SEVEN ESSENTIALS FOR SUCCESSFUL DRUG DEVELOPMENT

Jack G. Modell, M.D. and John H. Greist, M.D.

People need to be reminded more often than they need to be instructed.

– Samuel Johnson

In our many decades of combined experience in clinical medicine, academic research, and development of drugs and medical applications of information technology, we have had the privilege of being involved with dozens of drug and device trials and clinical development programs. Many were successful but too many failed, even when a drug or device likely had a favorable effect on disease outcome.

In reflecting on the differences between successful and unsuccessful clinical trials or programs, several factors consistently emerge as essential components of success and failure. Although all are largely within control of the developers, many are often inadequately considered or overlooked. We recognize that the factors involved in decision making may not be as clear cut as presented for the purpose of this article, and there is no doubt that many companies work diligently to design and conduct only the most scientifically sound clinical studies. Our goal herein is not to criticize any company or individual, but rather to remind us of the importance of these essentials so that effective therapies have a greater likelihood of reaching patients in need.

1. PRODUCT IMPORTANCE. *Develop a product that truly meets a therapeutic need.*

While this first essential may seem obvious, enthusiasm around developing a new product is sometimes so strong that companies can unintentionally deceive themselves about the true value of a new product, believing that small differences in their product over existing products will provide a superior choice for patients, even without the necessary evidence from head-to-head trials. Regulatory and reimbursement requirements for such products are usually higher than for products with demonstrable safety or efficacy advantages over existing products. Failure to gain approval or to reach expected sales is high. The "opportunity cost" of such development programs – millions of dollars, thousands of person-hours, and exposure of human subjects to clinical research that is unlikely to add substantially to medical or scientific knowledge – also detracts from resource allocation to products more likely to be approved and to improve quality of life or decrease suffering.

2. **KNOW WHEN TO QUIT.** ***Terminate a clinical development program quickly when data show a risk-benefit ratio unlikely to be favorable.***

Even when a product in development might genuinely benefit humanity were it to meet its expected potential, we must accept that efficacy and/or safety often turn out to be less than expected. When this is discovered or acknowledged only late in development, particularly at or near the time of regulatory approval, the news and financial consequences can be devastating. Too often, "discovery" of an unfavorable risk-benefit profile at a late stage of development was avoidable had product limitations been accepted as soon as they became evident. Evidence for an unfavorable risk-benefit ratio often occurs early in development when, for example, drug-placebo separation is marginal or occurs only for a subset of disease endpoints on post-hoc analysis, or adverse events or tolerability make compliance difficult or use inadvisable. Despite such mounting evidence, product developers may fail to understand the significance of these findings and soldier on – sometimes for years.

The most effective strategy to mitigate this risk is to develop a set of unambiguous "go/no-go" criteria at the start of product development, against which accumulating study data are assessed at all major milestones to determine whether the product remains likely to meet the required characteristics for successful registration and commercialization. Should predetermined "go/no-go" criteria fail to be met, it is usually best to terminate the particular development program. Tremendous amounts of time and money have been lost in programs that never bore fruit, thus depriving the sponsor of resources for more productive pursuits.

3. **STEP BY STEP.** ***Design each study to provide meaningful information to guide the next step(s) of product development.***

This complex problem involves many aspects of study design: incorporation of an appropriate and limited number of study endpoints; inclusion of a relevant comparison arm; inclusion/exclusion criteria that adequately and appropriately characterize the necessary study population; understanding the limitations of investigator judgment for subject selection; and proper study powering. Every study should be designed with one primary purpose: to provide in the most efficient manner possible, the information necessary to inform the next step in clinical development, including whether to proceed at all. By analogy, when crossing a river in a rowboat on a lengthy survival journey, if the boat is not adequately provisioned to enable the next phase of the journey, the crossing will be in vain. If, however, the boat is overburdened with unnecessary supplies, it may sink and the travelers must swim back to shore and start anew. Protocols with an excessive number of "nice-to-have" endpoints, procedures, or measurements "because we're doing the study anyway" often sink under their own weight, stalling or killing a development program.

Open-label or uncontrolled studies almost always show some treatment benefit, often of a magnitude that is seen as "unexpected" were the therapy to be truly placebo-like in efficacy. While the desire to conduct an open-label study "just to estimate the magnitude of possible treatment effect" is understandable, the risk that the outcome, even if robust, will fail to accurately predict the eventual success of the product in placebo- or active-control studies required for registration is very high. For this reason, we strongly recommend against the use of uncontrolled studies in early development programs particularly for efficacy, but also for safety where adverse events that occur with the investigational product can inappropriately taint the product with potential safety concerns that might also have occurred on placebo.

