BAYESIAN MODELS IN BIOSTATISTICS AND MEDICINE

1.1 Introduction
Biomedical studies provide many outstanding opportunities for Bayesian thinking. The principled and coherent nature of Bayesian approaches often leads to more efficient, more ethical and more intuitive solutions. In many problems the increasingly complex nature of experiments and the ever increasing demands for higher efficiency and ethical standards leads to challenging research questions.

In this chapter we introduce some typical examples. Perhaps the biggest Bayesian success stories in biostatistics are hierarchical models. We will start the review with a discussion of hierarchical models. Arguably the most tightly regulated and well controlled applications of statistical inference in biomedical research is the design and analysis of clinical trials, that is, experiments with human subjects. While far from being an accepted standard, Bayesian methods can contribute significantly to improving trial designs and to constructing designs for complex experimental layouts. We will discuss some areas of related current developments. Another good example of how the Bayesian paradigm can provide coherent and principled answers to complex inference problems are problems related to the control of multiplicities and massive multiple comparisons. We will conclude this overview with a brief review of related research.

1.2 Hierarchical Models
1.2.1 Borrowing Strength in Hierarchical Models
A recurring theme in biomedical inference is the need to borrow strength across related subpopulations. Typical examples are inference in related clinical trials, data from high throughput genomic experiments using multiple platforms to measure the same underlying biologic signal, inference on dose-concentration curves for multiple patients etc. The generic hierarchical model includes multiple levels of experimental units. Say

\[ y_{ki} \mid \theta_k, \phi \sim p(y_{ki} \mid w_{ki}, \theta_k, \phi) \]
\[ \theta_k \mid \phi \sim p(\theta_k \mid x_k, \phi), \]
\[ \phi \sim p(\phi) \]

(1.1)

Without loss of generality we will refer to experimental units \( k \) as “studies” and to experimental units \( i \) as “patients”, keeping in mind an application where \( y_k = (y_{ki}, i = 1, \ldots, n_k) \) are the responses recorded on \( n_k \) patients in the \( k \)-th study of a set of related biomedical studies. In that case \( w_k = (w_{ki}, i = 1, \ldots, n_k) \)
Fig. 1.1. Posterior estimates $E(\theta_k \mid \text{data})$ for each US state in a hierarchical model for mammography usage. The barplot shows the posterior means. States are ordered by posterior means. The barplot below the x-axis shows the posterior standard deviations $\text{SD}(\theta_k \mid \text{data})$. The solid line shows the data (empirical frequency of mammography usage). The dotted line shows the estimate $\hat{\theta}_k$ in a regression on state-specific demographic summaries.

$1, \ldots, n_k$ could be patient-specific covariates, $\theta_k$ are study specific parameters, typically including a study-specific treatment effect, and $x_k$ might be study-specific covariates. We will use these terms to refer to elements of the hierarchical model, simply for the sake of easier presentation, but keeping in mind that the model structure is perfectly general.

In a pharmacokinetic (PK) study $k$ could index patients and $y_k = (y_{ki}, i = 1, \ldots, n_k)$ could be drug concentrations for patient $k$ observed at $n_k$ time points, $w_{ki}, i = 1, \ldots, n_k$. In that case $\theta_k$ are the PK parameters that characterize how patient $k$ metabolizes the drug. In a pharmacodynamic (PD) study $y_{ki}$ could be repeat measurements on blood pressure, blood counts, etc.

In another example, $k$ could index different platforms for high throughput genomic experiments, for example $k = 1$ for RPPA data that records protein activation and $k = 2$ for microarray data that measures gene expression. In that case $i, i = 1, \ldots, n$, could index different genes and proteins and $\theta_k = (\theta_{ki}, i = 1, \ldots, n)$ could code differential gene expression and protein activation.

In [40] we use a hierarchical model for small area estimation. We borrow strength across states to estimate the rate of mammography usage $\theta_k$ in each state in the US, $k = 1, \ldots, K$. The data were collected at a national level, leaving very small or zero sample sizes in some states. By borrowing strength across states we can still report inference for all states. Figure 1.1 shows the posterior means $E(\theta_k \mid \text{data})$. 

Hierarchical Models

In [54] a hierarchical model is used to define a clinical trial design for sarcoma patients. Sarcoma is a very heterogeneous disease. The subpopulations $k = 1, \ldots, K$ index $K = 12$ different sarcoma types, and $i = 1, \ldots, n_k$ indexes enrolled patients who are diagnosed with sarcoma subtype $k$. The hierarchical model borrows strength across sarcoma types. Some sarcoma types are very rare. Even in a large cancer hospital it would be difficult to accrue sufficiently many patients for trial for one subtype alone. Only by borrowing strength across subtypes does it become feasible to investigate such rare subtypes. The other extreme of pooling all patients would be equally inappropriate, as it would ignore the heterogeneity and varying prognosis across different subtypes. The hierarchical model allows for a compromise of borrowing strength at a level between pooling the data and running separate analyses. One limitation, however, remains. The hierarchical model (1.1) assumes that all subtypes are a priori exchangeable. That is not quite appropriate for the sarcoma subtypes. There are likely to be some known differences. [38] develop a variation of hierarchical models that allows for exchangeability of patients across subsets of subpopulations. In the case of the sarcoma study this implies that patients within some sarcoma subtypes are pooled. The selection of these subsets itself is random, with an appropriate prior.

In all five examples the second level of the hierarchical model formalizes the borrowing of strength across the submodels. Most applications include conditional independence at all levels, with $\theta_k$ independent across $k$ conditional on $\phi$ and $y_{ki}$ independent across $i$ conditional on $\theta_k, \phi$. All five examples happen to use hierarchical models with two levels. Extensions to more than two levels are conceptually straightforward.

