Power Analysis and Data Augmentation for Real World Evidence with Uncertain Genetic Information

Wei (Vivian) Zhuang[1] and Joshua Xu[1]

[1]Division of Bioinformatics and Biostatistics, National Center for Toxicological Research, FDA

Abstract

- **Introduction:** Pharmacogenomics and genetic studies offer great promise in precision medicine to improve scientific knowledge on how genes affect a person's responses to certain drugs or exposures. In some real-world genetic studies, probands revealed interesting adverse events or phenotypes, but died before providing DNA (due to aging, rapidly progressing lethal diseases, or other reasons). The phenotypic and genotypic data of their relatives are often available or accruable.

- **Methods:** For a simple and realistic occurrence with probands missing genotypes completely at random, a power gain formula for a dichotomous outcome was derived with family data in the case of a single type of relative, such as a parent and a child. With the theoretical power gain result, we further used simulations to mimic more real-world scenarios and explore important factors in the power gain under more complex scenarios.

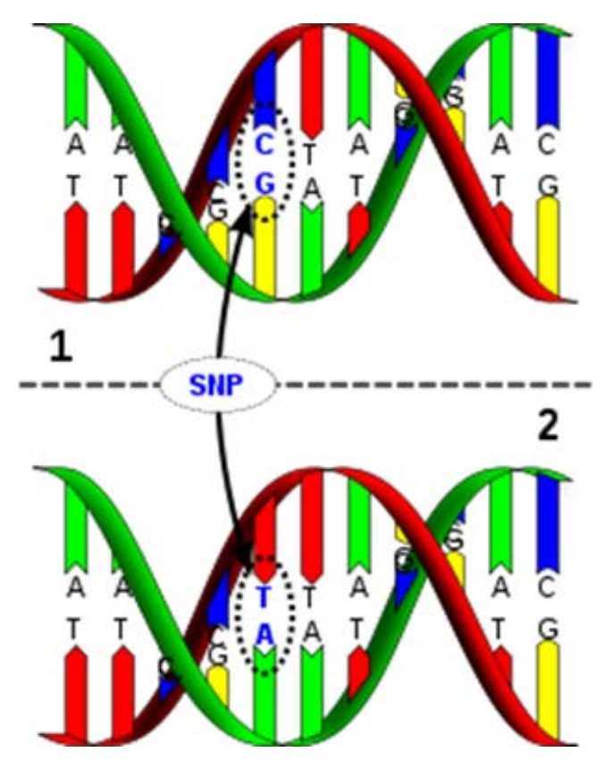
- **Results:** The power gain formula shows that statistical power can be increased when ungenotyped probands are included in analysis. The data augmentation method that includes ungenotyped probands in analysis helps improve power to discover real world evidence with uncertain genetic information. The missingness mechanism, study design, phenotypic heritability, genetic variation frequency, and genetic variation specific heritability are important factors in the changes in power and effect estimates.

- **Conclusion/Implications:** Our study joins the efforts to leverage real world data with uncertain genetic information for genetic and pharmacogenomic research. Our results show that the inclusion of ungenotyped probands in analysis can help uncover real world evidence on the effects of genetic variants on biological outcomes or responses, such as to toxicity and infectious agents.

Introduction

- Precision medicine
  - Genetic studies
  - Pharmacogenomics
    - Improve scientific knowledge
      - Genetic factors in a person's responses to
        - Drugs
        - Exposures
- Discover evidence in real-world genetic studies
  - Therapeutic treatment
    - Best-in-class drug
    - Optimal dose, optimal treatment length, etc.
  - Help prevent adverse events

Figure 1



Nucleotides: A, T, G, and C
Pair: A-T and G-C
Single-nucleotide polymorphism (SNP)

- In some real-world studies or clinical settings, probands revealed adverse events or interesting phenotypes, but died before providing DNA data, e.g. SNP data (Figure 1; the image source: https://isogg.org/wiki/Single-nucleotide_polymorphism)
  - Aging, rapidly progressing lethal diseases, etc.
  - The phenotypic and genotypic data of their relatives (e.g. siblings and offspring) are often available or accruable
    - Relatives' cell lines

- SNP genotyping methods
  - Sequencing, PCR-based methods, array-based hybridization, etc.
  - Used in real-world cross-sectional, case-control, longitudinal studies

