A Decision-Theoretic Comparison of Treatments to Resolve Air Leaks After Lung Surgery Based on Nonparametric Modeling

Yanxun Xu[*,†], Peter F. Thall[‡], Peter Müller[§], and Mehran J. Reza[¶]

Abstract. We propose a Bayesian nonparametric utility-based group sequential design for a randomized clinical trial to compare a gel sealant to standard care for resolving air leaks after pulmonary resection. Clinically, resolving air leaks in the days soon after surgery is highly important, since longer resolution time produces undesirable complications that require extended hospitalization. The problem of comparing treatments is complicated by the fact that the resolution time distributions are skewed and multi-modal, so using means is misleading. We address these challenges by assuming Bayesian nonparametric probability models for the resolution time distributions and basing the comparative test on weighted means. The weights are elicited as clinical utilities of the resolution times. The proposed design uses posterior expected utilities as group sequential test criteria. The procedure's frequentist properties are studied by extensive simulations.

Keywords: Bayesian nonparametric, clinical trial, mesothelioma, utility function.

1 Introduction

1.1 The motivating clinical trial

Intraoperative air leaks (IALs) occur in 48 to 75% of patients after pulmonary resection (Serra-Mitjans and Belda-Sanchís, 2005). Despite the routine use of intraoperative sutures and stapling devices, IALs remain a significant problem in the practice of thoracic surgery. IALs that persist beyond the immediate postoperative period of five days may result in longer chest tube drainage, greater postoperative pain, increased risk of infection, empyema, thromboemboli, and increased length of hospitalization (Merritt et al., 2010; Singhal et al., 2010). Air leaks are a particularly severe problem in patients with emphysematous lungs or who have undergone extensive visceral pleural denuding procedures, such as pleurectomy decortication. This is a surgical procedure in which the lining surrounding one lung first is removed (pleurectomy), and then any tumor masses that are growing inside the chest cavity are removed (decortication). In addition to the noted risks to the patient, the economic impact of a prolonged air leak is significant, primarily due to increased hospital stay. Because the standard procedure of suturing visible leaks and using staple reinforcement gives unpredictable results, an alternative

---

[*]Dept. of Statistics and Data Sciences, University of Texas at Austin
[†]Dept. of Applied Mathematics and Statistics, Johns Hopkins University
[‡]Dept. of Biostatistics, The University of Texas M.D. Anderson Cancer Center
[§]Dept. of Mathematics, University of Texas at Austin, pmueller@math.utexas.edu
[¶]Dept. of Thoracic Surgery, The University of Texas M.D. Anderson Cancer Center

technique to control IALs is the use of liquid sealants, which are thick fluids instilled in the areas of leaks. Progel (Neomend, Inc., Irvine, CA) is a polymeric biodegradable hydrogel sealant, that currently is the only FDA approved sealant to control IALs during pulmonary resection (Kobayashi et al., 2001).

Despite FDA approval, the true benefit of Progel in reducing the rate of occurrence or duration of IALs in lung resection patients has not been established, and therefore it is not used routinely. Researchers have conducted two studies comparing Progel (treatment group) with standard care (control group) to demonstrate the safety and efficacy of Progel (Allen et al., 2004; Klijian, 2012). Because the study of Allen et al. (2004) varied the application of Progel based on the size of the air bubbles seen in each patient, and the precise methodology of how this was done was not explained in sufficient detail to enable replication, the results of this trial are of limited use for a general comparison of Progel to standard care. The study of Klijian (2012) was retrospective and not randomized. Given these limitations of existing data, the desire to obtain a prospective randomized comparison of Progel to standard care motivated the clinical trial described in this paper. The trial has passed IRB (internal review board) approval and is scheduled to start accrual at The University of Texas M.D. Anderson Cancer Center.

1.2   Modeling considerations

Denote by T the days to resolve IALs, allowing the possibility that an air leak may not develop, represented by $T = 0$. Allen et al. (2004) and Klijian (2012) compared μ_0 and μ_1, the means of T in a control and treatment group, respectively, using a standard t-test, and concluded that Progel was superior to standard care in reducing IALs. Figure 1 plots the histogram of T obtained from non-randomized historical data in the clinical database of the Department of Thoracic and Cardiovascular Surgery at M.D. Anderson Cancer Center. The histogram suggests that a standard parametric model is inappropriate to describe air leak resolution time distributions. For example, a normal or log normal distribution would fail to allow for the observed multi-modality and late resolution times. Moreover, some patients treated with Progel after resection may be free of air leaks immediately following surgery, corresponding to a positive probability mass at $T = 0$.

