

The Triumph of Hope Over Experience: Why Johnny Can't Forecast Subject Enrollment

By Norman M. Goldfarb

94% percent of clinical studies are delayed because of slower-than-expected subject enrollment.¹ Is this the most incredible string of bad luck in history, or are there darker forces at play?

Slower-than-expected subject enrollment assumes the existence of subject enrollment predictions, presumably derived from subject enrollment plans, which are based on subject enrollment models, which may be explicit (e.g., in a spreadsheet) or implicit (i.e., in the project manager's head). This model includes assumptions related to investigator recruitment, subject enrollment, and sub-parameters such as screen failure rate. This article addresses slow enrollment by initiated sites, but the principles apply equally to investigator recruitment.

We know that sites, CROs, and project managers commonly over-estimate potential subject enrollment for two primary reasons: (a) not fully understanding the enrollment challenges and (b) knowing that conservative numbers will be unappetizing to the next bigger fish up the food chain. Why do the big fish repeatedly accept disappointment? It doesn't make a lot of sense, unless they want the smaller fish to create their own "stretch goals," and are confident that their leadership skills will offset the demotivation of repeatedly missing those goals.

Just for amusement, imagine that we actually want to predict subject enrollment accurately. How should we go about it?

The subject recruiting process involves numerous variables. If we can predict each variable with complete accuracy, we can predict enrollment with complete accuracy. For example:

Subjects randomized =
 Potential subjects contacted
 X % invited for first visit
 X % who show up
 X % who give consent
 X % who pass screen
 X % who are randomized

Assume we want to randomize 33 subjects in a study. If we contact 100 potential subjects, and there is a fallout rate of 20% at each step, randomized subjects = $100 \times 80\% \times 80\% \times 80\% \times 80\% \times 80\% = 33$. Mission accomplished.

However, since we cannot predict the exact percentage at each step, we might say that 20% fallout at each step is our most likely case, 10% our best case, and 30% our worst case. In the best case, we will randomize 59 subjects and in the worst case 17 subjects. Our level of optimism will dictate how many potential subjects we contact. We can cover the worst case by contacting $100 \times (33/17) = 194$ subjects.

In the real world, however, we will observe the likely case for some variables, the best case for others, and the worst case for others. For example, if the variables independently take on a range of values, we might enroll $100 \times 90\% \times 85\% \times 80\% \times 75\% \times 70\% = 32$ subjects. In other words, if the variables behave independently, we may have good luck with some and bad luck with others, and still achieve our goal of 33 subjects.

How likely is it that the variables will, in fact, behave independently, and not conspire to foil our plans? It depends on whether an underlying common factor influences them. In other words, what we think are independent variables may to various degrees be dependent on another variable. What variable could account for the industry's consistent record in over-estimating subject enrollment? Could it be "level of optimism", a key factor in forecasting activities? If we are uniformly optimistic and have no metrics for reference, most of the variables will have an error in the optimistic direction. For example, if are 10% optimistic on every variable, we will predict $100 \times 90\% \times 90\% \times 90\% \times 90\% \times 90\% = 59$ subjects. When we enroll 33 subjects, we will miss our forecast by $33/59=44\%$. On the other hand, if our ever-optimistic investigator always over-estimates the first variable, and our ever-pessimistic study coordinator always under-estimates the second variable, two bad estimates may yield one good combined estimate.

The Realistic Organization

An organization that values realism is more likely to generate realistic estimates than an organization that values optimism. As it turns out, most organizations – and people – tend to be optimistic. Effective organizations need to temper that optimism with some grounding in reality. Since few people want to be the bearer of bad news, depersonalizing the source of the realism with validated metrics and models can help organizations maintain their grasp on reality.

Statisticians use the "Monte Carlo" method, named after the famous casino, to determine the likelihood of various outcomes. They assign each variable an average value and a standard deviation (variability), and then observe what happens. If we simulate a study 1,000 times, about 68% of the values for each variable will fall within one standard deviation of the average; about 95% will fall within two standard deviations, and about 99.7% will fall within three standard deviations (assuming a "normal," bell-shaped distribution).

In the example above, 80% is the average for each of the variables. If our best and worst cases are actually our "reasonable" best and worst cases, 10% (10) of 100 subjects might equal one standard deviation. In other words, with a single variable, we can expect to recruit 80 of 100 potential subjects, with about 68% of the trials yielding between 70 and 90 subjects.

If we run our study 1,000 times with five variables, what is the likely range of outcomes? As mentioned above, it depends on the degree of independence of our variables. Let's look at two cases: (a) completely independent variables and (b) completely non-independent variables.

