Published in final edited form as:

Curr Gene Ther. 2009 October ; 9(5): 368–374.

Uncertainty in the Translation of Preclinical Experiments to Clinical Trials. Why do Most Phase III Clinical Trials Fail?

Pedro R. Lowenstein^{*} and **Maria G. Castro**

Board of Governors' Gene Therapeutics Research Institute, Departments of Medicine, and Biomedical Sciences, Cedars-Sinai Medical Center, and, Departments of Medicine, and Molecular and Medical Pharmacology, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, California, USA

Abstract

A large majority of Phase III, large scale, clinical trials will fail, including gene therapy trials. This paper attempts to address some of the causes that may have inadvertently led to such a high failure rate. After briefly reviewing the detailed and high quality work that goes both into the preparation and conduct of such large Phase III clinical trials, and the preclinical science that is used to support and originate such trials, this paper proposes a novel approach to translational medicine which would increase the predictability of success of Phase III clinical trials. We propose that a likely cause of such failures is the lack of “robustness” in the preclinical science underpinning the Phase I/II and III clinical trials. Robustness is defined as stability/reproducibility in the face of challenges. Many times preclinical experiments are tested in a very narrow set of experimental conditions. Thus, when such approaches are finally tested in the context of human disease, the challenge provided by the varied age of patients, the complex genetic makeup of human populations, and the complexities of the diseases to be treated provide challenges which were never tested or modeled. We believe that the introduction of revised approaches to preclinical science, including the use of the latest developments in statistical, scientific, mathematical, and biological models, ought to lead to more robust preclinical experimentation with its subsequent translation, to more robust Phase III clinical trials.

Keywords

Glioblastoma multiforme; phase III clinical trials; bayesian statistics; effect size; gene therapy; cancer

PART I: THE PLAYERS

“Don't worry; I know almost exactly what I'm doing.”

Z, in *Antz*, by Woody Allen

Transitions are always difficult as they force us to step from the predictably known into the uncertain. During and following a transition is when most failures occur. As a result, many times, instead of progress, we reach a stalemate. In business hierarchies this is summarized by the well-known Peter's Principle: “In a Hierarchy Every Employee Tends to Rise to His Level of Incompetence”, a transition thus followed by failure.

Does Peter's Principle apply in any way to the challenges we come across in translational medicine, and especially in gene therapy clinical trials? The cynic's version would read, "In the world of clinical trials, most preclinical approaches will eventually fail in translation". Though sad, this opinion is not too far from reality, depending on the particular therapeutic field examined; stroke being a particularly failure prone area, while cancer therapeutics in general would appear to be somewhat more successful. In gene therapy, given the relatively small numbers of Phase III trials that have been devoted to a single disease, it would be more difficult to comment on this accurately, although, as all other fields of translational medicine, it would be unexpected if the success rate in gene therapy would be much higher.

Given that much of basic biomedical science aims to be translated into novel clinical trials, such Phase III marginal benefits or failures, especially when subjected to stringent statistical meta-analyses[1-4], would be seen as a significant stumbling block in the overall translational medicine enterprise. Further, the delay to conduct randomized, double-blind controlled clinical trials, dramatically postpones a formal assessment of treatment bias. To begin to analyze the reasons for failures of large Phase III trials, one could in principle look for shortcomings in either the clinical trials themselves, or in the preclinical experiments that led to the clinical trials in the first place. If we decide to examine in detail the preclinical science underlying a particular clinical trial in gene therapy, we will be very likely to find many papers on the particular viral vectors used, the particular treatment strategy proposed, the efficacy of the proposed strategy in one or more preclinical experimental models, and usually there will be published work from different laboratories showing similar results in what appear to be comparable experiments.

However, until a given Phase III clinical trial is finally published many years (or even decades) after it was first proposed, very little information will be available on Phase III trials in progress. Nevertheless, it will usually be possible to find the results of the Phase I and II trials that precede the larger, and decisive, Phase III trials. In spite of the asymmetry in the available information from the basic and clinical camps, we will attempt a first dissection of how preclinical experiments become clinical trials [5], and how it would be possible to make this process more robust, such that we may increase the likelihood of success of Phase III trials.

Clinical Trials in Gene Therapy

An evaluation of how Phase III trials are conducted will point to the enormous amount of work that is required to prepare such large trials. Phase III trials are very complex, including large numbers of patients enrolled, they are multi-centric, involving many different hospitals across one country, or even across many different countries. Phase III trials are also double blinded, controlled trials, signifying that neither the patient, nor the doctor, knows who receives the active gene therapy, and who receives the placebo. The large scale of Phase III trials makes them very expensive, further supporting the eventual significance associated with such expensive ventures [6].

