


## MAIN PAPER

# Bayesian nonparametric statistics: A new toolkit for discovery in cancer research

Peter F. Thall<sup>1</sup>  | Peter Mueller<sup>2</sup> | Yanxun Xu<sup>3</sup> | Michele Guindani<sup>4</sup>

<sup>1</sup>Department of Biostatistics, University of Texas, MD Anderson Cancer Center, Houston, TX, USA

<sup>2</sup>Department of Mathematics, University of Texas at Austin, Austin, TX, USA

<sup>3</sup>Department of Applied Mathematics and Statistics, Johns Hopkins University, Baltimore, MD, USA

<sup>4</sup>Department of Statistics, University of California, Irvine, Irvine, CA, USA

## Correspondence

Peter F. Thall, Department of Biostatistics, The University of Texas, MD Anderson Cancer Center, 1515 Holcombe Boulevard, Houston, TX 77030-4009, USA.  
Email: rex@mdanderson.org

## Funding information

National Cancer Institute (NCI), Grant/Award Number: R01 CA157458-01A1, RO1 CA 83932.; NCI Cancer Center Core, Grant/Award Number: CA016672

Many commonly used statistical methods for data analysis or clinical trial design rely on incorrect assumptions or assume an over-simplified framework that ignores important information. Such statistical practices may lead to incorrect conclusions about treatment effects or clinical trial designs that are impractical or that do not accurately reflect the investigator's goals. Bayesian nonparametric (BNP) models and methods are a very flexible new class of statistical tools that can overcome such limitations. This is because BNP models can accurately approximate any distribution or function and can accommodate a broad range of statistical problems, including density estimation, regression, survival analysis, graphical modeling, neural networks, classification, clustering, population models, forecasting and prediction, spatiotemporal models, and causal inference. This paper describes 3 illustrative applications of BNP methods, including a randomized clinical trial to compare treatments for intraoperative air leaks after pulmonary resection, estimating survival time with different multi-stage chemotherapy regimes for acute leukemia, and evaluating joint effects of targeted treatment and an intermediate biological outcome on progression-free survival time in prostate cancer.

## KEYWORDS

Bayesian nonparametric statistics, clinical trial design, density estimation, dynamic treatment regime, targeted therapy

## 1 | INTRODUCTION

Many statistical methods commonly used for analyzing medical data or designing clinical trials rely on incorrect assumptions, or they assume an over-simplified framework that ignores important information. A common example is a statistical analysis based on the assumption that the data are normally distributed, despite the fact that the histogram is not bell shaped or has multiple modes. A second example is comparison of survival times for frontline anticancer treatments while ignoring subsequent salvage therapies, which fails to account for effects of salvage on survival. A third example is using the Cox model<sup>[1]</sup> to analyze survival data from 2 treatments having Kaplan-Meier plots<sup>[2]</sup> that cross, which implies that the proportional hazards assumption underlying the Cox model cannot be correct. Such statistical practices may lead to incorrect conclusions about treatment effects, which serve physicians and patients poorly.

In these and other examples, analysis of medical data often involves an unknown probability distribution, which we denote by  $G$ . Statistical inference usually proceeds by assuming that  $G$  is a member of some parametric family, such as normal or Weibull families of distributions, having unknown parameters. Similar assumptions are made for an unknown function,  $f$ , which may describe how the probability of response varies with dose or may be the longitudinal profile of a biological variable, such as prostate specific antigen, that is recorded repeatedly over time. Key questions are how the distribution  $G$  or function  $f$  depend

on treatment or other patient prognostic covariates. If one assumes a particular parametric form for  $G$  or  $f$ , and this assumed model is incorrect or overly restrictive, inference will be restricted to that assumed form, possibly leading to incorrect conclusions. In some applications, the conventional statistical practice of fitting several models to a dataset and choosing the model that gives the best fit may mitigate this problem, but it still is limited by the flexibility of models that are considered.

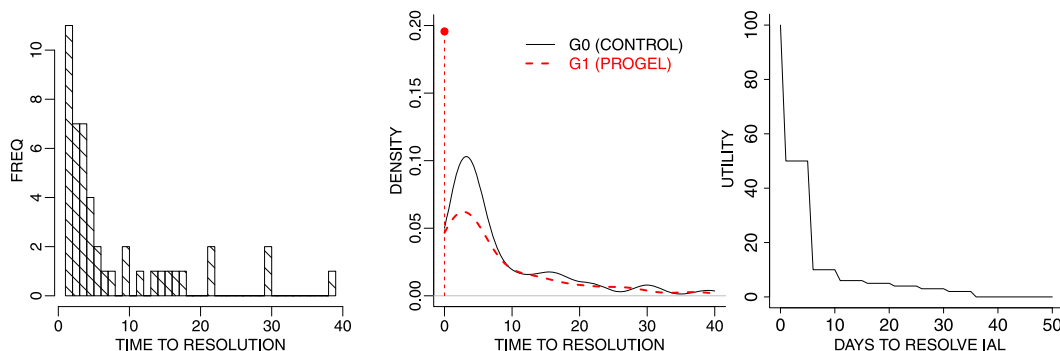
Bayesian nonparametric (BNP) models and methods (see, eg, Hjort et al,<sup>[3]</sup> Walker,<sup>[4]</sup> or Mueller et al<sup>[5]</sup>) are a new class of statistical tools that can overcome some limitations of conventional methods, including the specific issues in the examples described above. Bayesian nonparametric models are very flexible, and the family of BNP models can accommodate a broad range of different data structures and statistical problems. These include density estimation, regression, survival analysis, graphical modeling, neural networks, classification, clustering, population models, forecasting and prediction, spatiotemporal models, and causal inference.<sup>[6]</sup> A major advantage of BNP models is that, on the basis of the observed data, they can accurately approximate essentially any distribution or function, a property known as “full support.” For example, Barrientos et al<sup>[7]</sup> give a formal explanation of such full support properties for the widely used dependent Dirichlet process model.<sup>[8,9]</sup> For this reason, the class of BNP probability models and accompanying statistical methods can, in some cases, help investigators avoid making invalid statistical inferences due to assuming an overly restrictive model. An attractive feature of BNP models is that they often identify unexpected structures in a dataset that cannot be seen using conventional statistical models and methods. Examples of such structures include patient clusters, treatment-subgroup interactions, gene signatures, and complex patterns of biomarker change over time.

