

## Bench to Bedside with Quality-Speed

Many new drugs are the only effective treatments for some serious illnesses or unusual patients. The waiting patients may not survive to receive them. Halving the deaths of patients-in-waiting while enhancing the safety of newly-marketed drugs can be accomplished with strategies that are already tested and available. Ten principles encompass these re-engineering strategies.

### I. Decisions not publications

Therapeutics research is a series of decisions: How perfect a molecule is needed? Which species for preclinical? When ready for the first human? What pivotal trial doses? The quality and timing of each decision may be enhanced. Decision-makers should have profound knowledge of the project and their fields, assisted by informatics tools linking them with all pertinent public information globally, and with all pertinent developments internally. They should be diverse in education, role, age, gender, cultural heritage, and positional authority, as well as dedicated to this project with the authority of their managers. On-line decision discussions are optimal with seven active participants, although others may be *ad hoc* consultants or observers. Videoteleconferencing makes scheduling flexible and adjusts to urgency. Off-line bulletin boards may suffice, certainly for information and education. Co-located empowered teams focused on quality-speed are beautiful to behold. Powerful informatics tools should facilitate each

minute-data mining, knowledge generation, global networking, paperless, real-time, just-in-time. Digital clocks count down to the next milestone. Meetings are preceded by a Delphi questionnaire asking each participant to vote in advance on key questions, and the agenda is prepared in order of the contentiousness of these questions. In the meeting, anonymous commenting/brainstorming and anonymous voting on those questions guides the moderator as convergence is seen by all.

How different this is from publishing your first JCI paper when you replicated every experiment hundreds of times, controlling everything. To refine the probability distribution of the half-time of a new molecule in humans, how many mice need to be studied? What does the second inbred mouse of the same age, gender, diet, and activity add to the first? Heresy! The JCI needs hundreds of replicates with tiny standard error bars. The next decision may be facilitated by one mouse, not many. If the first mouse has enormous presystemic elimination or no absorption, do you need a second identical inbred mouse? You will never see a pharmacokinetics paper published on the basis of one animal from each of ten species, but would that be a paradigm that might optimize prediction of human kinetics? Would observation of one primate, *ex vivo* human microsomal metabolism, and *ex vivo* human blood element association provide clues to estimate human kinetics sufficient for the first human dose? Of new molecules tested in humans, 90% will never be marketed.

Do you really want a large definitive study of their kinetics in many members of many species? When you know you have a drug that is likely to be marketed, then you can return to perfect the publications.

### II. No perfect molecules!

Most pivotal new drugs begin with the conjunction of a biological innovation with chemical diversity, wedded through efficient screening. *In silico* simulation of this conjunction may provide a diversity index and permit more efficient testing of a cassette of molecules, or brute force testing of combinatorial chemistry batches may find something that “works.” But, there is always an improvement to be made in activity, specificity, kinetics, purity and manufacturing elegance. When do you stop fiddling with the molecule’s topiary?

Impotence is expensive. If you predict human half-time to be six hours, distribution to be uniform in body water, twice-daily dosing to be sufficient, and the minimum concentration in plasma to exceed at all times to be one millimolar of a drug with a molecular weight of 1000, then a just-sufficient dose in a 70 kg human will be 92 kg/year-expensive, toxic and awkward. Obviously micromolar (126 mg/dose) or nanomolar (transdermal, inhalation, etc.) are better levels of activity. Compare the probability distribution of your beliefs about activities that can be discovered with further molecular manipulation with the probability distribution of the time

to make those discoveries and make an informed patient-oriented decision about the level of perfection appropriate for the first-generation proof-of-principle molecule. As it is developed, continued molecular perfection can be employed to find a backup or to confirm the doodlings of the medicinal chemists.

Miniaturization is key. Molecules are expensive to synthesize, purify, analyze and test. Isolated receptors or cell organelles, *ex vivo* human microsomal enzymes, *ex vivo* human blood elements, CACO2 cell absorption, and screening association with >125 miscellaneous receptors all can be tested with a few milligrams of drug. A one millimolar concentration in a 20 gram mouse with the kinetics noted above requires an absorbed dose of 12 mg. A rhesus monkey requires a single dose of three gm. A two-year study of 100 rats and 100 mice given 100 times the human dose would require at least 145 kg. Minimizing the number of animals, duration of exposure, and the dose may be essential to an affordable preclinical plan.