Challenges in Data Analysis

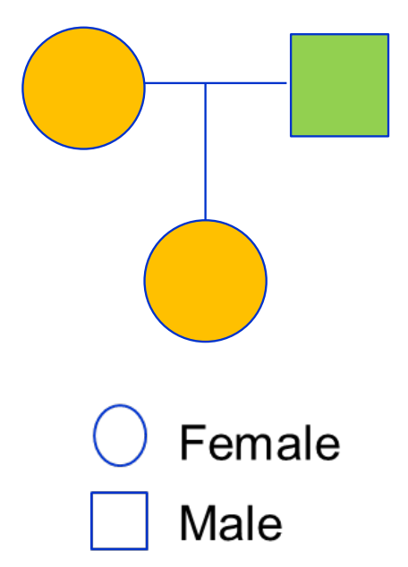
- Should we augment probands' genetic data?
  - How to properly handle missing genetic data?
  - In order to discover differential changes between groups in real-world genetic studies

- Important factors to power/error gain
  - Family-level parameters
  - Population-level parameters
  - Missingness mechanism
    - MCAR (missing completely at random)
      - Whether an observation is missing or observed is completely random
    - MAR (missing at random)
      - Bias may be removed by maximizing the likelihood using all observed data including those missing in the covariate (Little, 2019).
    - NMAR (not missing at random)
      - Neither MCAR nor MAR

Theoretical Proofs

- Simple and realistic data scenarios
  - A single type of relative per family, such as a parent and a child
    - Two subsets
      - Individuals are independent in each subset
        - One subject per family
          - For example, parent subset (e.g. Figure 2; Orange)
            - Phenotypes
            - Genotypes completely missing at random (MCAR)
          - For example, offspring subset (e.g. Figure 2; Orange)
            - Phenotypes
            - Genotypes

Figure 2



- Power analysis
  - Non-centrality parameter (NCP)
    - Related to power
      - e.g. power calculations for ANOVA in PASS 16
    - NCP can increase more than 20%
      - Include ungenotyped parents

  - Quantitative traits
    - Least square method
  - Dichotomous traits
    - Weighted Least square method
  - Pure genetic additive model

Table 1

Outcome Type	$\frac{NCP(Genotyped + Ungenotyped)}{NCP(Genotyped)}$
Quantitative traits*	$\frac{1-2\rho r+r^2}{1-\rho^2}$
Dichotomous traits	$\frac{1-2\rho r_w+{r_w}^2}{1-\rho^2}$

ρ is phenotypic correlation, r denotes relationship between relatives
r_w depends on the relationship between relatives, allele frequency, and variance of the outcome given covariants
*This equation was published by Visscher and Duffy (2006)

Simulation Design

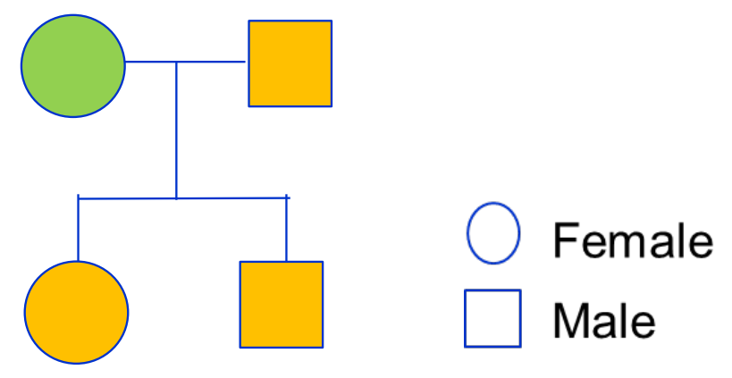
- Phenotypic heritability
  - Proportion of phenotypic variance that is accounted for by genetic factors
    - Analogous to R^2, proportion of phenotypic variance that is accounted for by predictor variables in linear regression
    - Relatives' cell lines
      - Chemotherapy cytotoxicity is heritable, varying with dose (Watters et al. 2004)
        - E.g. cytotoxicity to the mechanistically distinct chemotherapy agent 5-fluorouracil
          - 0.26 to 0.65