Let G_1 denote the distribution of T in the treatment (Progel) group and G_0 the distribution of T with the control (standard care). We will represent each G_j, $j = 0, 1$ as a mixture of a point mass δ_0 at 0 and a hypothesized distribution M_j for non-zero resolution times, with M_1 a left-shifted version of M_0 to formalize the assumption that, stochastically, IAL resolution times with Progel are no longer than with standard care. This order constraint is motivated by several medical considerations: Progel is inert, and thus it cannot react chemically with the patient's lung tissue, is not a potential source of infection, and does not slow down the healing process. Moreover, Progel cannot make an air leak worse because it does not contribute to air leak formation. These considerations motivate *a priori* stochastic ordering of G_1 and G_0, which effectively says that, in terms of time to resolve an IAL, Progel may be better than standard care, but it cannot be worse. Nevertheless, for comparison we later report also inference under an otherwise equivalent model without the stochastic ordering constraint.
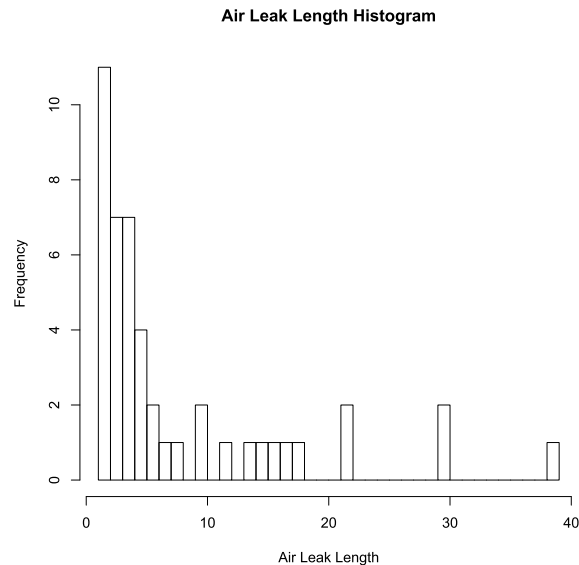
Air Leak Length Histogram

Figure 1: Histogram of times to resolve lung air leaks, from the historical dataset.

An important consideration in developing a trial design is that the use of an expected value as the target for a comparative test is inappropriate and inadequate, both because the historical distribution is skewed with a long right tail, and because a change in the early days after surgery is clinically more important than a comparable change in later days. Also, a standard test of $\mu_1 = \mu_0$ versus $\mu_1 < \mu_0$ would require an impractically large sample size to achieve any reasonable power. These complications are the principal reasons why designing a randomized trial to compare Progel to standard care is non-trivial, and why the use of Progel has not been widely accepted among surgeons who perform pulmonary resections.

The desire to obtain reliable confirmatory evidence to evaluate the comparative benefit of Progel motivates the randomized trial described in this paper. The goal of the trial is to assess the extent to which Progel is superior to standard care. The comparison also allows for the possibility no difference.

1.3   Stochastic ordering and Bayesian nonparametric priors

The time until resolution of air leaks for patients treated with Progel is *a priori* expected to be shorter than under standard care. This introduces a stochastic ordering constraint on G_0 and G_1. Formally, a distribution G_1 *is stochastically smaller than* G_0, denoted by $G_1 \preceq G_0$, if the corresponding cumulative distribution functions satisfy $F_1(t) \geq F_0(t)$ for all t. Lehmann and Romano (2006) and Randles and Wolfe (1979) have modeled stochastic ordering parametrically. Although straightforward, these approaches are limited by the requirement that a parametric family must be specified.

To compare distributions of air leak resolution times, detailed features (e.g., skewed or multi-modal) of the distributions are important, leading us to consider a Bayesian nonparametric (BNP) approach. Importantly, uncertainties about the inference on these details are critical, as posterior probabilities about comparisons drive the decisions about sequential continuation and the terminal decision. Such descriptions of uncertainties are best considered in the framework of a probability model on the unknown distributions, as they are implemented in BNP models.

Formally, BNP refers to prior models for infinite dimensional unknown quantities. Inference for random distributions, like G_0 and G_1 here, is a typical example. A common feature of BNP models is their large support, which allows one to approximate essentially arbitrary distributions (Ishwaran and James, 2001). For the proposed design, we use a model based on the Dirichlet process (DP) prior (Ferguson, 1973), which is by far the most commonly used BNP model for a random distribution. MacEachern (1999) introduced the dependent DP (DDP), which extended the DP to a probability model for a family $\{G_x, \ x \in X\}$ of random probability measures, indexed by some covariate x. The special case of a finite family, like $\{G_0, G_1\}$ in our application, was discussed in De Iorio et al. (2004). Several authors have considered BNP models for stochastically ordered distributions. Gelfand and Kottas (2001) started with two DP random probability measures G_0 and G_1, and used the product of the corresponding cumulative distribution functions to define a pair of stochastically ordered random probability measures. A general methodology for stochastic ordering by considering probability measures constrained to a convex set was proposed by Hoff (2003). Finally, Dunson and Peddada (2008) incorporated stochastic ordering constraints in the DDP prior. In this paper, we use a simple implementation of this finite DDP model with order constraint as the prior probability model for the distributions G_1 and G_0 of leak resolution times under Progel and control. Details are discussed in Section 3. For more extensive reviews of BNP methods, see, for example, Hjort et al. (2010).

Based on the proposed BNP model, we define a utility-based decision criterion to develop a clinical trial design. To our knowledge, there is no literature on using BNP stochastic ordering models in construction of clinical trial designs. The main novelties in the proposed approach are the successful use of utilities in a small scale clinical trial, a convincing case for the need of a full probabilistic description of uncertainties on random probability measures, and a simple and practicable construction of a BNP prior on stochastically ordered random probability measures with point masses. The trial will be conducted in The University of Texas M.D. Anderson Cancer Center, with a co-author of this paper (RM) its Principal Investigator.