Preclinical studies should be a braid of chemistry, pharmacology, toxicology, kinetics, and formulation studies. Impure but well characterized material may be suitable at the beginning. A few blood samples may define kinetics sufficiently early. The formulation does not have to be elegant. Assays may be assisted by use of deuterated materials. Too often, one component demands that another complete all studies with the QA seal of approval before it starts planning its own contributions. That is like the eighth oarsman waiting for the ripples to disperse before deciding on his next stroke.

### III. Safe first human dose

What is the acute toxicology in appropriate animals? How quickly will I eliminate the dose? Does it concentrate in any critical tissues? How will my kinetics differ from the species used in preclinical toxicology? Those are the questions the first human should

ask. It doesn't matter what the subchronic and chronic toxicity might be; the first human gets one small noncumulating dose. Human metabolism can be predicted from *ex vivo* human microsomal enzymes. Human distribution can be estimated in part from association with human blood elements. Renal excretion in humans is similar to that in other species. Use deuterated drug or other isotopes. Build a model *in silico*. Examine the probability distribution of renal excretion, distribution, etc. What is the plasma concentration associated with no effect, good effect, bad effect? What plasma concentrations will you seek in that first human dose?

Regulators, ethical review committees, and others now accept for many molecules acute toxicology in two relevant species preceding one-day dosing of informed consenting humans. Begin the day with a slow intravenous infusion of a deuterated dose, monitoring plasma concentrations and effects. If a problem arises, the infusion can be stopped immediately. This is safer than oral. From the preliminary kinetics plan, an oral unlabeled dose at the 23<sup>rd</sup> hour. You get absolute bioavailability, kinetics, and metabolism from one human. Do you need a second human? If you were testing three molecules from the structure activity relationship, why not select based on one or two humans?

If you test a second human, should that one resemble the first? Why not just test young, healthy, Caucasian, athletic, nonobese, nonsmoking men on controlled diets and activity (who are the least likely ever to use your drug)? Why not test conjoined Siamese twins? Try an obese older unfit couch potato for the second human. If the two are similar in kinetics and dynamics, terrific. If they are not, you've learned something valuable. If you can safely use patients, especially if there is a surrogate acute marker of the disease, so much the better.

The first single doses in the first human(s) serve to choose the doses for multiple dose studies. As soon as the probability distribution of the first dose for multiple dosing is sufficiently

narrow, stop single dosing and begin multiple dosing. As soon as you are confident of the kinetics on multiple dosing, to ensure there is no unexpected accumulation of drug or its metabolites, introduce variables on some days of multiple dosing. What will be the effect of suppressing stomach acid, or giving divalent cationic antacids, or eliciting diarrhea with hypertonic mannitol, or feeding a variety of meals, or giving activated charcoal? One or two patients with each intervention will not make a publication, but it will help plan the pivotal study. When the two-year-old child of a subsequent patient ingests a week's worth of drug, don't you want to know if charcoal diminishes absorption?

Where should first-human studies be done? There should be a rigorous intensive IRB examination, perhaps remote, even if the study is not in the US and not under an IND and a local ethical review committee as well. There should be facilities for performing preliminary kinetic analyses, using MS, within minutes of phlebotomy. There should be the equivalent of a hospital to provide totipotential emergency care if needed.

Some countries permit first-human studies with minimal government preapproval, especially in healthy subjects. After the preclinical scientists sign off on a 100-page clinical investigation brochure for ethical review, it requires about two months to compile and QA the IND document, and then FDA imposes a 30-day waiting period (that might be waived). Worst of all, the FDA may telephone at 4:55 pm on the 30<sup>th</sup> day to announce that a hold has been applied but it will take months to type the letter explain the hold. It is more expeditious to do the first-human studies outside the US and then file an IND on the basis of human data rather than mouse data.

Patients are good first-human subjects unless they are so frail that they are at unreasonable risk. They may promise entré to subsequent efficacy trials. They may have surrogate or real efficacy variables that can be assessed. Are elderly patients with age-associated memory impairment

“healthy?” Would cognitive testing provide an important dynamic variable? Would glucose clamp studies in diabetics be a reasonable first-dose variable?

First-human multiple dose studies should continue until there is a reasonable estimation of the doses to be employed in pivotal studies. These should encompass the best dose and include a distinctly less-than-best smaller dose and a large enough dose that its advantages and disadvantages compared to smaller doses can be assessed. As soon as these doses can be reasonably estimated, pivotal studies can begin. This may require four subjects, or very many subjects. In testing for activity in a dozen or more cancers, traditionally up to 30 patients with each different cancer are given doses before abandoning the new drug.

