Methodological issues

We are deeply grateful to GAPG / ICAP for your interest, concern, and insight as we further develop our proposal for a randomized trial of moderate alcohol intake. We respond to individual concerns below and would be happy to discuss any of these issues in more detail where our responses fail to address your primary concerns.

We would like to emphasize one important point, however. Many of these issues will ultimately be decided by a combination of NIAAA (as co-leader of a U01 or similar funding mechanism at NIH) and the final set of investigators. Our responses accurately reflect our efforts to date, which have developed in conjunction with NIAAA, but some of the smaller details will necessarily need to be adjusted based upon both internal and external review at NIH, thus ensuring that the trial is viewed as scientifically valid and unbiased and receives the widest possible attention. Nonetheless, the protocol that we submit to NIH will adhere closely to our suggestions below.

1. **Outcomes:**
   The slides include a large number of possible outcomes that could be measured. Will these be prioritized in some way into a few primary outcomes (e.g., CVD, diabetes, all-cause mortality) and possibly some secondary ones that could be derived from subsequent mining of the data?

Response:
We will follow the general strategy used in most clinical trials. In such trials, a single or two co-primary outcomes are specified, along with several secondary outcomes related to the primary outcome and one or more safety outcomes. For example, in the very large number of trials of new anticoagulants for acute coronary syndrome or atrial fibrillation, the primary outcome might be cardiovascular mortality, the secondary outcomes might include myocardial infarction or total mortality, and the safety outcomes would include major and minor bleeding.

We will use a similar strategy. Our co-primary outcomes are 1) incident cardiovascular disease, defined as non-fatal myocardial infarction, non-fatal ischemic stroke, coronary or carotid revascularization, or cardiovascular mortality, and 2) incident diabetes. Secondary outcomes include total mortality, 'hard' cardiovascular disease (excluding revascularization), and coronary heart disease (excluding stroke and carotid disease) - these are intended to demonstrate the robustness of the primary outcome. Safety outcomes will include cancer, trauma / injuries, all-cause hospitalization, and progression to excessive alcohol use. In conjunction with NIAAA, we will identify an independent data safety and monitoring board, which will ultimately have responsibility for identifying stopping rules, but we will recommend to the DSMB that stopping rules appropriately consider the severity of the outcome, so that, for example, we do not stop the trial early just because a significant benefit for alcohol in incident diabetes occurs, precluding further evaluation of cardiovascular disease.

As with any large randomized trial, once the primary outcomes are reported, several and often dozens of papers appear, either on other outcomes (e.g., change in cognition, change in weight, risk of congestive heart failure) or in interesting subgroups (e.g., risk of cardiovascular disease among diabetic participants). This ensures that the primary paper includes the primary outcomes and receives appropriate emphasis, while taking full advantage of the enormous effort that goes into completion of a clinical trial to produce a much richer body of science.

2. **Target population:**
   Is there an upper age limit cutoff? Narrowing the age band would give a tighter cohort. Is this worth doing?
   Is the sample intended to be nationally representative in each case, or is the focus specifically on at-risk groups?
   How will you control for other predisposing factors (e.g., family history), and also for socioeconomic and demographic variables that could influence the outcomes of drinking?

Response:
There is inevitably a trade-off when creating more stringent entry criteria, reducing the sample size (or prolonging enrollment) to gain a more homogeneous sample. In this case, excluding older individuals also excludes those at highest risk for events (who are desirable in a study like this). Nonetheless, because initiation of alcohol consumption in very old individuals is apt to be poorly received, we currently anticipate recruiting individuals 50-75 years of age, recognizing that few individuals near 80 are likely to comply with daily drinking.

We do not intend to enroll a nationally representative sample as that term is used epidemiologically (i.e., where the proportions of enrolled individuals can be tied back to the US population), although we do intend to enroll a geographically diverse sample that includes representative numbers of minorities and women; those are requirements of US participants in NIH-funded trials and hence we do intend to enroll a broad and diverse population within the US. By including individuals globally (and we now have collaborators identified in St. Petersburg and Hong Kong), we further ensure generalizability of our findings for policymakers such as the WHO.