The power of the Bayesian approach to inference in hierarchical models is the propagation of uncertainties and information across submodels. For example, when $k = 1, \ldots, K$ indexes related clinical trials then inference for the $k$-th trial borrows strength from patients enrolled in the other $K - 1$ trials. Let $y_{-k} = (y_{\ell}, \ell \neq k)$ denote all data excluding the $k$-th study. We can rewrite the implied model for the $k$-th study as

$$p(y_k \mid \theta_k, \phi) \text{ and } p(\theta_k, \phi \mid y_{-k})$$

This highlights the nature of borrowing information across studies. The original prior is replaced by the posterior conditional on data from the other studies. We could describe this aspect of hierarchical models as a principled construction of informative priors based on related studies. The important feature in the process is that this borrowing of strength is carried out in a coherent fashion as dictated by probability calculus, rather than ad-hoc plug-in approaches. The implication is a coherent propagation of uncertainties and information.

Besides the pragmatic aspect of borrowing strength, hierarchical models can also be introduced from first principles. Essentially, if observations within each subpopulation are judged as arising from an infinitely exchangeable sequence of random quantities, and the subpopulations themselves are judged to be ex-
changeable a priori, then model (1.1) is implied (Bernardo and Smith, 1994, chapter 4).

1.2.2 Posterior Computation

One of the early pathbreaking discussions that introduced Bayesian thinking for hierarchical models appears in [39]. The paper appeared long before the routine use of posterior Markov chain Monte Carlo simulation, when computational implementation of Bayesian inference in complex models was still a formidable challenge. One of the important contributions of Lindley and Smith’s paper was to highlight the simple analytic nature of posterior inference when all levels of the hierarchical model are normal linear models.

The restriction to models that allow analytic results severely limited the practical use of Bayesian inference in biomedical applications. This radically changed with the introduction of Markov chain Monte Carlo posterior simulation in [21]. In fact, hierarchical models were one of the illustrative examples in [21] and the companion paper [19] with more illustrative applications.

1.2.3 Related Studies and Multiple Platforms

One of the strengths of the Bayesian approach is the coherent and principled combination of evidence from various sources. A critical need for this feature arises in the combination of evidence across related studies. While many studies are still planned and published as if they were stand-alone experiments, as if they were carried out in total isolation from other related research, this is an increasingly unreasonable simplification of reality.

One simple approach to borrow information across related studies that investigate the same condition is post-processing of results from these studies. This is known as meta-analysis [11, chapter 3]. A typical example appears in [24], who analyze evidence from 8 different trials that all investigated the use of intravenous magnesium sulphate for acute myocardial infarction patients. The discussion in [24] shows how a Bayesian hierarchical model with a suitably sceptic prior could have anticipated the results of a later large scale study that failed to show any benefit of the magnesium treatment.

Multiple related studies need not always refer to clinical trials carried out by different investigators. An increasingly more important case is the use of multiple experimental platforms to measure the same underlying biologic signal. This occurs frequently in high throughput genomic studies. In a recent review paper [29] argue for the need of hierarchical modeling to obtain improved inference by pooling different data sources.

1.2.4 Population Models

In [63] the authors discuss population models as an important special case of hierarchical models and Bayesian inference in biostatistics. Typically each submodel corresponds to one patient, with repeated measurements $y_{ki}$, and a sampling model that is indexed by patient-specific parameters $\theta_k$ and perhaps additional fixed effects $\phi$. The first level prior in the general hierarchical model (1.1)
Bayes in Clinical Trials: Phase I Studies

5

now takes the interpretation of the distribution of patient-specific parameters \( \theta_k \) across the entire patient population.

In population PK/PD models the hierarchical prior \( p(\theta_k \mid x_k, \phi) \) represents the distribution of PK (or PD) parameters across the population. One of the typical characteristics of patient populations is heterogeneity. There is usually no good reason beyond technical convenience to justify standard priors like a multivariate normal. While the population distribution \( p(\theta_k \mid x_k, \phi) \) is usually not of interest in itself, good modeling is important, mainly for prediction and inference for future patients. Let \( i = n + 1 \) index a future patient and let \( Y = (y_1, \ldots, y_n) \) denote the observed data. Inference for a future patient is driven by

\[
p(\theta_{n+1} \mid x_{n+1}, Y) = \int p(\theta_{n+1} \mid x_{n+1}, \phi) \, dp(\phi \mid Y),
\]

assuming that the patient-specific PK parameters \( \theta_k \) are conditionally independent given \( \phi \). The expression for the posterior for \( \theta_{n+1} \) highlights the critical dependence of prediction on the parametric form of \( p(\theta_k \mid x_k, \phi) \). For example, assuming a normal distribution might severely underestimate the probability of patients with unusual PK parameters. Several authors have investigated the use of more general population models in Bayesian population PK/PD models. Let \( \mathcal{N}(x; m, S) \) denote a multivariate normal distribution for the random variable \( x \) with moments \( m \) and \( S \). For example, using a mixture of normals

\[
p(\theta_k \mid \phi) = \sum_{\ell=1}^{L} w_\ell \mathcal{N}(\theta_k; \mu_\ell, \Sigma_\ell) \tag{1.2}
\]

could be used to generalize a normal population model without substantially complicating posterior simulation. Here \( \phi = (L, w_\ell, \mu_\ell, \Sigma_\ell, \ell = 1, \ldots, L) \) indexes the population model. This and related models have been used, for example, in [43, 31] and others. The mixture model needs to be completed with a prior for the parameters \( \phi \). This is easiest done by writing the mixture as \( \int \mathcal{N}(\theta_k; \mu, \Sigma) \, dG(\mu, \Sigma) \) for a discrete probability measure \( G = \sum_\ell w_\ell \delta(\mu_\ell, \Sigma_\ell) \). Here \( \delta_x \) indicates a point mass at \( x \). The prior specification then becomes the problem of constructing a probability model for the random probability measure \( G \). We will discuss such models in more detail below, in §1.7.