#### IV. Global pivotal study

A single dose-response protocol pursued worldwide with many investigators and diverse patients is optimal. For example, 200 investigators in 24 countries treating 3000 patients, at least half on “appropriate” doses, is a good model. If a regulator demands two studies, one protocol can be divided statistically. The use of many countries is important if the product will be marketed in them. One would not market in Asia without testing Asian patients, whether in their home countries or expatriates. Similarly, one would wish to test all appropriate ages, genders, disease severity and duration, etc.

The Newtonian hypothesis-testing model is to test a single independent variable, such as dose of drug, while minimizing the effects of other variables through randomizing patients to different treatments. Unfortunately, randomizing may, by chance, allocate drug to the sickest patients and placebo to the least diseased. You would not be happy to find that all the old women patients had been randomly given drug and the young men given placebo. Assist randomizing by *prehoc*

stratification on **all** strata that might influence the outcome. Now that each container of doses is uniquely labeled and investigators can call into a central allocation computer, it is easy to stratify on age, gender, race, country, investigator, disease severity, disease duration, expected response, balancing random allocation of therapies among the strata. It is not required to analyze these strata in regard to outcomes, but *prehoc* stratification allows such analysis. If the strata have equivalent outcomes, pool them. If they do not, you have a valuable signal.

Should you try to limit patients to those with the greatest likelihood of a good response and no adverse effects? Should there be many exclusion and inclusion criteria? If the goal is to get  $P < .05$ , that might be a good strategy, but if you wish to market globally with a broad label, who will be testing those unusual patients? Better they are tested in a rigorous trial with good investigators than tested *ad hoc* by unsuspecting practitioners. Do not exclude the elderly. Extend downward in age as soon as safety is better defined. Ensure that patients have the target disease, but otherwise exclude only a few. If there is a special population that responds best, perhaps young women who have severe disease manifest for less than six months, *prehoc* stratification can identify and permit valid testing of such cohorts while the broader patient population can provide valued safety data and clues for treating the unusual patient.

Going directly from a few first-human subjects to 3000 patients in 24 countries requires three logistical elements. First, all data must be captured quickly, validated immediately, and transmitted promptly to a single skilled safety monitor, preferably by the World Wide Web. Core laboratories, probably one per continent, should report outlying data within 48 hours of sampling. Even less time should be required for reporting a serious unexpected adverse event.

Second, the doses chosen may not be optimal and there should be a strategy for extinguishing doses that are too small, and perhaps adding higher doses

when the largest chosen is proven safe but less effective than desired. This can be done *ad hoc* at an interim analysis, the adjustments can be made blindly following algorithms built into the protocol, or the doses allocated can be adjusted continuously by adaptive allocation. This third strategy will work if there is an efficacy variable that can be assessed early in recruiting patients. I prefer the Goldilocks variable—the patient and the investigator independently call the central data computer and report whether the dose in that patient is *too great*, *too small*, or *just right*. The allocation computer then adjusts the probability of assigning each dose to the next patient, increasing the probability of assigning a highly favored dose, and diminishing gradually, but not completely, the probability of assigning a less favored dose.

Third, interim analyses must quickly correct defects in the protocol. **Safety** must be monitored continually with stopping rules built into the protocol for predictable safety events such as death. Will you stop for safety if there are five bonafide deaths on drug and none on placebo? Six vs. one? Seven vs. two? The time to set these limits is before the trial begins, and then when the serious events are validated the computer can blindly apply the rules and raise a flag when a rule is true.

For some life-saving unique therapies it may be appropriate to build in **efficacy** rules that would stop early for outstanding efficacy, but even the best efficacy must also allow for establishing safety sufficient for approval for marketing. These analyses are infrequent.

**Power** must be calculated to determine when to stop the study and analyze results. It is estimated in advance from estimates of the mean difference and the variance. These estimates can be refined from data in the trial. Consider the consequences of premature unmasking when  $P = .06$ , versus unnecessarily extending a trial to  $P = .01$ . Power adjustments that should not require an alpha penalty may be done blindly.

**Futility** must be avoided by assuring

the study will stop as soon as it is clear that it won't work. A blinded analysis can be done frequently with the computer raising a flag only when futility must be considered. Because no efficacy can be declared, a futility analysis does not require an alpha adjustment.

If doses have been chosen properly, the pivotal study replaces a preliminary Phase II study. If doses are improper or other problems arise, the pivotal study can be stopped early with the result that it now constitutes a Phase II study from which better dose selection leads back to another pivotal study.

Special populations of patients such as those with kidney, liver, and heart dysfunction may be studied within the pivotal trial, but more usually the impact of organ dysfunction on kinetics and dynamics is assessed in smaller, more focused studies. Interaction of the new drug with established prescription and over-the-counter drugs, herbals, and diets may also be assessed in short, small studies.