As a consequence, Phase III trials take a long time in planning [7]. Not only do the necessary biologicals have to be produced (i.e., viral vectors), but the trial has to be coordinated across many different centers. The consequence of organizing such large scale trials, is that their preparation, including the coordination of patient population to be selected, treatment administration, standard of care to be administered, and discussions of the statistics and controls to be used has to be performed in much detail, and much in advance of the actual beginning of patient enrollment.

Further, during the initial planning stages investigators need to determine how results of the trial will be analyzed statistically. The clinical statisticians will make sure that the trial is sufficiently powered to be able to detect the percentage change in patient improvement that is

reasonably expected by the treating physicians. Thus, when aiming to obtain significance at equal p values, the number of patients needed to see an effect will vary inversely with the effect size. For example, the number of patients needed to be enrolled in a clinical trial to detect a 20% improvement, will be lower than the number of patients needed to reliably detect a 5% improvement in outcome measures. Thus, the smaller the potential therapeutic benefit to be detected, the larger the number of patients needed to detect the difference as statistically significant [7-14].

To provide an idea of what these expected changes actually mean, we will take a closer look at the situation with glioblastoma multiforme (GBM), a devastating primary brain cancer. As discussed at the recent ASCO meeting in Orlando in 2009, the likely median survival of patients diagnosed with GBM, of ages between 18 and 65 years of age, being treated at one of the advanced medical centers of the USA with best available standard of care, i.e. surgery, radiotherapy, chemotherapy, and salvage therapy, was considered to lie between 15-21 months. Thus, if we take 18 months as the potential median survival provided by the current standard of care at top university hospitals in the USA, a 10% improvement, would represent an extension of median survival by ~1.8 months, and 20%, an increased survival of ~3.6 months. Of course, more patients would be needed to demonstrate a smaller, albeit still statistically significant increase in median survival [15-17].

In summary, it would thus appear that clinical trials are properly planned, designed, and statistically powered to see expected differences. Given the time and effort that is devoted to the careful planning and performance of trials, the expectation is that everything is set to detect clinically relevant differences under stringent clinical, experimental and statistical conditions. Thus, *a priori* it is difficult to see how such carefully planned Phase III clinical trials would somehow be at fault for their eventual failure.

The Basic Science Aspects Underpinning the Clinical Trials

On the other hand, preclinical experiments represent the original science underlying the development and implementation of novel clinical trials. Although not articulated very clearly in the literature, preclinical experiments are mostly always necessary precursors of clinical trials. Similarities and/or differences in the physiology and anatomical organization between humans and experimental animals, as well as the availability of disease models that significantly recapitulate at least some aspect of the human disease in certain experimental animals, determine which animal models are used to test novel therapies for human diseases. Also, as a consequence of the Nuremberg Medical Trials' third item, animal experiments ought to precede human medical experiments [18-20].

A series of experiments are thus performed to determine the efficacy of a novel proposed treatment, data are stringently analyzed statistically, conclusions are drawn, and manuscripts are published in peer reviewed journals. As the exchange of therapeutic tools, e.g. viral vectors is straightforward, it is possible to exchange vectors between different laboratories, thus facilitating that identical therapeutic approaches are tested by different research groups. Alternatively, sometimes it is easy to redo the particular viral vector of interest, in such a way that the same potential therapeutic tool is produced in two different laboratories, and then tested independently.

The models in which novel treatments are tested will vary depending on the particular interests of the laboratory testing the novel therapeutic approaches, the target disease, and the general availability of a particular animal model. For brain tumors, models can be the relevant tumor cell line grown *in vitro*, obtained from a variety of brain tumors from various species, including humans; they can be primary cell lines outgrown from explanted tumors removed from patients; they can be xenogeneic grafts of human tumors (primary or cell lines) implanted into the brains

of immune-deficient animals; tumor cell lines implanted into syngeneic rodents[21-23]; or the models can be brain tumors induced through the over-expression of relevant oncogenes, or inhibition of tumor suppressors achieved directly through retroviral or lentiviral vectors or utilizing plasmids encoding the sleeping beauty transposon and the desired genetic lesions [24-27]. These are injected into the brains of rodents, adults or neonates. Also, transgenic models for GBM and other brain cancers have recently been developed [24,28].