1.3 Bayes in Clinical Trials: Phase I Studies

Few other scientific studies are as tightly regulated and controlled as clinical trials. However, most regulatory constraints apply for phase III studies that compare the new experimental therapy against standard of care and clinically relevant standards. For early phase studies the only constraint is that they be carried out in scientifically and ethically responsible ways. This is usually controlled by internal review boards (IRB) that have to approve the study.

Phase I studies aim to establish safety. In oncology a typical problem is to find the maximum tolerable dose (MTD) of a new chemotherapy agent (or
Bayesian Models in Biostatistics and Medicine

combination of agents). Let $z = (z_j, j = 1, \ldots, J)$ denote a grid of available doses. Many designs assume that the formal aim of the study is to find a dose with toxicity closest to a pre-determined maximum tolerable target level $\pi^*$. Typically the outcome is a binary indicator, $y \in \{0, 1\}$ for a dose-limiting toxicity. Let $\pi_j = p(y = 1 \mid z_j)$ denote the probability of toxicity at dose level $j$. One of the still most widely used designs is the so called 3+3 design. It is simply a rule for escalating the dose in subsequent patient cohorts until we observe a certain maximum number of toxicities in a cohort. The design is an example of a rule-based design. In contrast to model-based designs that are based on inference with respect to underlying statistical models, rule-based designs simply follow a reasonable but otherwise ad-hoc algorithm. There is no probabilistic guarantee that the reported MTD is in fact a good approximation of the unknown truth. Rule-based designs are popular due to the ease of implementation, but they are also known to be inefficient (Le Tourneau et al., 2009). Here, efficiency is judged by frequentist summaries under repeated use of a design. Summaries include the average sample size and the average true probability of toxicity at the reported MTD.

[47] proposed one of the first model-based Bayesian designs to address some of the limitations of traditionally used rule-based designs. The method is known as continual reassessment method (CRM). The underlying model is quite straightforward. Let $d_j$, $j = 1, \ldots, J$ denote a grid of available dose levels and recall that $\pi_j$ is the probability of toxicity at dose $j$. For a given skeleton $d = (d_1, \ldots, d_J)$ of doses the model is indexed with only one more parameter as $\pi_j = d_a^j$. Several variations with alternative one-parameter models are in use. The algorithm uses a target dose, say $\pi^* = 0.30$ and proceeds by sequentially updating posterior inference on $a$ and assigning the respective next patient cohort to the dose with $\hat{\pi}_j = d_a^j$ closest to the target toxicity $\pi^*$. Here $\hat{a} = E(a \mid data)$ is the posterior mean conditional on all currently available data and $\hat{\pi}_j$ is a plug in estimate of $\pi_j$. In a minor variation one could imagine to replace $\hat{\pi}_j$ by the posterior mean $\tilde{\pi}_j = E(d_a \mid data)$. The coding of the doses $d_j$ is part of the model, but is fixed up-front. The values $d_j$ are not raw dose values, but chosen to achieve desired prior means $\pi_j = E(d_a^j)$. Here the expectation is with respect to the prior on $a$. The initially proposed CRM gave rise to serious safety concerns. The main issue is that the algorithm could jump to inappropriately high doses, skipping intermediate and yet untried doses.

Several later modifications have improved the original CRM, including the modified CRM of [22] to address the safety concerns, and the TITE-CRM of [9] for time to event outcomes and many more. The TITE-CRM still uses essentially binary outcomes, but allows for weighting to accommodate early responses and enter patients in a staggered fashion. Let $U_i$ denote the event time for patient $i$, for example, time to toxicity. Let $y_i$ denote a binary outcome for patient $i$, defined as $U_i$ beyond a certain horizon $T$, i.e., $y_i = I(U_i > T)$. Let $y_{in}$ denote the toxicity status of patient $i$ just before the $n$-th patient enters the trial. When $U_i$ is already observed then $y_{in} = y_i$. Also when $U_i$ is censored at $T$, i.e., $U_i$
is known to be beyond the horizon $T$, then $y_{in} = y_i = 0$. Only when $U_i$ is censored before $T$, then $y_{in} = 0$, while $y_i$ would still be considered censored. The TITE-CRM replaces the binary response $y_i$ by $y_{in}$ and uses an additional weight $w_i = \min(U_i/T, 1)$ to replace $\pi_j$ in the likelihood by $g_j = \pi_j w_i$. This allows to make use of early responses implied by censored $U_i < T$ and can significantly reduce the trial duration. The approach of [55] goes a step further and uses a parametric model for the time to event endpoint. [3] generalize the TITE-CRM with a probit model for discretized event times to allow for lack of monotonicity. Some authors have suggested alternative model-based approaches. The EWOC (escalation with overdose control) method of [1] uses a cleverly parametrized logistic regression of outcome on dose. A common theme of these model-based methods is the use of very parsimonious models. This is important in the context of the small sample sizes in phase I trials.

Another common feature is the use of a target toxicity level $\pi^\star$. Several recent authors argue that the assumption of a single value $\pi^\star$ is unrealistic, and replace the notion of a target dose by target toxicity intervals. This view is taken, for example, in [30]. Keeping the model ultimately simple they use independent Beta/Binomial models for each dose. Only post-processing with isotonic regression ensures monotonicity. A related approach is proposed in [45] who go a step further and introduce ordinal toxicity intervals, including intervals for underdosing, target toxicity, excessive toxicity, and unacceptable toxicity. The underlying probability model is a logistic regression centered at a reference dose. Sequential posterior updating after every patient-cohort includes updated posterior probabilities for the target toxicity intervals at each dose. The respective next dose is assigned by trading off the posterior probabilities of these intervals.

