Ethical issues in oncology biostatistics

Peter F Thall Department of Biostatistics, University of Texas, USA

A medical statistician's routine professional activities are likely to have important ethical consequences. This is due in part to the fact that good medical practice and scientifically valid medical research both require as precursors high quality statistical design and data analysis. In this paper I discuss various ethical issues that I have encountered while working as a biostatistician at M.D. Anderson Cancer Center. I describe particular experiences and the ethical issues involved. Topics include medical decision making, benefit–harm trade-offs, safety monitoring, adaptive randomization, informed consent, and publication bias.

1   Introduction

In this paper I discuss a variety of ethical issues that I have encountered during my 11 years working as a biostatistician at M.D. Anderson Cancer Center (MDACC). In addition to recounting and discussing actual experiences, I explore some related hypothetical situations and generalities. The examples arose from the processes of designing cancer clinical trials, reviewing designs proposed by others, analyzing data, and collaborating on medical research projects. I have also drawn on my experiences making rounds at MDACC, which I do periodically at the invitation of a hematologist with whom I have long collaborated, Dr. Eli Estey. My focus here is on real situations because, like theoretical statistics, ethical concepts only matter if they are relevant to actual behavior.

It may seem that a statistician working in a medical center has few really difficult ethical choices, given that (s)he is not a physician and thus does not treat patients. This is not true. Medical decision making relies in a fundamental way on evaluating the potential benefits and risks associated with various therapeutic options. It follows that probability and statistics are essential components of medical decision making and hence of the ethical issues arising therein. In what follows, I will explain why a working medical statistician's actions often have a substantial impact on many medical decisions, and hence profound effects on the welfare of large numbers of patients. An implication is that medical statisticians should consider their day-to-day activities and decisions in an ethical as well as a scientific context. It is my intention here not just to relate my experiences, but also to influence the attitudes and behaviors of others involved in medical settings similar to those described here. Although the statistical concepts underlying medical ethics can be quite complicated, in order to address an

Address for correspondence: PF Thall, Department of Biostatistics, Box 447, University of Texas, M.D. Anderson Cancer Center, 1515 Holcombe Boulevard, Houston, Texas 77030, USA. E-mail: rex@mdanderson.org

audience including nonstatisticians I have minimized statistical notation as much as possible. While I have used simple probability statements to give substance to some of the examples, none are unduly complicated. Also, I have rounded numerical values to help simplify the illustrations.

The literature on medical ethics is substantial. Two useful books are Beauchamp and Childress[1] on biomedical ethics, and Kadane's[2] collection of articles on a wide variety of ethical issues in clinical trials. I have not attempted to place what follows into the context of the existing literature. Undoubtedly, I will reiterate many issues discussed elsewhere in much greater detail. Wherever I have committed this expository sin, I beg the reader's indulgence.

The examples that follow involve particular individuals, including physicians, nurses, statisticians, and patients. In many cases, there was strong disagreement with regard to decisions made or actions taken. The very fact that intelligent, highly skilled scientists and physicians may, in certain settings, hold very different views of what constitutes appropriate and ethical behavior underscores the complexity and subjectivity of medical decision making. While I have omitted names and, where reasonable, particulars of the disease and treatments, I have made no attempt to disguise my opinions. Most of the clinical trials described here either were designed at MDACC or originated elsewhere and it was formally requested that MDACC participate, and many of the trials involved multiple institutions.

2   Epistemology, in brief

Like any discussion of ethics, what follows is purely a matter of opinion. The question of whether or not a particular behavior is ethical hinges on what the individual engaging in that behavior knows, which in turn must rest on that individual's beliefs. It may be the case that, although two individuals engage in the same behavior, one is behaving ethically while the other is not. This could simply be due to the fact that the two individuals have different sets of rules for what constitutes ethical behavior. It also may be that they have different knowledge. For example, if two physicians both treat prognostically identical patients with treatment A, but one of the physicians is aware of a clearly superior treatment B while the other is not, then the physician with the greater knowledge may have behaved unethically. Of course, the argument that ignorance is a defense for apparently unethical behavior has its limits. A competent physician must have at least a reasonable knowledge of current practice.

On a deeper level, knowledge rests on belief. That is, one 'knows' something to be true if one believes it to be true, one has good reason to believe that it is true and, furthermore, it is true. A difficulty inherent in evaluating knowledge to determine whether a behavior is ethical is that, scientific method and all available empirical evidence notwithstanding, people believe what they want to believe. In the abstract, many particular actions may be categorized unambiguously as 'ethical' or 'unethical' within a given value system. In real life, aside from extreme cases where the distinction is obvious, the situation is often more complex than this simple dichotomy. This is especially true in oncology. In my experience, difficult circumstances and practical imperatives often force compromises in the design and conduct of oncology clinical

trials. These compromises, while far from ideal, are all that can be achieved in an imperfect world. So, ethics is a matter of belief, knowledge, circumstance, and practical necessity. Not an easy thing.

3   Clinical trials

A clinical trial is essentially a statistical device for evaluating medical treatments with the aim of developing improved therapies. The idea of conducting a medical experiment with human subjects only makes sense if one accepts the premise that the data obtained from the experiment may benefit future patients, and if it can be maintained plausibly that the benefit/harm ratio is favorable for the patients enrolled in the trial. Most people have little understanding of clinical trials, and many have never heard of them. People place themselves in the hands of a physician in order to obtain treatment for whatever malady afflicts them. So the first purpose of a clinical trial must be to treat the patients in the trial, otherwise it is clearly unethical. At the same time, there are many medical conditions for which no effective treatment exists or for which the existing treatments are not the best that one might hope for. That is, there is always plenty of room for improvement in medicine. From both a practical and an ethical perspective, then, it makes sense for physicians to learn from their experiences while treating patients. In modern medicine, it is neither necessary nor appropriate for a physician to rely on individual experience alone. The medical literature provides an essential shared resource in which physicians and scientists pool their experiences. The quality of this resource is determined by the quality and validity of the scientific method underlying what is reported. Statistics is the basis of scientific method. Consequently, it is essential that medical statisticians carefully consider the ethical issues involved in their professional activities.

While the outcomes of any medical procedure are not certain, the statistical notion of average behavior provides a basis for evaluating and comparing treatment effects within populations of similar patients. The rapidly developing array of modern statistical methods for reducing uncertainty through designed experiments and data analysis are increasingly being brought to bear in the process of developing improved treatments. The fact that human disease and medical practice are often chaotic, messy, and difficult to control notwithstanding, clinical trials are currently the best scientific device available for placing this process into the context of an inferentially useful experiment. So, the ethical imperative of learning as we go in medicine motivates experimentation in the form of clinical trials.