Hastily written inclusion and exclusion criteria that rely excessively on investigator judgment often allow subjects into the trial for whom the product is less likely to be safe or efficacious, as well as making it more difficult to replicate results across trials. For each and every inclusion and exclusion criterion written, the study team should justify why that criterion is necessary to define the study population. Inclusion and exclusion criteria should be written with careful attention to their desirability, necessity, and potential unintended consequences. Ambiguous, duplicative, overlapping, and potentially contradictory criteria should be avoided. For safety criteria, it is important to recognize the risk of leaving important assessments to "investigator judgment" alone as investigators vary considerably in this regard. If the presence of a certain risk factor is important, specific diagnoses and/or laboratory cutoff values should be employed.

Study powering is also often misunderstood. We frequently see product developers who "power" a study based on unrealistic estimates of product benefit. Unrealistic power estimates may simply be overly optimistic thinking, but may also result from unquestioned reliance on results from a previously conducted "successful" study. Because of publication bias and greater drug-placebo separations often seen in smaller or less stringently conducted trials than which occur in regulated programs with many sites and investigators, reliance on such studies for powering runs a high risk of unintentional underpowering. On the other hand, studies are too often intentionally underpowered, "just to see if we have a signal and how much that might be." While an underpowered study can give an estimate of possible signal size, the confidence one can have in that estimate falls rapidly with dwindling sample sizes; and as that confidence decreases, so does the likelihood that the estimate will accurately predict effects seen in subsequent studies. Put simply, if a trial is worth doing it is worth doing well. This includes appropriate selection of study endpoints, carefully selected subjects, quality sites, and an adequate sample size.

4. **USE EXPERT INPUT** ***adequately and appropriately.***

"But the key opinion leaders (KOLs) told us this!" We should not forget that valuable as expert opinion is to our development programs, the "O" in "KOL" stands for "opinion," not fact. While many KOLs give carefully considered excellent advice, some give advice that they believe whoever hired them wants to hear (and some companies may have a complementary proclivity to discount disagreeable KOL advice), or to give advice about which they may feel passionately but that ultimately proves inaccurate. But no matter how honorable or knowledgeable the KOL may be, relatively few have broad or deep enough experience in clinical care, clinical development, research methodology, regulatory requirements, and market access considerations to be able to give advice that is unequivocally beneficial for product development. A sponsor should seek and consider expert opinion from consultants with differing backgrounds and without known biases or conflicts of interest, and take care not to overvalue these opinions. Conversely, failure to seek expert input based on belief that the development team already knows what is best can also seriously jeopardize a product development program.

One other consideration with use of KOLs bears mention – that of acceding to the desires of the KOL primarily to please him or her. A common example of this includes using a KOL as an investigator despite that person not randomizing many subjects or producing high-quality data. While diplomacy is important, it is unwise to compromise any aspect of study quality by saying "yes" to a consultant – "key opinion leader" or otherwise – simply for fear of upsetting that person. Most KOLs will understand that not everyone can agree with them or give them what they might want; and for those who do not, we do our development programs a far greater disservice by keeping them onboard than by respectfully declining to accept their input or grant their wishes.

5. **WHAT THE HEART LOVES BECOMES TRUTH.** ***Avoid false assumptions about trial requirements based on what worked or didn't work in the past.***

In designing our clinical trials, we must take care not to fall prey to the common error in logic, "post hoc ergo propter hoc" (after this, therefore because of this). Previous experience should be considered carefully in designing new studies, but we often see sponsors insist on a study design that is heavily driven by something perceived in earlier trials, positively or negatively, without strong evidence, to have been critical in the observed outcome. Many studies designed on this basis fail because of unanticipated consequences of a newly introduced factor. Some of these studies cannot be completed successfully because inclusion/exclusion criteria unnecessarily limit the study population. These studies may also run into regulatory troubles because the study population or design was constrained or redefined in ways that reduce applicability of study results to the target patient population. While careful consideration and incorporation of information from previous clinical trials is essential, there should be clear and objective evidence that this information is relevant and important to the current trial design, and possible unintended consequences of introducing protocol components based on this information must be considered.