This paper presents 3 examples that illustrate how BNP models and methods can be applied to obtain insights that may not be provided by conventional statistical methods. The examples include a clinical trial to compare treatments for intraoperative air leaks (IALs) after pulmonary resection, estimation of expected survival time with different multistage chemotherapy (chemo) regimes for acute leukemia, and evaluation of the joint effects of a targeted treatment and an intermediate biological outcome on progression-free survival (PFS) time in prostate cancer. We focus on describing how each BNP method addresses critical aspects of the problem that conventional statistical models and methods may not accommodate. In describing each analysis, we minimize technical details, include graphical illustrations, and provide references to methodological papers. Recent reviews of BNP models and methods are given, for example, by Mueller et al<sup>[5]</sup> and Dunson,<sup>[10]</sup> and specifically for biomedical problems by Mitra and Mueller.<sup>[6]</sup>

## 2 | COMPARING TREATMENTS FOR AIR LEAKS AFTER LUNG SURGERY

Xu et al<sup>[11]</sup> apply a BNP model to construct a clinical trial design for patients who experience IALs after pulmonary resection. The goal of the trial is to compare the use of a novel hydrogel sealant, Progel, to the standard procedure of using sutures and staples. Although Progel has been approved by the US Food and Drug Administration, its efficacy for reducing IALs in lung resection patients has not been established, and it is not used routinely.

Historical data from 46 mesothelioma patients who underwent lung resection and received standard care for IALs show that the distribution of  $Y =$  time to resolve an air leak is not bell-shaped and does not follow any standard parametric form, like a normal, lognormal, or Weibull distribution. Figure 1 (left) shows that the histogram of  $Y$  values has an irregular shape, with multiple modes. The trial's hypothesis is that Progel may reduce the average time to resolve an IAL, and it possibly may prevent



**FIGURE 1** Left side, time to resolution of intraoperative air leaks (historical data). Middle, density estimates for time to resolution of intraoperative air leak (IAL) under control ( $G_0$ , black) and hypothesized density under Progel ( $G_1$ , red). Right side, plot of elicited utility as a function of time to resolve an IAL

IALs in some patients ( $Y = 0$ ). The multimodality and possibility that  $Y = 0$  present technical complications that standard statistical distributions and methods cannot accommodate. To construct a clinical trial design to compare Progel to standard care, conventional methods that compare mean times of the two treatments based on a parametric distribution for computing power and sample size are not appropriate. First, a conventional test generally ignores the spike at 0. Given the irregular shape of the distribution in Figure 1 (left), and the possibility that  $Y = 0$  with Progel, summarizing the two IAL distributions in terms of their means is misleading.

Conventional approaches also lead to designs requiring sample sizes that were not considered feasible in the single-institution setting where the trial was planned. For example, assuming that  $Y$  is normal, based on the historical mean of 8 days and standard deviation of 8.76, a 2-sample 1-sided .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 476 patients.

Alternatively, using a log scale, one could compute sample size based on a historical mean of 1.61 and standard deviation of 0.97, and a drop from  $\log(8)$  to  $\log(6)$  days. This design would require a sample size of 280 patients, which still would be an infeasible accrual target for this medical condition and treatment in a single institution trial.

For later comparison, we briefly outline how a conventional parametric analysis that accounts for the possibility that  $Y = 0$  might be set up. As a conventional comparator, we can assume a zero-enriched Weibull distribution, that is, a 2-component mixture of a point mass at 0 and a Weibull. Inference under this model can be formulated easily based on maximum likelihood estimation. The main limitation of this setup is that actual differences between the distributions under Progel and standard care might not be well characterized by shifts of the entire distribution, which is all that the Weibull model can accommodate.

Taking a BNP approach to constructing a design, the first step is to apply BNP density estimation (Mueller and Mitra<sup>[12]</sup>) to estimate the distribution, denoted by  $G_0$ , of the IAL resolution times for standard care based on the historical data. This estimate is used as a basis for comparison, and for constructing a hypothesized distribution,  $G_1$ , for Progel. The BNP estimate of  $G_0$  and hypothesized  $G_1$  are given in Figure 1 (middle), which shows that  $G_0$  is a smoothed version of the historical data histogram (black curve) and  $G_1$  (red dashed curve) is a left-shifted version of  $G_0$ . The red spike at 0 for  $G_1$  formalizes the hypothesis that air leaks might not develop at all in some patients given Progel, although the model is formulated generally to allow the possibility that the standard distribution,  $G_0$ , also may have a spike at 0. A major advantage of BNP density estimation is that it naturally accommodates irregular shaped distributions, including the spike at 0.