Many medical practices have been based on assumed common knowledge later shown to be incorrect by a randomized clinical trial (RCT). Beta-blockers were not given to patients with heart disease until a RCT showed that this treatment could increase survival in some of these patients. Autologous bone marrow transplantation (BMT) was considered by many oncologists to be a superior treatment for breast cancer until a RCT showed survival with autologous BMT to be the same as that achieved with other treatments. Granulocyte-colony stimulating factor (G-CSF) was considered a desirable treatment for myelodysplastic syndromes (MDS). However, a RCT of G-CSF

versus no therapy showed that the MDS patients given G-CSF actually had shorter survival. There are many other examples.

The main difficulty in the design and conduct of a clinical trial is that, ideally, it must provide the patients in the trial with the best available treatments while also generating data that will provide a valid basis for making inferences aimed at developing improved therapies. These two goals, each ethically motivated, are often in conflict. The statistical idea of treating patients in a clinical trial designed to generate data that will be applicable to similar future patients embodies a viewpoint that is different from actual medical practice. Physicians treat patients one at a time, not in groups. Necessarily, a clinical trial protocol imposes mandates or limitations on how the physician should or may treat each patient. A physician who finds these requirements unacceptable may feel ethically obligated to avoid enrolling patients in the trial. Another individual, either statistician or physician, may consider withholding eligible patients from a clinical trial to be unethical because it reduces the amount of information in the trial and thus the probability of benefiting future patients; it also may deprive a patient of the opportunity to receive a promising investigational treatment.

A pervasive difficulty for proponents of clinical trials is that statistical theory and practise are not so well established that it is always clear how to design a given trial or analyze a given data set. Statisticians argue endlessly about which methodology is most appropriate in a particular setting. This includes the Bayesian-versus-frequentist controversy, competing methods for dose-finding in phase I trials or safety monitoring in phase II trials or interim decision making in phase III trials, whether and how to account for multiple testing and data-driven model selection, the use of model-based versus nonparametric methods, and so on. These statistical controversies may undercut the argument that it is unethical not to conduct clinical trials since a skeptical physician may be disinclined to participate in an enterprise that, at its scientific basis, is a source of disagreement within the community of statisticians. So the conflicts and limitations of clinical trials are substantial, both ethically and methodologically. Nonetheless, the limitations of clinical trials should not be used as a basis for not conducting trials, since the benefits of clinical trials, if properly designed and conducted, greatly outweigh their drawbacks. Innovation is always controversial, and I believe that ethical and scientific disagreement will always be intrinsic parts of medical research. As a statistician, I regard this not as an impediment but as a strong motivation for statisticians to continue developing improved clinical trial designs and methods of data analysis, just as physicians continue to search for improved therapies.

4   Treatment choice and statistical inference

Therapeutic decision making often involves choosing one of several available treatments. For example, if a patient with acute leukemia is not treated, (s)he is almost certain to die very soon. Alternatively, any treatment combination will certainly cause harm in the form of one or more toxicities, varying in severity from nausea and vomiting to regimen-related death. In the comparatively happy cases where, for example, it is clear that surgery plus some chemotherapy or radiation therapy has a

high likelihood of providing the patient with years of high quality, disease-free life, there is no problematic ethical choice. The difficult cases are those where the likelihood of this outcome is not high, or where the expected disease-free survival interval is less than a few years, or where the only available treatment with any antidisease effect is known to be very harmful. The ethical choices that must be made in such settings can be very complex. They hinge not only on the physician's knowledge of available treatments and their possible desirable and undesirable effects but, because this knowledge is imperfect and uncertainty is substantial in any oncology setting, the probabilities of these effects. Thus, probability and statistical inference are necessary components of routine oncology practice. At the scientific level, where treatment advances are achieved, experimental design and data analysis are necessary tools. The role of the biostatistician is thus intrinsic to both good oncology practice and clinical research.

An oncologist has three basic therapeutic choices for each patient: treat with an established therapy, treat with an investigational therapy, or do not treat the patient with anything other than palliative drugs. The first avenue is appropriate when the patient's diagnosis is relatively straightforward and when an established therapy has a high chance of extending the patient's life. In contrast, if there is very little chance of improving the patient's survival with any established treatment, then it is the physician's ethical responsibility to convey this to the patient and possibly to their spouse or family members. This situation often arises with patients for whom one or more previous courses of treatment have failed. Invariably, such patients have suffered the adverse effects of these treatments, and the prospect of repeating this sort of suffering with little hope of extending their life is unattractive. Patients with highly advanced disease at diagnosis are in a similar circumstance therapeutically. This is often the case with pancreatic cancer because it usually is discovered only after reaching an advanced stage. Similarly, patients with early symptoms who delay the process of undergoing appropriate diagnostic tests may unwittingly allow their disease to progress to an advanced stage with poor prognosis. In many such cases the patient, if properly informed, will choose to decline treatment other than palliative medication and spend what little time (s)he has left at home or in a hospice. In contrast, some patients in such circumstances cling to the small hope of success with an investigational therapy and ask for the treatment even after they understand the small probability of therapeutic success. It is ethically appropriate to treat such patients in a phase I or phase II clinical trial. This requires, however, that the physician honestly and effectively explain the chances of success and of adverse events to the patient or their family members. A physician who knowingly misrepresents the chance of success with an investigational therapy in order to populate an early phase clinical trial is behaving unethically. In the absence of empirical data on an untried new agent, there is a great difference between saying that achieving a disease remission with the agent is possible and claiming that it has a 50% remission rate.

Consider a 30-year-old patient diagnosed with acute myelogenous leukemia (AML) or MDS who has favorable cytogenetic prognostic variables. Without treatment, such a patient has a median survival of about nine months, with very poor quality of life. With aggressive chemotherapy, which typically involves very unpleasant side effects, the patient is very likely to survive a year, has a median survival of about 30 months and, if

(s)he survives three years, has a very high probability of surviving many years thereafter. So treatment is clearly more desirable than no treatment. Now consider a 75-year-old AML/MDS patient who has unfavorable cytogenetic prognostic factors and is bedridden. With either no treatment or standard chemotherapy this patient has a median survival of about one month, with very poor quality of life, and less than a 2% chance of surviving nine months. So there is no real advantage to treatment with standard chemotherapy. As noted above, many such patients choose to be treated with an investigational therapy. In this case, the real choice is between an experimental treatment and no treatment. More difficult decisions arise in circumstances somewhere between these two extremes, or where there are qualitative differences between possible treatment modalities and their potential effects.

The above trichotomy, although a simplification, provides a useful basis for thinking about therapeutic choices. Importantly, each decision requires that the physician estimate the patient's survival time distribution, or at least portions of it such as the probabilities of surviving specified time periods, for each treatment option. In general, this estimation will depend greatly on the patient's prognostic covariates, such as age, disease stage, or previous treatment history, with the list of relevant covariates depending on the particular type of cancer. Such estimation relies on previous statistical analyses of one or more data sets arising from groups of similar patients in which the joint effects of treatments and patient covariates on survival time have been evaluated. That is, the decisions that an oncologist must make routinely and repeatedly in the course of patient diagnosis and treatment, ideally, rely on a good working knowledge of the results of detailed statistical analyses. A competent oncologist must be familiar with the medical literature on both established and newer, more recently studied treatments. Of course, medical papers summarizing such statistical analyses are only useful to the extent that the analyses, and the clinical trials from which the underlying data arise, are of good quality. Simply put, good statistical practice is a necessary basis for good medical practice. By the same token, bad statistics may lead to bad medicine, and in this regard physicians are at the mercy of incompetent or irresponsible medical statisticians. Consequently, it is ethically imperative that medical statisticians endeavor to provide the best design and data analysis they can, and that they pay careful attention to the manner in which the substantive conclusions of these analyses are summarized in any resulting medical papers.