6. **WHEN ALL ELSE FAILS, ASK THE PATIENT.** ***Recognize limitations of rater objectivity.***

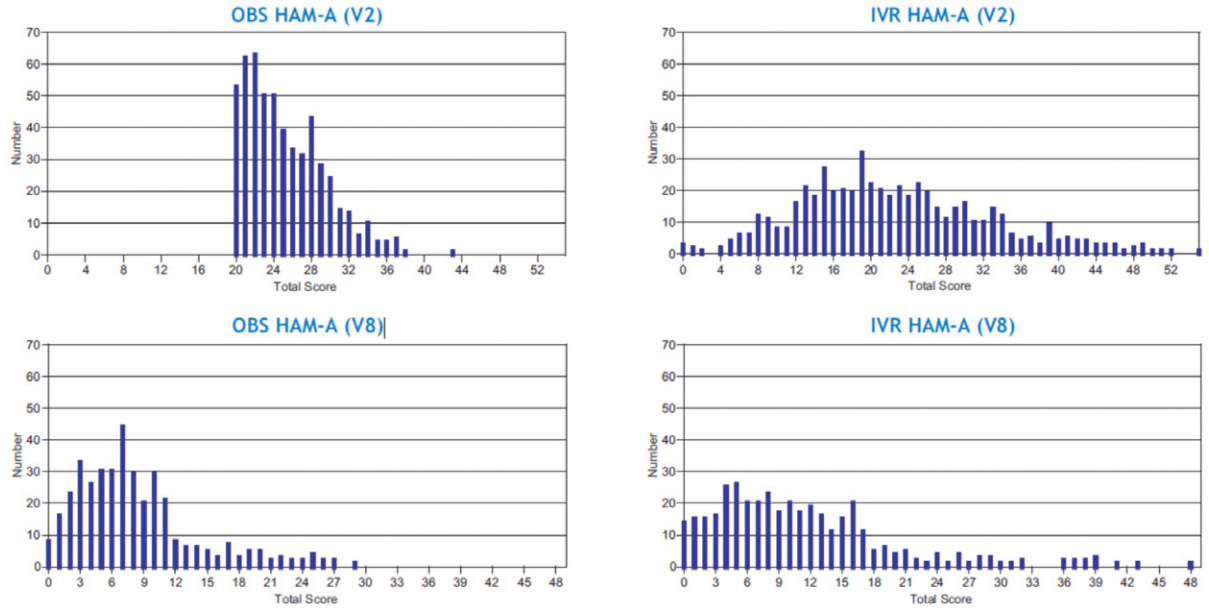
Consciously or unconsciously, clinical raters may inflate or deflate clinical ratings to enable subjects to gain entry or

remain in a clinical trial.[1,2] Subtly coaching subjects on their answers or rounding up or down when subject responses or findings seem to fall between scale severity ratings may allow entry or retention of unqualified subjects in the trial. Additionally, subjects and investigators, expecting that the onset of treatment should coincide with at least some clinical improvement, may bias early ratings to reflect this expectation even though signs and symptoms of the illness show no true change. While this early "improvement" in rating scores for subjects in clinical trials may appear to be a placebo effect and is frequently confused with it, the apparent improvement is often the result of artificially inflated or deflated rating scores regressing back toward original true (and more normally distributed) values, in combination with whatever actual treatment and placebo effects may have occurred.[3,4]

An excellent example of rater bias and unreliability observed in an FDA registration trial is informative.[2] In a study of pregabalin in generalized anxiety disorder (GAD), both score inflation and deflation occurred with the same clinician raters during the course of the trial compared with computer ratings. Clinician and computer (interactive voice response; IVR) ratings of the Hamilton Anxiety Rating Scale (HAM-A) were made at Week 2 (the open-label baseline) and Week 8 (after open-label pregabalin treatment), but only the clinician-rated scores determined subject selection. As seen in the figure, the clinician raters, knowing that a minimum HAM-A score of 20 was required at Visit 2 for study participation, rated all subjects as having a HAM-A score of 20 or greater (OBS HAM-A (V2)), while the computer independently assessed a sizable proportion of this same population as having a non-qualifying score of < 20 (IVR HAM-A (V2)). At Week 8, however, when a HAM-A of ≤ 11 was required for continuation into a 24-week double-blind pregabalin vs. placebo relapse-prevention phase, clinician-rated HAM-A scores were skewed towards lower scores (OBS HAM-A (V8)) compared with independently assessed computer ratings (IVR HAM-A (V8)). During the subsequent 24-week double-blind treatment phase, all HAM-A ratings were made by clinicians, yet differences in relapse rates between active drug and placebo were always numerically larger using the IVR HAM-A qualifying scores.