Inference, based on the trial data, comparing  $G_1$  to  $G_0$  under the BNP model characterizes uncertainty about the fitted distribution  $G_0$ , rather than treating it as fixed. The estimated density of  $G_0$  in Figure 1 (middle) illustrates how the BNP model, known as a dependent Dirichlet process (DDP),<sup>[8,9]</sup> can be applied to estimate the distribution of IAL resolution times based on the historical data. The DDP estimate of  $G_0$  is essentially a weighted average, or “mixture,” of normal distributions, with mixture components corresponding to the peaks in the histogram in Figure 1 (left). The joint model for the two distributions requires  $G_1$  to be shifted to the left of  $G_0$ , formally; they are “stochastically ordered”; and it also allows  $G_1$  or  $G_0$  to have a spike at 0, corresponding to immediate resolution of air leaks. Although this did not occur with standard care in the historical data, it is clinically possible and is accommodated by the probability model. Therefore, we assume that each of the distributions  $G_1$  (Progel) and  $G_0$  (standard care) of time to resolve an IAL is an average (mixture) of a point mass at 0 (no air leak) and a weighted sum of log normal distributions having different means. These BNP distributions are extremely flexible and can closely fit virtually any observed data, which in turn leads to a very reliable clinical trial design.

The stochastic ordering in the joint model of  $G_1$  and  $G_0$  formalizes the assumptions that, compared to the standard, Progel is more likely to have no air leak and also has shorter expected resolution times for any air leaks that occur. It also implies that  $G_1$  has smaller means than  $G_0$  for the normal distributions in the weighted sum. These model assumptions reflect the facts that Progel is inert and thus cannot react chemically with the patient's lung tissue, is not a potential source of infection, does not slow down the healing process, and does not contribute to air leak formation.

The trial design based on the BNP model uses decision rules based on utilities of possible  $Y$  values, described below, that were elicited from the trial's principal investigator. This design requires a maximum of 48 patients, much smaller than 476 or 280 patients required by conventional methods. In simulations, we found this to be a sufficient sample size for meaningful inference under the proposed BNP model under a set of realistic scenarios.

To quantify the clinical relevance of each possible IAL resolution time  $Y$ , numerical utilities quantifying the desirabilities of the possible values of  $Y$  were elicited from the oncologist planning the trial. These are shown in Table 1 and plotted in the right side of Figure 1. The numerical utility,  $u(Y)$ , of each resolution time  $Y$  quantifies its relative desirability.

The highly nonlinear nature of the clinical preferences in  $u(Y)$  reflects the large clinical utility of early resolution of IALs, with short resolution times far more preferable than larger times beyond 10 days. An average reduction of 1 day is far more important for short resolution times, eg, from 6 to 5 days, than the same reduction from a much larger resolution time, say from 30 to 29 days. The values of  $u(Y)$  show why any conventional 2-sample test based on comparing the means of  $Y$  is inappropriate.

**TABLE 1** Elicited utilities for intraoperative air leak (IAL) resolution times

Days to resolve IAL	0	1-5	6-10	11-15	16-20	21-25	26-30	31-35	>35
Utility	100	50	10	6	5	4	3	2	0

Aside from these issues, the sample sizes of 476 or 280 required by conventional  $t$ -tests to compare means are far too large for running a single institution a trial in a practicable time frame.

Given the general BNP model setup, the inferential targets used by the comparative test are the mean utilities, denoted by  $U_1$  and  $U_0$ . These are computed with respect to  $G_1$  and  $G_0$  based on the observed data. Below, we describe a group sequential test that concludes Progel is superior to standard care if  $Pr(U_1 > U_0 + 18 \mid \text{data})$  is sufficiently large. This says that, given the observed data, it is likely that Progel provides a mean advantage over standard care of at least 18 points on the utility scale. The targeted improvement 18 for the mean utility was elicited from the principal investigator by showing him several different desirable distributions  $G_1$  for Progel and their corresponding mean utilities.

During the trial, this posterior probability is compared to test cutoffs after cohorts of 16, 32, and a maximum of 48 patients, with the test cutoffs derived to control the overall false positive probability at .05. In the protocol design, a threshold of 90% was used for stopping for efficacy and 5% for stopping for futility.

Extensive simulations, given in tables 2 and 3 of Xu et al,<sup>[11]</sup> showed that this design has high power to detect true mean utility differences under a wide range of possible cases where Progel provides an actual benefit over the control. Since the advantage of the design over a design based on a 2-sample  $t$ -test may be ascribed to both the BNP model and the use of  $u$  ( $Y$ ) in place of  $Y$ , a design using  $u(Y)$  but based on a zero-enriched Weibull model also was studied. Simulations for comparing the designs based on the BNP model and the zero-enriched Weibull model, which were presented in table 4 of Xu, et al,<sup>[11]</sup> showed that the BNP model-based design compares quite favorably and has much higher power and smaller type 1 error. In 5 simulated scenarios, the difference in probabilities of correct decision for the BNP design and conventional designs ranged from .09 to .46. This large advantage is due to the fact that the BNP model can accommodate, for example, increased probabilities of short resolution time. The expected utility criterion is very sensitive to such changes, since early days are given much larger utility weights than later days.

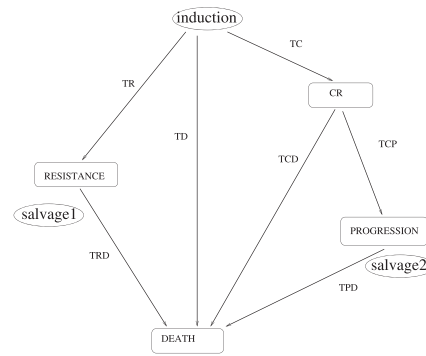
### 3 | ADJUSTING FOR LACK OF RANDOMIZATION

Estey et al<sup>[13]</sup> reported results of a 4-arm randomized clinical trial designed to compare the effects of remission induction chemo combinations on 210 patients with acute myelogenous leukemia (AML) or myelodysplastic syndrome (MDS). Patients with AML or MDS have blast cell percentages far above the normal range of 3% to 5%, and low platelet and white cell counts. Thus, the first goal of induction chemo for AML/MDS is to achieve a complete remission (CR), which means that these three types of blood cells have returned to normal functional ranges. Complete remission is defined as the patient having less than 5% blast cells, a platelet count of at least  $10^5/\text{mm}$ , and a white blood cell count of at least  $10^3/\text{mm}$ . Complete remission is a necessary but not sufficient condition for long term survival, however, since patients who achieve CR may relapse or die.

