Reviewer's report

Title: Statistical design of personalized medicine interventions: The Clarification of Optimal Anticoagulation through Genetics (COAG) trial

Version: 4 Date: 27 September 2010

Reviewer: Garnet Anderson

Reviewer's report:

Major Compulsory Revisions:

This paper describes an interesting trial designed to determine the value of genotype-driven warfarin dosing relative to a dosing schedule determined by clinical factors. The objective of the manuscript is not entirely clear. In places it is written as a methods paper attempting to define approaches to trial design for clinical trials of personalized medicine using the COAG trial as the motivating example. More often it is a description of key statistical design features of the COAG trial. Either goal is worthy but the current manuscript seems to miss the mark for either. Given the nature of the trial, I would recommend the authors focus more on a detailed description of the COAG trial, including the motivation and design.

A thorough discussion of the rationale for selecting this design is needed. In COAG, all eligible patients are randomized to either genotyped-guided or clinically-guided warfarin dosing. While reasonable at first glance, the justification is less obvious when it becomes clear that treatment effect may be null for ~40% of patients, a genotypically-defined subgroup. The authors account for this null effect in the design by acknowledging that this will dilute the overall observed treatment effect and therefore inflate the sample sizes to assure power to detect a much smaller overall difference between groups. At first, I thought the problem was one of effect modification with this subgroup (possibly) experiencing a different treatment effect, a problem many trials must consider. However, in this case, the intervention itself apparently does not differ for this large subset (“subjects who possess a single genetic variant . . . would not benefit from clinical-guided dosing because previous data suggest that the genotype-guided algorithm will predict essentially the same dose as the clinical-guided algorithm” p. 6). Thus it appears not to be true effect modification but rather no effective intervention in this substantial subgroup. A fuller discussion of the pros and cons of this design relative to a targeted one, randomizing only patients for whom genotype guided dosing differs, is needed. The arguments for and against an inclusive versus a targeted design are quite different when one is expecting differential intervention effects across subgroups. The sample sizes (and hence cost) here should be very different. There are no obvious logistical concerns that make an inclusive design necessary. The statement that the inclusive design allows the results to be more generalizable does not seem appropriate; if these design assumptions are correct, the most valid interpretation will be specific to
the genetically defined subgroups.

I don’t understand the distinction the authors (are trying to make p. 5) between this “personalized medicine intervention” and a traditional clinical trial. In many traditional trials there are covariates thought to be effect modifiers. Usually one does not know in advance the prevalence of these covariates and one must guard against an unfavorable representation of them in the recruited population by either increasing the sample size or imposing some constraints on the distribution. How is the distribution of allelic variants different?

I was not able to reproduce these power calculations using some validated calculators. I get somewhat smaller required samples sizes for all of the settings I checked. (E.g., For overall test alpha = 0.04, beta = .2, delta = 5.49, SD = 20, the required sample size is 446 instead of the 550 shown in Table 1), Perhaps I misinterpreted some of the parameters—is the SD for the total sample or within strata (one genetic variant vs 0 /multiple)? Were the weights (wi) assumed to be fixed or random? Please check the calculations.

Discretionary Revisions
There are some smaller points that would be worth amplifying in a fuller trial design paper. Probably the most important issue is describing the minimal detectable effect that is worth finding. Specifically, what does a 15% relative difference in PTTR translate into, in terms of clinical endpoints (morbidity, hospital admissions, etc. )

This is to be a double-blinded trial for the first 4 weeks of the program. How will double-blinding be implemented? Will a dosing schedule that varies from what is clinically recommended cause automatic unblinding? Further, since the authors indicate that there may be cross-overs in the trial (because genetic information may not available at the time of initial dosing) won’t this also introduce both unblinding and further dilution of the observable difference. Are there estimates of how often this is expected to happen?

Some description of the dosing schedules for the two arms by strata would also be helpful.

**Level of interest:** An article whose findings are important to those with closely related research interests

**Quality of written English:** Acceptable

**Statistical review:** Yes, and I have assessed the statistics in my report.