Important practical advantages of the proposed approach are that it allows meaningful borrowing of information from historical data (by centering the BNP model), borrowing across treatments (by constructing correlated priors on G_0 and G_1), and the exploitation of stochastic ordering constraints, if warranted and approved in IRB reviews. The use of utility weighting for the outcomes, as in this application, is particularly natural under a BNP model because it allows inference about all aspects of the event time distribution, without constraint to parametric families. Together, these features allow the investigators to plan a much smaller sample size than what would be required

T (days)	0	5	10	15	20	25	30	35	≥ 40
Utility	100	50	10	6	5	4	3	2	0

Table 1: Elicited utility $u(T)$ for $T =$ days to resolve intra-operative air leak.

by a conventional trial design. For example, based on the historical mean of 8 days and standard deviation 8.76, a two sample one-sided 0.05-level t-test with power 0.80 to detect a 25% drop in the mean, from 8 to 6 days, would require a sample of $n = 476$ patients. This is impossible for this single-institution trial. Given the realistic maximum target accrual of 48 patients, the question is whether a design can be constructed that has reasonably high power to conclude that Progel is superior to standard care under clinically meaningful alternatives. We will show that the proposed design, based on differences in mean utilities evaluated from the posterior under the BNP model, has very desirable operating characteristics with $n = 48$ patients. The impact is the opportunity to establish what is expected to be a greatly superior treatment option for patients, with reasonable cost and effort.

2 Utilities and trial design

The primary outcome is T, the time (in days) to resolve an air leak in the lungs following surgery, and we define $Y = \log(T+1)$. The possibility that an air leak may not develop is represented by $T = Y = 0$. The mean, or any other single measure of central tendency of Y, is not an appropriate summary for treatment comparisons. Instead, we take a utility-based approach. Utility-based decision criteria have been used recently in clinical trial designs (Thall and Nguyen, 2012; Lee et al., 2015). We use utilities to weight the importance of air leak resolution times after surgery. For example, a difference of a few days in time to resolution of air leaks in the days immediately after surgery is far more important than a comparable difference in later days. We performed a formal utility elicitation with our clinical collaborator RM. The rationale of the utility elicitation includes: 1) the most desirable resolution time is $T = 0$ (free of air leaks immediately after surgery, although this ideal outcome is almost never seen with standard care); 2) early $(1 \leq T \leq 5)$ resolution of air leaks is very desirable and therefore the interval $[1, 5]$ received a relatively high utility; 3) the utilities drop off steeply for later resolution times $(T > 5)$. These considerations are both medical and economic, and they motivated the elicited utilities $u(Y)$ for $Y = \log(T + 1)$ in Table 1. In the table, the numerical utility 50 assigned to the outcome $T = 5$ days corresponds to the subjective assessment of RM that this comparatively favorable outcome is half as desirable as the ideal outcome of having no air leak at all. Similarly, the utility for $T = 10$ days reflects that this outcome, which involves a long hospital stay and the complications described earlier, is 1/5 as desirable as the outcome $T = 5$ days. We now are ready to define the expected utility for each group as

$$\overline{U}_j = \int u(Y)G_j(dY), \ j = 0, 1, \tag{1}$$

where G_j is a sampling model for the outcome in treatment group j. This expectation is over the distribution of the outcome T, and is conditional on the unknown distribution

G_j. We do not need to make any specific assumptions about G_j yet, except for the existence of such a distribution.

Based on the probability model and the utilities of Table 1, we now define a design for the Progel trial. There are two types of decisions to be considered. At each interim test in the group sequential procedure, we make a stopping decision $d_i \in \{0,1\}$ to stop ($d_i = 0$) or continue ($d_i = 1$). If we reach a predetermined maximum sample size, N, we set $d_N = 0$ by definition. Upon stopping, a terminal decision $a \in \{0,1\}$ reports the final recommendation, with $a = 1$ denoting a recommendation for Progel and $a = 0$ for standard care. A decision-theoretic optimal solution would require backward induction (Bellman, 1957) to solve the full sequential decision problem. We stop short of carrying out this computationally prohibitive solution. Instead, we propose to conduct the trial as follows.

Sequential stopping rule   Patients are enrolled in the trial sequentially in cohorts of size $m = 16$ until a maximum of $N = 48$ patients is reached or early stopping is indicated. All patients are randomized equally to control group and treatment group, with the restriction of perfect balance after each cohort, when the continuation decisions d_i are made.

Denote $\boldsymbol{Y}_n = \{Y_{ji}, \ i = 1, \ldots, n/2, \ j = 0, 1\}$, the observed data for the first n treated patients, that is $n/2$ patients in each group under the restricted equal randomization (rounding $n/2$ for odd n). The proposed decision criterion is the posterior probability

$$\eta(\epsilon_U, \boldsymbol{Y}_n) = p(\overline{U}_1 > \overline{U}_0 + \epsilon_U \mid \boldsymbol{Y}_n), \tag{2}$$

where $\epsilon_U \geq 0$ is a minimum clinically meaningful difference in expected utility. Because the sequential rule makes multiple decisions, as with any group sequential procedure the decision boundaries must be calibrated to control the design's overall false positive error rate. This is similar to the use of so-called alpha-spending functions in conventional frequentist group sequential designs. Like other frequentist summaries, false positive error rate (type-I error) is a probability under an assumed truth, with respect to repeated simulations of the entire trial. In the context of clinical trial designs such summaries under repeated simulations are also known as (frequentist) operating characteristics (OCs). Because evaluating the design's OCs analytically is far too complex, we do this by repeated computer simulations of the design, under an array of different possible scenarios. This follows routine practice in evaluating the behavior of sequentially adaptive clinical trial designs. In the present setting, the OCs are the type I error, mean sample size, and probabilities of different possible decisions (correct decision, stop due to futility, stop due to superiority). Details are reported in Tables 3 and S1 (Xu et al., 2016).