Denoting FAI = fludarabine + ara-C + idarubicin, GCSF = granulocyte colony stimulating factor, ATRA = all-trans retinoic acid, and HDAC = high-dose ara-C, the 4 chemos were FAI, FAI + ATRA, FAI + GCSF, and FAI + ATRA+ GCSF. The goal was to assess the effects on patient outcome of adding ATRA, G-CSF, or both to the baseline chemo FAI. On the basis of conventional Cox model, logistic regression, and logrank test analyses, Estey et al<sup>[13]</sup> concluded that G-CSF had a beneficial effect on the probability of CR, but that neither G-CSF nor ATRA had any significant effects on either overall survival (OS) or event-free survival among patients who achieved CR.

To account for additional effects of salvage treatments on OS time, Wahed and Thall<sup>[14]</sup> reanalyzed this trial's data. Salvage chemo was given at 2 different possible stages, if the patient's disease was resistant to induction (salvage 1) or at disease progression following CR achieved by induction (salvage 2). The possible combinations of treatment sequences and patient outcomes are illustrated in Figure 2. The numbers of patients with outcomes of each type following induction, and following salvage therapy, are given in Table 2.

A dynamic treatment regime (DTR) in the trial consisted of the triple (frontline, salvage 1, and salvage 2). Categorizing salvage by whether or not it contained HDAC, there were a total of 16 possible DTRs. While all patients received 1 of the 4 induction chemo combinations, depending on their induction outcome (CR, disease resistant to induction, and died during induction),



**FIGURE 2** Flow diagram showing the possible combinations of treatment (ellipses, with lower case tags) and patient outcomes (rectangles, with upper case tags) in the trial comparing chemotherapies for acute myelogenous leukemia. CR, complete remission; OS, overall survival

$$OS = \begin{cases} T^D, & \text{if death during induction} \\ T^R + T^{RD}, & \text{if death after salvage for resistant disease} \\ T^C + T^{CP} + T^{PD}, & \text{if death after salvage for progression after CR} \\ T^C + T^{CD}, & \text{if death in CR} \end{cases}$$

**TABLE 2** Numbers of patients with outcomes of each type following induction therapy and following salvage therapy

Induction	Death	Resistant	CR	
All patients	69	39	102	
FAI	17	17	20	
FAI + ATRA	15	13	26	
FAI + GCSF	20	4	28	
FAI + GCSF + ATRA	17	5	28	
Salvage	Death after resistance	Death in CR	Progression after CR	Death after progression
All patients	37	9	93	83
HDAC	25	-	53	-
Non HDAC	12	-	40	-

Abbreviations: ATRA, all-trans retinoic acid; CR, complete remission; FAI, fludarabine + ara-C + idarubicin; GCSF, granulocyte colony stimulating factor; HDAC, high-dose ara-C.

each patient could receive either salvage 1 or salvage 2 but not both. The flow diagram in Figure 2 shows how each patient's OS time was 1 of 4 possible sums: (1) time to death during induction chemo, (2) time to achieve CR + time to death in CR, (3) time to CR + time to progression + subsequent survival time, or (4) time to resistance + subsequent survival time.

When accounting for the DTRs, a key problem is that patients were not randomized between different types of salvage chemos, which may introduce selection bias into the estimates. Wahed and Thall<sup>[14]</sup> dealt with this by using inverse probability of treatment weighting (IPTW<sup>[15] [16] [17]</sup>). Wahed and Thall<sup>[14]</sup> focused on estimating mean OS for each DTR, using both IPTW- and likelihood-based methods. Their results, summarized in table 6 of Wahed and Thall,<sup>[14]</sup> showed that the two methods gave numerically very different estimates, but they gave the same conclusions regarding the comparative effects of the 16 DTRs. They concluded that mean OS time was smallest for the 4 strategies with FAI alone as frontline regardless of salvages therapies, with the 2 exceptions that (FAI + GCSF + ATRA, HDAC, HDAC) was slightly inferior to (FAI, Other, HDAC) and (FAI, Other, Other). They also concluded that FAI + ATRA was the best remission induction therapy, the type of salvage therapy was irrelevant for patients with disease resistant to FAI + ATRA, and HDAC was a superior salvage therapy for patients who achieved CR with FAI + ATRA but later relapsed.

Xu et al<sup>[18]</sup> reanalyzed this dataset, with the aim to obtain improved estimates of mean OS time by exploiting the flexibility of BNP models. Inference using the more general augmented inverse probability of treatment weighting (AIPTW<sup>[19]</sup>