5   Benefit–harm trade-offs

Several hundred years ago, bleeding a patient was in many medical settings the correct and ethical thing for a physician to do. Today, in many oncology settings, injecting poisons into a cancer patient's body is the correct and ethical thing. Unfortunately, at present the three most widespread cancer treatment modalities are cut, burn, and poison. These are known euphemistically as surgery, radiation, and chemotherapy. A fourth, somewhat more modern modality is blood or marrow cell transplantation, which carries with it a long list of potential adverse effects, including treatment-related death. The use of biological agents comprises a fifth therapeutic category. At present, some combination of these approaches is the best that the oncology community can

provide any given cancer patient. For most cancers, and there are many different kinds, no complete cure is available. Realistically, the best that can be hoped for therapeutically within a given cancer disease category and prognostic subgroup is a substantive improvement in the patient's chance of long-term survival, or in the quality of the patient's life during the time (s)he has remaining.

The general ethical principal 'First, do no harm' that is taught to young physicians must be set aside routinely in many oncology settings. This is because this approach will in many cases result in rapid death. Unfortunately, in our present state of technology, the only treatments that have any substantive effect against many cancers are often themselves harmful. Surgery for bone sarcoma may extend a patient's life, but it occasionally involves removal of an arm or a leg, or saving the patient's limb but with a lifelong functional impediment. Prophylactic irradiation of a lung cancer patient's brain may reduce the likelihood that the cancer will metastasize from the lung to the brain, but it also may cause permanent brain damage. Treatment of soft tissue sarcoma with ifosfamide may bring the disease into remission, but it also may permanently damage the patient's kidneys. Long-term treatment of chronic myelogenous leukemia with interferon may extend the patient's life, but it is also very likely to continually degrade motor function and cognitive abilities. Chemotherapy of acute leukemia or lymphoma kills cancer cells, but it also severely damages the ability of the patient's immune system to fight life-threatening infections. Allogeneic bone marrow transplantation from a matched donor may eradicate a patient's cancer, but the engrafted cells may attack and kill the patient. All of these examples suggest that trade-offs between potential desirable and undesirable treatment effects play an important role in cancer therapeutics.

Oncologists often must choose among a set of available treatment alternatives, each of which is very likely to have undesirable consequences but possibly also desirable effects. The most simple example is the most common, namely that of an oncologist choosing among several available chemotherapies. I will assume that the patient is relying entirely on the physician's recommendation, which is typically the case. Each choice might be described as infusing the patient with poisons, each likely to cause any of a variety of adverse effects, including transient but terrible effects such as nausea and vomiting, low white blood cell count leading to infection, transient or permanent organ damage, or death. Considered *per se*, using chemotherapy appears to be the behavior of a criminal. It is only seen as ethical, and in fact desirable, in light of the facts that (1) the patient is very likely to die of the disease if it is untreated, (2) the patient's survival time is, on average, more likely to be longer if (s)he is treated than if not, and (3) no treatment with life-extending capacity and without adverse effects exists.

Roughly speaking, in evaluating trade-offs a distinction may be made between events occurring within the same time frame and settings where the good and bad events are separated in time. For rapidly fatal diseases where only aggressive therapy is effective, either disease remission or regimen-related death will occur within a few weeks. Here, the trade-off is between the probabilities of these two outcomes. Late onset toxicities or secondary cancers caused by chemotherapy or radiation are a different matter. In these cases, the risk of the much later adverse effect must be weighed against the benefit of treatment obtained by eradicating or bringing into remission the initial cancer. For example, many leukemias are 'downstream' consequences of chemotherapy given years earlier to treat an initial cancer. This latter situation arises in the treatment of Hodgkin's

or non-Hodgkin's lymphoma, where there is a substantial cure rate but the treatment may be highly associated with a secondary cancer of a completely different type that appears many years later. Unfortunately, the rates of initial cure and downstream secondary cancer both increase with the aggressiveness of the initial therapy.

Most oncologists assess the trade-offs between such possibilities for each treatment in a primarily subjective manner by synthesizing their knowledge of the medical literature and their personal experience. Formal statistical methods for evaluating benefit–harm trade-offs do exist, however. Working in an oncology environment naturally leads one to think about methods for decision making based on trade-offs. A general Bayesian strategy for constructing designs to monitor multiple events in phase II trials, including both good and bad outcomes, is described by Thall, Simon, and Estey[3,4] and Thall and Sung. [5] Stallard, Thall, and Whitehead[6] provide a formal decision–theoretic basis for conducting this sort of trial. Statistical tests based on bivariate efficacy and toxicity outcomes are given by Bryant and Day[7] and Conaway and Petroni[8] for single arm trials, and by Willan and Pater, [9] Jennison and Turnbull, [10] Cook, [11] and Thall and Cheng[12,13] for randomized trials. Thall, Simon, and Shen[14] provide a Bayesian method for evaluating multiple treatment effects. Thall, Sung, and Estey[15] describe a method for quantifying the trade-off between the probabilities of complete remission and death in a clinical trial of several treatments for advanced acute leukemia. There is a large literature on quality of life indices, including Feeney and Torrance's[16] utility-based measures, the idea of quality-adjusted time without symptoms described by Goldhirsch, *et al.*,[17] and quality-adjusted survival introduced by Glasziou, Simes, and Gelber[18] and formalized by Zhao and Tsiatis.[19] The use of such formal methods in the day-to-day practice of oncology may provide oncologists with an improved basis for ethical decision making.

Consider therapy of AML within a particular prognostic subgroup that, if untreated, has a median survival of about six months. With treatment, the primary therapeutic goal is achieving a complete remission (CR), because without a CR the patient's survival is not much better than if (s)he is not treated. If a CR is achieved, then the patient has a chance of long-term survival. If the patient achieves a CR but later suffers a relapse, the chance of achieving a subsequent CR is much smaller and hence the chance of long-term survival is greatly reduced. Median survival post CR decreases with the length of time, or with the number of courses of chemotherapy, required to achieve the CR. Thus, initial prognosis consists of the probability of CR and the distribution of survival time, adjusted for the patient's baseline prognostic covariates. Survival should be considered overall at the start of therapy, but once therapy has begun the relevant survival distribution is conditional on whether or not the patient has achieved a CR and, once CR is achieved, how long it took to achieve CR and whether the patient has relapsed. Once a patient's disease stays in remission for three years, the chance of relapse declines precipitously, so a CR duration of three years is considered a 'cure.'