Although a detailed comparison of all available models will necessitate more space than the one allotted to this paper, a highlight of some advantages and disadvantages as they related to models of human brain tumors as necessary examples to support the later discussion herein. For example, cell lines will allow the determination of direct killing mechanisms, especially as these may differ in different species; however, cell lines grown *in vitro* lack the *in vivo* geometrical organization, and cellular complexity of *in vivo* models. *In vivo* models, especially if the tumor is implanted or induced orthotopically in the brain itself, will recapitulate the dynamics of a real tumor, and the challenges in making the therapeutic agent available to the tumor cells. Nevertheless, if the model is a rodent one, there will remain significant differences because of the species, the way of inducing the tumor, and thus, the overall complex amount of different cell types and anatomical organization that constitute the real human tumors. Induction of tumors through the introduction of genetic lesions similar to those present in human tumors recapitulates at least partially the genetic, and many times the anatomical organization of brain tumors, as well as some of the typical infiltrative behavior of these tumors. Thus, each model has particular advantages and disadvantages and it is easy for the scientist to justify the model selected for each particular application, as no model is ideal[21,29-33].

A further complication in evaluating the models in which the novel therapies are tested is the fact that novel therapies in humans are usually tested as additions to the 'standard of care' treatments. In the case of a primary de novo glioblastoma multiforme, the patient would be treated with surgery, followed by radiotherapy, and the chemotherapeutic drug temozolomide. The novel therapy, depending on the particular therapeutic approach, will then be administered to the patients either concurrently with standard of care, or thereafter. It is practically impossible to mimic this in experimental models, as the combination of surgery, radiotherapy, and temozolomide all together is realistically almost impossible to apply reasonably and logically in the context of the animal models. Nevertheless, even if some aspects of standard of care could be added to the experimental paradigms tested in animals, because of the differences between the patients' diseases and the animal models, it remains unclear how well we can model the situation that will be encountered in a given human patient.

Preclinical experimental studies are also analyzed statistically, whether to determine the existence of any treatment-induced toxicity, or the efficiency of a novel therapy[31]. The number of animals is usually planned in advance, and in the case for brain tumors, scientists will look for life extension, or the existence of long term survivors after tumor implantation followed by the treatment. Various statistical tools will be used to analyze the data, with the accepted cut-off-point of $p < 0.05$ being the limit of accepted statistical significance which deserves further experimental exploration, or more usually, publication.

An important limitation of all experimental studies is that so far, genetic heterogeneity has been difficult to model in animals, although in the future it should become possible to do so if this would be considered of high priority in the preclinical testing of novel cancer therapies. At this time, all animal used in preclinical experimentation are all usually genetically identical, a situation quite unlike the human patient population.

In summary, if the clinical trials are prepared at the highest clinical and statistical standards, and the experimental science supporting the clinical trials is also of the highest experimental

standards, where are the unseen limitations that eventually cause the failure of most Phase III clinical trials. It appears that the answer cannot be found in the clinical trials themselves. These are large enterprises, not dominated by a single person or laboratory; they are randomized, double blind, and are analyzed through stringent statistical methods. And yet, the basic science is usually of similar high standards, performed by internationally renowned laboratories and published by in good quality peer-reviewed scientific journals.

And yet, a central question that concerns us here, is why a multitude of clinical Phase III trials are unsuccessful because they fail to reach statistical significance [34].¹

PART II: IN SEARCH OF EXPERIMENTAL ROBUSTNESS IN TRANSLATIONAL MEDICINE

“A time to search and a time to give up, a time to keep and a time to throw away”
Ecclesiastes 3: 1, New International version

“We occasionally stumble over the truth, but most of us pick ourselves up and hurry off as if nothing had happened.”

Winston Churchill

One thing that strikes us about the famous passage from Ecclesiastes is its reassuring description of there being a time for everything. And still, Ecclesiastes alerts us to “a time to search and a time to give up”. As scientists, doctors, physicians, we are taught never to give up; failure is not an option. As scientists we learn how to calculate the statistics of significance. Thus, we are taught to calculate an estimate of the error of our scientific measurements, but we are not taught to calculate whether a given result matters biologically, or when does it not. When do we have an effect that might translate into a successful human clinical trial and when not? There is no easy way to define failure, and we have so many mechanisms in place to detect success, that failure appears invisible to us, or, as Churchill would have said, “if we see it, we hurry off as if nothing has happened”.

Yet, how will we be able to truly advance, if we cannot describe when, if, and how, our experimental or clinical trials have truly failed. Put in a slightly different context, how do we reach the conclusion that our results are truly of broad relevance to the problem under study, beyond and in addition to what the outcome of the statistical analysis. Thus, how could we have a statistically meaningful difference in an experiment, and yet, consider the effects not important enough to either publish them as they are, or progress towards a clinical trial. Paraphrasing Fisher, the famous statistician, the difference between $p < 0.05$ and $p < 0.06$ should not be ‘a paper’, but an indication to continue our investigations towards, in our case, clinical effectiveness [8,9,34-38].