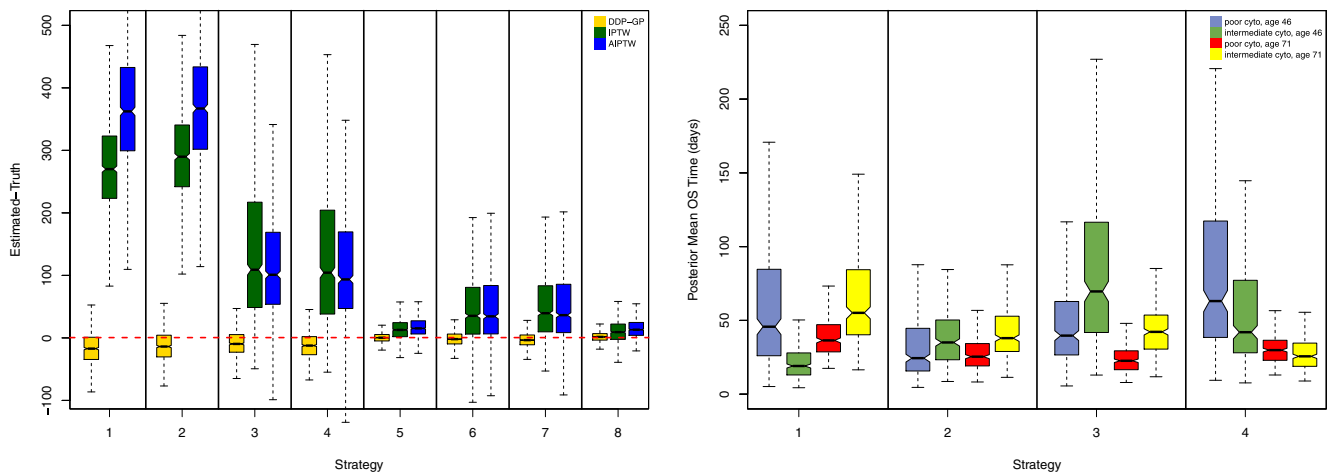


<sup>[20]</sup> <sup>[21]</sup>) hinges on the correctness of the assumed model or, in the case of AIPTW, the correctness of at least one of the assumed models for either the propensity score (treatment assignment probability) or observed outcome. Reliable inference is not possible when neither model is known. Bayesian nonparametric inference can fill this gap, because BNP models always provide a good fit to the data, due to the fact that the prior model has full support. Xu et al<sup>[18]</sup> applied a DDP with a Gaussian process prior (DDP-GP model), to account for possible patient covariate effects on treatment selection at any stage of the DTR.

While we briefly summarize the BNP model, these technical details are not needed in the following discussions and can safely be skipped, unless a reader is interested in the specific model. The diagram in Figure 2 includes 7 possible transition times, including, for example,  $T^c$  for the transition from the initial state to CR. We constructed a BNP survival regression model by modeling each possible transition time based on the patient's current history, including baseline covariates and preceding transition times. Let  $T^k$  denote the  $k$ th transition time in the flow chart of Figure 2, for  $k = 1, \dots, 7$ . The BNP model assumes that each  $\log(T^k)$  follows a DDP mixture (similar to the model in §2), with the mean of each normal distribution in the mixture following a Gaussian process (GP) prior that is a function of the patient's history up to the  $k$ th transition, the DDP-GP model noted earlier.

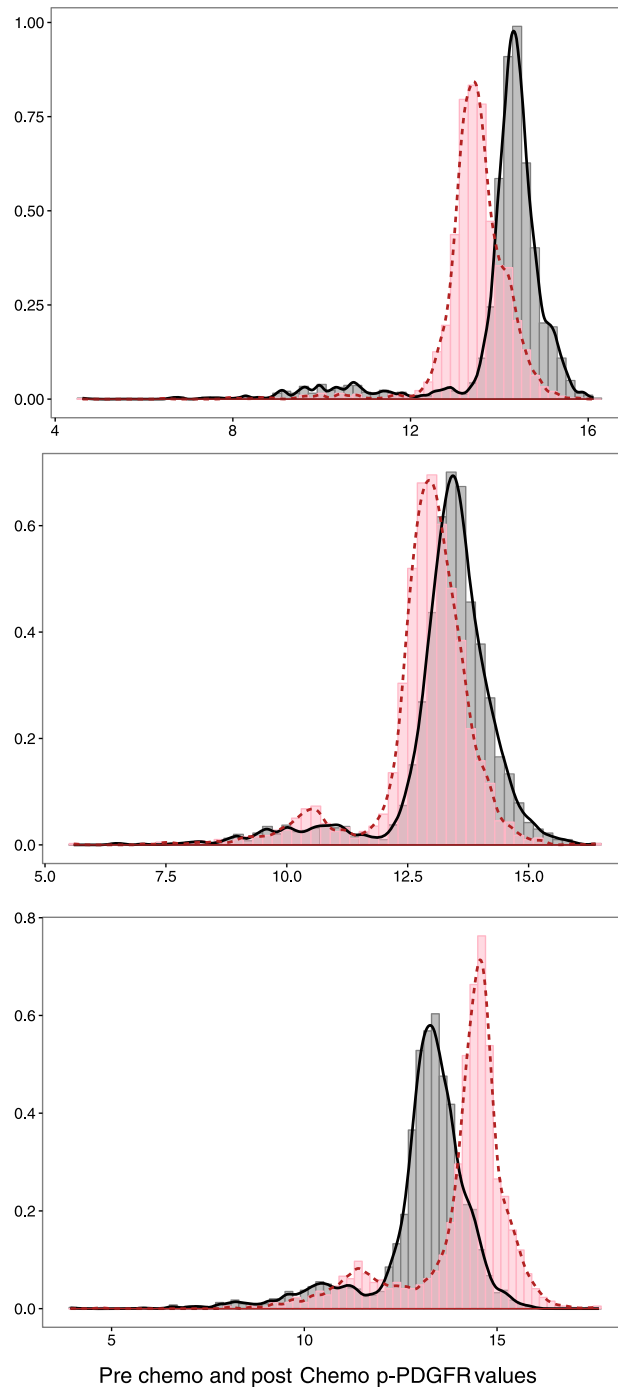
The overall BNP model is constructed from 7 such DDP-GP models, 1 for each possible transition time. The relevant summary is OS time, which is determined by sums of transition probabilities along 4 possible paths through the diagram. A patient's OS time (summarized in Figure 2) could be (1) the time to death during induction; (2) the time to achieve CR plus the time to death in CR; (3) the time to resistance plus the subsequent time to death following salvage; or (4) the sum of the times to CR, progression after CR, and death after progression. The BNP model flexibly accounts for these 4 possibilities, covariate effects, and the relevant frontline and salvage treatment effects on each transition time. See Xu et al<sup>[18]</sup> for technical details.