Here are two ways to quantify the trade-off between the probabilities of CR and death and incorporate this into a clinical trial design. Suppose that among, say, 300 historical AML/MDS patients treated with an established, 'standard' treatment, S, two months from the start of therapy 50% had achieved a CR, 20% had died, and the remaining 30% were alive but not in CR. Suppose that the aim of a clinical trial of an experimental therapy, E, is to improve the CR rate while controlling the death rate.

Denoting the probabilities of CR and death with S by θ_{CR} and θ_{DEATH} and the corresponding probabilities with E by λ_{CR} and λ_{DEATH}, a pair of Bayesian rules that together allow a 0.05 increase in the death rate as a trade-off for a 0.20 improvement in the CR rate are to stop the trial of E if either $\text{Prob}(\theta_{DEATH} + 0.05 < \lambda_{DEATH} \mid \text{data})$ is unacceptably large or $\text{Prob}(\theta_{CR} + 0.20 < \theta_{CR} \mid \text{data})$ is unacceptably small. Technical details are given in Thall, Simon, and Estey.[3,4] Another approach combines λ_{CR} and λ_{DEATH} into a single CR–death trade-off function, $\phi = \phi(\lambda_{CR}, \lambda_{DEATH})$, as follows. One may assign the desirability index, say, 0 to the historical probabilities (0.50, 0.20) with S, and the desirability index 1 to the physician's desired trade-off target (0.70, 0.25). The physician is then asked to specify other probability pairs (λ_{CR}, λ_{DEATH}) that have the same desirability as the trade-off target. The numerical parameters of the function ϕ corresponding to these probability targets are then derived algebraically. The physician may be shown contour plots of this function and, if necessary, its parameters are calibrated interactively on that basis. Decisions and inferences are then based on ϕ. Technical details are given in Thall, Sung, and Estey.[15]

6   Early stopping in single-arm trials

The nominal goal of a phase II cancer clinical trial is to determine whether a new treatment is sufficiently safe and promising to warrant further study in a large-scale, randomized phase III trial with survival or disease-free survival as the outcome. The usual outcomes in phase II are relatively early events such as disease remission and toxicity. There are many phase II designs. Most phase II trials are single-arm, in which case the comparison to the standard must be based on historical data, and consequently all experimental-to-standard parameters are confounded with trial effects. Because the treatment is new, and because anticancer treatments generally have adverse effects, safety is a major concern. Consequently, an essential design component consists of formal rules to stop the trial early if the interim data show a high likelihood that the treatment is either unsafe or inefficacious, compared to the event rates of standard treatments. If the trial is not stopped early, then the design should provide reasonably reliable estimators of the event rates and it should declare the new treatment promising with reasonably high probability if it is in fact superior to the standard.

A phase II trial was proposed to determine whether a new combination chemotherapy, E, for breast cancer was 'promising.' The protocol specified that the chemotherapy was to be used in conjunction with surgery, which is typical, with 'response' defined as complete or partial remission of the disease at six weeks. The statistical section of the protocol provided details of a Simon[20] optimal two-stage design to test the null hypothesis that the response rate of E was 20% against the alternative that it was at least 40%. The specified Simon design having type I error 0.05 and power 0.90 to detect this alternative requires 19 patients to be treated and evaluated in a first stage. If four or fewer responses are observed in the 19 patients then the trial is terminated; if there are five or more responses then 35 additional patients are treated and evaluated. While this seemed like a reasonable design for this trial, a closer look at the protocol showed that the historical response rate with standard treatment was 60%, not the assumed 20% null rate in the hypotheses underlying the design. A simple computation

shows that if the experimental treatment's response rate is actually 40%, which is 20% below the established rate with standard treatment, then the above design would stop the trial early with probability only 0.07. Another way of putting this is that if the experimental treatment actually *decreases* the response probability by 20% compared to the standard, then the design has probability 0.93 of proceeding to a second stage and using this inferior experimental treatment on 35 more patients.

When confronted with this glaring deficiency, the physician who had written the protocol and proposed this particular design for the trial, YoungDoc, explained that he was very excited about the experimental treatment and wanted to give it as good a chance as possible of getting through phase II. When asked why he had not used a design with a null rate of 60%, the standard therapy response rate, and targeted an improvement to 75% or 80%, YoungDoc explained that he had initially considered this but that the resulting Simon design would be 'too likely to stop the trial early.' Even after it was explained to YoungDoc that the protocol design had a very small probability of protecting the second cohort of 35 patients from being treated with a greatly inferior treatment, YoungDoc maintained that the design was appropriate. YoungDoc further argued that the Simon design is used very widely for phase II trials and that, moreover, the particular numerical design parameterization had been used in a previous phase II clinical trial in breast cancer. A number of questions regarding other important details of the protocol were raised, including the absence of any formal early stopping rule for adverse events. My proposal that such a safety monitoring rule be included in the protocol was addressed by pointing out that the protocol included a provision for the principal investigator to stop the trial if an unacceptably high adverse event rate was observed.

Given that response was the only outcome considered in the design, it obviously did not protect the patients in the trial from an inferior treatment. The null and alternative values, 20% and 40%, were a complete fiction in that they represented nothing more than the investigator's desire to treat as many patients as possible with the new regimen. In formal terms, YoungDoc cared a great deal about the risk of a false negative (Type II error) but nothing at all about a false positive (Type I error). In common terms, YoungDoc felt that it was much more important to protect the new treatment than to protect patients in the trial should the new treatment in fact turn out to be substantially inferior to the standard.

Considering YoungDoc's claimed optimism regarding the experimental treatment, and taking into account his apparent inability to think quantitatively, we must ask whether his behavior was ethical. Sadly, the answer is 'Yes.' As noted earlier, where ethics are concerned, ignorance is a perfectly valid defense. The argument YoungDoc put forth for the design's validity was based on the optimistic viewpoint that the experimental treatment was very likely to be superior to the standard, and on the fact that the same design had been used for the same type of trial in the same disease. Of course, this viewpoint regards everything qualitatively and completely ignores any quantitative issues. It also replaces empirical evidence, both past and future, with immovable prior belief. A deeper question is whether a physician so implacably ignorant of basic concepts of hypothesis testing and safety monitoring should be permitted to conduct a clinical trial.

