (the prior means used in cases 1-5), respectively. This configuration with fixed C leads to power of 81% and 94% at $n=50$ and $n=100$, which is 3% and 1% greater than the corresponding values from case 4. By also fixing the prevalence parameter at 0.75, the power was further increased to 83% and 96% at these sample sizes.

3.2. Two conditionally independent tests

Consider a population where the prevalence is known to be approximately 30% (modeled under H_0 and H_1 as π~beta(15,35)) and the sensitivity and specificity of the standard test are both known to be low (S_1≈50% and C_1≈60%), as described in the bottom half of Table 2. Because the prevalence is relatively low, we consider a setting where a future study will be designed to evaluate the specificity of a new test 2. The null and alternative models specify S_1~beta(25,25) and C_1~beta(30,20) for the standard test. The null hypothesis states that the new test is no better than the standard test: S_2~beta(25,25) and C_2~beta(30,20). For illustration, we considered three alternative hypotheses (details in the bottom half of Table 2). All 3 alternatives posit the same slightly higher sensitivity for the new test (S_2≈60%). For comparison, we considered alternatives with increasingly greater specificity for the new test (C_2≈80% in case 1, C_2≈86% in case 2, and C_2≈95% in case 3).

Results for ω and power at different sample sizes are presented in Table 4. The threshold ω was determined from the distribution of $\ln BF_{01\cdot y_0}$ by setting α = 0.05. Case 3 yields the highest power, with a sample size of 75 corresponding to over 90% power. Note that the power is low for case 1 for all sample sizes considered.

We also calculated power when using fixed parameter inputs. Compared to 66% and 74% power when n=75 and n=100 in case 2, replacing beta priors on π, S_1, C_1, S_2 with constants (equal to the prior means in case 2) resulted in slightly larger power values of 70% and 79%.

3.3. Toxoplasmosis

Toxoplasmosis is a parasitic disease that can be transferred from animals to humans (a zoonotic disease). One common mode of transmission is consumption of undercooked meat. Symptoms are generally mild, but the disease can be lethal for people with compromised immune systems. Data from the 1999-2004 waves of the National Health and Nutritional Examination Survey (NHANES) estimate the seroprevalence of toxoplasmosis in humans at approximately 11% in the United States (Jones et al., 2007); global estimates have been as high as 33% (Montoya and Liesenfeld, 2004).

In a previous study, the sensitivities and specificities of a microscopic agglutination test (MAT) and an enzyme-linked immunosorbent assay (ELISA) for toxoplasmosis in pigs were estimated using conditional independence and conditional dependence models (Georgiadis et al., 2003). Here we show that only the conditional dependence model should be used. Both diagnostic tests were applied to 999 pigs with the following data breakdown:

		ELISA	
		+	-
MAT	+	164	58
	-	77	700

We used Bayes factors to compare 3 conditional dependence models to the conditional independence model. Using the same priors as Georgiadis et al. (2003), log marginal likelihoods were: -10.7 (conditional independence), -8.6 (conditional dependence between sensitivities), -8.5 (conditional dependence between specificities), and 1.7 (conditional dependence between sensitivities and between specificities); the corresponding Bayes factors in support of conditional dependence models over the independence model are 9, 10, and e^{12}, respectively. Estimates of MAT and ELISA test accuracy should therefore be adjusted for both types of correlation, namely dependency between test sensitivities and between test specificities.