The designs and models mentioned so far are all exclusively phase I designs. Inference is only concerned with toxicity outcomes, entirely ignoring possible efficacy outcomes. The EffTox model introduced in [56] explicitly considers both, toxicity and efficacy. Thall and Cook develop a design that trades off target levels in both endpoints. The probability model is based on two marginal logistic regressions for a binary toxicity and a binary efficacy outcome, and one additional parameter that induces dependence of the two outcomes. The design includes sequential posterior updating and a desirability function that is used like a utility function to select acceptable doses and eventually an optimal dose.

1.4 Phase II Studies

Phase II studies aim to establish some evidence for a treatment effect, still short of a formal comparison with a control or standard of care in the following final phase III study. Larger sample sizes and more structure, compared with phase I, allow for more impact of model-based Bayesian design. Opportunities for innovative Bayesian designs arise in sequential stopping, adaptive design and the use of problem-specific utility functions.
1.4.1 Sequential stopping

Some sequential stopping designs use posterior predictive probabilities to decide upon continuation versus termination of a trial. Let $PP$ denote the posterior predictive probability of a positive result at the end of the trial. For example, assume that the efficacy outcome is a binary indicator for tumor response, and that a probability of tumor response $\pi > \pi_0$ is considered a clinically meaningful response. Let $y$ denote the currently available data, i.e., outcomes for already enrolled patients, and let $y^o$ denote the still unobserved responses for future patients if the trial were to run until some maximum sample size. Also assume that evidence for efficacy is formalized as the event $\{ p(\pi > \pi_0 \mid data) > 1 - \epsilon \}$. The posterior predictive probability $PP = p( p(\pi > \pi_0 \mid y, y^o) > 1 - \epsilon \mid y)$ is the probability of a successful trial if the study continues until the end. Continuous monitoring of such predictive probabilities facilitates the implementation of flexible stopping rules based on the chance of future success. This is implemented in [28] who define a sequential stopping rule based on the posterior predictive probability of (future) conclusive evidence in favor of one of two competing treatments. Similarly, also [37] argue for the use of posterior predictive probabilities to implement sequential stopping in phase II trials. See also the discussion in [7].

Similar in spirit, many recently proposed clinical trial designs use continuously updated posterior probabilities of clinically meaningful events to define stopping rules. The use of posterior predictive probabilities can be seen as a special case of this general principle, using a particular type of posterior inference related to future posterior probabilities. In general, one could consider any event of interest, usually related to some comparison of success probabilities or other parameters under one versus the other therapy. [11, chapter 6] refers to such approaches as “proper Bayes” design. A class of such designs that use continuously updated posterior probabilities for sequential stopping for futility and efficacy was introduced in [58] and [57]. For example, Let $p_n(\theta_E > \theta_S + \delta \mid data)$ denote the posterior probability that the response rate $\theta_E$ under the experimental therapy is larger than the response rate $\theta_S$ under standard of care by more than $\delta$. A design could stop for futility when $p_n < L$ and stop for efficacy when $p_n > U$, where $U$ and $L$ are fixed thresholds. [58] include a model for multiple outcomes, including indicators for toxicity and tumor response. In that case the posterior probabilities of appropriate combinations of outcomes under competing treatments can be used to define stopping. Figure 1.2 summarizes a possible trial history of a trial using such stopping rules. The two curves $p_n(EFF)$ and $p_n(TOX)$ plot the continuously updated posterior probabilities $p_n(\cdot) = p(\cdot \mid data)$ of a toxicity event (TOX) and an efficacy event (EFF). The two horizontal lines are thresholds. When $p_n(TOX) > U$ the trial stops for toxicity. When $p_n(EFF) < L$ the trial stops for futility, i.e., lack of efficacy. This particular trial had no sequential stopping for efficacy. The figure shows a simulated trial history under a very favorable assumed truth. Neither curve crosses the corresponding stopping boundary.
1.4.2 Adaptive Allocation

Sequential stopping is one (important) example of outcome adaptive designs. Those are clinical trial designs that use interim analysis during the course of the trial to change some design element as a function of already observed outcomes. The implementation and careful planning of such experiments is far more natural and straightforward under a Bayesian approach than under the classical paradigm (Berry, 1987).

Besides sequential stopping another commonly used type of adaptive design is adaptive treatment allocation for multi-arm trials, i.e., trials that recruit patients for more than one therapy. The intent of most adaptive allocation designs is to favor allocation to the better therapy, but keep some minimum level of randomization. Adaptive allocation is thus different from deterministic rules like play-the-winner. The designs are usually rule-based, i.e., there is no notion of optimality. The adaptive allocation is carried out following some reasonable, but ad-hoc rule. The idea goes back to at least [59]. A recent review appears in [53].

Consider a study that allocates patients to two treatments, say $T_1$ and $T_2$, and let $\theta_1$ and $\theta_2$ denote treatment-specific parameters that can be interpreted as efficacy of treatment $T_1$ and $T_2$, respectively. For example, $\theta_j$ might be the probability of tumor response under treatment $T_j$. Let $r_j(y)$ denote the probability of allocating the next patient (or patient cohort) to treatment $T_j$. Importantly, $r_j(\cdot)$ is allowed to depend on the observed outcomes $y$. A popular rule is

$$r_j(y) \propto \left\{ p(\theta_j = \max_i \theta_i \mid y) \right\}^c.$$

The two probabilities, the randomization probability $r_j$ and the posterior probability of $T_j$ being the best treatment, are defined on entirely different spaces, and are entirely unrelated in the probability model. Only the adaptive allocation...
rule links them, following the vague notion that it is preferable to assign more patients to the better treatment. The power \( c \) is a tuning parameter. It is usually chosen as \( c = 1 \) or \( c = 1/2 \). \cite{53} recommend \( c = n/2N \) for current sample size \( n \) and maximum sample size \( N \). The recommendation is based on empirical evidence only.