To validate the DDP-GP model-based approach, Xu et al<sup>[18]</sup> performed an extensive simulation study in which treatments were chosen based on the values of patient covariates, as a physician would do in practice. Thus, selection bias was built into the simulation, and its magnitude was known. Figure 3 (left) shows the results for 8 two-stage regimes, based on the simulated data, comparing the BNP method, IPTW, and AIPTW. In Figure 3 (left), the vertical axis gives the difference between each estimated posterior mean OS time and the “true” mean used in the simulations, so values closer to 0 are more desirable. Each case was simulated 1000 times. In each notched box-whisker plot, the box gives the interquartile range (IQR) from the 25th percentile (Q1) to 75th percentile (Q3), the mid-line is the median, the top whisker limit is  $Q3 + 1.5 \text{ IQR}$ , and the bottom whisker limit is  $Q1 - 1.5 \text{ IQR}$ . The plots show that, across 8 different strategies, the BNP method (yellow boxes) gives estimates that are both more accurate and more reliable than those provided by the 2 conventional methods (green and blue boxes) for bias correction. This may be attributed to the fact that the complex simulated survival time distributions were fit very accurately by the BNP model.



**FIGURE 3** Left side, box plots of estimated mean overall survival (OS) for the Bayesian nonparametric (BNP), inverse probability of treatment weighting (IPTW), and augmented inverse probability of treatment weighting (AIPTW) methods, based on simulated data for 8 strategies with treatments chosen based on patient covariates. Right side, box plots of estimated covariate-specific mean OS, based on the fitted BNP model for the acute myelogenous leukemia data, for 4 strategies and 4 combinations of age and cytogenetics. DDP-GP, dependent Dirichlet process with a Gaussian process

Figure 3 (right) presents similar box plots, based on the data reported by Estey et al,<sup>[13]</sup> for the regimes indexed by 1 = (FAI, HDAC, HDAC); 2 = (FAI + ATRA, HDAC, Other); 3 = (FAI + GCSF, HDAC, HDAC); and 4 = (FAI + ATRA + GCSF, Other, Other). Covariate-specific posterior mean OS estimates obtained by the BNP-model-based method are given for 4 combinations of patient age and cytogenetic abnormality. While the posterior estimates have substantial variability, they suggest that, for younger patients, regime 4 is best for poor cytogenetics and regime 3 is best for intermediate cytogenetics. They also suggest that, for older patients, regime 1 is best for either cytogenetic subgroup. While, unfortunately, no regime provides a substantive improvement in OS, this illustrates how BNP model-based methods can be used to optimize personalized treatments or multistage regimes.



**FIGURE 4** Histograms of the phosphorylated platelet-derived growth factor (p-PDGFR) values  $C_{post}$  (pink) and  $C_{pre}$  (gray) for 3 patients, along with the corresponding Bayesian nonparametric density estimates of the distributions

## 4 | ESTIMATING TARGETED AGENT EFFECTS ON SURVIVAL

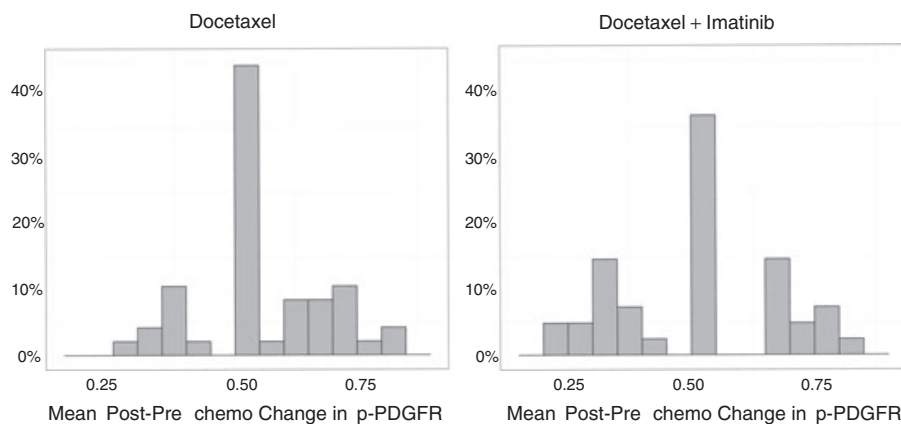
Mathew et al<sup>[22]</sup> reported results of a randomized trial of Docetaxel + Imatinib (D + I) versus Docetaxel (D) in men with advanced prostate cancer. The trial's primary goal was to assess the effect on PFS time of adding Imatinib to Docetaxel. It was hypothesized that Imatinib would reduce the concentration of phosphorylated platelet-derived growth factor (p-PDGFR) in the blood and that this in turn would inhibit tumor angiogenesis and reduce the incidence of bone metastases and thus improve PFS. A total of 88 patients were enrolled in the study (47 in the D arm and 41 in the D + I arm). For each patient, p-PDGFR was measured for each of a large number of cells in each of 2 blood samples, one taken pre-chemo and the other post-chemo. A key parameter was the effect of the within-patient change,  $C = C_{\text{post}} - C_{\text{pre}} = [\text{post-chemo p-PDGFR concentration}] - [\text{pre-chemo p-PDGFR concentration}]$ , on PFS. The motivating ideas were that, on average, smaller (or ideally, negative) values of  $C$ , corresponding to a drop in p-PDGFR after chemo, would be associated with longer PFS and that D + I would yield smaller values of  $C$  compared to D alone. Assessing this required accurate estimates of the distributions of  $C_{\text{post}}$  and  $C_{\text{pre}}$  for each patient.