After each cohort, we carry out Markov chain Monte Carlo posterior simulation and evaluate the posterior estimates $\widehat{\eta}(\epsilon_U, \boldsymbol{Y}_n)$. Let ξ_U be an upper probability boundary for which the trial will be terminated early and the treatment arm declared superior if $\widehat{\eta}(\epsilon_U, \boldsymbol{Y}_n) \geq \xi_U$. Similarly, let ξ_L be a lower boundary for which the trial will be terminated early due to futility, with the null hypothesis accepted, if $\widehat{\eta}(\epsilon_U, \boldsymbol{Y}_n) \leq \xi_L$.

The bounds ξ_L and ξ_U are chosen by preliminary simulations to obtain a design with desirable frequentist OCs. In Section 4, we will illustrate how one may calibrate these bounds. In summary, the sequential stopping decision at any point of the trial is: $d_n = 1$ if $\xi_L < \widehat{\eta}(\epsilon_U, \boldsymbol{Y}_n) < \xi_U$; $d_n = 0$ otherwise.

Terminal decision rule Upon stopping, we record the terminal decision $a = 1$ if $\widehat{\eta}(\epsilon_U, \boldsymbol{Y}_n) > \frac{1}{2}$ and $a = 0$ otherwise. Assuming that $\xi_L < 0.5 < \xi_U$, the rule simply records whether we stop due to crossing either the upper or lower bound, respectively. If the trial reaches the maximum number of patients, $N = 48$, the terminal decision uses the threshold $\widehat{\eta} > 0.5$ to determine a recommendation for Progel.

3 Probability model

3.1 Model and properties

We now construct a prior probability model for G_j, the sampling model for Y_{ij} for patients under control ($j = 0$) and Progel ($j = 1$). Because some patients may be free of air leaks immediately following surgery, we allow a point mass at $Y_{ji} = 0$ by defining G_j, $j = 0, 1$, as mixtures

$$
\begin{aligned}
G_j &= \nu_{j0}\delta_0 + \sum_{h=1}^{\infty} \nu_{jh} N(\theta_{jh}, \sigma^2) = \nu_{j0}\delta_0 + (1 - \nu_{j0}) \sum_{h=1}^{\infty} w_h N(\theta_{jh}, \sigma^2) \\
&= \nu_{j0}\delta_0 + (1 - \nu_{j0}) M_j,
\end{aligned}
\tag{3}
$$

where $\sum_{h=1}^{\infty} w_h = 1$. Also, we impose a constraint $\nu_{10} \geq \nu_{00}$ on the probabilities ν_{10} and ν_{00}, and $M_1 \preceq M_0$, formalizing the prior belief that patients are more likely to be free of an air leak in the treatment group than in the control group. For $M_j = \sum_{h=1}^{\infty} w_h N(\theta_{jh}, \sigma^2)$, $j = 0, 1$, we use a DDP prior with common weights and dependent atoms. The common weights w_h have the DP stick-breaking prior, $w_h = v_h \prod_{\ell < h}(1 - v_\ell)$ with $v_h \sim \text{Beta}(1, \alpha)$. The dependent prior on the atoms is constructed as follows, to ensure $M_1 \preceq M_0$. We assume $\boldsymbol{\theta}_h = (\theta_{0h}, \theta_{1h}) \sim M^\star$, where $M^\star$ is a truncated multivariate normal base measure, including a positive probability κ for ties $\theta_{0h} = \theta_{1h}$:

$$
M^\star(\boldsymbol{\theta}_h) = N(\theta_{1h} \mid \mu_1, \sigma_1^2) \left(\kappa I(\theta_{0h} = \theta_{1h}) + (1 - \kappa) N_+(\theta_{0h} \mid \theta_{1h}, \tau^2) \right),
\tag{4}
$$

where $N_+(x \mid m, V)$ refers to a truncated normal random variable x subject to $x \geq m$, and $\kappa = p(\theta_{0h} = \theta_{1h})$. For comparison we will also consider inference under a variation of model (4) without the order constraint, replacing the N_+ kernel by an unconstrained normal $N(\theta_{1h}, \tau^2)$.

Denote $\widetilde{M}_j = \sum_h w_h \delta_{\theta_{jh}}$, where $\delta_{\theta_{jh}}$ denotes a point mass at θ_{jh}, $j = 0, 1$. It is straightforward to show that $\widetilde{M}_1 \preceq \widetilde{M}_0$, which implies $M_1 \preceq M_0$, and this in turn implies $G_1 \preceq G_0$, as desired. Barrientos et al. (2012) study the support properties of various DDP models. Applying Theorem 2 of Barrientos et al. (2012), it follows that the proposed model has full support over all pairs of stochastically ordered random probability measures.

For reference we state the complete model,

$$Y_{ji} \mid \nu_{j0}, \nu_{j1}, M_0, M_1 \quad \sim \quad G_j = \nu_{j0}\delta_0 + (1 - \nu_{j0})\sum_{h=1}^{\infty} w_h N(Y_{ji} \mid \theta_{jh}, \sigma^2)$$

$$(\theta_{0h}, \theta_{1h}) \mid \mu_1, \sigma_1, \kappa, \tau \quad \sim \quad M^{\star}. \tag{5}$$

We complete the model specification with choices for the hyperparameters $\nu_{j0}, \nu_{j1}, \sigma^2$, $\kappa, \mu_1, \sigma_1, \tau$. In the context of clinical trial design, the hyperparameters should not introduce inappropriately strong information into the prior. To ensure this, we provide the following guidelines.