Rater unblinding by observed or reported side effects or by unintended revelation of treatment assignment further undermines outcome accuracy.[5] Inclusion of non-qualified subjects and rater bias hampers detection of actual drug-placebo differences throughout the study. Despite a common misconception, this problem cannot be mitigated by "increasing power by increasing the

FIGURE



Clinician (OBS, observer) - versus computer (IVR)-rated HAM-A scores for study qualification at Visit 2 (V2) and Visit 8 (V8);see body for details.[2] Reproduced with permission from author.

number of subjects" because lower effect sizes can result from pressure to enroll more but less qualified subjects and to include more study sites, thus negating original power assumptions and leaving studies short of statistical significance despite larger sample size.

Part of the reason behind rater bias is that investigators and site staff often do not fully understand the true objective of the clinical trial: it should not, for example, be "to show treatment efficacy" or to show that a product is "safe and well tolerated," but rather, to test the null hypothesis of no treatment difference or to estimate likely treatment effect, as well as to objectively assess and record all adverse effects that may emerge during treatment. Thus, investigators and site staff must fully understand the importance of complete objectivity and consistency in performing clinical ratings, the rationale for each inclusion and exclusion criterion, and the deleterious effect that even well-intended efforts to include subjects who are not fully qualified can have on the outcome and scientific integrity of the trial.

While thorough and meticulous investigator and rater training can be helpful in mitigating some of these problems, humans simply cannot match computers in objectivity and consistency in making assessments based on subject responses to questions in clinical trials. Unless programmed to do so, a computer cannot coach a subject how to respond, nor would it inflate or deflate ratings based on feelings, expectations, response interpretations, or desired outcomes. A computer faithfully asks the same questions every time, following the same algorithm, and records responses exactly as provided by the subject. Several studies have shown that computerized assessments (electronic patient-reported outcomes, [ePRO]) of entry criteria and outcome measures in clinical trials provide data quality for signal detection that exceeds that obtained by human raters.[4,6-8] In a careful study of variability in conduct of Hamilton Depression Scale ratings, for example, 92% of variability was attributable to clinical raters, and 8% to patients.[9] Another study confirmed loss of clinical rater reliability after rigorous training during a year of rating in registration trials.[10]

For these reasons, strong consideration should be given to using ePRO systems for assessing study entry criteria and endpoints,[11] particularly for endpoints that "[measure] a concept best known by the patient or best measured from the patient perspective."[12] The use of computer-interview ePRO assessments in clinical research can increase measurement precision and objectivity, while freeing investigators to manage the overall safe and effective conduct of the clinical study. This advance in clinical research has a parallel in modern aviation, where computerized flight systems and pilot checklist standardization have greatly enhanced the capabilities and safety of modern commercial aircraft by automating tasks that require integration of information, precision, and invariability that unaided humans simply cannot match.

7. **IN CLINICAL STUDY REPORTS,** ***regulatory submissions, and regulatory interactions, we must tell the truth, the whole truth, and nothing but the truth.***

The FDA is not our enemy. Their goal is not to create unnecessary hurdles to product approval. FDA is responsible for protecting public health by ensuring the safety, efficacy, and security of human and veterinary drugs, biological products and medical devices. This objective means that FDA is tasked with the difficult job of thoroughly and objectively assessing all preclinical and clinical product data presented to them. This sometimes includes separating fact from aspirational content to determine whether the overall risk-benefit of the product is favorable, the proposed product label fully and accurately provides information necessary for prescription in the intended patient population, and whether the product is likely to be used as labeled in clinical practice. Although few medications or devices have unequivocal and robust efficacy along with few or no risks, clinical study reports and documents submitted for regulatory review often suggest an inappropriately favorable risk-benefit in the way data and arguments are presented within the documents or in face-to-face meetings. FDA reviewers are and should be put off by product portrayals that are not entirely factual, complete, and balanced. Impatient frustration is the best outcome; without complete transparency on the sponsor's part, the opportunity for a productive relationship with FDA is compromised and failure to approve a marketing application becomes more likely. Non-approval outcomes determined by FDA are neither vindictive nor reflect reviewer opposition; rather, they are an appropriate response of wanting to do what is best for patients but not being given the objective information to do so.