3.4. Diagnostic accuracy

The one-test and two-test designs we consider in this paper are primarily used to study medical test sensitivity and specificity. Although uncommon, they can also be used when the goal is to evaluate diagnostic accuracy, namely positive and negative predictive values. For instance, the parameters in the one-test setting are the positive predictive value, PPV=Pr(D=1|T=1), negative predictive value, NPV=Pr(D=0|T=0), and the probability of a positive test, p=Pr(T=1). Let (a_{ppv}, b_{ppv}), (a_{npv}, b_{npv}), and (a_p, b_p) denote the beta hyperparameters for PPV, NPV, and p, respectively, and let Z_1 and Z_2 be defined as in Section 3.1. Then the joint posterior distribution is proportional to $PPV^{Z_1+a_{ppv}-1}(1-PPV)^{n_1-Z_1+b_{ppv}-1}\times$ $NPV^{n_2-Z_2+a_{npv}-1}(1-NPV)^{Z_2+b_{npv}-1}p^{n_1+a_p-1}(1-p)^{n_2+b_p-1}$, with Gibbs sampling and calculation of Bayes factors implemented as described in Section 3.1. As an example, consider a study designed to evaluate the diagnostic accuracy of an exercise stress test for coronary artery disease (Pepe, 2003). The null model posits H_0: PPV~beta(,), NPV~beta(,), and p~beta(,). The alternative is H_1: PPV ~ beta(88,12), NPV ~ beta(61,39), p~beta(,), which is consistent with values from a previous similar study. Then, a sample size even as high as 1000 people yields a relatively low power of 65%.

4. Conclusion

We addressed design and inference procedures for studies of medical test accuracy. In terms of design, we presented a new method to aid in sample size determination for studies involving one medical test and comparative studies of two tests. The procedure requires informative prior distributions and involves simulating Bayes factors under two competing models. The procedure returns a value of "power," a term we used throughout this paper to mean a measure of strength of an alternative hypothesis over a null hypothesis. We also used Bayes factors for the important task of deciding whether inference for the sensitivities and specificities of two tests should be adjusted for conditional correlations.

References

1. Black MA, Craig BA. Estimating disease prevalence in the absence of a gold standard. Statistics in Medicine. 2002, 21:2653-2669.

2. Branscum AJ, Gardner IA, Johnson WO. Estimation of diagnostic-test sensitivity and specificity through Bayesian modeling. Preventive Veterinary Medicine. 2005, 68:145-163.

3. Chib S. Marginal likelihood from the Gibbs output. Journal of the American Statistical Association. 1995, 90:1313-1321.

4. Chib S, Jeliazkov I. Marginal likelihood from the Metropolis-Hastings output. Journal of the American Statistical Association. 2001, 96:270-281.

5. Dendukuri N, Joseph L. Bayesian approaches to modeling the conditional dependence between multiple diagnostic tests. Biometrics. 2001, 57:158-167.

6. Friel N, Pettitt AN. Marginal likelihood estimation via power posteriors. Journal of the Royal Statistical Society B. 2008, 70:589-607.

7. Gelfand AE, Smith AFM. Sampling-based approaches to calculating marginal densities. Journal of the American Statistical Association. 1990, 85:398-409.

8. Georgiadis MP, Johnson WO, Gardner IA, Singh R. Correlation-adjusted estimation of sensitivity and specificity of two diagnostic tests. Applied Statistics. 2003, 52:63-76.

9. Green PJ. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika. 1995, 82:711-732.

10. Gustafson P. On model expansion, model contraction, identifiability and prior information: two illustrative scenarios involving mismeasured variables. Statistical Science. 2005, 20:111-140.

11. Hui SL, Walter SD. Estimating the error rates of diagnostic tests. Biometrics. 1980, 36:167-171.

12. Johnson WO, Gastwirth JL, Pearson LM. Screening without a gold standard: the Hui-Walter paradigm revisited. American Journal of Epidemiology. 2001, 153:921-924.

13. Jones G, Johnson WO, Hanson TE, Christensen R. Identifiability of models for multiple diagnostic testing in the absence of a gold standard. Biometrics. 2010, 66:855-863.

14. Jones JL, Kruszon-Moran D, Sanders-Lewis K, Wilson M. *Toxoplasma gondii* infection in the United States, 1999–2004, decline from the prior decade. American Journal of Tropical Medicine and Hygiene. 2007, **77**: 405–410.

15. Joseph L, Gyorkos TW, Coupal L. Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. American Journal of Epidemiology. 1995, 141:263-272.

16. Kass RE, Raftery AE. Bayes factors. Journal of the American Statistical Association. 1995, 90:773-795.

17. Montoya J, Liesenfeld O. Toxoplasmosis. Lancet**.** 2004, 363:1965–1976.

18. Neil RM. Annealed importance sampling. Statistics and Computing. 2001, 11:125-139.

19. Newton MA, Raftery AE. Approximate Bayesian inference by the weighted likelihood bootstrap. Journal of the Royal Statistical Society B. 1994, 56:3-48.