This episode also raises the question of what actually constitutes a phase II clinical trial. Many regard it as a scientific experiment to evaluate the efficacy of a new

treatment, with the important condition that it should be terminated early with reasonably high probability if the experimental treatment is inferior to the standard. The physician who proposed the above protocol apparently regarded a phase II trial as a device to allow him to treat a large number of patients with an experimental regimen, with little attention paid to patient safety during the trial. This sort of behavior is not uncommon, in my experience, and I attribute it to the fact that research oncologists are rewarded for conducting clinical trials, and for accruing a large number of patients into a trial. The institution may receive a fixed amount of money for each patient accrued. The oncologist's career will benefit greatly if a trial of an investigational agent is 'successful,' which in particular requires that it not be stopped early for an unacceptably high adverse event rate. Attitudes with regard to trial conduct often reflect a large previous investment of preliminary laboratory research, time, and money by a pharmaceutical company supplying the investigational agent. Many pharmaceutical company representatives, and physicians who work with them, regard phase I and phase II trials not as scientific experiments but as bureaucratic hurdles that stand in the way of drug approval. Essentially, optimism and the hope of professional advancement or financial gain may overcome consideration of the consequences of a false positive conclusion and, all too frequently, honest consideration of patient safety.

In fairness to clinical oncologists who propose and conduct early phase trials of new treatments, there is a duality of clinical optimism and empirical skepticism. The clinician must be very optimistic to propose the trial in the first place, since in fact most investigational cancer treatments do not provide an improvement. It is essential that oncology statisticians understand that oncologists must deal with many failures, both with individual patients and with investigational treatments. The ability to confront this on a daily basis rests on a foundation of optimism that may seem unrealistic to those unfamiliar with the harsh realities of clinical oncology. Certainly, the statistician must provide a basis for objectively evaluating new treatments, which includes designs with explicit early stopping rules. This is an ethical *sine qua non* in clinical trial design. However, I feel that it is also the statistician's responsibility to provide encouragement, since the research oncologist's optimism is necessary for the entire enterprise of developing improved therapies. In my experience, many oncologists rely on their statisticians in much the same way that patients rely on their physicians.

7   Randomization

A lot of thought has been devoted to both the technique and ethics of randomization in clinical trials. Royall[21] and the comments and references therein provide a very thorough account. Kadane[22] proposes a radical Bayesian alternative to randomization. The ECMO trial described by Ware[23] and the object of much discussion, shows how highly skilled statisticians with the best of intentions may find themselves in difficult ethical circumstances when proposing a randomization scheme.

The idea of randomizing human beings between different medical treatments in order to compare the treatments' effects fairly is actually pretty strange. Explaining this modern, exotic statistical procedure to a nonstatistician is always an interesting process. The explanation requires casting some rather deep statistical ideas in simple

terms, as well as explaining the ethical concept of equipoise and, inevitably, appealing to personal, often emotional evaluation. Equipoise between two treatments simply says that the physician is as willing to treat the next patient with one treatment as with the other. Many physicians hate this idea and are unwilling to allow their patients to be randomized. Their usual argument is that explaining and proposing the randomization to a patient is an admission that they do not know what they are doing, since a truly competent physician should be able to choose the better of the two treatments for any given patient. Moreover, many physicians feel that randomization allows a statistician to rob them of the authority to choose, or at least to recommend, each patient's treatment. Unfortunately, the beliefs with regard to treatment effects that a physician develops based on individual experience may be quite different from the conclusion reached by objective analysis of existing data. This is due in large part to the facts that individual experience is necessarily limited, and memory is highly subjective and biased.

Several years ago I was asked to design a clinical trial in soft tissue sarcoma to compare surgery plus radiation, which was standard therapy, to surgery alone. The motivating rationale was that radiation increases the likelihood of limb dysfunction, and the hypothesis was that radiation could be dropped without a reduction in disease-free survival (DFS). So the goal was to improve limb function while maintaining DFS. I developed a design with both DFS and an index of limb function as the bivariate outcome, based on the test described in Thall and Cheng.[12,13] The surgeons liked the idea very much, the radiation oncologists hated the idea, and the trial was never run. People believe what they want to believe.

Bone marrow or peripheral blood stem cell transplantation (tx) and chemotherapy (chemo) are two very different treatment modalities for leukemia, lymphoma, and advanced breast cancer. Both involve very aggressive therapy that entails severe adverse effects, including regimen-related death. While it might seem that oncologists who use such therapies would be very interested in randomizing patients between these two modalities, this has seldom been done. Most clinical trials of experimental tx or chemo are either single-arm phase I or phase II trials of one of the two modalities. When randomized trials are conducted, the different treatments are usually two versions of chemo, such as different combinations of agents or different dose schedules, or two versions of tx, usually using different preparative regimens. This is not surprising if one realizes that the physicians who do tx and those who do chemo come from two very different groups, each with its own subculture and therapeutic orientation, sometimes residing within different sections or even different departments within an institution. In addition, a complication in designing a randomized trial of allogeneic tx versus chemotherapy is that a tx donor must first be found, and it would be necessary to tell the donors of all patients randomized to chemotherapy that their cells were not needed. There has been a recent advent of so-called 'mini tx' in which a nonablative chemo or radiation dose precedes the injection of cells, with reliance to some extent on the ability of the injected cells to kill cancer cells, the so-called 'graft-versus-disease' effect. This has lessened the cultural divide between the tx and chemo oncologists since, at least in theory, both the chemo and the injected cells play prominent roles in fighting the patient's disease. The two subcultures are still distinct, however, and randomized trials comparing chemo to tx are not common.

I once reviewed a large randomized clinical trial aimed at comparing two chemotherapy combinations, A and B. Some rather disturbing historical data, available before the trial, were available. These data came from a preliminary, small-scale randomized trial of A versus B in a prognostically homogenous patient group. Severe toxicity was observed in 4/20 (20%) of patients given A and in 20/22 (91%) of patients given B. Aside from my feeling that the preliminary trial should have been stopped early due to this observed extreme difference, given the historical data it seemed unethical to propose a new, much larger randomized trial. Starting with uninformative beta(0.50, 0.50) priors on the toxicity probabilities associated with A and B, given the historical data it is virtually certain *a posteriori* that B was more toxic than A. Moreover, the observed response rate with B was actually slightly *lower* than that of A. Who would want to be treated with B?

Even in situations where the physicians involved in designing a trial have equipoise between two treatments, they are often concerned that this will disappear if one treatment outperforms the other. Strict adherence to equipoise allows randomization for the first patient, but once that patient's outcome is observed equipoise disappears. Depending on one's viewpoint, equipoise might return at a subsequent point if the observed success rate of each treatment is exactly the common value believed *a priori*. For example, suppose that one starts with the belief that on average both treatments A and B have a 20% success probability for some binary outcome, treats the first patient with A as a result of a 50:50 randomization, and that patient's treatment is a success. Then equipoise is gone and one would necessarily treat the second patient with A. However, if A fails with each of the next four patients, then its observed success rate is back down to the 20% value initially believed for both A and B, so it seems sensible to again randomize. However, the evidence supporting the 20% rate for A is now stronger than that for B, so it is not clear that a 50:50 randomization is best. To address this problem, a Bayesian analysis that begins with, say, a beta(0.20, 0.80) prior on each success probability would yield a beta(1.20, 4.80) posterior on the success rate of A. So the posterior mean success rates are again both 20%, but the posterior of the success rate with A is more informative than the beta(0.20, 0.80) distribution of the success rate with B. If one accounts for the value of what is learned from each new piece of data, more can be learned by treating the next patient with B rather than with A. If one looks further into the future than one patient, then this sort of analysis rapidly becomes quite mathematically complex. It has been dealt with extensively under the rubric of 'bandit problems.' A basic reference is Berry and Fristedt.[24] The seminal paper on this seems to be Thompson.[25]