We first standardize the data by subtracting the sample mean $\bar{Y}_1$ of the Y_{1i}'s of the treatment group and scaling with the sample standard deviation s_1, mapping $Y_{ji} \to (Y_{ji} - \bar{Y}_1)/s_1$. This is done to mitigate sensitivity to the measurement scale. We fix $\mu_1 = 0$ and $\sigma_1 = 1$, to reflect the standardization. For σ^2, we assume $p(1/\sigma^2) = \mathrm{Ga}(0.001, 0.001)$ to ensure that the prior is not too informative, where $\mathrm{Ga}(a, b)$ denotes a gamma distribution with mean a/b. To allow for a wide range of shifts in the response density, we specify $p(1/\tau^2) = \mathrm{Ga}(0.5, 0.5)$. This implies a Cauchy distribution for θ_{0h}, which often is used as a robust choice in parametric models. To satisfy the constraint $\nu_{10} \geq \nu_{00}$, we let $\zeta_0 = \nu_{00}$, $\zeta_1 = \nu_{10} - \nu_{00}$, and assume $p(\zeta_0, \zeta_1) = \mathrm{Dirichlet}(0.1, 0.1, 0.1)$. Finally, we assume $p(\kappa) = \mathrm{Beta}(1, 1)$ and $p(\alpha) = \mathrm{Ga}(1, 1)$. The conjugacy of the implied normal on $\boldsymbol{\theta}_h$ in (4) and the normal kernel in (5) greatly simplify posterior inference. Any Markov chain Monte Carlo (MCMC) scheme for DP mixture model as described, for example, in Neal (2000), can be applied. In our implementation, we used an implementation based on the finite DP (Ishwaran and James, 2001), which truncates the infinite sum in the DP mixture model after a finite number of terms. We used $H = 10$, following a recommendation based on Theorem 1 in Ishwaran and James (2002) that gives tight bounds on the approximation error, well below what is clinically relevant in this application. Details of the MCMC implementation are presented in Supplement A.

We carried out a preliminary simulation study to better understand the nature and accuracy of posterior inference under the proposed model for a reasonable sample size. The simulation setup and results are summarized in Supplement B. The inference under the proposed method incorporating the stochastic ordering constraint performed well, indicating small bias even with moderate sample size.

4   Trial simulation study

To assess average behavior of the proposed BNP trial design, we performed an extensive simulation study under a variety of scenarios that were constructed to mimic the Progel trial. For the proposed stopping and decision rules, we fixed the parameters as $\xi_U = 0.9, \xi_L = 0.05$, based on preliminary studies (described later) and examining the OCs of the proposed BNP design. In all scenarios, we set the maximum number of patients to be $N = 48$, randomized equally between the control and treatment group, with cohorts of 16 patients. The smallest clinically meaningful improvement used to define the decision criterion $\eta(\epsilon_U, \boldsymbol{Y}_n)$ was determined by our clinical collaborator (RM) to be $\epsilon_U = 18$, given the numerical utilities of IAL resolution times in Table 1.

Scenario	$U_1^o - U_0^o$	Group	U_j^o	Simulation truth G_j^o
1	0	Progel	-	Resample from historical data
		Control	-	Resample from historical data
2	0	Progel	23.44	$0.1\delta_0 + 0.63N(2,\bar{\sigma}^2) + 0.27N(3,\bar{\sigma}^2)$
		Control	23.44	$0.1\delta_0 + 0.63N(2,\bar{\sigma}^2) + 0.27N(3,\bar{\sigma}^2)$
3	41.64	Progel	57.25	$0.3\delta_0 + 0.49N(1,\bar{\sigma}^2) + 0.14N(2,\bar{\sigma}^2) + 0.07N(3.5,\bar{\sigma}^2)$
		Control	15.61	$0.1\delta_0 + 0.63N(2.5,\bar{\sigma}^2) + 0.18N(3,\bar{\sigma}^2) + 0.09N(4.5,\bar{\sigma}^2)$
4	19.15	Progel	48.94	$0.2\delta_0 + 0.56N(1.5,\bar{\sigma}^2) + 0.24N(2,\bar{\sigma}^2)$
		Control	29.79	$0.1\delta_0 + 0.63N(1.8,\bar{\sigma}^2) + 0.27N(3,\bar{\sigma}^2)$
5	29.68	Progel	64.82	$0.4\delta_0 + 0.48N(1,\bar{\sigma}^2) + 0.12N(2.5,\bar{\sigma}^2)$
		Control	35.14	$0.1\delta_0 + 0.54N(1.5,\bar{\sigma}^2) + 0.18N(2.5,\bar{\sigma}^2) +$
				$0.18N(3.5,\bar{\sigma}^2)$
6	34.15	Progel	60.80	$0.4\delta_0 + 0.36N(1,\bar{\sigma}^2) + 0.12N(2,\bar{\sigma}^2) + 0.12N(3,\bar{\sigma}^2)$
		Control	26.65	$0.1\delta_0 + 0.36N(1.5,\bar{\sigma}^2) + 0.54N(3.5,\bar{\sigma}^2)$
7	43.47	Progel	55.33	$0.3\delta_0 + 0.3Exp(1) + 0.4Exp(0.5)$
		Control	11.86	$0.1\delta_0 + 0.4N(3,0.2^2) + 0.5N(4,0.2^2)$
8	8.13	Progel	45.08	$0.2\delta_0 + 0.4\text{Weib}(1,2) + 0.4\text{Weib}(0.7,2)$
		Control	36.95	$0.2\delta_0 + 0.5N(1.8,0.2^2) + 0.3N(2.5,0.3^2)$
9	10.25	Progel	55.33	$0.3\delta_0 + 0.3Exp(1) + 0.4Exp(0.5)$
		Control	45.08	$0.2\delta_0 + 0.4\text{Weib}(1,2) + 0.4\text{Weib}(0.7,2)$

Table 2: In each scenario, the models G_j^o in the right column are the simulation truths. Here $\bar{\sigma} = 0.3$ and $\text{Exp}(\cdot), \text{Weib}(\cdot,\cdot)$ denote an exponential distribution and a Weibull distribution, respectively. U_j^o reports the expected utilities under the simulation truth G_j^o. The second column reports the true difference $U_1^o - U_0^o$.