Working with FDA regulators as allies in development programs is essential. Listen to their input and incorporate it into the development program unless there is a compelling medical or scientific reason not to. In that case, the reason(s) for the disagreement should be respectfully presented and include supporting rationale with data or precedent, the goal being a mutually agreeable path forward. FDA welcomes science-based interactions with sponsors in efforts to implement robust and efficient product development programs throughout the development process. Important components of any communication with the FDA include thorough and accurate descriptions of demonstrated product benefits, observed and potential risks and how serious risks will be identified and mitigated, why the risk-benefit for the intended patient population is favorable, and any questions that remain unanswered with a clear description of how the sponsor will address outstanding questions. Anything short of providing a complete, balanced, and fully objective picture of all data from the development program only hinders chances of a successful program outcome and ultimate product approval.

The path to product development is complex, time consuming and expensive; but with careful attention to these essentials, chances for a successful outcome that will bring new and needed products to patients are greatly increased. If there is one overriding lesson inherent in all these considerations it is that the "First Law of Disney" – "Wishing will make it true" – has its place only in The Magic Kingdom. Whatever our aspirations for a product and its inherent potential, success is proportional to objectivity and transparency in clinical trial design, conduct, analysis, reporting, and disclosure to regulatory agencies.

REFERENCES

1. DeBrota DJ, Demitrack MA, Landin R, et al. A comparison between interactive voice response system-administered HAM-D and clinician administered HAM-D in patients with major depressive episode. Paper presented at: 39th Annual Meeting of the NIMH New Clinical Drug Evaluation Unit; June 1–4, 1999; Boca Raton, FL.
2. Feltner DE, Kobak KA, Crockatt J, Haber H, Kavoussi R, Pande A, Greist JH. Interactive Voice Response (IVR) for Patient Screening of Anxiety in a Clinical Drug Trial. NIMH New Clinical Drug Evaluation Unit, 41st Annual Meeting, 2001, Phoenix, AZ.
3. Greist JH, Mundt JC, Kobak K. Factors contributing to failed trials of new agents: can technology prevent some problems. J Clin Psychiatry 2002;63[suppl 2]:8-13.
4. Mundt JC, Greist JH, Jefferson JW, Katzelnick DJ, DeBrota DJ, Chappell PB, Modell JG. Is it easier to find what you are looking for it you think you know what it looks like? J Clin Psychopharm 2007;27:121-125.
5. Marcus SM, Gorman JM, Tu X, Gibbons RD, Barlow DH, Woods SW, Katharine Shear M. Rater bias in a blinded randomized placebo-controlled psychiatry trial. Stat Med 2006;25:2762-70.
6. Kobak KA, Kane JM, Thase ME, Nierenberg AA. Why do clinical trials fail. The problem of measurement error in clinical trials: time to test new paradigms? J Clin Psychopharmacol 2007;27:1-5.
7. Greist J, Mundt J, Jefferson J, Katzelnick D. Comments on "Why Do Clinical Trials Fail?" The problem of measurement error in clinical trials: time to test new paradigms? J Clin Psychopharmacol 2007;27:535-536.
8. Moore HK, Mundt JC, Modell JG, Rodrigues HE, DeBrota DJ, Jefferson JJ, Greist JH. An Examination of 26,168 Hamilton Depression Rating Scale Scores Administered via Interactive Voice Response (IVR) Across 17 Randomized Clinical Trials. J Clin Psychopharmacol 2006;26:321-324.
9. Kobak KA, Brown B, Sharp I, Levy-Mack H, Wells K, Okum F, Williams JBW. Sources of unreliability in depression ratings. J Clin Psychopharmacol 2009;29:82-85.
10. Kobak KA, Lipsitz J, Williams JBW, et. al. Are the effects of rater training sustainable? Results from a multicenter clinical trial. J Clin Psychopharmacol 2007;27:534-535.
11. Marder SR, Laughren T, Romano SJ. Why Are Innovative Drugs Failing in Phase III? Am J Psychiatry 2017;174:829-831.
12. Guidance for Industry. Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims. U.S. Department of Health and Human Services Food and Drug Administration Center for Drug Evaluation and Research (CDER), Center for Biologics Evaluation and Research (CBER), Center for Devices and Radiological Health (CDRH). December 2009, page 2.