Finally, for a fair discussion of adaptive trial design we should note that Bayesian adaptive designs are far from routine. In fact, in a recent guidance document the US Food and Drug Administration (FDA) does not encourage the use of Bayesian methods for adaptive designs.

### 1.4.3 Delayed Response

One of the big white elephants in drug development is the high failure rate of phase III confirmatory studies, even after successful phase II trials. In cancer more than 50\% of phase III studies fail (Kola and Landis, 2004). One of the possible causes for this is that phase II studies typically use a binary indicator for tumor response, for example an indicator for tumor shrinkage, as a proxy for survival, which is a typical phase III endpoint. Ideally one would want to see the same endpoint being used in phase II and III. But the delayed nature of a survival response would make phase II trials unacceptably slow. See \cite{26} for a discussion. They use model-based Bayesian inference to develop a phase II design that mitigates the problem. The solution is simple and principled. Let \( S_i \in \{1, 2, 3, 4\} \) denote an indicator for tumor response for the \( i \)-th patient. They code tumor response as a categorical outcome with 4 levels. Let \( T_i \) denote survival time. They use a joint sampling model \( p(S_i, T_i \mid \lambda, p) \), where \( p \) are the probabilities of \( S_i \) under different treatments and \( \lambda \) indexes an exponential regression of \( T_i \) on \( S_i \) and treatment assignment. Under this model it becomes now meaningful to report posterior probabilities for events related to survival. As a welcome side-effect inference about tumor response is improved by borrowing strength from the survival outcome.

### 1.5 Two Bayesian Success Stories

One of the big opportunities and challenges for Bayesian approaches in clinical trial design is the possibility to match treatments with patients in a coherent fashion. Two recent successful studies illustrate these features.

#### 1.5.1 ISPY-2

ISPY-2 (Barker et al, 2009) uses a Bayesian adaptive phase II clinical trial design. The trial considers neoadjuvant treatments for women with locally advanced breast cancer. The study simultaneously considers five different experimental therapies. All five treatments are given in combination with standard chemotherapy, before surgery (neoadjuvant). ISPY-2 defines an adaptive trial design, i.e., design elements are changed in response to the observed data. Adaptation in ISPY-2 includes changing probabilities of assigning patients to the treatment arms and the possibility of dropping arms early for futility or efficacy.
In the latter case the protocol recommends a following small phase III study. The treatment is “graduated.”

In addition to these more standard adaptive design elements, the most innovative and important feature of the trial is explicit consideration of population heterogeneity by defining subpopulations based on biomarkers and a process that allows different treatments to be recommended for each subpopulations. The important detail is that both, the identification of subpopulations and the treatment recommendation happen in the same study. With this feature ISPY-2 might be able to break the so-called biomarker barrier. Many previous studies have attempted to identify subpopulations, but only very few stood the test of time and proved useful in later clinical trials. In ISPY-2, the process of biomarker discovery starts with a list of biomarkers that define up to 256 different subpopulations, although only about 14 remain as practically interesting, due to prevalence and biologic constraints. For each patient we record presence or absence of these biomarkers, including presence of hormone receptors (estrogen and progesterone), human epidermal growth factor receptor 2 (HER2) and MammaPrint risk score. The recorded biomarkers determine the relevant subgroup, which in turn determines the allocation probabilities.

The trial is a collaboration of the US National Cancer Institute, the US Food and Drug Administration, pharmaceutical companies and academic investigators. Please see http://www.ispy2.org for more details.

1.5.2 BATTLE

The BATTLE (Biomarker-integrated approaches of targeted therapy of lung cancer elimination) trial is in many ways similar to ISPY-2. The study is described in [66]. BATTLE is a phase II trial for patients with advanced non-small cell lung cancer (NSCLC). The design considers five subpopulations defined by biomarker profiles, including EGFR mutation/amplification, K-ras and B-raf mutation, VEGF and VEGFR expression and more. The primary outcome is progression free survival beyond 8 weeks, reported as a binary response. We refer to the binary outcome as disease control. Similar to ISPY-2, the design allows to allocate treatments with different probabilities in each subpopulation, and to report subpopulation-specific treatment recommendations upon conclusion of the trial. Treatment allocation is adaptive, with probabilities proportional to the probabilities of disease control. Let \( \gamma_{jk} \) denote the current posterior probability of disease control for a patient in biomarker group \( k \) under treatment \( j \). The next patient in biomarker group \( k \) is assigned to treatment \( j \) with probability proportional to \( \gamma_{jk} \). Posterior probabilities are with respect to a hierarchical probit model. The probit model is written in terms of latent probit scores \( z_{jki} \) for patient \( i \) under treatment \( j \) in biomarker group \( k \). The model assumes a hierarchical normal/normal model for \( z_{jki} \). The model includes mean effects \( \mu_{jk} \) of treatment \( j \) in biomarker group \( k \), and mean effects \( \phi_j \) for treatment \( j \).

The model is also used to define early stopping of a treatment arm \( j \) for the \( k \)-th disease group. An arm is dropped for futility when the posterior predictive
probability of the posterior probability for disease control being beyond $\theta_1$ is less than $\delta_L$. Here the latter posterior probability refers to inference conditional on future data that could be observed if the treatment were not dropped. And $\theta_1$ would naturally be chosen to be the probability of disease control under standard of care. Similarly, treatment $j$ is recommended for biomarker group $k$ if the posterior predictive probability of (future) posterior probability of disease control being greater than $\theta_0$ is greater than $\delta_U$. Here $\theta_0 > \theta_1$ would be some clinically meaningful improvement over $\theta_1$.