20. Pan-ngum W, Blacksell SD, Lubell Y, Pukrittayakamee S, Bailey MS, de Silva HJ, Lalloo DG, Day NP, White LJ, Limmathurotsakul D. Estimating the true accuracy of diagnostic tests for dengue infection using Bayesian latent class models. PLoS One. 2013, 8(1).

21. Pepe MS. The Statistical Evaluation of Medical Tests for Classification and Prediction. 2003. Oxford University Press.

22. Raftery AE. Approximate Bayes factors and accounting for model uncertainty in generalized linear models. Biometrika. 1996, 83:251-266.

23. Robert CP, Casella G. Introducing Monte Carlo Methods with R. 2010, Springer.

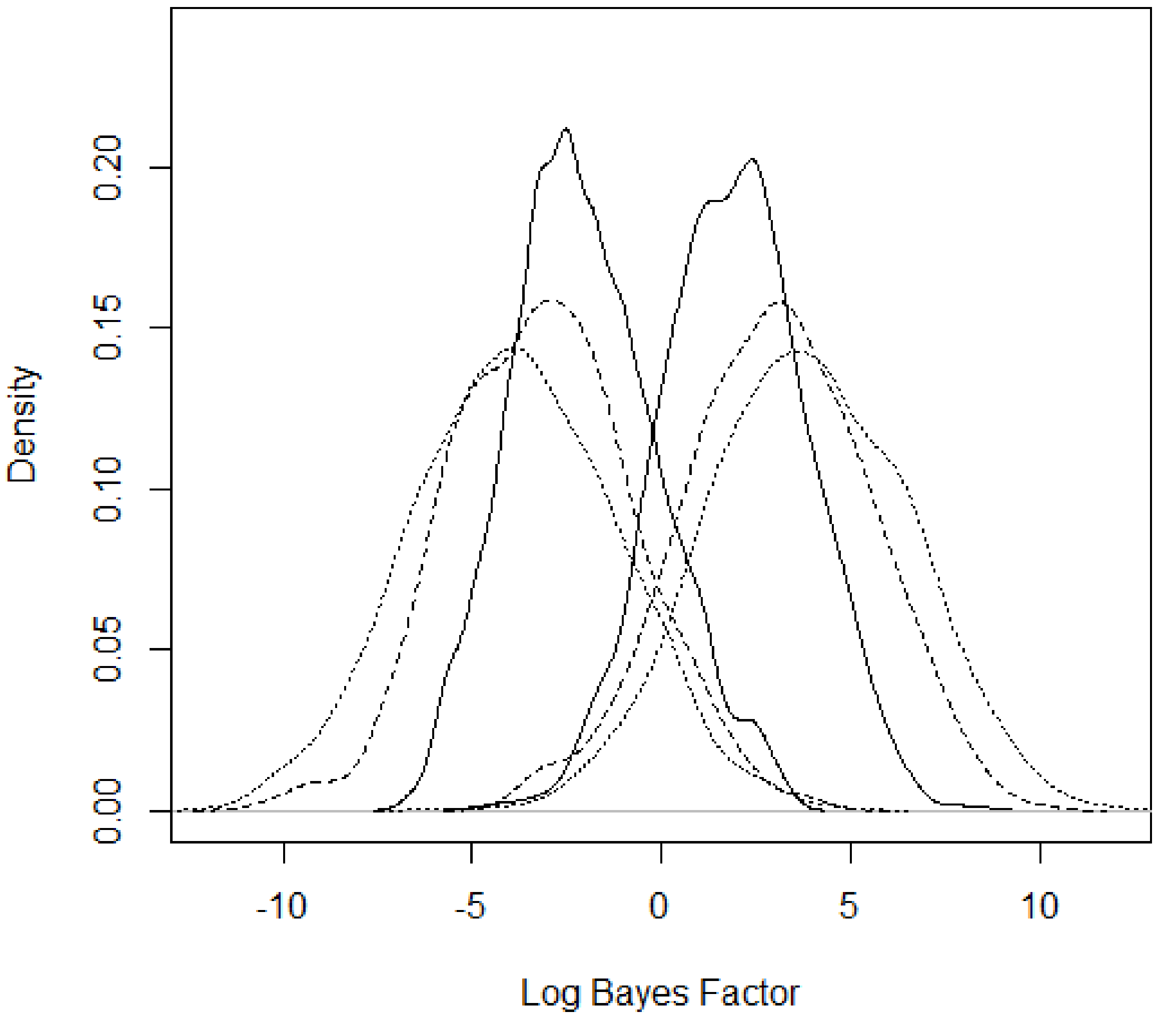

Figure 1: Distributions of log Bayes factor for the one test example in Section 3.1. The solid lines are for n=25 under H_0 (right side) and H_1 (left side), while the dashed lines are for n=50 and the dotted lines are for n=75.

Table 1: The full conditional distributions used in a Gibbs sampler of the augmented data posterior for the case of two conditionally independent tests. The Rao-Blackwellized (RB) estimator is used in the calculation of Bayes factors.

Variable	Distribution	Parameter 1	Parameter 2	RB Estimator
Z_{11}	binomial	n_{11}	$\frac{\pi S_1 S_2}{\pi S_1 S_2+(1-\pi)(1-C_1)(1-C_2)}$	N/A
Z_{10}	binomial	n_{10}	$\frac{\pi S_1(1-S_2)}{\pi S_1(1-S_2)+(1-\pi)(1-C_1)C_2}$	N/A