We considered nine scenarios, and simulated 100 trials for each scenario. The response outcomes Y_{ji} were generated from the simulation truth G_j^o shown in the last column of Table 2. Other columns in the same table show the true utilities $U_j^o = \int u(y) \, dG_j^o(y)$ for each arm and the differences $U_1^o - U_0^o$.

To calculate type I error and power, we define the null hypothesis $H_0 : G_1 = G_0$. Under the proposed design, the test rejects H_0 in favor of Progel if $\widehat{\eta}(\epsilon_U, \boldsymbol{Y}_n) \geq \xi_U$ interimly with early stopping at n=16 or 32, and if $\widehat{\eta}(\epsilon_U, \boldsymbol{Y}_N) \geq 0.5$ for the terminal rule at $N = 48$. Similarly, the test fails to reject H_0 if $\widehat{\eta}(\epsilon_U, \boldsymbol{Y}_n) \leq \xi_L$ ($n = 16, 32$), with early stopping for futility, and if $\widehat{\eta}(\epsilon_U, \boldsymbol{Y}_N) < 0.5$ at $N = 48$.

We fixed the hyperparameters as described earlier in Section 3 and fit the proposed BNP model (5) to each simulated data set. Table 3 summarizes the OCs of the proposed BNP utility-based design for nine scenarios. The OCs include the average number of patients treated, type I error, the probabilities of making the correct decision (PCD), stopping the trial early due to either superiority, Pr(EarS), or futility, Pr(EarF) and, in a final analysis without early stopping, declaring superiority, Pr(FinS), or futility Pr(FinS). For comparison we also implemented inference under a variation of the model without the stochastic ordering constraint, that is, model (4) with an unconstrained normal $N(\theta_{0h} \mid \theta_{1h}, \tau^2)$ replacing the truncated normal in (4). Scenarios 1a and 2a

Scenario	$U_1^o - U_0^o$	MSS	TIE	PCD	$Pr(\text{EarS})$	$Pr(\text{FinS})$	$Pr(\text{EarF})$	$Pr(\text{FinF})$
1	0	16.80	0.00	1.00	0.00	0.00	1.00	0.00
1a	0	16.32	0.00	1.00	0.00	0.00	1.00	0.00
2	0	28.80	0.02	0.98	0.01	0.01	0.64	0.34
2a	0	28.00	0.01	0.99	0.00	0.01	0.69	0.30
3	41.64	29.12	-	1.00	0.81	0.18	0.01	0.00
4	19.15	40.16	-	0.63	0.15	0.48	0.15	0.22
5	29.68	31.68	-	0.93	0.60	0.33	0.04	0.03
6	34.15	29.92	-	0.94	0.65	0.29	0.05	0.01
7	43.47	28.64	-	0.96	0.77	0.19	0.04	0.00
8	8.13	34.08	-	0.79	0.04	0.17	0.45	0.34
9	10.25	34.88	-	0.74	0.10	0.13	0.34	0.43

Table 3: Trial simulation results. MSS = mean sample size, TIE = type I error, PCD = probability of making the correct decision, $Pr(\text{EarS})$ = probability of stopping early due to superiority, $Pr(\text{EarF})$ = probability of stopping early due to futility, $Pr(\text{FinS})$ = probability of declaring superiority in a final analysis without early stopping, and $Pr(\text{FinF})$ = probability of declaring futility in a final analysis without early stopping. All probabilities are computed by repeated simulations.

show summaries of inference under this unconstrained version of the model, using the same simulation truths as in scenarios 1 and 2.

Details of the simulation results are discussed in Supplement C. Scenarios 1 and 2 are null scenarios; in Scenario 3 we assumed a large treatment effect of $U_1^o - U_0^o = 41.6$, far beyond $\epsilon_U = 18$; Scenario 4 has a small treatment effect of $U_1^o - U_0^o = 19.2$, barely beyond ϵ_U; under Scenarios 5 and 6 we assumed a moderate treatment effect; and the last three Scenarios 7, 8, and 9 have simulation truths different from the assumed mixture of normal distributions.

For reference we also evaluated summaries related to estimation. Denote the true utility difference under the simulation truth by $\Delta U^\star = U_1^o - U_0^o$, and $\Delta U = \overline{U}_1 - \overline{U}_0$. For each scenario, we computed the estimation bias $E\{\Delta U - \Delta U^\star\}$ and root mean squared error (RMSE) $\sqrt{E\{(\Delta U - \Delta U^\star)^2\}}$, where the expectation is over repeated simulations under each scenario. The results are given in Table S1.

The OCs under all nine scenarios are given in Table 3, and show a favorable evaluation of the proposed design. The results under scenarios 1a and 2a, compared to scenarios 1 and 2, show that the design's frequentist OCs appear to be robust with respect to the inclusion or not of the constraint $G_0 < G_1$ in the prior. The simulation truths in scenarios 7, 8, and 9 are different from the assumed mixture of normal distributions with equal variance. The results under these scenarios demonstrate the flexibility of BNP mixture models with a common variance parameter. In summary, inferences under the proposed BNP model and trial monitoring rules exhibit desirable OCs across all nine scenarios.

