Multi-Variant Risk Reporting
in the Coriell Personalized Medicine Collaborative (CPMC®) Research Study

Section 1: Background, Purpose, and Executive Summary

Background and Purpose

On May 9, 2013 the Coriell Personalized Medicine Collaborative (CPMC) research study's independent expert advisory panel, the Informed Cohort Oversight Board (ICOB), approved the incorporation of the multiplicative, multi-variant risk estimation method implemented by the R package REGENT (Crouch et al, 2013) into the CPMC research study. As part of that decision-making process, ICOB Chair David Pellman, MD requested that criteria for single nucleotide polymorphism (SNP) inclusion and other details of implementation be provided to ICOB members and the wider research community for evaluation and comment. Here, we detail both the process of calculating reported risk values based on multiple ICOB approved variants associated with a given complex condition using REGENT and the unique issues related to the construction of multi-variant risk reports. This document is meant as a supplement to the CPMC technical paper also posted on the CPMC website (http://www.cpmc.coriell.org) which describes standard reporting procedures within the project.

Executive Summary

Since ICOB-approved variants used in multi-variant reports are CLIA-validated, well-replicated risk factors showing consistent effect sizes for their particular condition(s), each contributes valuable information regarding the underlying risk distribution. Therefore, and in consultation with the authors of the method, the CPMC research team finds insufficient theoretical and/or statistical justification to exclude any approved variants from their relevant condition-specific multi-variant models. Nevertheless, practical issues such as the violation of model assumptions (e.g. independence, no epistasis, Hardy-Weinberg equilibrium) or the pattern of missing data in the cohort, which do not factor into ICOB decisions regarding the validity of a disease-SNP association, may dictate that an approved variant be excluded from the final model. In practice, the largest subset of independent ($R^2 < 0.20$) ICOB-approved SNPs that ensures >95% of the cohort will receive a multi-variant risk estimate will be used to generate the values returned to all participants with complete data. Individuals with missing genotypes will not receive an aggregate risk score. However, all CPMC participants will receive a clinical report detailing all genotypes and risk values for the individual SNPs in the model which were successfully typed in their sample. Next, as the method systematically identifies genotypic combinations that significantly differ from average and those that do not, all individuals possessing a combination conferring average risk will receive a value of 1.00 as the genotype relative risk. Individuals possessing combinations showing either reduced or increased (elevated or high) risk will receive the actual genotype relative risk value associated with their combination rounded to two digits. However, because REGENT is multiplicative in nature, it has the potential to produce very large genotype relative risks as well as values near zero; estimates in both of these extremes can be misleading. Therefore, when raw genotype relative risk values produced by a condition-specific model exceed 10.00 or are less than 0.10, individuals with combinations assigned such values will be informed that their value is greater or less than these figures. Finally, given that REGENT includes two simulation steps involving confidence intervals (CI), the consistency of both will be assessed prior to reporting and relevant metrics will be made available as condition-specific appendices to this document.

Section 2: Variants and Model
SNP Exclusion Criteria

All variants considered for use in multi-variant risk reporting are ICOB-approved, CLIA-validated markers showing replicated effects of meaningful size (i.e. homozygote Relative Risk (RR) >= 1.20). The goal is to incorporate as much ICOB-approved information as possible from the literature regarding the underlying distribution of genetic risk. Given this objective, and after consultation with the authors of the method (Goddard and Lewis 2010; Crouch et al, 2013), there appears to be no theoretical justification for excluding any variants that have received ICOB approval unless their use results in a violation of model assumptions or substantially impacts the reporting process. For example, REGENT requires all risk factors to be statistically independent. Even if not physically linked, if two predictor variables are correlated beyond a certain threshold, the underlying assumption of independence is violated. Note that the shared variance among markers (R^2) and not a more traditional measure of linkage disequilibrium (e.g. D') is the relevant value in this context. The CPMC research team finds no published analysis justifying an acceptable threshold for independence; however an informal value of $R^2 < 0.20$ for identifying unlinked variants is anecdotally common in the GWAS literature (C. Lewis, personal communication with J. Jarvis, June 17, 2013). For the purposes of CPMC research and reporting this value will be applied until a published, authoritative figure is identified. In practice, such high correlations are expected to be relatively rare among ICOB-approved variants.