Herein, we would like to make the argument that the roots of most Phase III clinical trials' failures are to be found in our poor understanding of statistics, and a lack of robustness of much preclinical science. This leads to clinical trial designs which thus fail when confronted with the challenges of the human disease conditions. At the end of the article we will also suggest ways forward to develop a blueprint for a novel approach to the development of robust translational medicine, and its development into Phase III clinical trials of increased resilience to clinical challenges.

¹The related problem of potentially promising Phase I or II trials that never progress to a decisive Phase III trial constitutes a different cause for failure which will not be addressed in this article.

In Search of Preclinical Robustness

A main limitation of the basic science which constitutes the basis for future clinical trials is the lack of comprehensive understanding of which variables being examined are actually significant and/or rate limiting parameters that are relevant to the study of human disease, and predictive of novel treatments' efficacy in human patients. Although genetics has provided many animal models in which to study the consequences of individual gene mutations, it is clear that beyond the homology of genes between different species, there are many variables that are not mimicked by simply introducing the same genetic mutation responsible for a human disease into a rodent. Thus, the unpredictability of directly establishing a mouse model of disease by introducing the same mutation that occurs in human patients should be taken into account. For example, mouse models of cystic fibrosis, or muscular dystrophy, only have had minimal predictive power vis-à-vis the development of novel therapies.

Further, issues of scaling, i.e., how tumor size varies across species, or how disease time course is affected by the species being studied, remain very poorly understood, addressed, or researched. Thus, in the absence of answers to the scaling problem, it is difficult to assess how predictive it is to study the genesis or progression of human brain tumors in the much smaller mouse brain? If we were to identify how cell cycle, tumor size doubling, etc., scale from mouse to humans, it would be easier to determine the important parameters that need to be assessed for test novel therapies to be predictive of the human disease. Equally, tumors in rodents usually grow much faster than in humans; this is likely to affect the growth patterns and relationship to blood vessels and patterns of tumor neo-vascularization. Understanding the fundamental mechanisms of tumor growth, cell division, vascularization, necrosis, immune response, etc., and how this scales across to human tumors is an important area of research.

Although mathematical models of tumors are available, it appears that what is lacking are mathematical models detailed and powerful enough to be able to predict the effect of perturbations induced by pharmacological agents and designed to interfere selectively with individual tumor components, whether cellular components, i.e., blood vessels, or biochemical pathways, i.e. growth factor signaling [39-42]. Thus, for example, the effect of VEGF inhibitors on brain tumor growth has been somewhat unexpected. While there is an improvement in patients' symptomatology and MRI images, the tumors continue to grow, and it is unlikely that the VEGF antagonists tested so far; extend patients' survival. A mathematical model that could predict the response of brain tumors to these and other pharmacological manipulations would be extremely helpful in the development of novel therapies.

In translational medicine, we need to be able to report that the results obtained have a certain degree of robustness [43]. A given system is said to be robust if it is able to maintain its central functions in the face of challenges. In basic science, experimental robustness has not been explored, defined, or studied. The robustness of a new preclinical procedure should indicate its treatment efficiency, in spite of variations in the age, size, and genetic composition of individual tumors from different patients. For example, if we wish to implement a novel treatment for brain tumors, we should be able to demonstrate that this treatment is effective in brain tumors in animals of different genetic backgrounds, in tumors of varying genetic makeup, in tumors of different sizes, and in animals of different ages. The efficiency of the new treatment under the different conditions will provide us with an indication of the treatment's robustness. For example, if a particular treatment only works in C57Bl/6 mice, but fails in mice of three other genetic backgrounds, the robustness of the treatment is lower when compared to one which is effective in all four strains of mice. Therefore, it would be possible to provide the scientific community with an indication of the robustness of a given treatment, before clinical trials are initiated.

However, many potential clinical treatments are only tested in one animal model of a given genetic background, where treated and control animals are genetically identical. Based on such data, clinical trials are initiated. This is misleading as the potential of such a single model to predict the human paradigm is very low. On the other hand, if tumor treatments were to be tested in a variety of models in different genetic backgrounds, ages, species, to show whether efficiency seen in a homogenous genetic background is robust vis-à-vis genetic heterogeneity, the predictive robustness of the preclinical science underlying clinical trials would be increased [5,44].