At the same time, we will focus on individuals at higher cardiovascular risk, as has been done in most trials of primary cardiovascular prevention (whether of lifestyle or pharmacological therapy). Highly influential trials like PREDIMED (of Mediterranean diet) or JUPITER (of rosuvastatin) focused upon higher risk individuals for several reasons. First, the higher risk of cardiovascular disease in such individuals allows for a smaller sample size; a comparable trial in low-risk individuals might well need to exceed 40,000 individuals. Second, the higher risk also ensures events occur sooner, so that the trial need not exceed 5 years. Third, a focus on individuals at higher cardiovascular risk tends to mimic the population of patients for whom questions about the benefits of alcohol consumption are most salient. That is, we intend to focus on those patients in whom the likelihood of benefit (and, usefully, the desire to remain

adherent) is greatest, maximizing chances for benefit while minimizing risk, exactly as individual clinicians are likely to do.

The advantage of a large randomized trial is that the randomization itself controls for, at least in theory, ALL factors that differ across individuals, including factors we have not yet even identified. Therefore, our primary method of controlling for predisposing factors is randomization. Fortunately, there are also statistical methods (e.g., multivariable proportional hazards regression) that allow us to adjust statistically if, by chance, they are not evenly distributed across the two groups. As with most such trials, the baseline evaluation is typically the most extensive of the entire study, ensuring that we have detailed information on each group for which to adjust. In addition, that information will allow us to conduct subgroup analyses, in which we evaluate whether the effect of alcohol intake is greater among certain types of individuals. Although such analyses must be performed cautiously (because an almost unlimited number could be performed), we will particularly focus on evaluating the consistency of the effect of alcohol across groups defined by factors known to influence cardiovascular disease (i.e., cardiovascular risk factors) or alcohol use (e.g., sex, age).

We will employ centrally-determined randomization with permuted blocks that ensure each site has comparable numbers of individuals in each arm and that random imbalances do not skew our results. In addition, we have explored adaptive designs that allow flexibility in recruitment to ensure adequate numbers of individuals in primary demographic groups and that reduce the final sample if event rates exceed expectation.

3. **Sample size:**
   The total sample size is 10-20,000 individuals over 13 sites, which, in the best case scenario, allows only for about 1500 per site (750 per group), not taking into account attrition. This is a very small sample size.
   Would it be worth reducing the number of sites (e.g., 6) and increasing the cohort size in each?

Response:
It is the total study sample size that determines the power of the study, not the number of sites nor the number of individuals per site (as we propose individual and not cluster randomization). As a result, the effective sample size remains the same, whether we recruit 10 individuals from 1000 sites or 1000 individuals from 10 sites.

There are important reasons not to limit to a very small number of sites, and indeed many trials in the last several years have used the opposite approach (recruiting from literally hundreds or thousands of individual offices). First, it is not feasible to recruit thousands of individuals from a single site in the time period of this study, and recruitment is inevitably a key rate-limiting step. Second, a smaller number of sites would result in less generalizability (i.e., would necessarily not include participants worldwide). Third, it poses a risk of more bias, because problems at a single site would factor more heavily and could ruin an entire trial.

For these reasons, our preference is to expand beyond 13 sites, although still within conservative limits. The conduct of the study itself will ideally occur at a larger number, albeit with at least 250 individuals at each site to ensure a manageable number of sites, such that the total would likely run closer to 25. That ensures a trial that remains easy to monitor but large enough to enroll on time.

4. **Adherence**
   How likely is it that subjects will adhere to the regimen and the trial groups over several years? An alternative might be to run a pilot and assess likely compliance.

Response:
We believe that, with appropriate screening and evaluation, adherence (and retention) should be high, although we will certainly maximize this in several ways. This belief is borne out by our experience with dozens of clinical trials of a host of interventions, including physical activity, dietary change, and pharmacological therapies. For example, our collaborator [redacted], who has led trials across the US, has studied an intensive lifestyle intervention in LOOK AHEAD and a focused physical activity for elders in the LIFE Trial with clear evidence of adherence – typically 80% of participants achieve high levels of adherence, with variable adherence among the remainder. Similarly, our collaborators from PREDIMED and the Women's Health Initiative demonstrably altered individual's entire diets over several years.