Table 2

Factors/Parameters for Simulation	**Values Based on Real-world Studies**
Phenotypic heritability (H^2)	0.1, 0.3, or 0.5
Coefficient of relationship (r)	0.5
Risk allele frequency (f)	0.05, 0.2, or 0.4
Effect size of risk allele (h^2)	0.01 – 0.09 by 0.02
Probability of dichotomous traits (P_d)	0.1, 0.3, or 0.5

- Realistic data scenarios
  - MCAR
  - MAR
    - Pedigree of 4 persons
    - A nuclear family
      - The parent with the disease of interest (e.g. a fatal/serious adverse event)
        - Tends to be ungenotyped (e.g. Figure 3; Green)

Figure 3



❖ $\Pr(G_i=. \mid D_i=1) = 0.5P_d + P_g(1-P_d)$ for parent i
❖ P_g: from 0.6 to 1 by 0.1

Acknowledgements
The authors wish to thank their Division Director Dr. Weida Tong for his great support on this work. This work integrated the authors' research at NCTR with Dr. Wei (Vivian) Zhuang's previous dissertation work in the School of Public Health at Boston University, Boston, Massachusetts.

Financial competing & interests disclosure
All authors are currently employed by NCTR, US FDA. The authors have no other financial relationships or relevant affiliations with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the poster other than those disclosed.

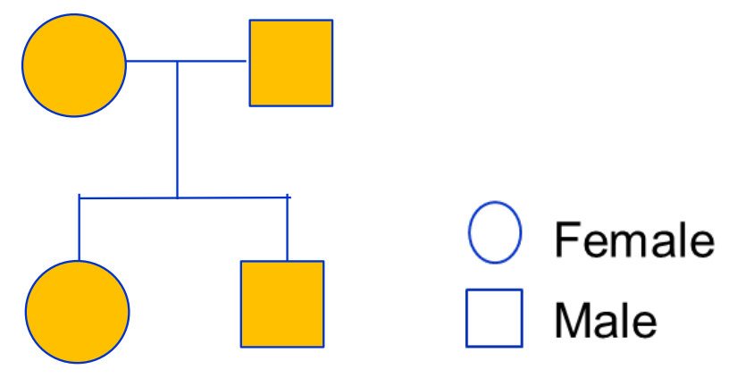
FDA disclaimer
The information presented in this poster is not a formal dissemination of information by the US Food & Drug Administration and does not represent agency position or policy.

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)
  - Partial data (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**
  - The nominal type I error rates are well maintained
  - MCAR in genetic data
    - Proportion with the ratio of test statistics ≥1
      - 82.2%-100%
        - 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
          - PD versus DA
            - H^2=0.3, f=0.2, h^2=0.09, P_d=0.3, and P_g=1 (Table 2; Figure 3)
    - The test statistic ratios
        - Mean and Median > 1
          - DA versus PD and FD versus PD
            - Statistical 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)
              - H^2=0.3, f=0.2, h^2=0.09, P_d=0.3, and P_g=1 (Table 2; Figure 3)

Figure 4



Conclusions and Implications

- This study joins the efforts to address concerns with bias and limited power in real-world data
  - Helping mitigate the barriers in real-world genetic data
    - Data augmentation with genetic inheritance laws
  - Discovering real-world evidence with scientific computing
    - More statistically significant discoveries for biologically relevant consideration

- 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 events
    - 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

1. Little R. Statistical analysis with missing data (3rd Edition). John Wiley & Sons, Inc, Hoboken, NJ (2019).
2. Liu, J. Z., Erlich, Y., & Pickrell, J. K. (2017). Case-control association mapping by proxy using family history of disease. *Nature Genetics*, 49(3), 325-331. doi:10.1038/ng.3766
3. Marchenko, O., Russek-Cohen, E., Levenson, M., Zink, R. C., Krukas-Hampel, M. R., & Jiang, Q. (2018). Sources of Safety Data and Statistical Strategies for Design and Analysis: Real World Insights. *Therapeutic Innovation & Regulatory Science*, 52(2), 170-186. doi:10.1177/2168479017739270
4. Watters, J. W., Kraja, A., Meucci, M. A., Province, M. A., & McLeod, H. L. (2004). Genome-wide discovery of loci influencing chemotherapy cytotoxicity. *Proceedings of the National Academy of Sciences of the United States of America*, 101(32), 11809-11814. doi:10.1073/pnas.0404580101