IMPORTANCE OF UNCERTAINTY IN MEAN SCORE ESTIMATES OF ONE SURVEY QUESTION ABOUT TEST ACCURACY: AMERICAN COLLEGE OF MEDICAL GENETICS (ACMG) RECOMMENDATIONS FOR NEWBORN SCREENING (NBS) IN MEDIUM CHAIN KETOACYL-COA THIOLASE DEFICIENCY (MCKAT)

Brian Rittenhouse , PhD
MCPHS University, Boston, MA

Background & Objectives

In 2006 the American College of Medical Genetics (ACMG) surveyed stakeholders for their opinions to make recommendations for a set of 84 rare conditions as Core, Secondary Target and Not Recommended for NBS. For each condition 19 aspects were assessed (attributes of the condition, screening test and treatment). The ACMG-ascribed weight assigned to a given answer for each of the 19 questions was multiplied by the percentage of respondents agreeing to the presence of the condition surveyed, determining the attribute's mean score. The total of all scores resulted in a score that determined the entry point to an algorithm (EPA), containing additional questions that varied by the EPA and that determined the final recommendation (Primary, Secondary or Not Recommended).

A controlling question for even gaining entry to the algorithm was: "Was a sensitive and specific (SS) test for the condition available?" Scoring was 0/200 for No/Yes. With a mean less than 100, the condition would not even enter the algorithm and would be Not Recommended. This question is arguably the most important question in the survey as a score adequate to move the condition into the algorithm can be sufficient to gain a positive recommendation almost regardless of the total score on all other questions. We sought to assess the importance of uncertainty around the SS survey score (score reported, "65% of the maximum" of 200 points (i.e. a rounded value of 130).

This research examines the robustness of the ACMG recommendation for one condition, MCKAT, by considering unacknowledged issues of uncertainty around AMCG scoring, focusing on this one survey question.

Uncertainty in research may come from multiple sources. Among these, two are examined here. ACMG treated scores as objective truth rather than as estimates, ignoring issues of uncertainty involving both sampling variability and missing survey data. We address the potential influences on confidence in ACMG scoring with MCKAT as an example. With an original total score of 1170, MCKAT was recommended as a Secondary Target. Given the algorithm controlling importance of the SS question, we focused on this one question here.

Methods

Our analysis of missing data and sampling variability concerns respondents who answered the MCKAT survey, but who failed to answer the SS question. ACMG wholly ignored sampling variability and missing data, treating survey scores as if they were population means (instead of sample means with uncertainty) and assuming any missing data were ignorable.

Missing data are often treated as ignorable, implicitly assuming what Manski (1989) calls an "invariance assumption" (equal means of observed/unobserved data). He suggests using boundary estimates to capture implications of uncertainty without making such assumptions which may lack credibility. Using Manski's general notation, the following decomposition of the element of interest, E(y|x) can be made. Here, x are respondent characteristics, z is an indicator variable for whether responses are observed or not and P, the frequency of observed and unobserved responses. E(y|x) is composed of a weighted average of known and unknown responses, where the weights are the proportions of the survey respondents answering or not answering the particular question. The weights are applied to the means of the known and unknown responses. In the equation below, all values are known except the elements in **bold type.** However, with known boundaries on the response mean score for the unknown, missing responses, boundaries may be established on the element of ultimate interest **E(y|x).**

$$\mathbf{E(y|x)} = E(y|x, z=1)*P(z=1) + \mathbf{E(y|x, z=0)*} P(z=0)$$

The boundary is established by knowledge of the boundary for the mean score of the missing responses. If the minimum (K_0) and maximum (K_1) scores for the question are known, those represent the minimum and maximum possible values for **E(y|x, z=0).** These values, along with the other known values in the equation establish the bounds on the element of interest **E(y|x).** The lower bound substitutes K_0 (0) for E(y|x, z=0) in the equation above, while the upper bounds substitutes K_1 (200).

The ACMG did not report the number of missing (n_M) responses for the SS question (thus also P(z=0) is unknown). However, the mean score for that question, given the number of respondents to any MCKAT question (n = 23) was consistent with three possible values for n_M (0, 3 or 6). That is we can obtain a percentage of maximum that rounds to the reported 65% [64.5, 65.5) with various possible numbers of scores (0 or 200) out of a possible n = 23 sample size. The valid score range consistent with the reported 65% is [129, 131). If there are no missing responses ($n_M = 0$), then a score in the valid range may be obtained with 15 responses of 200 and the remainder (8) responses of 0 (score = 130.43). With 3 missing ($n_M = 3$) then a score in the valid range may be obtained with 13 responses of 200 and the remainder (7) responses of 0 (score = 130). With 6 missing ($n_M = 6$) then a score in the valid range may be obtained with 11 responses of 200 and the remainder (6) responses of 0 (score=129.41). No other score combinations were possible that could yield a score for SS in the valid range [129, 131).

Also ignored in the ACMG analysis is the influence of sampling variability. While the mean response may be accurately estimated, that estimation is subject to standard qualifications on its accuracy, frequently associated with a confidence interval. We explore this uncertainty without invoking a normality assumption, by using bootstrapping techniques, resampling (with replacement) from the original data to estimate influences of sampling uncertainty. We estimate (separately and then together) the influence of boundary estimates and sampling variation.