Using the Microsoft Excel data analysis tool, we can do the experiment. With completely independent variables and 1,000 trials, we obtain an average of 32.97 subjects with a standard deviation of 8.95 subjects. (With more trials, the average will approach even closer to the 33 subjects predicted by our original formula.) In other words, about 68% of the time, we will enroll between 24.02 and 41.92 subjects. But note the interesting part: With five independent variables, each with an average of 80% and standard deviation of 10%, the combined standard deviation drops from 10 subjects to 8.95 subjects. In other words, errors in estimating one variable offset errors in estimating another. As the number

of independent variables grows, the accuracy of our combined estimate improves. Adding a sixth variable, for example, further reduces the standard deviation from 8.95 subjects to 7.77 subjects.

In our example, although the accuracy of the estimate improves, it's still not that great: 32.97 +/- 8.95 subjects, or +/- 27% of 32.97. As the number of variables under 100% increases, the mean estimate declines substantially, but the standard deviation declines only slightly, causing the relative likely error to almost triple. This result confirms the obvious: increasing the number of uncertain variables increases the potential total uncertainty. The good news is that the uncertainty increases less than might be expected, because errors in one direction offset errors in another. Would we get more accurate results by collapsing multiple variables into a single consolidated variable? We might if we are lucky, just as betting on a single number in roulette is potentially highly lucrative but also highly risky.

Now let's do the same experiment, but assume that the variables are completely non-independent. We can perform this experiment by applying the exact same error to each variable in each Monte Carlo run (e.g., +5% in run 1 to all variables and -5% in run 2 to all variables). This approach amplifies the errors in a multiplication effect, with no offsetting of overestimates by underestimates. In this case, we obtain an average result of 38.05 subjects. (This result is biased high because big numbers multiplied together are bigger than smaller numbers multiplied together are small. For example, 150% X 150% = 225%; 50% X 50% = 25%; and (225%+25%)/2 = 125%. Averaging our original best case of 59 subjects and worst case of 17 subjects gives us the same result of 38 subjects. We can fix this bias by using geometric instead of arithmetic averages, but there are already enough statistics in this article.)

With all the errors piling one on top of another, the standard deviation does not shrink from 10 subjects to 8.95 subjects. Instead, it grows to 23.37 subjects. Our result will be 38.05 +/- 61% subjects. If we remove the high bias with geometric averaging, the result will be 33 +/- 71% subjects, or a likely error 2.6 times larger than with independent variables.

To summarize, as the number of variables increases, independent errors balance out, gradually improving the accuracy of our combined estimate. If, however, there is a systematic error affecting some or all of the variables, our estimate drifts higher and the accuracy of our estimate declines rapidly. If all of our estimates are 1 standard deviation too high, we will estimate enrollment of 61 subjects (38.05 + 23.37), 86% higher than our likely yield of 33 subjects. Sound familiar? To make matters worse, one standard deviation is not an extreme case. Most clinical trials are powered to an uncertainty level of 0.05, which is two standard deviations.

So, assuming that we actually want to estimate subject recruiting more accurately, here's how to go about it:

- Create a prediction model with as many variables as practical.
- Adjust the model with the equivalent of geometric averaging.
- Estimate each variable as accurately as possible.
- Estimate the high and low estimates (standard deviation) of each variable as accurately as possible.
- Accumulate data on actual results.
- Use the data to refine the estimates and correct systematic errors.
- Repeat until as accurate and reliable as practical (i.e., estimates are unbiased and standard deviation of combined estimates is minimal).

Our estimates won't become perfect, but their accuracy will improve from study to study, and the bias will gradually disappear. We can also apply the same method within a study, refining our estimates over time. As an added bonus, more granular models are more likely to include variables that can be applied across multiple sites and studies.

Not only can we improve the accuracy of our estimates, but, by identifying and addressing the problem variables, we can also accelerate enrollment. For example, if a high percentage of initial contacts are with unqualified candidates, we may want to clarify the eligibility criteria in our ads.

We can also accelerate identification of investigators that are unlikely to meet their enrollment targets, especially if we track their performance over multiple studies. For example, if an investigator's failure mode is often a high no-show rate at initial visits, we can watch that variable more closely. Sites can monitor their own performance, and focus their resources on the most promising studies.

The current system is designed for failure 94% of the time. It's a challenge to improve the process while the alligators demand our full attention, but the return on investment will be rapid and substantial.

When the big fish want accurate estimates, they will insist on accuracy, which can only be generated with a systematic approach such as that described in this article. Is Johnny read to learn how to forecast subject enrollment?

The enrollment simulation used for this article is at http://www.firstclinical.com/journal/2005/0507_EnrollmentSimulation.xls.

Reference:

1. Thomson CenterWatch 2003 Survey of 308 U.S. Investigative Sites

Norman M. Goldfarb is Managing Partner of First Clinical Research, a provider of a clinical research best practices consulting, training, implementation and research services. Contact him at (650) 465-0119 or ngoldfarb@firstclinical.com.