Existing experience in alcohol supports this view. Smaller trials in the range of months to years have now been conducted with alcohol, with proven effects on biomarkers such as HDL-cholesterol of the expected magnitude and direction. Thus, both the published experience with alcohol and the much larger experience in clinical trials suggest that, with sufficiently intense monitoring and feedback, adherence should be more than acceptable. In this case, we plan repeated study visits and frequent study contact (by telephone, email, or web-based methods, based upon local resources) to ensure that participants remain motivated and excited. We will also have a washout period to ensure a common baseline for all participants; this will further serve as a run-in to exclude individuals unable to abstain from alcohol. Further, our power calculations (described further below) already include estimates of adherence.

It is important to note that we anticipate a biological effect of alcohol at a lower dose than that chosen. That is, the bulk of the epidemiological evidence suggests that even intake 3-4 days per week is likely to be beneficial. As a result, adherence in the alcohol arm need not be perfect to demonstrate benefit and even 50% adherence might yield a full result. Because we anticipate less-than-full adherence and because a recommendation for daily drinking is easiest to implement, our intervention includes 7 drinks weekly, thus giving us a wide margin of potential effectiveness.

5. **Retention**
   Attrition over 5 years is likely to be high. How can this be addressed? For example, by increasing individual cohort size?

Response:
Attrition is a very important limitation, and one that experienced trialists take great pains to avoid. We will only use experienced clinical trial sites specifically to ensure that they have demonstrated an ability to retain participants. Our sample size overall also reflects the degree of attrition seen in large clinical trials of lifestyle. However, attrition will not necessarily be high. In the Dietary Modification arm of the Women's Health Initiative, for example, which required wholesale dietary changes to reduce dietary fat and most closely approximates this trial, nearly 49,000 women were randomized and ~2,000 women were lost to follow-up or withdrew (i.e., ~4%).

To address these issues more explicitly, it is useful to review our proposed power calculations. In these, we test a variety of scenarios, some intentionally conservative, to ensure that our sample size needs will be meet. We make the following assumptions: 1) a multi-center individually randomized controlled trial; 2) subject enrollment during 3 years (see below) for at least 5 years of follow-up; 3) our primary statistical measure will be a log-rank test of survival curves free of the primary endpoint; 4) the relative risk for the intervention is in the range of 0.7-0.8; 5) possible scenarios of loss to follow-up of 0%, 1%, 3% or 5%; 6) a 5-year risk of developing diabetes of 8.2%; 7) a 5-year risk of CVD of 5.0% (annual incidence rate is 0.0103); and 8) scenarios of true non-adherence (i.e., non-adherence to the point of no effect) of 0%, 5%, 10%, 12.5% or 20%. Within these scenarios, the estimated sample size requirement ranges from ~2,500-12,500 for diabetes and ~4,500-20,000 for CVD, even with the most conservative assumptions. Thus, our estimated sample size (10-20,000, as noted) should account for both lack of perfect adherence and attrition.

6. **Monitoring**
   How will monitoring be conducted? What are some of the biomarkers that will be measured and how confident are you that they will accurately be able to reflect drinking levels and patterns?

Response:
Monitoring in a clinical trial like this takes many forms. As noted, participants will be contacted repeatedly electronically, using varying approaches to ensure that they remain novel and capture participants' attention, independent of study visits. In-person visits will be most frequent early in the study, when adherence is most challenging. We will conduct timeline followbacks to capture drinking that occurs beyond participant-specific limits (i.e., one drink daily in the intervention arm and any drinking in the control arm), a gold-standard method of evaluation of alcohol use. We are adapting web-based methods to ensure easy and reliable data capture for participants and investigators between visits. Thus, we will have ample information on individual participant's drinking habits using standard clinical methods, and these will form our primary basis for monitoring the safety and adherence of participants.

Biomarkers are attractive methods to supplement clinical interviews, although no biomarker appears to be perfect in assessing adherence to moderate drinking, and we have serious concerns about relying extensively upon them for monitoring. For alcohol, existing biomarkers

have been designed to capture heavy drinking, and the level of intake recommended here will alter them variably. More generally, biomarkers are not commonly used in clinical trials to alter participant behavior. That is, process measures (such as timeline followbacks) are used to improve adherence, while biomarkers are used as secondary outcome measures. As an example, bottle openings might be used captured to reflect adherence to medication, while change in biomarkers like blood pressure or cholesterol would be used as secondary outcomes.