Safety needs to be shown in at least two species, yet, efficacy appears to be tested under less stringent criteria. Equally, genetic diversity is usually not addressed specifically during preclinical toxicity or effectiveness; i.e., whether a particular treatment's effectiveness is affected by the genetic background of animal models. Usually this results in potential treatments which have many times only been tested within a restricted genetic background, quite unlike the genetic heterogeneity of human patient populations. Thus, we should be prepared, for the fact that many such treatments may fail when tested against genetically diverse and complex human populations. It is not that the treatments do not work; it is that they are not robust enough. And their lack of robustness is not a consequence of false premises, careless science, lack of controls, flawed statistics or experimental designs. It is just that treatment robustness is never tested. Thus, when challenged, as the novel treatments get tested in the clinic, they fail to show robust responses in the presence of perturbations, i.e., genetic heterogeneous patient populations.

A final item that will increase the preclinical robustness of experiments is an analysis of effect size, beyond statistical significance [34,45,46]. Currently any difference that provides a statistical difference of $p < 0.05$ will be taken to be sufficient to publish a paper and provide underlying rationale for a clinical trial. Effect size is not considered in the regular analysis of basic science experiments, and it is also unclear how much importance it should be given when proposing to move from animal experimentation to clinical trials [34,45,46]. However, effect size should contribute to our estimation of preclinical robustness. Eventually, a mathematical simulation could also be developed to calculate how robust a particular preclinical approach is, e.g., how much individual variations in disease state, genetic heterogeneity, patient age, etc., will affect the efficiency of the therapy. Needless to say, such a rational and rigorous approach to translational science could be implemented with available computer, mathematical, statistical, and genetic knowledge.

In Search of Large and Clinically Significant Therapeutic Effects

Much of the edifice of clinical trials is built on safety and a strict compliance with detailed safety rules, while clinical efficacy is only introduced at a much later stage of the clinical trials program (Phase III trials). It is understandable that past abuses have led to a thorough emphasis on safety assurance in the evaluation of clinical trials. The dilemma is that the ultimately goal of the clinical trials is treat diseases effectively and safely. However, as currently performed, therapeutic efficacy determinations are relegated to Phase III trials. Therefore, the sole emphasis for performing a Phase I and II trial is to assess the safety of the proposed therapy. Phase I and II trials usually provide a strong indication about the right dose to use, and the overall safety, they are nevertheless imprecise in predicting Phase III results. As large funds have been invested at the time a Phase III trial is initiated, the power of the statistical analysis is calculated to be able to detect relatively small differences. As p values depend on the number of events measured, many Phase III trials employ a very large number of patients [8,9].

It is known and understood that successful commercialization following approval by FDA needs a minimum of statistical significance. Trials testing literally hundreds if not thousands of patients are sometimes used to detect differences of 10 day increased survival with a new

drug added to standard of care. As such small differences require further trials, it becomes very difficult to determine the true worth of such large trials. Because of so many constraints on clinical trials and the pressures for economic success, the emphasis in the overall analysis of the trial is on the exactness of the measurements, rather than on the size of the therapeutic effect.

Why do clinical trials rarely yield large, biologically meaningful effects? Are the improvements achieved by small increments worthwhile? Has it provided the expected benefits, i.e., will small incremental benefits eventually lead to substantial clinical benefits? Are there any studies that have demonstrated this, or is it just an interesting unproven concept that allows pharmaceutical companies to devote resources into introducing small chemical changes on currently used drugs to try to carve up financially rewarding niches, rather than attempting to solve any of the big problems that remain unsolved [10,11,34].

These are difficult problems to address and even more difficult to overcome. However, alternative statistical methods, i.e. Bayesian analysis, and especially statisticians working within the Bayesian framework, have tried to challenge the dissociation of safety and efficacy in early clinical trials [8-11,47,48].

Equally important in the discussion of size effect is the problem of “pathophysiological ethics”, i.e., the fact that experimental models might test stringent stages of the disease, but also, stages of the disease where the treatment has a chance of working [49,50]. Unfortunately, and due to a misunderstanding of the power of certain treatments to work best at certain stages of disease compared to others, many times trials are restricted to treating patients with advanced disease. Although it might be easier to apply rigid ethical standards to the choice of patients to be enrolled in clinical trials (i.e., adults can give consent for a quality of life treatment, which would not be allowed to go ahead in children) this could lead to the wrong choice of patients. Any treatment should be primarily tested in those patients having a highest chance of benefitting the most, rather than in those patients who can comply with all the bureaucratic legalities. If a particular treatment would work best in young children, it is up to the investigator, the clinical trialists, the ethicists and the regulatory agencies to make this implementation possible, rather than selecting the patient who can fill out all the forms, instead of the patient who can benefit the most. In the absence of potential clinical benefit the benefit vs. risk ratio becomes indeed a very, very small number [49-52].