Scenario	True Diff	Group	MSS	PCD	Pr(EarS)	Pr(FinS)	Pr(EarF)	Pr(FinF)
3	41.64	BNPSO	29.12	1.00	0.81	0.18	0.01	0.00
		Z-Weib	44.32	0.65	0.10	0.52	0.04	0.34
4	19.15	BNPSO	40.16	0.63	0.15	0.48	0.15	0.22
		Z-Weib	42.40	0.17	0.03	0.14	0.18	0.65
6	34.15	BNPSO	29.92	0.94	0.65	0.29	0.05	0.01
		Z-Weib	40.00	0.85	0.28	0.57	0.02	0.13
7	43.47	BNPSO	28.64	0.96	0.77	0.19	0.04	0.00
		Z-Weib	41.92	0.87	0.20	0.65	0.04	0.11
8	8.13	BNPSO	34.08	0.79	0.04	0.17	0.45	0.34
		Z-Weib	45.28	0.42	0.06	0.50	0.05	0.39

Table 4: Comparisons in selected scenarios under the proposed BNP model with stochastic ordering (BNPSO) versus an alternative parametric model with a zero-enriched Weibull (Z-Weib).

Parametric models and sensitivity analyses For comparison, we implemented alternative inference under a parametric model assuming a zero-enriched Weibull distribution, that is, a mixture of point mass at 0 and Weibull distribution. We assumed $T_{ij} \overset{\text{i.i.d.}}{\sim} G_j^W$, $i = 1, \ldots, n_j$ for groups $j = 0$ and $j = 1$, using $G_j^W = \pi_j \delta_0 + (1 - \pi_j)\text{Weib}(\lambda_{1j}, \lambda_{2j})$. We completed the model with a prior $p(\pi_j) = \text{Beta}(0.1, 0.1)$ and a conjugate prior $p(\lambda_{2j}) = \text{InvGa}(b_{1j}, b_{2j})$. The hyperparameters b_{1j} and b_{2j} were determined by matching the prior mean of λ_{2j} with a maximum likelihood estimate and assuming a prior variance of 10. Finally, for λ_{1j} there is no conjugate prior. We followed Fink (1997) by assuming $p(\lambda_{1j}) \propto \lambda_{1j}^{a_{1j}} \exp(-a_{2j}\lambda_{1j} - \frac{a_{3j}^{\lambda_{1j}}}{\lambda_{2j}})$, with $a_{1j} = 1$, $a_{2j} = \log(\prod_{i=1}^{n_j} T_{ij}) + 2$, and $a_{3j} = 2$, $j = 0, 1$.

Table 4 shows the OCs comparing the inferences under the proposed model versus the zero-enriched parametric Weibull model in some (arbitrarily) selected scenarios. The proposed BNP model with stochastic ordering compares quite favorably, with much larger probabilities of making a correct decision and correctly stopping early for superiority.

Finally, we carried out an alternative analysis to understand how much the results might change if different utilities $u(t)$ were elicited. Table S2 in the supplement presents the results of a sensitive analysis using different utilities. In summary, while the actual decisions naturally change, the frequentist OCs change only slightly. Different decisions are desirable, under different utilities that reflect different clinical preferences.

A final set of simulations explored robustness with respect to the decision boundaries ξ_U and ξ_L for the continuation decision. Table S3 summarizes OCs under Scenarios 2, 3, and 4. Again, while some summaries, like the probability of early stopping for futility, change in the expected direction, the nature of the overall comparison across scenarios remains unchanged under different criteria.

5   Conclusions and discussion

We developed a Bayesian nonparametric (BNP) utility-based group sequential design to compare Progel with standard care in resolving air leaks after lung surgery. In this setting, standard statistical tests or parametric models are not appropriate for trial designs or to describe air leak resolution time distributions. We solved the problem by developing a BNP model with a stochastic ordering constraint and proposing a trial design based on expected utility, computed from elicited utility values. The model assessment and trial simulation studies show unbiased results and desirable OCs.

Beyond the application discussed in this paper, the proposed BNP utility-based method can be extended to many other contexts. For example, in applications that involve multiple groups, one may replace the truncated bivariate normal base measure M^* in (5) with a truncated multivariate normal distribution that incorporates the desired stochastic ordering constraints. Furthermore, the hypothesis testing framework discussed in Section 4 can be extended easily to testing equalities in multiple distributions that are stochastically ordered.

Finally, we note that the BNP model could be replaced by a sufficiently flexible parametric model without any substantial change in the performance of the proposed design. For example, one could use a mixture of $H = 5$ normals as the model. However, the computational effort for posterior simulation in any finite mixture of normal model is nothing less than in the proposed DDP model. We prefer the BNP model for reasons of conceptual clarity and, in principle, natural scaling to larger sample sizes and greater precision.

Supplementary Material

Supplement: A Decision-Theoretic Comparison of Treatments to Resolve Air Leaks After Lung Surgery Based on Nonparametric Modeling (DOI: 10.1214/16-BA1016SUPP; .pdf). It includes all the MCMC implementation details, model assessment simulation, more trial simulation results, and sensitivity analyses.