Consistency in reported multi-variant risk estimates is also critical from both a research perspective and to accommodate those limitations inherent in the CPMC reporting infrastructure. Thus, the project reports genotype relative risk values based on complete marker data using the single most inclusive set of ICOB-approved SNPs that meet inclusion criteria at the time of report release. Should additional variants and/or an alternative multi-variant model be submitted and approved by the ICOB, the CPMC research team fully intends to revisit and update the relevant report. To minimize the impact of missing data, every effort will be made to ensure that > 95% of the cohort receives a multi-variant estimate based on observed genotypes. Markers with the poorest call rates and smallest effect sizes will be preferentially removed should missing data prove an issue. **Every** CPMC research participant will receive individual genotypes and risk values for those individual SNPs included in the model which were successfully typed in their sample as part of a CLIA-approved clinical report.

In summary, non-independence, violations of Hardy-Weinberg equilibrium, a problematic pattern of missing data, and/or compelling evidence for epistatic interaction may cause one or more variant to be removed from the model used in risk reporting.

Input Data

Calculation of multi-variant genotype relative risk values using REGENT requires a condition-specific estimate of prevalence. We obtain these values from appropriate primary literature and/or databases such as those hosted at the Surveillance, Epidemiology, and End Results (http://seer.cancer.gov) and Center for Disease Control and Prevention (http://www.cdc.gov) websites. Multi-variant risk estimation also requires SNP-specific data including risk values, the risk allele frequency and the number of cases and controls used to estimate risk values. We utilize the input file format that allows specification of the risk values associated with heterozygous and homozygous genotypes independently. When only a single, allele-specific value is reported, homozygote and heterozygote values are calculated using the model type stated in the primary source (e.g. additive, dominant, or multiplicative). Allele frequencies are taken directly from the source analysis. When odds ratios (OR) are provided in the primary source, the following equation is used to convert them to relative risk (RR) scores for heterozygote and homozygote genotypes in REGENT.

RR = OR / ((1-prevalence) + (prevalence)(OR))

This method is employed when the prevalence rate of the condition exceeds 10% within the general U.S. population and/or when the OR diverges greatly from 1.00, since these scenarios may lead to an over-estimation of the risk estimate. When relative risk scores are provided or calculated for an allele in which the control population frequency is greater than 0.5, we use 1/RR as the relative risk value for the alternate allele (which will have a control population frequency of < 0.50). In this manner, the minor allele (i.e. less frequent) is always the one contributing the effect to the model.

Finally, we use the standard default REGENT.model parameters for the following: cv=0.05, alpha=0.05, sims=10^5, indsims=10^5, SmallSampAdjust=0.5, BaseRange =0.01, PlotMax=5.

Section 3: Model Characterization Metrics

A variety of metrics for evaluating the predictive ability of risk models have been proposed in the literature, including the area under the receiver operating characteristic curve (AUC) and several reclassification measures. However, the best approach remains unclear, and methods continue to evolve (Janssens and Khoury 2010). Additionally, the empirical data required to measure the "true" predictive value of multi-variant genetic risk estimates is seldom available. For rare conditions, obtaining a sufficient sample of cases can be infeasible. For conditions with a complicated clinical presentation, the true status of "control" individuals may be ambiguous with some developing the condition at a later date. Finally, even when such data do exist, they may be unavailable for use within the CPMC and so cannot be universally relied upon to benchmark the performance of genetic risk estimates for all conditions on which we intend to report. However, the standardized assessment of models used in multi-variant risk reporting and the publication of their attributes is clearly an important component of the CPMC research project and its aims. Thus, characterization of each condition-specific multi-variant model including the values described below will be provided to the wider research community as appendices to this document as new reports are released. When appropriate, these will also be submitted for publication to appropriate peer-reviewed journals**.**