The study of bandit problems leads to the notion of outcome-adaptive randomization, which provides a compromise between ethical concerns and the scientific goal of obtaining an unbiased treatment comparison. An advantage of outcome-adaptive randomization is that it directly addresses the problem that equipoise is lost as interim data become available during the trial. Some useful survey papers are Rosenberger and Lachin,[26] Berry and Eick,[27] and Rosenberger.[28] The idea is to use the observed patient outcome data available at any interim point in the trial to compute randomization probabilities that are biased in favor of the treatment having more favorable outcomes. For example, if there have been 6/10 (60%) successes with A and 8/12 (67%) successes with B, the difference between these two observed rates might reflect actual superiority

of B over A or, alternatively, they may be due entirely to the play of chance. Based on these data, one adaptive Bayesian rule would randomize the next patient between A and B in a 3:5 ratio rather than 1:1. Specifically, the patient would receive A with probability $3/8 = 0.375$ and B with probability $5/8 = 0.625$. Moreover, the observed difference might become extreme enough to lead one to the conclusion that one treatment is clearly superior. For example, if there are 7/14 (50%) successes with A and 30/36 (83%) successes with B, then the randomization probability for B is 0.99. In practice, one would either terminate the trial or treat all remaining patients with B in this case. In my experience, many oncologists find this type of treatment assignment procedure very attractive. At MDACC, we have recently developed computer programs that perform the computations necessary to implement a Bayesian outcome-adaptive randomization, simulate the trial to obtain its operating characteristics and calibrate parameters on that basis, and a user-friendly front-end necessary for trial conduct. The front-end essentially looks like a 'point-and-click' patient log that tells the user, typically a research nurse or physician, each new patient's treatment assignment. Once this software became available, our oncologists rapidly began to use this methodology in numerous settings. These have included trials of surgery plus chemotherapy for gastrointestinal sarcoma, bone marrow transplantation, chemotherapy of acute leukemia, interferon for renal cell cancer, and treatment of acute respiratory distress syndrome. In several cases, a physician who initially planned to conduct a single-arm phase II trial of an experimental treatment decided to do an adaptively randomized trial of the experimental versus standard therapy. A remarkable aspect of this development at MDACC is that quite a few surgeons and tx oncologists, both historically disinclined to randomize patients, have found adaptive randomization to be ethically very attractive. Thus, the development of computer software that enabled us to implement adaptive randomization in the clinic has resulted in what I regard as a substantial cultural change in our community of oncologists.

8   The logistics of adaptive decision making

Adaptive randomization is an example of an outcome-adaptive decision rule, which is one that uses the outcomes of patients treated previously in the trial to make treatment decisions for new patients. A common example of outcome-adaptive decision making in oncology is a phase I dose-finding trial in which successive cohorts of patients are treated at specified doses. Once it is observed whether or not each patient in a given cohort experiences toxicity, the dose for the next cohort is chosen based on these outcomes and, ideally, also the dose—outcome data from all previous cohorts. Another example is a phase II trial with a rule saying that if there are three or fewer responses in the first 19 patients then the trial must be stopped; otherwise 30 additional patients should be treated. A very common logistical problem arising in such settings is that the outcome may not be observed immediately but rather is defined over a particular time interval. 'Toxicity' may be defined over the first month of therapy, so that a patient can be scored as 'No toxicity' only after waiting a month. Similarly, 'response' may be defined as the event that the patient is alive with disease in remission at two months. In such settings, formal application of an adaptive decision rule may require waiting

this time period in order to obtain the data and apply the rule to determine the next patient's treatment. This may be logistically inconvenient or impossible. For example, formally applying a safety monitoring rule after each fifth patient in a trial where 'response' is defined over a three-month time period in a 50-patient trial would require up to $3 \times 9 = 27$ months during which accrual is suspended. While this is simply not feasible, the ethical question of how to conduct the trial safely remains, since it should be terminated with high probability if the six-month response rate is indeed unacceptably low. One seemingly obvious approach is to continue accruing patients before the outcomes of previously treated patients are observed and make all decisions based on the outcomes that have been observed. This may have a high risk of treating new patients ineffectively, especially if the accrual rate is high relative to the three-month outcome window. In the case of dose-finding, the risk is treating a large number of patients with a current dose that later turns out to be overly toxic. So dealing with issues that seem to be purely logistical actually has serious scientific and ethical implications. Most of the statistical literature completely ignores this pervasive logistical problem, instead assuming the mathematically convenient fantasy that outcomes are observed immediately. Two papers dealing with this issue in dose-finding trials are Thall *et al.*[29] and Cheung and Chappell.[30] Eick[31] gives a formal treatment of this problem for randomized trials with a survival time outcome. Cheung and Thall[32] propose a method for monitoring composite binary events in phase II trials that does away with the need to suspend accrual.

9   Informed consent

The idea of providing an easily understood description of the possible benefits and adverse effects of the treatments to be studied in a clinical trial seems like a sensible, ethical way to inform prospective patients about what they might be getting into. Patients who satisfy a trial's entry criteria are asked to read this description, or have it explained to them, and a patient is not entered into the trial unless (s)he signs an 'Informed Consent' form. This seems to makes sense. Any reasonable person would agree that patients should have the right to decide whether or not to participate in a clinical trial based on an honest description of what it entails.

While this may sound wonderfully ethical, in practice it is far from this ideal. Cancer patients seldom choose a treatment. Instead, they choose a physician and then rely entirely on their chosen physician's advice. Most cancer patients are so emotionally upset by the discovery that they have cancer, or that a previously treated cancer has recurred and requires further therapy, that they are incapable of making objectively rational decisions. Family members or friends are not much help in this regard, either, since they are also emotionally involved. Only a patient who is calmly rational and also intelligent enough to understand the intricacies of the treatments and the probabilities of the various outcomes can make such a decision effectively. Meisel and Roth[33] provide a review of how informed consent actually works. In reality, in almost all cases it falls on the physician honestly and accurately to present options, possible consequences, and probabilities to the patient. In order to do this effectively the physician must have a rather deep understanding of the possible treatment outcomes and their

probabilities for the particular patient. As explained above, this in turn relies on inferences from previous statistical data analyses. So, in reality, the nominal goal of informed consent can only be achieved if high quality statistical analyses of high quality data have been performed and the results of these analyses have been well understood by the oncologist and communicated in a readily interpretable way. Once again, effective application of statistical methods is a precursor to the ethical practice of medicine.

