Precision medicine and real-world evidence

Our study joins the efforts to leverage real-world data with uncertain genetic information. Inclusion of ungenotyped probands in analysis can help uncover real-world evidence on the factors in the power gain under more complex scenarios. Further use of simulations to mimic more real-world scenarios and explore important factors in the power gain under more complex scenarios.

Results

- Simulated and augmented data
  - Full data (FD)
  - Simulated benchmark
  - Fully simulated data with the true values of the subjects who would be unavailable for genotyping in reality
  - For example, 4 subjects per family have phenotypic and genotypic data (Figure 4)
  - Porable (PD)
  - Partial data, excluding the subjects who would be unavailable for genotyping
  - For example, 3 subjects per family have phenotypic and genotypic data (Figure 3)
  - Data augmentation (DA)
    - Augmented data with the inferred genotypic data of the subjects who are ungenotyped in reality
    - For example, 3 subjects per family have phenotypic and genotypic data and 1 subject per family has phenotypic data

- Statistical results
  - Some mistakes in error rates are well maintained
  - MCAR in genetic data
    - Proportion with the ratio of test statistics
      - 92% ± 10%
      - All simulated scenarios (Table 2)
  - MAR in genetic data
    - Bias in effect-estimations based on PD was completely removed or mitigated by DA
    - Quantitative or dichotomous outcomes
      - Bias is defined as the difference between the true effect size and the effect estimator based on PD, DA, or FD (the smaller the difference, the less the bias).
      - The median difference between the true effect size and the effect estimator was reduced from 0.03 to 0.01
  - DA versus PD
    - Power is increased with DA
      - e.g. the median test statistic ratio of DA versus PD is 1.22; the corresponding median of FD versus PD is 1.44 (simulated benchmark)

Conclusions and Implications

- This study joins the efforts to address concerns with bias and limited power in real-world data
  - Phenotypic or dichotomous outcomes
    - Data augmentation with genetic inheritance laws
    - Discovering real-world evidence with scientific computing
    - This statistically significant discoveries for biologically relevant considerations

- Power gain formulas
  - To facilitate design real-world genetic studies where clinical trials are practically infeasible
    - Cost-effectiveness
    - Embrace the advancements in biomedical technologies, e.g. cell line developments to help completely address safety concerns, therapeutic treatment optimization and the prevention of disease and adverse effects
  - The additional inclusion of ungenotyped probands in study design and analysis can help uncover the effects of genetic variants on biological outcomes, responses, such as to toxicity and infectious agents
  - Simulations showed the importance to incorporate ungenotyped probands in study design and analysis, especially for MAR genetic data
  - For unbiased or less biased effect estimates
    - For increased test statistics and statistical power
  - This power analysis and data augmentation study complements standard and state-of-the-art power and sample size software, e.g. PASS (https://www.ncss.com/)

References


Figure 1: Nucleotide A, T, G, and C

Figure 2: Table 1

Figure 3: Table 2

Figure 4: Table 3