REGENT.model Evaluation

First, we evaluate the performance and consistency of the REGENT.model function (Goddard and Lewis, 2010) in establishing the categorical boundaries (the range in confidence intervals of a given set of genotypes that is assigned to a given category). This step requires the prevalence of the condition, the RR for each variant, the case and control sample sizes used in the studies from which the RR values were derived, and the allele frequency of each risk/protective variant in the control sample. We use the standard default REGENT.model parameters for the following: cv=0.05, alpha=0.05, sims=10^5, indsims=10^5, SmallSampAdjust=0.5, BaseRange =0.01, PlotMax=5. However, we raise the Block value to ensure that the total number of potential genotype combinations can be retained in memory for the analysis. To evaluate the stability of the multi-variant model, we plot variation in categorical boundary assignments as well as the proportion of simulated individuals assigned to a given category across at least 100 runs of REGENT.model. If these metrics are consistent across runs of REGENT.model, we proceed to the evaluation of REGENT.predict.

REGENT.predict Evaluation

Next, we evaluate the performance and consistency of the REGENT.predict function (Goddard and Lewis, 2010). While point estimates of genotype relative risk are invariant, the confidence intervals and category assignments produced by REGENT.predict vary slightly due to stochasticity across internal simulations. For research and reporting purposes, it is critical that these values show relative consistency for a given multi-variant model. Thus, we developed a pipeline to assess

model stability based on multiple case/control datasets generated using the PredictABEL R package (Kundu et al. Eur J Epidemiol. 2011 April; 26(4): 261–264)).

Specifically, input case/control data (10^6 individual genotypes) are derived via PredictABEL using identical input values for prevalence et cetera as are used in REGENT.model for the condition of interest. Simulating datasets allows for the systematic generation of data for all possible health conditions regardless of prevalence, number of genetic variants involved or their effect sizes and so facilitates quantitative assessment of consistency. However, we acknowledge that some degree of systematic bias and potential circularity may be introduced as a result of simulation procedures and input data.

Using the appropriate prevalence, risk allele frequencies, estimated RR values, case/control sample sizes, and the following default parameters: cv=0.05, alpha=0.05, sims=10^5, SmallSampAdjust=0.5, we execute at least 100 runs of REGENT.predict (which generates the individual genotype relative risk scores and corresponding CI) for each reported condition using simulated data. In all cases, we raise the Block value to ensure that the total number of potential genotype combinations can be retained in memory for the analysis. Each run involves at least 100,000 individuals. However, for extremely rare conditions (e.g. prevalence < 0.001), simulated samples of up to 1,000,000 may be required to obtain adequate numbers of cases for evaluation. Since data simulation approaches probabilistically assign case/control status from similar input data as the risk estimation procedure (prevalence, number of variants, variant effect sizes and frequencies), using them to calculate performance metrics such as the AUC and the median relative risk for cases and controls becomes potentially circular and so results from such analyses must be interpreted with great caution. However we note that a similar approach has been shown to accurately replicate AUC values from some empirical studies (Kundu et al., 2012).

Empirical Data

When feasible, the CPMC will also utilize multiple sources of empirical data to assess various features of condition-specific models. First, when a sufficient number of self-reported cases (N >= 25) can be identified in data from the CPMC cohort itself, these will be used to calculate the AUC, the median genotype relative risk, and risk distributions for cases and non-cases, and the proportions of each across risk categories. The empirical distribution of AUC values under the null model will also be assessed by randomly permuting case/non-case status. Similar values will also be produced for additional empirical datasets (e.g., dbGaP) that are reasonable to obtain for a given condition.

Section 4: Reporting Considerations

Risk Categories and Messaging

A unique feature of REGENT – when compared to other multiplicative approaches – is the systematic calculation of categorical boundaries (for *reduced, average, elevated*, and *high risk*) in its modeling step and confidence intervals surrounding the genotype relative risk estimates in its prediction step. These may be used to situate a calculated risk score in the context of the full distribution defined by the variants included in the model. For example, a specific genotypic combination may show a calculated relative risk of 1.2 but with such a wide confidence interval that it is algorithmically assigned to the "average" category. While confidence intervals will not be reported to participants, this allows messaging that will help participants better understand their genetic risk. In order to maintain consistency with previous reports, individuals in the *"elevated"* and *"high"* categories will be considered to be at "*increased risk*." Likewise, the CPMC will report that individuals in the *"reduced"* category are at *"decreased risk."* Individuals determined to be at *"average risk"* will be messaged accordingly. Reports for all genotype combinations that are assigned to either the *decreased* or *increased* risk category will contain the genotype relative risk point estimate rounded to two decimal places

(however, see possible truncation step below). When a given genotype combination is assigned to the average risk category, we will report a multigenic relative risk of 1.00.