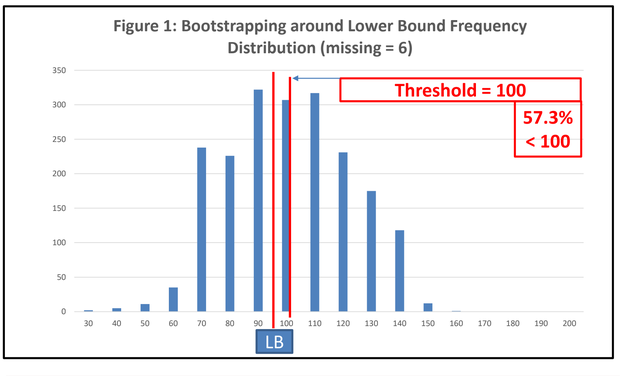
Results

Manski bounds are reported in Table 1 for the three possible missing scenarios. The bound width is zero if no missing observations. In the other two cases the lower bound **does not and does** cross the threshold value of 100 for n_M values of 3 and 6.

For the bootstrapping examination of sampling variability combined with the lower bounds analysis with no missing cases, 5.9% of the bootstraps fell below the 100 point threshold which would only cast modest doubt on the gateway to the algorithm being open for MCKAT. However, in the other 2 cases a considerable percentage of the bootstrap values failed to reach the threshold point value of 100 (27.9 and 57.3% for missing 3 and 6 observations respectively). **Figure 1** shows the frequency distribution for the latter case. Scores of less than 100 would have led to a Not Recommended conclusion.

Table 1: Original Scores, Manski Upper and Lower Bounds and Bootstrapping around LB for 3 Valid Numbers of Missing Observations: SS Question

Survey Question	Lower Bound	Original Mean	Upper Bound	% of Scores < 100 Bootstrapping LB
SS ($n_M = 0$)	130.4	130.4	130.4	5.9
SS ($n_M = 3$)	113	130	139.1	27.9
SS ($n_M = 6$)	95.7	129.4	147.8	57.3



Discussion

A previous similar MCKAT analysis of two questions for which ACMG reported the score distribution and the numbers of missing observations indicated uncertainty about the placement within the algorithm. Interestingly, this uncertainty could not affect the final recommendation due to the nature of the algorithm questions. Virtually regardless of total score, MCKAT would be a Secondary Target because of its being possible to screen as part of a multiplex platform (multiple conditions in one sample). For MCKAT, this raises questions about ACMG methods. The survey appears to have been wasted effort.

The current research suggests that there was one place where the survey could have made a difference – the algorithm's gateway SS question – Does a Sensitive and Specific test exist? We have shown there may be a lack of confidence in the SS estimated score and, therefore, in the recommendation. A full reporting of the extent of missing responses would have provided additional clarity.

We also suggest that the question itself lacks validity. While sensitivity and specificity have clear meanings, it is not clear how to answer the question - what sensitivity/specificity values imply an affirmative answer? Each respondent determines that herself, rendering the question likely worthless. Moreover, instead of opinions, why not use facts? Using means of opinion-based data can only obfuscate by comparison. Lastly, all of our analysis (and ACMG's) relies on random sampling from a population of interest While the sample frame employed is vague, clearly ACMG failed to use a random sample, making its results even less credible.

Conclusions

Famously, there are things we know, things we know we know we do not know (known unknowns) and things we do not know that we do not know (unknown unknowns). Some of the known unknowns are the ACMG targeted population for its survey, its sampling frame, its survey response rate and known respondents' answers to questions they chose not to answer. Even assuming no issues with any of those except the last, we have seen some doubt on the confidence we can have in ACMG survey scoring, even without questioning other aspects of its approach. These are known unknowns and in the latter case we are able to account for a possible range of responses. What this research has uncovered in its latter stages is that there was an unknown unknown – that the ACMG process was meaningless in terms of its recommendation for MCKAT which was destined for Secondary Target status once the ACMG set its algorithm – with one exception where the SS question response mean was less than 100. There is non-ignorable uncertainty on the validity of MCKAT entering the algorithm.

Importantly, we do not here question the correctness of the MCKAT recommendation. Rather, we question the basis of it by raising questions about the validity of the ACMG approach. Even when using the likely flawed ACMG process, the MCKAT recommendation is fixed almost regardless of survey responses. The one place where there is a possible change is with respect to the SS question and, depending on unknown missing responses to that question, MCKAT might have been labelled Not Recommended. We suggest that ACMG and others use accepted practices of technology assessment (Cost-Effectiveness Analysis or Multi-Criteria Decision Analysis[2]) instead of apparently ad hoc and highly questionable methods.

References

1. Watson, MS, Lloyd-Puryear, MA, Mann, MY, et al. Newborn screening: toward a uniform screening panel and system. *Genet. Med.*. 2006; 8 Suppl 1:1S-252S.
2. Keeny RL and Raiffa H *Decision with Multiple Objectives* (J Wiley & Sons, 1976)