10   Designed confounding

A particular new chemotherapy, which I will call Newdrug, was studied in a series of clinical trials for treatment of AML/MDS. Newdrug was always used in combination with a standard drug, which I will call Standard. So, some patients received this combination while others received Standard. Some of the trials were single-arm phase II, others were small randomized trials with the different arms arising from other agents being added to Newdrug + Standard. At the insistence of a particular scientist, LabGuy, involved in the development and promulgation of Newdrug, it was decided that, whenever used in this combination to treat AML/MDS, Standard should be administered as a bolus. This was important because the mode of administration for Standard without Newdrug was continuous infusion. Another complication was that none of the trials included randomization of Newdrug + Standard versus Standard; consequently, the Newdrug effect was always confounded with trial effects. A statistical analysis of the resulting data showed that, with the exception of patients with very good prognosis, after accounting for prognostic covariates and possible trial effects, the Newdrug + Standard combination appeared to result in much worse survival compared to Standard alone. While data from multiple trials were combined in this analysis, the between-trial effects within each treatment group were small in relation to the treatment effects. The negative conclusion regarding the efficacy of Newdrug was at odds with what was reported previously in a series of papers based on each successive trial, and many physicians worldwide have treated AML/MDS patients with Newdrug on that basis.

   Since the Newdrug effect was confounded with bolus administration of Standard, LabGuy argued that the relatively poor performance of Newdrug may have been due to the fact that Standard was given as a bolus in the Newdrug + Standard combination and by continuous infusion when Standard was given without Newdrug. LabGuy also argued strongly and repeatedly that the paper describing these analyses should not be submitted for publication, citing the need for collegiality as an additional motivation. The discussion became acrimonious and unproductive. The paper was written, submitted, and published in the medical literature.

   How might this unfortunate series of events have been avoided? There are several ethical and scientific issues involved. While each clinical trial was designed in a reasonable manner, the overall evaluation of Newdrug by the collection of trials was scientifically flawed in two important ways. The first flaw was that Newdrug + Standard was never randomized against Standard, which resulted in a confounding of the Newdrug effect with trial effects. Only a randomized comparison can really

resolve the issue. This may be addressed by estimating the trial effects using external data, although this approach requires additional assumptions that would not be necessary had a randomized trial been conducted. The second flaw was that the real comparison became Newdrug + Standard given as bolus versus Standard given as continuous infusion, so the poor performance of the former could be attributed to the bolus effect. Both of these problems were design flaws, and with 20/20 hindsight I see that I should have paid greater attention to the overall process of evaluating Newdrug, and for certain trials I should have argued more forcefully for randomization of Newdrug + Standard versus Standard. In fact, I lost this argument. The question of bolus versus continuous infusion of Standard seemed like a purely medical decision beyond my authority as a statistician, since LabGuy insisted that Standard be given as a bolus when combined with Newdrug on fundamental scientific grounds. I had no idea that it would arise as an important statistical issue. All of this suggests to me that, to help avoid this sort of mess in the first place, application of fundamental statistical design principles is essential. At a deeper level, it is the statistician's ethical responsibility to argue clearly and, if necessary, forcefully for the use of appropriate design in clinical trials. Of course, sometimes you just cannot win.

Another trial that I was asked to review aimed to study the effect of a drug, FeelGood, when injected into the patient prior to surgery. The trial design randomized patients between surgery alone and surgery + injected FeelGood. The primary outcome was quality of life (QOL) as measured by a questionnaire administered shortly after recovery from the surgery, and the motivating hypothesis was that FeelGood would improve QOL. Importantly, the FeelGood injection was given prior to surgery while the patient was conscious. Patients receiving surgery alone received no injection. Thus, based on this and the informed consent describing the treatments, the patient knew whether (s)he had received FeelGood. Consequently, because QOL was the primary outcome and this is of course an entirely subjective variable, the FeelGood effect and the prior injection effect were completely confounded. The physicians also explained that FeelGood could not be given effectively as part of the intravenously administered mixture that patients receive prior to and during surgery.

I strongly suggested that patients in the surgery alone arm be given a sham injection to avoid this confounding. I was not alone in this viewpoint. The discussion of this protocol, which continued over several weeks and became highly politicized, ended with the protocol being approved in its original form. Of course, the resulting data from the trial will be completely useless for evaluating the primary outcome, since any difference in QOL is easily attributable to the patient's knowledge of whether (s)he received an injection. Had the primary outcome been survival time, then the injection effect would have been irrelevant. One might argue that it is unethical to give a sham injection, since it does not benefit the patient. If one adopts this viewpoint, however, then it becomes impossible to conduct the experiment in any manner but the fundamentally flawed way that was agreed upon. The question then becomes whether it is ethical to devote time, resources, and, above all, patients to such a fruitless enterprise. I think not.

This episode also illustrates an important behavioral component of the process whereby groups of people within an institutional setting are required to reach a consensus. In my experience, many individuals would much rather avoid disagreement

and the possibility of discord rather than hold to a viewpoint that they actually believe is correct. That is, they would rather be wrong along with everyone else than be right by themselves. In many settings over the years, I have watched individuals first assess group opinion without expressing their own view and then vote with the majority. Since people believe what they want to believe, it is easy to rationalize such behavior simply by saying that one has acted in a manner concordant with the viewpoints of respected colleagues. The point is that ethical behavior may be unpopular, especially within settings involving many individuals and one or more institutions. The issue for each individual is whether one has the courage to act in accordance with one's convictions.

11   Some simple criteria

A long-term, large-scale randomized trial was designed to assess the ability of a certain drug, Preventol, to reduce the risk of breast cancer. This was not a trial of treatments for cancer, but rather it was a prevention trial aimed at healthy women who satisfied a particular risk profile. It happened that a female physician involved in organizing the trial satisfied the entry criteria. She decided not to participate.

I have no idea why this individual declined to enter the trial, and I consider it inappropriate for me to ask. One relevant fact is that, at the time the trial was organized, there was thought to be a small probability that Preventol might increase the risk of ovarian cancer. One might assume that this did not motivate the physician's decision, since in recommending the trial to other women she must have considered this risk an acceptable trade-off for the potential benefit from Preventol. Perhaps she felt that participating would have impaired her objectivity. In any case, her actions seem to have a logical inconsistency. The question is whether it is ethical to propose that others take a medication, either preventive or therapeutic, while considering this medication unacceptable for oneself. I wonder how the decisions of women who were offered participation in the trial would have been affected had they known of this physician's personal decision.

I use the following simple criterion when designing a clinical trial, which I highly recommend. Ask yourself, 'If I were an eligible patient, would I want to participate in this trial?' If the answer is 'No,' then the next step is to modify the statistical design to change the answer to 'Yes.' In the rare instances where a physician has imposed limitations that made me unable to get to 'Yes,' we parted company. If one is reviewing someone else's design, then it is one's responsibility not only to carefully explain its deficiencies, but to make it clear where ethical issues are involved.