CONCLUSIONS: A BLUEPRINT FOR ROBUST TRANSLATIONAL MEDICINE

It is clear that more work is needed to reduce the high failure rate of translational medicine, and especially the high failure rate of Phase III clinical trials. We would like to propose that the concept of ‘preclinical robustness’ should be further developed, and that ‘robust preclinical experiments’ should be the ones that are effectively and selectively chosen to be translated into clinical trials. ‘Robust preclinical experiments’ are described ideally as those in which a new potential treatment is tested in various different experimental settings/models. Not only should such new therapeutic approaches be tested using cell lines *in vitro*, but also in a variety of *in vivo* animal models. Especially the efficacy ought to be challenged by testing the therapeutic approach in multiple animal models of varied genetic backgrounds, and experiments selected for translation ought to provide not only statistically significant differences between treated and control groups, but those differences should indeed be large, of definite clinical significance, clearly beyond a statistically calculated $p < 0.05$ [8-11,47,48].

Intuitively, any differences obtained between treated and controls, should be inversely proportional to the genetic homogeneity of the sample; i.e., in an inbred model, the differences aimed at should be higher than when using an outbred model. If experiments only provide

small differences in a genetically homogeneous background, minor differences in the genetic makeup of the animals would likely abolish those effects. Thus, as originally proposed by Fischer, small differences with a $p < 0.05$, should be taken as incentives to work to further increase the difference between control and treated animals [8,9,34–38], and concomitantly increase the robustness of the therapeutic approach, rather than as an incentive to jump blindly into clinical trials.

Specifically, what would be of much use would be a standard blueprint to move preclinical experiments into clinical trials. The safety data are not enough, since they fail to provide any information on the potential effectiveness of the novel therapy. If safety standards only get satisfied through stringent safety tests, similar rigorous standards should also be applied to the proposed efficacy. In summary, so far the early clinical trials are designed to detect potential toxicity; we propose that early trials should be designed to simultaneously target safety and treatment efficacy. If this requires us to develop novel tests, including novel statistical tools, this should be attempted. If it also requires rethinking of how we evaluate the effects in our preclinical models, and our statistical analyses, this should take priority in order to increase the efficiency of the translation of effective pre-clinical models into clinically effective novel treatments. To provide robust preclinical translational medicine and robust clinical trials is our scientific, clinical, and ethical responsibility towards our patients.

Acknowledgments

Our experimental work on novel therapies for brain tumors is supported by National Institutes of Health/National Institute of Neurological Disorders & Stroke (NIH/NINDS) Grant 1R01 NS44556.01, Minority Supplement NS44556.01; 1R21-NS054143.01; 1U01 NS052465.01, 1R03 TW006273-01; 1R01-NS 057711 to M.G.C.; NIH/NINDS Grants 1R01 NS 054193.01; R01 NS 42893.01, U54 NS045309-01 and 1R21 NS047298-01 to P.R.L. The Bram and Elaine Goldsmith and the Medallions Group Endowed Chairs in Gene Therapeutics to PRL and MGC, respectively, The Linda Tallen & David Paul Kane Foundation Annual Fellowship and the Board of Governors at CSMC.

REFERENCES

1. Garside R, Pitt M, Anderson R, et al. The effectiveness and cost-effectiveness of carmustine implants and temozolomide for the treatment of newly diagnosed high-grade glioma: a systematic review and economic evaluation. *Health Technol Assess* 2007;11:1–221.
2. Amarasingh S, Macleod MR, Whittle IR. What is the translational efficacy of chemotherapeutic drug research in neuro-oncology? A systematic review and meta-analysis of the efficacy of BCNU and CCNU in animal models of glioma. *J Neurooncol* 2009;91:117–25. [PubMed: 18813876]
3. Metcalfe SE, Grant R. Biopsy versus resection for malignant glioma. *Cochrane Database Syst Rev* 2001;CD002034. [PubMed: 11687008]
4. Hart MG, Grant R, Garside R, et al. Temozolomide for high grade glioma. *Cochrane Database Syst Rev* 2008;CD007415. [PubMed: 18843749]
5. Perel P, Roberts I, Sena E, et al. Comparison of treatment effects between animal experiments and clinical trials: systematic review. *Bmj* 2007;334:197. [PubMed: 17175568]
6. Edelstein ML, Abedi MR, Wixon J. Gene therapy clinical trials worldwide to 2007--an update. *J Gene Med* 2007;9:833–42. [PubMed: 17721874]
7. Piantadosi, S. *Clinical trials: a Methodologic perspective*. 2nd ed.. Wiley-Interscience; London, New York: 2005.
8. Goodman SN. Toward evidence-based medical statistics. 1: the P value fallacy. *Ann Intern Med* 1999;130:995–1004. [PubMed: 10383371]
9. Goodman SN. Toward evidence-based medical statistics. 2: the Bayes factor. *Ann Intern Med* 1999;130:1005–13. [PubMed: 10383350]
10. Micksad RA, Gonen M, Lynch TJ, Roberts TG Jr. Interpreting trial results in light of conflicting evidence: a Bayesian analysis of adjuvant chemotherapy for non-small-cell lung cancer. *J Clin Oncol* 2009;27:2245–52. [PubMed: 19307513]