### 1.6 Decision Problems

Some biomedical inference problems are best characterized as decision problems. A Bayesian decision problem is described by a probability model $p(\theta, y)$ on parameters $\theta$ and data $y$, a set of possible actions $d \in A$, and a utility function $u(d, \theta, y)$. The probability model is usually factored as a prior $p(\theta)$ and a sampling model $p(y | \theta)$. It is helpful to further partition the data into $(y_o, y)$ for data $y_o$ that is already observed at the time of decision making, and future data $y$. The utility function describes the decision maker’s relative preferences for actions $d$ under assumed values of the parameters $\theta$ and hypothetical data $y$. In this framework the optimal decision is described as

$$d^* = \arg \max_{d \in A} U(d, y^o) \text{ with } U(d, y^o) = \int u(d, \theta, y) \, d p(\theta, y | y^o). \quad (1.3)$$

In words, the optimal decision maximizes the utility, in expectation over all random quantities that are unknown at the time of decision making, and conditional on all known data. The expectation $U(d, \cdot)$ is the expected utility.

In [48] we find an optimal schedule for stem cell collection (apheresis) for high-dose chemo-radiotherapy patients under two alternative treatments, $x \in \{1, 2\}$. We develop a hierarchical model to represent stem cell counts for each patient over time. Stem cell counts are measured by CD34 antigen levels per unit volume. We do not need details of the model for the upcoming discussion. The model includes a sampling model $p(y_{ij} | t_{ij}, \theta_i)$ for the CD34 counts $y_{ij}$ that is recorded for patient $i$ at (known) time $t_{ij}$, a random effects distribution $p(\theta_i | x_i, \phi)$ for patient-specific parameters of a patient under treatment $x_i \in \{1, 2\}$, and a hyperprior $p(\phi)$. The decision is a vector of indicators $d = (d_1, \ldots, d_N)$, with $d_j = 1$ when an apheresis for a future patient is scheduled on day $j$ of an $N$ day period. The action space $A$ is the set of all binary $N$-tuples. The utility function $u(\cdot)$ is a combination of sampling cost for $n_d = \sum_j d_j$ stem cell collections and a reward for collecting a target volume $y^*$ of stem cells. Let $L$ denote the volume collected at each collection and let $y = (y_1, \ldots, y_n, y_{n+1})$ denote the observed CD34 counts, including the future patient $n + 1$. Then

$$u(d, \theta, y, x) = n_d + \lambda I(\sum_j y_{n+1,j} L < y^*),$$

for a future patient $i = n + 1$ assigned to treatment $x_i = x$. The optimal design $d^*_x$ is found by maximizing
Decision Problems

Table 1.1. Optimal (and 2nd and 3rd best) design for two future patients. The first column reports the estimated value $\hat{U}(d^*_x, x)$.

<table>
<thead>
<tr>
<th>$d$</th>
<th>$d_1$</th>
<th>$d_2$</th>
<th>$d_3$</th>
<th>$d_4$</th>
<th>$d_5$</th>
<th>$d_6$</th>
</tr>
</thead>
<tbody>
<tr>
<td>treatment $x = 1$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.03</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3.00</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4.00</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>treatment $x = 2$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5.14</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5.18</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5.47</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

$U(d, x) = \int u(d, \theta, y) p(\theta, y_{n+1} | x_{n+1} = x, y_1, \ldots, y_n).$

Another interesting example of the benefit of decision theoretic thinking occurs in phase II clinical trials with sequential stopping decisions. After each cohort of patients we make a decision $d \in \{0, 1, 2\}$, where $d = 0$ indicates stopping enrollment and recommending against further development (failure), $d = 2$ indicates stopping enrollment recommending for a following phase III confirmatory trial (victory), and $d = 1$ indicates continued enrollment. A useful utility function in this context formalizes the considerations that feature in this decision. Enrolling patients is expensive, say $c$ units per patient. If the following confirmatory trial shows a statistically (frequentist) significant effect then the drug can be marketed and the developer collects a substantial reward $C$. A significant effect in a follow-up trial is usually easy to consider. The beauty of frequentist tests is that they are usually quite simple. Let $z$ denote the (future) responses in the follow-up trial. All that is needed for the utility function is the posterior predictive probability of the (future) test statistic $S(z)$ falling in the rejection region $R$, using a sample size $n_3$ based on the usual desiderata of type-I error and power under a current estimate of the treatment effect.

For the statement of the utility function we do not need any details of the probability model, short of an assurance that there is a joint probability model for data and parameters. Let $y = (y_1, \ldots, y_n)$ denote the currently observed data. Let $n_3(y)$ denote the sample size for a future phase III trial. The sample size can depend on current estimates of the treatment effect and other parameters. Let $z$ denote the (future) data in the phase III study, let $S(z)$ denote the test statistic that will be evaluated upon conclusion of the phase III study, and let $R$ denote the rejection region for that test. Also, without loss of generality we assume that patients are recruited in cohorts of size 1, i.e., we consider stopping after each new patient. Formally, the utility function becomes
Bayesian Models in Biostatistics and Medicine

\[ u(d, \theta, y) = \begin{cases} 
  c + E\{U(d^*, y, y_{n+1}) \mid y\} & \text{if } d = 1 \\
  0 & \text{if } d = 0 \\
  cn_3(y) + C p\{S(z) \in R \mid y\} & \text{if } d = 2 
\end{cases} \]

Here \( U(d^*, \ldots) \) is the expected utility under the optimal action in the next period. This appears in the utility function because of the sequential nature of the decision problem. Discussing strategies for the solution of sequential decision problems is beyond the scope of this discussion. Here we only want to highlight how relatively straightforward it is to incorporate a stylized description of the following phase III study in the utility function. Variations of this utility function are used in [49, 62, 41].