Truncation of the risk distribution

Because REGENT is a multiplicative model, it has the potential to produce very large genotype relative risks as well as values near zero. As raw values in both extremes can be seriously misleading, the CPMC will employ a "truncated" approach to reporting when necessary. Specifically, when raw genotype relative risk values produced by a condition-specific model exceed 10.0 or are less than 0.1 for a given genotypic combination, individuals possessing it will be informed that their relative risk value is greater or less than this figure. This provides the added benefits of both ensuring a reasonable y-axis on graphical depictions of risk and avoiding inappropriate emphasis on the range of possible genetic risk over non-genetic sources.

Genotypic Relative Risk Table

Once the consistency and performance of a given set of variants and parameter values is established and all other reporting issues have been taken into consideration, our final step is to construct a table of multi-variant genotype relative risk values for every possible combination of genotypes used in the model. To accomplish this, we input all possible combinations of genotypes into REGENT.predict, and appropriately format the output data elements critical for reporting. All considerations discussed above are thus represented in this single, authoritative table which is incorporated into the standard CPMC reporting infrastructure and used to deliver participant-specific estimates.

How common?

As part of the educational materials for each complex condition reported within the CPMC, estimates of the frequency of specific observations (e.g. the percentage of the population homozygous for the risk allele) will be provided to participants. In multi-variant reports these values are derived from REGENT.model's proportion of population assigned to each risk category (increased risk, average, decreased risk).

Section 5: Summary and Conclusions

Given the nature of CPMC goals and the attendant operational and theoretical constraints, we begin with the premise that approved genetic variants are valid risk factors for the condition of interest and that the peer-reviewed and ICOB-approved method for combining their effects is sound. In applying the method, the CPMC will include as many independent, ICOB-approved loci as possible and utilize several sources of data to characterize each condition-specific model. We will calculate and report model performance metrics including the AUC, median risk scores for cases and non-cases and percent category membership from CPMC, dbGaP derived datasets, and/or other sources when feasible. Minimally, the stability of each model will be assessed via simulation. Individuals will always receive all genotype data and risk scores for those individual variants from the model which were successfully typed in their sample, but only those with complete data (ensured to be >= 95% of the cohort at release) will be given a multi-variant risk estimate. Further, scores from the multi-variant model will be contextualized using category membership ("*decreased risk*", "*average risk*", and "*increased risk*") and relative risk values in the extremes of the distribution will be expressed as greater than 10.00 or less than 0.10 when appropriate. Finally, the details provided here and other relevant information on multi-variant risk calculation and reporting will be incorporated into condition-specific appendices which will be added to this document to allow for evaluation and comment by the wider research community.

Works Cited

Crouch, Daniel JM, Graham HM Goddard, and Cathryn M Lewis. REGENT: a risk assessment and classification algorithm for genetic and environmental factors. *Eur J of Hum Genet* 2013; 21, 109–111.

Goddard GH, Lewis CM, Risk categorization for complex disorders according to genotype relative risk and precision in parameter estimates. *Genet Epidemiol* 2010; 34: 624–632.

Kundu, Suman, Yurii S Aulchenko, Cornelia M van Duijn, and Cecile JW Janssens. PredictABEL: an R package for the assessment of risk prediction models. *Eur J Epidemiol* 2011; 26(4): 261–264.

Kundu S, Karssen LC, Janssens AC. 2012. Analytical and simulation methods for estimating the potential predictive ability of genetic profiling: a comparison of methods and results. Eur J Hum Genet. 20(12):1270-1274.

Janssens, A Cecile JW and Muin J Khoury. Assessment of improved prediction beyond traditional risk factors: when does a difference make a difference? *Circ Cariovasc Genet* 2010; 3:3-5.