11. Miksad RA, Schnipper L, Goldstein M. Does a statistically significant survival benefit of erlotinib plus gemcitabine for advanced pancreatic cancer translate into clinical significance and value? *J Clin Oncol* 2007;25:4506–7. author reply 8. [PubMed: 17906218]
12. Ioannidis JP. Why most published research findings are false. *PLoS Med* 2005;2:e124. [PubMed: 16060722]
13. Goodman S, Greenland S. Why most published research findings are false: problems in the analysis. *PLoS Med* 2007;4:e168. [PubMed: 17456002]
14. Goodman, S.; Greenland, S. Assessing the unreliability of the medical literature: a response to “Why most published research findings are false”. Johns Hopkins University, Department of Biostatistics; 2007. Available from: <http://www.bepress.com/jhubiostat/paper135>
15. Buckner JC. Central nervous system tumors (Discussion) - Central nervous system tumors track. ASCO Annual Meeting. 2009
16. Wen PY, Kesari S. Malignant gliomas in adults. *N Engl J Med* 2008;359:492–507. [PubMed: 18669428]
17. Grossman SA, Ye X, Piantadosi S, et al. Current survival statistics for patients with newly diagnosed glioblastoma treated with radiation and temozolomide on research studies in the United States. ASCO Annual Meeting. 2009
18. Markman JR, Markman M. Running an ethical trial 60 years after the Nuremberg Code. *Lancet Oncol* 2007;8:1139–46. [PubMed: 18054883]
19. Weindling P. The origins of informed consent: the International Scientific Commission on Medical War Crimes, and the Nuremberg code. *Bull Hist Med* 2001;75:37–71. [PubMed: 11420451]
20. Ernst E, Weindling PJ. The nuremberg medical trial: have we learned the lessons? *J Lab Clin Med* 1998;131:130–5. [PubMed: 9488495]
21. Candolfi M, Curtin JF, Nichols WS, et al. Intracranial glioblastoma models in preclinical neuro-oncology: neuropathological characterization and tumor progression. *J Neurooncol* 2007;85:133–48. [PubMed: 17874037]
22. Kitange GJ, Carlson BL, Mladek AC, et al. Evaluation of MGMT promoter methylation status and correlation with temozolomide response in orthotopic glioblastoma xenograft model. *J Neurooncol* 2009;92:23–31. [PubMed: 19011762]
23. Ashley DM, Kong FM, Bigner DD, Hale LP. Endogenous expression of transforming growth factor beta1 inhibits growth and tumorigenicity and enhances Fas-mediated apoptosis in a murine high-grade glioma model. *Cancer Res* 1998;58:302–9. [PubMed: 9443409]
24. Marumoto T, Tashiro A, Friedmann-Morvinski D, et al. Development of a novel mouse glioma model using lentiviral vectors. *Nat Med* 2009;15:110–6. [PubMed: 19122659]
25. Shih AH, Dai C, Hu X, et al. Dose-dependent effects of platelet-derived growth factor-B on glial tumorigenesis. *Cancer Res* 2004;64:4783–9. [PubMed: 15256447]
26. Assanah M, Lochhead R, Ogden A, et al. Glial progenitors in adult white matter are driven to form malignant gliomas by platelet-derived growth factor-expressing retroviruses. *J Neurosci* 2006;26:6781–90. [PubMed: 16793885]
27. Wiesner SM, Decker SA, Larson JD, et al. De novo induction of genetically engineered brain tumors in mice using plasmid DNA. *Cancer Res* 2009;69:431–9. [PubMed: 19147555]
28. Huse JT, Holland EC. Genetically engineered mouse models of brain cancer and the promise of preclinical testing. *Brain Pathol* 2009;19:132–43. [PubMed: 19076778]
29. Ghulam Muhammad AK, Candolfi M, King GD, et al. Anti-glioma immunological memory in response to conditional cytotoxic/immune-stimulatory gene therapy: humoral and cellular immunity lead to tumor regression. *Clin Cancer Res* 2009;15(19):6113–27. [PubMed: 19789315]
30. Candolfi M, Yagiz K, Foulad D, et al. Release of HMGB1 in response to proapoptotic glioma killing strategies: efficacy and neurotoxicity. *Clin Cancer Res* 2009;15:4401–14. [PubMed: 19570774]
31. Curtin J, Liu N, Candolfi M, et al. HMGB1 mediates endogenous TLR2 activation and brain tumor regression. *PLoS Med* 2009;6:e10. [PubMed: 19143470]
32. Ali S, King GD, Curtin JF, et al. Combined immunostimulation and conditional cytotoxic gene therapy provide long-term survival in a large glioma model. *Cancer Res* 2005;65:7194–204. [PubMed: 16103070]