There is a caveat about the use of decision theoretic arguments in biomedical research problems. Many investigators are reluctant to use formal decision theoretic approaches in biomedical problems. The main reason is perhaps the nature of the optimal decision \( d^* \) as an implicitly defined solution. There is no guarantee that the solution \( d^* \) looks intuitively sensible. Technical details of the probability model and the typically stylized utility function determine the solution in a very implicit way, as the solution to the optimization problem (1.3). However, many details are usually chosen for technical convenience rather than based on substantive prior information. Of course, when counter-intuitive solutions are found, one could always go back and include additional constraints in \( \mathcal{A} \). But this is a post-hoc fix, and not all counter-intuitive features are readily spotted.

1.7 Nonparametric Bayes

1.7.1 BNP Priors

Nonparametric Bayesian (BNP) models are priors for random probability models or functions. A common technical definition of BNP models are probability models for infinite dimensional random quantities. However, traditionally most applications of BNP involve random distributions only. So for this discussion we will restrict attention to BNP priors for random probability measures. Let \( G \) denote an (unknown) probability model of interest. For example, \( G \) could be the distribution of event times in a survival analysis, or the random effects distribution in a mixed effects model, or the probability model for residuals in a regression problem. Generically, let \( p(y \mid G, \eta) \) denote the sampling model for the observable data given \( G \) and possibly other parameters \( \eta \). To be specific we assume a density estimation problem for a random sample \( y = (y_1, \ldots, y_n) \) with

\[ p(y \mid G) = \prod_{i=1}^{n} G(y_i). \]  

Bayesian inference requires that the model be completed with a prior for the unknown quantities \( G \) and \( \eta \). When the prior \( p(G) \) restricts \( G \) to a family of
Nonparametric Bayes

probability models \{G_{\theta}, \theta \in \mathbb{R}^p\} that is indexed by a finite dimensional parameter vector \(\theta\), then we are back to standard parametric inference. For example, assuming a normal sampling model \(G(\cdot) = N(\cdot; \mu, \sigma)\) reduces the problem to inference for the unknown normal moments \(\hat{\theta} = (\mu, \sigma)\). If, however, the investigator is not willing or not able to restrict attention to a parametric family then we require a prior for the infinite dimensional quantity \(G\),

\[ G \sim p(G). \]

Good recent reviews of BNP appear in [65] and [25]. Figure 1.3 illustrates the flexibility of BNP in a simple density estimation problem. For a good recent discussion of nonparametric Bayes in biostatistics and bioinformatics we refer to [13].

1.7.2 Survival Analysis

One of the most common applications of BNP in biostatistics is to survival analysis, i.e., density estimation (1.4) when the data are event times, and typically include extensive censoring. [52] and [18] developed BNP estimates of survival functions in the presence of censoring, still relying on analytic results. They used the Dirichlet process (DP) prior (Ferguson 1973, 1974). The DP prior and variations is to date by far the most commonly used BNP model. With the introduction of Gibbs sampling the modeling options greatly improved. [34] discuss Gibbs sampling for right, left and interval censored data. They assume that data is observed at discrete times only, reducing the DP prior to a Dirichlet distribution of the quantiles over finitely many time intervals. The mentioned approaches all use variations of the basic model

\[ T_i \sim F \text{ and } F \sim \text{DP}(F^*, \alpha), \]
Bayesian Models in Biostatistics and Medicine

i.e., a DP prior on the unknown event time distribution. The DP requires two parameters. The base measure $F^*$ and the total mass parameter $\alpha$. The base measure $F^*$ fixes the prior mean, $E(F) = F^*$, Among other implications, $\alpha$ determines the variation, $F(A) \sim \text{Be}[\alpha F^*(A), \alpha(1 - F^*(A))]$. An important property of the DP prior is that it generates a.s. discrete random probability measures. This is awkward and often avoided by using an additional convolution with a continuous kernel $k(x; \theta)$, for example, a normal kernel $N(x; \theta, \sigma)$ with known scale.

$$F(T_i) = \int k(T_i; \theta) dG(\theta) \text{ and } G \sim \text{DP}(G^*, \alpha). \quad (1.5)$$

[10] propose an accelerated failure time model using a Dirichlet process prior. Similarly, [33] and [20] develop accelerated failure time models based on DP mixtures. These models are semiparametric. The model component that implements the regression on baseline covariates remains parametric. A fully nonparametric survival regression based on DP priors is proposed in [12]. In principle any nonparametric density regression model, i.e., a nonparametric prior for families of random distributions $\{G_x\}$ indexed by covariates $x$, could be used. Many such models have been proposed in the recent literature, including, for example, [14] and [15].

Besides the DP model and variations, many other BNP models have been proposed for survival data. Many approaches are based on the Polya tree (PT) prior. See [35] for a discussion of these priors and references. A fully Bayesian semiparametric survival model using the PT was proposed by [64]. More recently, [23, 60] proposed mixtures of Polya Tree priors for use in semiparametric proportional hazards, accelerated failure time and proportional odds models. They assume mixture of PT priors for a baseline survival distributions. A fully nonparametric extension of the PT model for survival data to include a nonparametric regression on covariates, is proposed in [61]. We refer to [27] and the following chapter in this volume for a thorough review of nonparametric Bayesian methods in survival analysis.

1.8 Multiplicities and Error Control

1.8.1 Posterior inference accounts for multiplicities

Many important scientific problems naturally lead to massive multiple comparisons. Typical examples are experiments that record gene expression under different biologic conditions, simultaneously for massively many genes. The problem is to compare for each gene relative gene expression across biologic conditions and report those genes that show significant differences across conditions. The number of comparisons can be thousands or tens of thousands.