As such, we will measure a variety of biomarkers, some to detect complications of drinking (e.g., GGT, AST, CDT) and others to follow adherence (e.g., HDL-C, ethyl glucuronides, tartaric acid). We believe that some of these (e.g., ethyl glucuronides) should provide useful population-level discrimination between those who drink and those who don't, although they are unlikely to capture exact amounts of drinking and will be of at best uncertain value for individuals. We are working with our partners to evaluate additional biomarkers (e.g., metabolites of hops for beer use) and will add those during the trial if they appear to work well. Because these biomarkers will be used to supplement much more nuanced clinical interviews, however, they need not discriminate perfectly to offer a valuable addition. Furthermore, they will be useful from a publication standpoint to show that the groups differ, even if they are less useful for monitoring individual participants.

7. **Beverage variation**
   Given that beer and spirits have different alcohol contents, and that beers vary in strength, how will you ensure that these variations do not influence the strength of the effect.

Response:
This is a somewhat complex issue that bears on the question of what the intervention itself is. Several factors bear on our approach. First, the differences in alcohol content between beverages, while real, are much, much smaller than the difference between drinking and non-drinking. As a result, the two groups will differ meaningfully in their ethanol intake, regardless of what beverage is used. That reflects current clinical practice, in which physicians ask individuals about their servings of alcohol but rarely query the type or strength of that serving. Second, we have found few consistent differences across beverages in their effect on heart disease or diabetes. This suggests that the small differences in alcohol content are indeed modest relative to the similarities across beverages. Third, the key question to consider is what intervention we aim to test. To our group, the intervention is "a daily drink of alcohol", recognizing that it will be operationalized differently by different individuals at different times. Ultimately, the intervention we will test will represent the blended average of all of the types and strengths of alcohol, reflecting the best epidemiological evidence and providing the most useful information to clinicians who might recommend alcohol.

With that said, we will provide repeated, standardized information on portion sizes to attempt to minimize heterogeneity in actual ethanol intake and carefully collect information on the type of beverage (including its alcohol by volume) monthly. With that information, we can, at the end of the trial, examine the 'actual' ethanol intake (i.e., reported frequency of use x the

participant-specific alcohol average) of each participant and relate that to outcomes. We anticipate that the variability across all drinkers will still be far less than the difference between drinkers and non-drinkers, but our analyses will allow us (within the limits of the observed variability) to estimate 'dose-response' curves for our primary outcomes.

8. **Incentives**
   What incentives will be offered to individuals being encouraged to abstain from drinking? Is it only advice? If so, what kind of advice?

Response:
All participants will receive generic health advice, typical for long-term clinical trials (e.g., information on a healthy diet, smoking cessation, etc.), along with periodic health examinations and regular attention; experience demonstrates that these are key factors in ensuring adherence and retention. Individuals in the abstention arm will receive modest monetary compensation, similar to (but likely less than) the cost of alcohol in the other arm. For ethical reasons, it is important to ensure that individuals receive similar compensation in both arms.

9. **Feasibility study**
   Would it be useful to consider a feasibility study for a year to iron out some of the issues before launching into the full 5-year study?

Response:
In essence, all of the previous experience of our team represents a vast storehouse of feasibility studies; we have included the leaders of some of the most difficult, complex clinical trials ever conducted. Nonetheless, we recognize the difficulties inherent in any clinical trial. Rather than a separate feasibility study, our preferred approach mimics that of the Women's Health Initiative (a much larger and much more complex trial). We will establish a small number of vanguard sites, chosen for their experience, generalizability, and ability to scale up quickly. These will represent 'beta' test sites, where the trial will be implemented but can be altered and updated to reflect actual trial conduct. In the unlikely event that these sites indicate that the trial is simply not feasible, we will have the opportunity to end the experiment quickly, ensuring savings of resources, time, and effort to NIH, funders, and investigators. This has the very large advantage, however, of using the data from the vanguard sites if the trial is indeed feasible, so that no time or (worse) data are wasted, and ensuring that other sites can be added expeditiously. This approach helps to ensure that the full trial can be completed in the original time allotment (and within budget). As noted above, our sample size calculations reflect this approach, with stepped entry of sites into the recruitment phase over three years as we gain insight from our initial set of vanguard sites.