33. Dewey RA, Morrissey G, Cowsill CM, et al. Chronic brain inflammation and persistent herpes simplex virus 1 thymidine kinase expression in survivors of syngeneic glioma treated by adenovirus-mediated gene therapy: implications for clinical trials. *Nat Med* 1999;5:1256–63. [PubMed: 10545991]
34. McCloskey, DN.; Ziliak, S. The cult of statistical significance: how the standard error costs us jobs, justice, and lives (Economics, Cognition, and Society). University of Michigan Press; Ann Arbor, MI: 2008.
35. Curran-Everett D. Explorations in statistics: hypothesis tests and P values. *Adv Physiol Educ* 2009;33:81–6. [PubMed: 19509391]
36. Curran-Everett D. Explorations in statistics: confidence intervals. *Adv Physiol Educ* 2009;33:87–90. [PubMed: 19509392]
37. Fisher, RA. Statistical methods for research workers. Oliver and Boyd; London: 1925.
38. Fisher RA. The arrangement of field experiments. *J Minist Agric* 1926;GB 33:503–13.
39. Bearer EL, Lowengrub JS, Frieboes HB, et al. Multiparameter computational modeling of tumor invasion. *Cancer Res* 2009;69:4493–501. [PubMed: 19366801]
40. Macklin P, McDougall S, Anderson AR, et al. Multiscale modelling and nonlinear simulation of vascular tumour growth. *J Math Biol* 2009;58:765–98. [PubMed: 18781303]
41. Harpold HL, Alvord EC Jr. Swanson KR. The evolution of mathematical modeling of glioma proliferation and invasion. *J Neuropathol Exp Neurol* 2007;66:1–9. [PubMed: 17204931]
42. Szeto MD, Chakraborty G, Hadley J, et al. Quantitative metrics of net proliferation and invasion link biological aggressiveness assessed by MRI with hypoxia assessed by FMISO-PET in newly diagnosed glioblastomas. *Cancer Res* 2009;69:4502–9. [PubMed: 19366800]
43. Carlson JM, Doyle J. Complexity and robustness. *Proc Natl Acad Sci USA* 2002;99(Suppl 1):2538–45. [PubMed: 11875207]
44. Sandercock P, Roberts I. Systematic reviews of animal experiments. *Lancet* 2002;360:586. [PubMed: 12241927]
45. Nakagawa S, Cuthill IC. Effect size, confidence interval and statistical significance: a practical guide for biologists. *Biol Rev Camb Philos Soc* 2007;82:591–605. [PubMed: 17944619]
46. Kraemer HC, Kupfer DJ. Size of treatment effects and their importance to clinical research and practice. *Biol Psychiatry* 2006;59:990–6. [PubMed: 16368078]
47. Diamond GA, Kaul S. Bayesian classification of clinical practice guidelines. *Arch Intern Med* 2009;169:1431–5. [PubMed: 19667308]
48. Goodman SN. Building a Bayesian bridge from evidence to guidelines. *Arch Intern Med* 2009;169:1436–7. [PubMed: 19667309]
49. Lowenstein PR. A call for physiopathological ethics. *Mol Ther* 2008;16:1771–2. [PubMed: 18956015]
50. Lowenstein PR. Clinical trials in gene therapy: ethics of informed consent and the future of experimental medicine. *Curr Opin Mol Ther* 2008;10:428–30. [PubMed: 18830917]
51. Kimmelman J. The ethics of human gene transfer. *Nat Rev Genet* 2008;9:239–44. [PubMed: 18278058]
52. Kimmelman J. Ethics of cancer gene transfer clinical research. *Methods Mol Biol* 2009;542:423–45. [PubMed: 19565916]