Formally, let $\delta_i \in \{0, 1\}$, $i = 1, \ldots, n$ denote the unknown truth for $n$ comparisons. A stylized example of differential gene expression experiments could involve a sampling model $p(y_i \mid \theta_i) = N(\theta_i, \sigma)$, $i = 1, \ldots, n$, for a gene-specific difference score $y_i$. The sampling model is indexed with a gene-specific parameter
that can be interpreted as the level of differential expression. The variance \( \sigma \) is assumed to be common across all genes. The \( i \)-th comparison is \( \theta_i = 0 \) versus \( \theta_i \neq 0 \), i.e., no differential expression versus non-zero differential expression. In this context \( \delta_i = I(\theta_i \neq 0) \) is an indicator of (true) differential expression. The model is completed with a hierarchical prior on \( \theta_i \). For the moment, we need not worry with the details of this prior model. For a meaningful discussion we only need to assume that the prior includes a positive prior probability \( p(\delta_i = 0) > 0 \) of non-differential expression. We focus on inference of \( \delta_i \). Under a Bayesian approach one might report, for example, \( p_i = p(\delta_i = 1 \mid y) \), the posterior probability of differential expression for the \( i \)-th gene.

A popular folk theorem asserts that Bayesians need not worry about multiplicities. Posterior probabilities are already (automatically) adjusted for multiplicities. This is illustrated, for example in [50]. Under some assumptions on the hierarchical prior for \( \delta_i \), the statement is correct. The posterior probabilities \( p_i \) adjust for multiplicities, in the following sense. Focus on one particular comparison, say the \( i \)-th comparison with a particular difference score \( y_i \) and consider two scenarios with the same value \( y_i \), but different values for other \( y_j \), \( j \neq i \). When all other observed difference scores \( y_j \), \( j \neq i \), are closer to zero, then \( p_i \) will be shrunk more towards zero than what would be the case for larger \( y_j \). In other words, posterior probabilities adjust for multiplicities by increased shrinkage in \( p_i \) when the data suggest high level of noise in the data. Of course, this statement depends on details of the model, but it holds for any reasonable hierarchical model. See [50] for details.

However, reporting posterior probabilities \( p_i \) “is only half the solution” (Berry and Berry, 2004). The remaining part of the solution is the decision to select which genes should be reported as differentially expressed. The magic automatic adjustment for multiplicities only happens for the probabilities, not for the decision. Let \( d_i(y) \in \{0, 1\}, i = 1, \ldots, n \), denote this inference summary. In a classical framework, this could simply be a test of \( H_{0i} : \theta_i = 0 \). Under a Bayesian perspective one could consider, for example, \( d_i = I(p_i > c) \). This seems an innocent and reasonable decision rule. It will turn out to actually be the Bayes rule under certain assumptions. In any case, it is a plausible rule to consider. In both cases, frequentist and Bayesian, we are left with the decision of where to draw the line, i.e., how to choose the threshold \( c \)? Clearly, traditional type I error control for each comparison is meaningless. In most applications one would end up reporting way too many false positives. The other extreme of controlling experiment-wide error rates, on the other hand, is way too conservative in most applications.

### 1.8.2 False discovery rate (FDR)

A commonly used compromise between the two extremes of comparison-wise and experiment wide error control is the control of false discovery rate, the relative fraction of false positives, relative to the number of positives. Let \( D = \sum d_i \) denote the number of reported genes. The false discovery rate is defined as
\[
\text{FDR} = \frac{1}{D} \sum_{i=1}^{n} (1 - \delta_i) d_i.
\]

Often the denominator is replaced by \((D + \epsilon)\) to avoid zero division. At this moment the FDR is neither frequentist nor Bayesian. It is a function of both, the data, indirectly through \(d_i(y)\), and the parameters \(\delta_i\). Marginalizing with respect to the data leads to a frequentist expectation of the FDR. Most references to FDR in the literature refer to this frequentist expectation. We refer to it as \(\hat{\text{FDR}}\), to distinguish it from the posterior expected \(\text{FDR} = E(\text{FDR} | y)\), conditional on the data.

A clever procedure proposed and popularized by [4] allows straightforward control of \(\hat{\text{FDR}}\). In contrast, the evaluation of \(\text{FDR} \) requires no clever tricks. Conditional on the data the only unknown remains \(\delta_i\) and we find \(\text{FDR} = 1/D \sum_{i=1}^{n} (1 - \bar{p}_i) d_i\). One could use a desired bound on \(\text{FDR}\) to fix the threshold \(c\) in the decision rule \(d_i = I(\bar{p}_i > c)\). This is proposed, for example, in [46]. Under the Bayesian paradigm, FDR control turns out not only to be easy, it is even the correct thing to do. The decision rule \(d_i = I(\bar{p}_i > c)\) can be justified as Bayes rule under several loss functions combining false discovery counts (or proportions) and false negative counts (or proportions). See [42] for a discussion.

An important special case of multiplicity adjustments arises in subgroup analysis. Many clinical studies fail to show the hoped for effect for the target patient population. It is then tempting to consider subgroups of patients. Often it is possible to make clinically or biologically reasonable arguments of why the treatment should be more specifically suitable for certain subgroups. With the increased use of biomarkers it is becoming more common to consider interesting subpopulations defined by various relevant biomarkers. For example, a targeted therapy that aims to address a pathological disruption of some molecular network can only be effective when the disease is in fact caused by a disruption of this network. Naturally, the investigation of possible subgroup effects needs to proceed in a controlled fashion to avoid concerns related to multiplicities. The problem is naturally approached as a decision problem, with a probability model for all unknowns, including the unknown subgroup effects, and a utility function that spells out the decision criterion. Recent discussions of a Bayesian decision theoretic approach appear in [51] and [44].

1.9 Conclusion

We discussed some prominent applications of Bayesian theory in biostatistics. The judgment of what is prominent, of course, remains subjective. Surely many important applications of Bayesian methods have been missed. A good recent discussion of some applications of more sophisticated Bayesian models in biostatistics appears in [13]. An extensive review of Bayesian methods specifically in clinical trials and health care evaluations appears in [11].
REFERENCES


References