Z_{01}	binomial	n_{01}	$\frac{\pi(1-S_1)S_2}{\pi(1-S_1)S_2+(1-\pi)C_1(1-C_2)}$	N/A
Z_{00}	binomial	n_{00}	$\frac{\pi(1-S_1)(1-S_2)}{\pi(1-S_1)(1-S_2)+(1-\pi)C_1C_2}$	N/A
π	beta	$Z_{11}+Z_{10}+Z_{01}+Z_{00}+a_\pi$	$n-Z_{11}-Z_{10}-Z_{01}-Z_{00}+b_\pi$	$\frac{Z_{11}+Z_{10}+Z_{01}+Z_{00}+a_\pi}{n+a_\pi+b_\pi}$
S_1	beta	$Z_{11}+Z_{10}+a_{S_1}$	$Z_{01}+Z_{00}+b_{S_1}$	$\frac{Z_{11}+Z_{10}+a_{S_1}}{Z_{11}+Z_{10}+Z_{01}+Z_{00}+a_{S_1}+b_{S_1}}$
C_1	beta	$n_{01}+n_{00}-Z_{01}-Z_{00}+a_{C_1}$	$n_{11}+n_{10}-Z_{11}-Z_{10}+b_{C_1}$	$\frac{n_{01}+n_{00}-Z_{01}-Z_{00}+a_{C_1}}{n-Z_{11}-Z_{10}-Z_{01}-Z_{00}+a_{C_1}+b_{C_1}}$
S_2	beta	$Z_{11}+Z_{01}+a_{S_2}$	$Z_{10}+Z_{00}+b_{S_2}$	$\frac{Z_{11}+Z_{01}+a_{S_2}}{Z_{11}+Z_{10}+Z_{01}+Z_{00}+a_{S_2}+b_{S_2}}$
C_2	beta	$n_{10}+n_{00}-Z_{10}-Z_{00}+a_{C_2}$	$n_{11}+n_{01}-Z_{11}-Z_{01}+b_{C_2}$	$\frac{n_{10}+n_{00}-Z_{10}-Z_{00}+a_{C_2}}{n-Z_{11}-Z_{10}-Z_{01}-Z_{00}+a_{C_2}+b_{C_2}}$

Table 2: Beta prior distributions used in Sections 3.1 (one test) and 3.2 (two conditionally independent tests). Five simulation scenarios that hypothesized different alternative values for test sensitivity were considered for the one test setting. Three simulation scenarios that hypothesized different alternative values for the specificity of test 2 were considered for the two test setting.