References

Allen, M. S., Wood, D. E., Hawkinson, R. W., Harpole, D. H., McKenna, R. J., Walsh, G. L., Vallieres, E., Miller, D. L., Nichols, F. C., Smythe, W. R., et al. (2004). "Prospective randomized study evaluating a biodegradable polymeric sealant for sealing intraoperative air leaks that occur during pulmonary resection." *The Annals of Thoracic Surgery*, 77(5): 1792–1801.   640

Barrientos, A. F., Jara, A., Quintana, F. A., et al. (2012). "On the support of MacEachern's dependent Dirichlet processes and extensions." *Bayesian Analysis*, 7(2): 277–310.   MR2934952. doi: http://dx.doi.org/10.1214/12-BA709.   645

Bellman, R. E. (1957). *Dynamic Programming.* Princeton University Press. MR0090477. 644

De Iorio, M., Müller, P., Rosner, G. L., and MacEachern, S. N. (2004). "An ANOVA model for dependent random measures." *Journal of the American Statistical Association*, 99(465): 205–215. MR2054299. doi: http://dx.doi.org/10.1198/016214504000000205. 642

Dunson, D. B. and Peddada, S. D. (2008). "Bayesian nonparametric inference on stochastic ordering." *Biometrika*, 95(4): 859–874. MR2461216. doi: http://dx.doi.org/10.1093/biomet/asn043. 642

Ferguson, T. S. (1973). "A Bayesian analysis of some nonparametric problems." *The Annals of Statistics*, 209–230. MR0350949. 642

Fink, D. (1997). "A compendium of conjugate priors." See http://www.leg.ufpr.br/lib/exe/fetch.php/projetos:mci:tabelasprioris.pdf, 46. 649

Gelfand, A. E. and Kottas, A. (2001). "Nonparametric Bayesian modeling for stochastic order." *Annals of the Institute of Statistical Mathematics*, 53(4): 865–876. MR1880817. doi: http://dx.doi.org/10.1023/A:1014629724913. 642

Hjort, N. L., Holmes, C., Müller, P., and Walker, S. (2010). *Bayesian Nonparametrics.* Cambridge, UK: Cambridge University Press. MR2722987. doi: http://dx.doi.org/10.1017/CBO9780511802478. 642

Hoff, P. D. (2003). "Bayesian methods for partial stochastic orderings." *Biometrika*, 90(2): 303–317. MR1986648. doi: http://dx.doi.org/10.1093/biomet/90.2.303. 642

Ishwaran, H. and James, L. F. (2001). "Gibbs sampling methods for stick-breaking priors." *Journal of the American Statistical Association*, 96(453). MR1952729. doi: http://dx.doi.org/10.1198/016214501750332758. 642, 646

Ishwaran, H. and James, L. F. (2002). "Approximate Dirichlet Process computing in finite normal mixtures." *Journal of Computational and Graphical Statistics*, 11(3). 646

Klijian, A. (2012). "A novel approach to control air leaks in complex lung surgery: a retrospective review." *Journal of Cardiothoracic Surgery*, 7: 49–54. 640

Kobayashi, H., Sekine, T., Nakamura, T., and Shimizu, Y. (2001). "In vivo evaluation of a new sealant material on a rat lung air leak model." *Journal of Biomedical Materials Research*, 58(6): 658–665. 640

Lee, J., Thall, P. F., Ji, Y., and Müller, P. (2015). "Bayesian dose-finding in two treatment cycles based on the joint utility of efficacy and toxicity." *Journal of the American Statistical Association*, 110(510): 711–722. MR3367259. doi: http://dx.doi.org/10.1080/01621459.2014.926815. 643

Lehmann, E. L. and Romano, J. P. (2006). *Testing Statistical Hypotheses.* Springer Science & Business Media. MR2135927. 641

MacEachern, S. (1999). "Dependent nonparametric processes." In *ASA Proceedings of the Section on Bayesian Statistical Science.* Alexandria, VA: American Statistical Association.   642

Merritt, R. E., Singhal, S., and Shrager, J. B. (2010). "Evidence-based suggestions for management of air leaks." *Thoracic Surgery Clinics*, 20(3): 435–448.   639

Neal, R. M. (2000). "Markov chain sampling methods for Dirichlet process mixture models." *Journal of Computational and Graphical Statistics*, 9(2): 249–265. MR1823804. doi: http://dx.doi.org/10.2307/1390653.   646

Randles, R. H. and Wolfe, D. A. (1979). *Introduction to the Theory of Nonparametric Statistics*, volume 1. New York: Wiley. MR0547836.   641

Serra-Mitjans, M. and Belda-Sanchís, J. (2005). "Surgical sealant for preventing air leaks after pulmonary resections in patients with lung cancer." *The Cochrane Library.*   639

Singhal, S., Ferraris, V. A., Bridges, C. R., Clough, E. R., Mitchell, J. D., Fernando, H. C., and Shrager, J. B. (2010). "Management of alveolar air leaks after pulmonary resection." *The Annals of Thoracic Surgery*, 89(4): 1327–1335.   639

Thall, P. F. and Nguyen, H. Q. (2012). "Adaptive randomization to improve utility-based dose-finding with bivariate ordinal outcomes." *Journal of Biopharmaceutical Statistics*, 22(4): 785–801.   MR2931071. doi: http://dx.doi.org/10.1080/10543406.2012.676586.   643

Xu, D., Thall, P. F., Müller, P., and Reza, M. J. (2016). "Supplement: A decision-theoretic comparison of treatments to resolve air leaks after lung surgery based on nonparametric modeling." *Bayesian Analysis.* doi: http://dx.doi.org/10.1214/16-BA1016SUPP.   644

Acknowledgments