Often, an investigator at MDACC is asked to participate in a multi-institution trial where the options are to 'take it or leave it' in that the design has already been negotiated and will not be changed in any case. Participation in multi-institution clinical trials is an essential part of the scientific activity in any large research hospital. The research clinicians in each disease area belong to a larger community of oncologists who work in other medical institutions, pharmaceutical companies, or regulatory agencies. The decision by MDACC to not participate in a multi-institution trial thus has many consequences involving relationships on both the personal and institutional level. Consequently, a 'take it or leave it' protocol typically is disapproved for MDACC

participation only if either it is considered to be clearly unethical or the science is just plain wrong. Otherwise, even if the science is considered to be not the best, typically the protocol is approved.

12 Publication bias

Some years ago, I presented a physician colleague with the results of a statistical analysis of data from a recent clinical trial of an investigational treatment regimen, NewRx. The analyses clearly showed that NewRx provided no improvement whatsoever over the standard therapy. It was the statistical equivalent of pointing out that a dog has four legs. The physician then told me that, given such negative results, (s)he intended not to write a paper describing the trial's results. After a short but extremely interesting discussion, the oncologist agreed to write the paper, and it was subsequently published in the medical literature.

   The issue here is that it is essential to convey negative results to the scientific community. Otherwise, other investigators may waste time and resources pursuing NewRx when, if they were but aware of the negative results, they would instead devote their time elsewhere, say investigating NewNewRx. In scientific terms, the failure to publish negative results produces inflated false positive rates and misleads colleagues. A related issue is that it is the ethical responsibility of the statistician to make sure that any scientific enterprise in which (s)he is involved is honestly reported in the scientific literature. This includes not only insisting that negative studies be published, but also making sure that the specific language in any scientific paper clearly and accurately reflects the conclusions of any statistical analyses. That is, the statistician's responsibility does not end with data analysis.

Acknowledgements

References

1 Beauchamp T, Childress, J. *Principles of biomedical ethics*, 3rd edn. New York: Oxford University Press, 1989.

2 Kadane JB. (Editor) *Bayesian methods and ethics in a clinical trial design.* New York: John Wiley, 1996.

3 Thall PF, Simon R, Estey EH. Bayesian sequential monitoring designs for single-arm clinical trials with multiple outcomes. *Statistics in Medicine* 1995; **14**: 357–79.

4 Thall PF, Simon R, Estey EH. New statistical strategy for monitoring safety and efficacy in single-arm clinical trials. *Journal of Clinical Oncology* 1996; **14**: 296–303.

5 Thall PF, Sung H-G. Some extensions and applications of a Bayesian strategy for monitoring multiple outcomes in clinical trials. *Statistics in Medicine* 1998; **17**: 1563–80.

6 Stallard N, Thall PF, Whitehead J. Decision theoretic designs for phase II clinical trials with multiple outcomes. *Biometrics* 1999; **55**: 971–77.

7 Bryant J, Day R. Incorporating toxicity considerations into the design of two-stage

phase II clinical trials. *Biometrics* 1995; **51**: 1372–83.

8   Conaway MR, Petroni GR. Designs for phase II trials allowing for a trade-off between response and toxicity. *Biometrics* 1996; **52**: 1375–86.

9   Willan AR, Pater JL. Hypothesis testing and sample size for bivariate binomial response in the comparison of two groups. *Journal of Chronic Diseases* 1985; **38**: 603–608.

10 Jennison C, Turnbull BW. Group sequential tests for bivariate response: Interim analysis of clinical trials with both efficacy and safety endpoints. *Biometrics* 1993; **49**: 741–52.

11 Cook RJ. Coupled error spending functions for parallel bivariate sequential tests. *Biometrics* 1996; **52**: 442–50.

12 Thall PF, Cheng S-C. Treatment comparisons based on two-dimensional safety and efficacy alternatives in oncology trials. *Biometrics* 1999; **55**: 746–53.

13 Thall PF, Cheng S-C. Optimal two-stage designs for clinical trials based on safety and efficacy. *Statistics in Medicine* 2001; **20**: 1023–32.

14 Thall PF, Simon R, Shen Y. Approximate Bayesian evaluation of multiple treatment effects. *Biometrics* 2000; **56**: 213–19.

15 Thall PF, Sung H-G, Estey EH. Selecting therapeutic strategies based on efficacy and death in multi-course clinical trials. *Journal of the American Statistical Association* 2002; **97**: 29–39.

16 Feeney DH, Torrance GW. Incorporating utility-based quality-of-life assessment analysis. *Medical Care* 1989; **27**: S190–S204.

17 Goldhirsch A, Gelber RD, Simes RJ *et al.* for The Ludwig Breast Cancer Study Group. Costs and benefits of adjuvant therapy in breast cancer: a quality-adjusted survival analysis. *Journal of Clinical Oncology* 1989; **7**: 36–44.

18 Glasziou PP, Simes RJ, Gelber RD. Quality adjusted survival analysis. *Statistics in Medicine* 1990; **9**: 1259–76.

19 Zhao H, Tsiatis AA. A consistent estimator for the distribution of quality adjusted survival time. *Biometrika* 1997; **84**: 339–48.

20 Simon R. Optimal two-stage designs for phase II clinical trials. *Controlled Clinical Trials* 1989; **10**: 1–10.

21 Royall RM. Ethics and statistics in randomized clinical trials. *Statistical Science* 1991; **6**: 52–88.

22 Kadane JB. Progress toward a more ethical method for clinical trials. *Journal of Medicine and Philosophy* 1986; **11**: 385–404.

23 Ware JH. Investigating therapies of potentially great benefit: ECMO (with discussion). *Statistical Science* 1989; **4**: 298–340.

24 Berry DA, Fristedt B. *Bandit Problems.* New York: Chapman and Hall, 1985.

25 Thompson WR. On the likelihood that one unknown probability exceeds another in view of the evidence in two samples. *Biometrika* 1933; **25**: 275–94.

26 Rosenberger WF, Lachin JM. The use of response-adaptive designs in clinical trials. *Controlled Clinical Trials* 1993; **14**: 471–84.

27 Berry DA, Eick SG. Adaptive assignment versus balanced randomization in clinical trials: A decision analysis. *Statistics in Medicine* 1995; **14**: 231–46.

28 Rosenberger WF. New directions in adaptive designs. *Statistical Science* 1996; **11**: 137–49.

29 Thall PF, Lee JJ, Tseng C-H, Estey EH. Accrual strategies for phase I trials with delayed patient outcome. *Statistics in Medicine* 1999; **18**: 1155–69.

30 Cheung Y-K, Chappell R. Sequential designs for phase I clinical trials with late-onset toxicities. *Biometrics* 2000; **56**: 1177–82.

31 Eick SG. The two-armed bandit with delayed responses. *Annals of Statistics* 1988; **16**: 254–64.

32 Cheung Y-K, Thall PF. Monitoring the rates of composite events with censored data in phase II clinical trials. *Biometrics* 2002; **58**: 89–97.

33 Meisel A, Roth L. Toward an informed discussion of informed consent. *Arizona Law Review* 1983; **25**: 265–346.