The critical element in completing the pivotal study is rapid enrollment of appropriate patients. This requires a large pool of investigators, intensive education of potential patients and their physicians, prequalification by phone or web, and referral to convenient sites. HMOs would be optimal sites, as all investigators could be trained in advance and they could select appropriate patients, all of whom could be scheduled to enroll on the first day. To familiarize investigators with the protocol, one may initiate a placebo-treatment initial qualification period to be followed by the actual protocol in qualifying patients.

## V. Data capture from patients

History, symptoms, adverse events, cognitive function, mood, psychiatric scales and quality of life are all derived from patients who surf the web and play palmtop computer games. Use computer-assisted instruction techniques of context-sensitive on-line help, graphical data entry, videos,

rewards, reinforcement, recursive validation, etc. to capture data more reliably. Recently, paper diaries have been shown to be fudged most of the time, while palmtop diaries are accurate. All data can be transmitted immediately *via* the web to the central data computer. Investigator-acquired signs and interpretations should be validated and transmitted immediately. We watch wars on-line as the bombs burst; why must we wait for clinical data?

## VI. Proactive pivotal submissions

For pivotal drugs, regulators will work with sponsors to define requirements, accept portions of the submission in advance, define optimum formats, and prepare for rapid review. Before the last blind is broken, the sponsor should prewrite the entire document with all tables, graphics, and conclusions. Of course it should be paperless. When the data are locked and validated, actual treatment assignments will provide the final tables and graphics and the text can be edited appropriately. As serious safety concerns will have been addressed already and most clinical trials will have been completed, the final document should be ready no more than two weeks after the final data lock point. Of course, if the data are far different than expected, it may set you back further.

How important is speed? For a product that will sell US \$1 billion at peak worldwide, every one-second delay costs \$31.69! How many seconds represent the life of a patient-in-waiting? How much pain is endured for that second?

The regulatory submissions should be user-*seductive* for the reviewers. If possible, submit a DVD. Start with the global summary of 200 pages that begins with the package insert and has each label statement, the studies and supporting data hyperlinked to it. Include photos of patients where

appropriate or key preclinical, toxicological, or clinical laboratory results. Include interviews with opinion leaders, investigators and patients.

## VII. Electronic publishing

Rapid implementation of evidence-based medicine saves lives! Clinical trial results are highly standardized. An electronic journal with tough peer-review could publish, in standardized format, key results from clinical trials with brief introductions and discussions, appropriate references and links to sponsors, regulators, and investigators for additional information. If rigorously peer-reviewed, the FDA should permit promotion based on this type of journal, even if the information expands pre-approved labeling.

## VIII. Global public adverse event database

Infrequent serious adverse events usually appear after marketing. They are reported to sponsors and regulators and find their way to the FDA who makes them available to anyone, redacting the identity of the reporter and patient. As patients differ in adverse events, and pharmacogenomics is the popular buzzword, it is clear that your patient, though unusual, might resemble one who has had a similar adverse event. Details of that event are very valuable. What were patient characteristics, dose and duration, concomitant therapies, treatment of the event and outcome? Why can't you access the FDA database directly? Why can't NLM build an excellent search engine? Why can't you file an ADE report electronically and interactively view similar reports? Why doesn't the FDA maintain a drug-drug and drug-behavior database?

## **IX. Clinical trial registry**

Often discussed and achieved in fragments, why isn't there a comprehensive guide for physicians and patients to planned and active clinical trials with qualification and referral information?

## **X. Numerical simulation and planning: Be blind not dumb**

Desktops and laptops can perform the most complex Monte Carlo simulations overnight. Probability distributions can be refined, and profound insight gained, about proposed studies. As actual data arrive, estimates can be refined. Algebraic statistics, simulating results from defined distributions, have given

way to permutation tests where all actual data are redistributed many times to see how often the observed result, or one more extreme, occurs by chance. Why do regulator's statisticians demand thirty-year-old frequentist tests? The plethora of thoughtful statistical scholarship by those who perform clinical trials seems not to have penetrated regulatory thinking.

### **Proposal**

NIH could increase support of postgraduate training in clinical research with centers of excellence that, like Duke Cardiology, actually coordinate exemplary multicenter clinical trials. NLM could focus more on trial, regulatory, and safety uses of

informatics. NIH could reform its own clinical trials to achieve quality-speed. NIGMS, or a new NICMS, could sponsor consensus groups to examine aspects of clinical trial and regulatory science. The NIH Director could lead a stellar group of clinical trial scholars to guide such efforts.