It might seem that simply computing the means of  $C_{\text{post}}$  and  $C_{\text{pre}}$  would be sufficient. However, histograms of these variables based on the within-patient blood cell samples show that the distributions have multiple modes in some patients. Figure 4 gives histograms of  $C_{\text{post}}$  (pink) and  $C_{\text{pre}}$  (gray) cell measurements for 3 patients. The bimodality implies that reducing each sample to its mean misrepresents the  $C_{\text{pre}}$  and  $C_{\text{post}}$  values.

To assess association of change from  $C_{\text{pre}}$  to  $C_{\text{post}}$  with treatment and PFS, a useful parameter is  $\Delta = Pr(C_{\text{pre}} < C_{\text{post}})$ , the probability that the patient's p-PDGFR went up after chemo. A value  $\Delta > .50$  corresponds to p-PDGFR going up,  $\Delta < .50$  to p-PDGFR going down, and  $\Delta = .50$  to no change in p-PDGFR. Mathew et al<sup>[22]</sup> estimated  $\Delta$  very reliably for each patient from the large numbers of cells in each blood sample. A conventional parametric setup may consider a Cox model for PFS, but this was shown to provide a very poor fit to this data set. Alternative parametric time-to-event regression models still rely on a single point estimate of  $\Delta$  for use as a covariate, which does not take into account the complexity and multimodality of the p-PDGFR samples (see Mathew et al<sup>[22]</sup>).

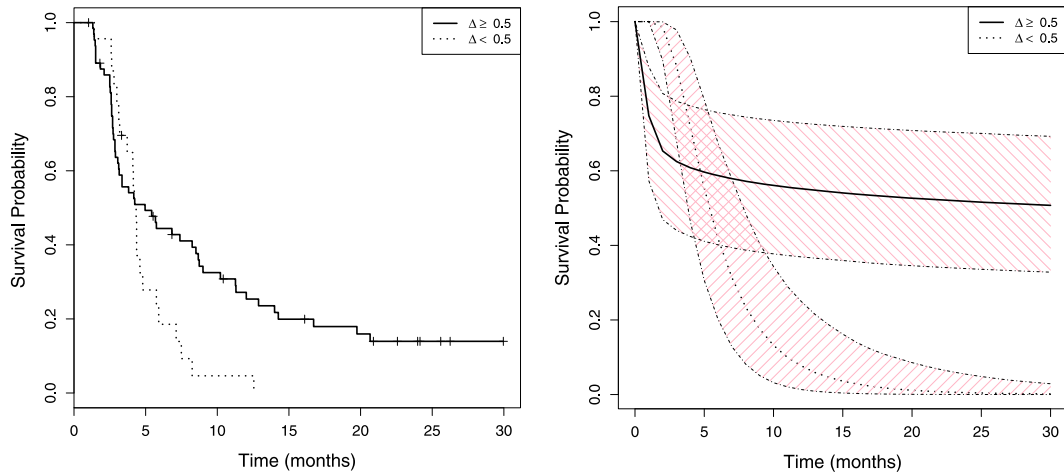
Motivated by this problem, Graziani et al<sup>[23]</sup> reported a BNP analysis of the trial data. Figure 4 gives BNP density estimates, plotted as dashed lines for  $C_{\text{post}}$  and solid lines for  $C_{\text{pre}}$ , corresponding to each histogram. These density estimates were obtained by assuming BNP models (Dirichlet process mixtures of normals) for the distributions of  $C_{\text{post}}$  and  $C_{\text{pre}}$ . As shown by Figure 4, the BNP density estimates identified the multiple modes in the distributions of  $C_{\text{post}}$  and  $C_{\text{pre}}$  for each patient, allowing the numbers and locations of the modes to vary from patient to patient.

The BNP formulation used a survival time regression model to quantify the effects of treatment on the levels of the p-PDGFR values in terms of  $\Delta$ . The model also provided a framework for estimating how interactions between  $\Delta$  and treatment may affect PFS, formalized in equation 7 of Graziani et al.<sup>[23]</sup> The BNP analysis identified clusters of  $\Delta$  values, shown in Figure 5. This figure illustrates that, contrary to the hypothesis of Mathew et al,<sup>[22]</sup> p-PDGFR went down more often for patients in the D + I arm. Also, contrary to the hypothesis, the BNP analysis of the joint effects of treatment,  $\Delta$ , and prognostic covariates on PFS showed that an increase in p-PDGFR was associated with longer rather than shorter PFS. This was the case both



**FIGURE 5** Clusters of posterior means of  $\Delta = Pr(C_{\text{pre}} < C_{\text{post}})$ , the probability of an increase in phosphorylated platelet-derived growth factor (p-PDGFR) values from pre-chemotherapy to post-chemotherapy





**FIGURE 6** Left side, Kaplan Meier estimates of progression-free survival (PFS) for patients experiencing a drop (dashed line) or increase (solid line) in phosphorylated platelet-derived growth factor (p-PDFGR). Right side, median of the distribution of the posterior predictive survival curves for PFS, with corresponding 95% credible intervals, for 2 new representative patients characterized by similar covariate values and PDGFR samples as in Figure 4, top ( $\Delta \geq 0$ ) and bottom ( $\Delta < 0$ ), respectively

overall (Figure 6, left side) and predictively for individual patients (Figure 6, right side). After accounting for these effects, there was little difference in PFS between the D + I and D arms.

## 5 | CONCLUSIONS

The illustrations show how BNP models and methods may be applied in settings where conventional statistical methods are inadequate or inappropriate. For the intraoperative air-leak clinical trial, BNP density estimation accounted for the complex multimodal structure of the historical resolution time distribution. This provided a basis for stating hypotheses in terms of entire distributions, rather than in terms of means based on unimodal distributions, which clearly were inappropriate. By exploiting elicited utilities and this density estimate, the BNP approach accounted for the fact that a given reduction in resolution time had different desirabilities for large versus small resolution times. This gave a design with a practical sample size, while conventional approaches did not.

