Hindawi Publishing Corporation
Journal of Probability and Statistics
Volume 2012, Article ID 805426, 3 pages
doi:10.1155/2012/805426

Editorial

Advanced Designs and Statistical Methods for Genetic and Genomic Studies of Complex Diseases

Yongzhao Shao,[1] Wei Pan,[2] and Xiaohua Douglas Zhang[3]

[1] *Division of Biostatistics, New York University School of Medicine, 650 First Avenue, No. 538, New York, NY 10016, USA*
[2] *Division of Biostatistics, School of Public Health, University of Minnesota, A460 Mayo Building, MMC 303 Minneapolis, MN 55455, USA*
[3] *Biometrics Research, Merck Research Laboratories, WP53B-120, West Point, PA 19486, USA*

Correspondence should be addressed to Yongzhao Shao, shaoy01@nyumc.org

Received 21 August 2012; Accepted 21 August 2012

The completion of the Human Genome Project and the International HapMap Project, coupled with rapid advancement of high-throughput biotechnologies including next-generation sequencing (NGS), has facilitated the discovery of genetic and genomic variants linked to many human diseases. Massive amounts of data from genetic and genomic studies provide a great opportunity for researchers to investigate and propose novel statistical methods and algorithms that can effectively identify disease-associated or causal genetic/genomic markers while avoiding an abundance of false positive results.

Despite many recent advances in statistical designs and methods for the analysis of genetic and genomic data on complex diseases, numerous challenges remain. For example, complex diseases including many cancers are heterogeneous in both disease phenotypes and disease etiology. The specification of disease phenotype and measurement of risk factors or environmental exposures are often subject to missing data, measurement errors, or over-simplification. Disease susceptibility is often affected by heterogeneous genetic or genomic factors including rare variants and further altered by various environmental exposures. Therefore, novel study designs and analysis methods are essential for proper adjustment of latent heterogeneity, and for robust inferences using data with possible misspecification of disease phenotypes, incompletely measured exposures, or other complexities.

This special issue is devoted to original research articles as well as overview papers that propose and discuss innovative study designs, novel probabilistic and statistical models, and analysis methods and/or algorithms for genetic and genomic studies of complex diseases. In particular, genome-wide association studies (GWAS) provide an important screening approach to identify single nucleotide polymorphisms (SNPs) and pathways that

underlie complex diseases and traits without requiring prior knowledge about disease-associated chromosomal loci or genetic functions. There are several papers in this special issue that contribute novel and innovative statistical methods for the design, analysis, and prioritization of GWAS results. The paper by, J. Zhao and Z. Chen introduces a two-stage penalized logistic regression approach to case-control genome-wide association studies. While the common practice is to examine each SNP separately, ignoring correlation among the SNPs, the proposed method takes into account correlations among the vast number of SNPs to select etiologically important SNPs. The paper by J. H. Zhao and J. Luan provides an indepth review of mixed models with whole genome data which can deal with complex dependencies introduced by known or unknown familial relationships. Controlling false discovery rate (FDR) or false discovery proportions (FDPs) is one of the most fundamental statistical issues for GWAS and other genetic and genomic studies involving testing a large number of hypotheses. Controlling FDR, as suggested by Benjamini and Hochberg among many others, has been the most widely studied approach; however, the FDR is only the mathematical expectation of the false discovery proportion (FDP) which can be more directly relevant to specific studies. Direct control of the random variable FDP has recently attracted much attention. The paper by Y. Ge and X. Li proposed an upper bound to directly control the FDP under the assumption of independence among some test statistics. The paper by S. Shang et al. develops statistical designs including sample size calculations that can control the FDP to a prescribed level and achieve some desired overall power of making a desired percentage of true discoveries under some semiparametric assumptions of weak dependence between test statistics. For design studies involving testing a vast number of hypotheses, the paper by C.-H. Tseng and Y. Shao evaluates the growth rate of the required sample size as the number of hypotheses to be tested grows rapidly from 10 to 10 billion. The paper by P. Liu and C. Wang introduces a semiparametric optimal testing (SPOT) procedure for high-dimensional data with a small sample size as arising in microarray and RNA-seq experiments. The SPOT procedure is robust because it does not depend on any parametric assumption for the alternative means. The problem of high-dimensional data with a small sample size is also tackled by the paper of S. Kwon et al. where a multinomial ordinal probit model with singular value decomposition is developed for testing a large number of single nucleotide polymorphisms (SNPs) simultaneously for association with a multidisease status or multinomial trait. Indeed, several groups of researchers have been developing statistical methods that can effectively deal with multivariate outcomes, these novel methods and algorithms are important for genetic and genomic studies and are reviewed in the paper by Q. Yang and Y. Wang. Motivated by studying the genetic basis of Huntington's diseases, T. Chen et al. propose methods for the prediction of disease onset from mutation status using proband and relative data. Using prostate cancer as a prototype example, the paper by Pearlman A. et al. provides a case study of translational research, where genomic copy number alterations (CNA) are being clustered to build a metastatic potential score towards the development of statistical prediction models for the risk of metastasis at the time of primary tumor diagnosis. The paper of H. Jia and J. Li proposes a novel computational method that combines sequence overrepresentation and cross-species sequence conservation to detect transcriptional factor binding sites (TFBSs) in the upstream regions of a given set of coregulated genes. In modeling heterogeneity for many applications, testing homogeneity is an interesting and challenging question even in parametric context. The paper by Liu et al. develops an empirical likelihood-based method for the problem of testing homogeneity in a semi-parametric two-sample problem. Missing parental genotype data is quite common for linkage analysis

particularly for late-onset diseases, the paper by J. Han and Y. Shao introduces a method for reconstruction of parental genotypes and a transmission/disequilibrium heterogeneity (RC-TDH) test for fine mapping of complex diseases. The RC-TDH test extends the current classic transmission/disequilibrium test (TDT) or RC-TDT. The paper by J. Xu reviews the high-dimensional Cox regression analysis in genetic/genomic studies with censored survival outcomes. Gene-environmental interactions are important for studying complex diseases as evidenced by the well-known fact that about 80% of lung cancer patients are smokers. However, measurements on environmental factors are often misclassified or with measurement error. The paper by I. Lobach and R. Fan proposes methods for genotype-based Bayesian analysis of gene-environment interactions with multiple genetic markers and misclassifications in environmental factors. Next-generation sequencing (NGS) has been increasingly used in genetic/genomic studies to investigate roles of rare genetic variants. The paper by T. Wang et al. introduces some novel statistical designs and analysis methods for analyzing pooled sequencing data for rare variants.

The editors of this special issue would like to thank the large number of authors who have shared with them their research achievements. They sincerely thank the large number of professional and diligent referees whose great efforts resulted in rapid reviews and useful feedbacks incorporated into the revisions of the herein published papers. Without their generous contributions this special issue would not be possible.

Yongzhao Shao
Wei Pan
Xiaohua Douglas Zhang