One Test	H_0				H_1			
	a	b	Mean	SD	a	b	Mean	SD
π	37.5	12.5	0.75	0.06	same as π in H_0			
S^1	25.5	24.5	0.51	0.07	35	15	0.7	0.06
S^2	same as S^1 in H_0				37.5	12.5	0.75	0.06
S^3	same as S^1 in H_0				40	10	0.8	0.06
S^4	same as S^1 in H_0				42.5	7.5	0.85	0.05
S^5	same as S^1 in H_0				47	3	0.94	0.03
C	25	25	0.5	0.07	32.5	17.5	0.65	0.07
Two Tests	H_0				H_1			
	a	b	Mean	SD	a	b	Mean	SD
π	15	35	0.3	0.06	same as π in H_0			
S_1	25	25	0.5	0.07	same as S_1 in H_0			
C_1	30	20	0.6	0.07	same as C_1 in H_0			
S_2	25	25	0.5	0.07	same as S_2 in H_0			
C_2^1	30	20	0.6	0.06	40	10	0.8	0.06
C_2^2	same as C_2^1 in H_0				43	7	0.86	0.05
C_2^3	same as C_2^1 in H_0				47.5	2.5	0.95	0.03

Table 3: Power for the one test setting. In each case, H_0: π ~ beta(37.5,12.5), S ~ beta(25.5,24.5), and C ~ beta(25,25). Under H_0, the mean of $p=S\pi+(1-\pi)(1-C)$ is $E_{H_0}(p)=0.51$.

Sample Size	Power[1]	Power[2]	Power[3]	Power[4]	Power[5]
50	0.34	0.52	0.61	0.79	0.96
100	0.43	0.65	0.76	0.93	0.98
150	0.45	0.65	0.84	0.94	1.00
200	0.46	0.66	0.86	0.96	1.00
250	0.47	0.74	0.91	0.99	1.00

[1]: H_1: π ~ beta(37.5,12.5); S ~ beta(35,15); C ~ beta(32.5,17.5); $E_{H_1}(p)=0.61$

[2]: H_1: π ~ beta(37.5,12.5); S ~ beta(37.5,12.5); C ~ beta(32.5,17.5); $E_{H_1}(p)=0.65$

[3]: H_1: π ~ beta(37.5,12.5); S ~ beta(40,10); C ~ beta(32.5,17.5); $E_{H_1}(p)=0.69$

[4]: H_1: π ~ beta(37.5,12.5); S ~ beta(42.5,7.5); C ~ beta(32.5,17.5); $E_{H_1}(p)=0.73$

[5]: H_1: π ~ beta(37.5,12.5); S ~ beta(47,3); C ~ beta(32.5,17.5); $E_{H_1}(p)=0.79$

Table 4: Power and threshold cutoff (ϖ) for three alternative hypotheses in the two conditionally independent test setting of Section 3.2. In each case, the null model is H_0: π ~ beta(15,35); S_1 ~ beta(25,25); C_1 ~ beta(30,20); S_2 ~ beta(25,25); C_2 ~ beta(30,20), with $E_{H_0}(p_{11}, p_{10}, p_{01}, p_{00})$=(0.19, 0.24, 0.24, 0.33).

Sample Size	ϖ^1	Power[1]	ϖ^2	Power[2]	ϖ^3	Power[3]
25	-1.41	0.30	-1.50	0.38	-1.32	0.65
50	-1.38	0.35	-1.44	0.53	-0.98	0.83
75	-1.27	0.37	-1.13	0.66	-0.49	0.91
100	-1.23	0.47	-0.92	0.74	-0.21	0.93
$E_{H_1}(p_{11})$		0.15		0.12		0.10
$E_{H_1}(p_{10})$		0.28		0.30		0.33
$E_{H_1}(p_{01})$		0.17		0.15		0.11
$E_{H_1}(p_{00})$		0.40		0.42		0.46

[1]: H_1: π ~ beta(15,35); S_1 ~ beta(25,25); C_1 ~ beta(30,20); S_2 ~ beta(30,20); C_2 ~ beta(40,10).

[2]: H_1: π ~ beta(15,35); S_1 ~ beta(25,25); C_1 ~ beta(30,20); S_2 ~ beta(30,20); C_2 ~ beta(43,7).

[3]: H_1: π ~ beta(15,35); S_1 ~ beta(25,25); C_1 ~ beta(30,20); S_2 ~ beta(30,20); C_2 ~ beta(47.5,2.5).