For analysis of DTRs in chemotherapy of acute leukemia, the BNP model provided superior estimation of overall mean survival time and covariate-specific survival estimates to identify optimal individualized multistage treatment regimes. For the prostate cancer trial data, the BNP model facilitated cluster identification and inferences about key relationships between targeted treatment, biomarker change, and PFS time. Methodological and technical details and online tools for implementation of the 3 examples that we have reviewed are given in the original publications.

While BNP models often are mathematically complex, their properties are intuitively clear in the sense that distributions are represented as mixtures. Consequently, BNP inference is usually more robust against moderate changes in the data and outliers, because the mixture structure of typical BNP models explicitly account for heterogeneity and irregular patterns in the distribution. Quantifying such uncertainties is important even with small or moderately large datasets, as illustrated by the clinical trial design example.

Finally, some words of caution are needed. Bayesian nonparametric inference is no panacea. Implementation requires special purpose computer software. Such software is freely available (see, for example, Jara<sup>[24]</sup>), but it requires statistical expertise for proper model construction. Bayesian nonparametric methods are most advisable for settings where the scientific inference problem calls for the level of generality and flexibility that they provide.

## ACKNOWLEDGMENTS

This work was supported by National Cancer Institute (NCI) Cancer Center Core grant CA016672. Peter Thall's research was supported by NCI grant RO1 CA 83932. Peter Mueller and Yanxun Xu's research was supported by NCI grant R01 CA157458-01A1. The authors thank 2 reviewers for their constructive comments.

## REFERENCES

- [1] D. R. Cox, *J. R. Stat. Soc. B. Methodol.* **1972**, *34*(2), 187.
- [2] E. L. Kaplan, P. Meier, *J. Am. Stat. Assoc.* **1958**, *53*(282), 457.
- [3] N. L. Hjort, C. Holmes, P. Mueller, S. G. Walker (Eds), *Bayesian Nonparametrics*, New York: Cambridge University Press **2010**.
- [4] S. Walker, Bayesian nonparametrics, in *Bayesian Theory and Applications*, (Eds: P. Damien, P. Dellaportas, N. G. Polson, D. A. Stephens), Oxford, United Kingdom: Oxford University Press **2013**, 249.
- [5] P. Mueller, F. Quintana, A. Jara, T. Hanson, *Bayesian Nonparametric Data Analysis*, London: Springer-Verlag **2015**.
- [6] R. Mitra, P. Mueller (Eds), *Nonparametric Bayesian Inference in Biostatistics*, London: Springer-Verlag **2015**.
- [7] A. F. Barrientos, A. Jara, F. A. Quintana, *Bayesian Anal.* **2012**, *7*(2), 277.
- [8] S. MacEachern, *Dependent Nonparametric Processes*, *ASA Proceedings of the Section on Bayesian Statistical Science*, American Statistical Association, Alexandria, VA **1999**.
- [9] M. De Iorio, P. Mueller, G. L. Rosner, S. N. MacEachern, *J. Am. Stat. Assoc.* **2004**, *99*, 205.
- [10] D. Dunson, in *Nonparametric Bayes Applications to Biostatistics*, (Eds: N. L. Hjort et al.), London: Cambridge University Press **2010**, 223.
- [11] Y. Xu, P. F. Thall, P. Mueller, R. J. Mehran A Bayesian nonparametric utility-based design for comparing treatments to resolve air leaks after lung surgery. ArXiv e-prints 1506.07687, **2016a**, 111.
- [12] P. Mueller, R. Mitra, *Bayesian Anal.* **2013**, *8*(2), 269.
- [13] E. H. Estey, P. F. Thall, S. Pierce, J. Cortes, M. Beran, H. Kantarjian, M. J. Keating, M. Andreeff, E. Freireich, *Blood* **1999**, *93*, 2478.
- [14] A. S. Wahed, P. F. Thall, *J. R. Stat. Soc.: Ser. C: Appl. Stat.* **2013**, *62*, 67.
- [15] S. Murphy, M. Van Der Laan, J. Robins, *J. Am. Stat. Assoc.* **2001**, *96*(456), 1410.
- [16] M. J. van der Laan, M. L. Petersen, *Int. J. Biostat.* **2007**, *3*(1).
- [17] J. Robins, L. Orellana, A. Rotnitzky, *Stat. Med.* **2008**, *27*, 4678.
- [18] Y. Xu, P. Mueller, A. Wahed, P. Thall, *J. Am. Stat. Assoc.* **2016b**, 921.
- [19] A. Tsiatis, *Semiparametric Theory and Missing Data*, New York: Springer **2007**.
- [20] E. E. Moodie, T. S. Richardson, D. A. Stephens, *Biometrics* **2007**, *63*(2), 447.
- [21] Y.-Q. Zhao, D. Zeng, E. B. Laber, R. Song, M. Yuan, M. R. Kosorok, *Biometrika* **2015**, *102*, 151.
- [22] P. Mathew, P. F. Thall, D. Jones, C. Perez, C. Bucana, P. Troncoso, S.-J. Kim, R. Millikan, I. J. Fidler, C. Logothetis, *J. Clin. Oncol.* **2004**, *22*, 3323.
- [23] R. Graziani, M. Guindani, P. F. Thall, *Biometrics* **2015**, *71*, 188.
- [24] A. Jara, *Rnews* **2007**, *7*, 17.

**How to cite this article:** Thall PF, Mueller P, Xu Y, Guindani M. Bayesian nonparametric statistics: A new toolkit for discovery in cancer research. *Pharmaceutical Statistics*. 2017. <https://doi.org/10.1002/pst.1819>