Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Robust tests for gene–environment interaction in case-control and case-only designs

Yong Zang [a,b,*,1], Wing Kam Fung [c], Sha Cao [a,b], Hon Keung Tony Ng [d], Chi Zhang [b,e]

[a] *Department of Biostatistics, Indiana University, Indianapolis, IN, USA*
[b] *Center of Computational Biology and Bioinformatics, Indiana University, Indianapolis, IN, USA*
[c] *Department of Statistics and Actuarial Science, University of Hong Kong, Hong Kong, China*
[d] *Department of Statistical Science, Southern Methodist University, Dallas, TX, USA*
[e] *Department of Medical and Molecular Genetics, Indiana University, Indianapolis, IN, USA*

ARTICLE INFO

Article history:
Received 25 May 2017
Received in revised form 22 May 2018
Accepted 20 August 2018
Available online xxxx

Keywords:
gene–environment interaction
Robust test
Genetic model
Case-control design
Case-only design

ABSTRACT

The case-control and case-only designs are commonly used to detect the gene–environment (G–E) interaction. In principle, the tests based on these two designs require a pre-specified genetic model to achieve an expected power of detecting the G–E interaction. Unfortunately, for most complex diseases the underlying genetic models are unknown. It is well known that mis-specification of the genetic model can result in a substantial loss of power in the detection of the main genetic effect. However, limited effort has been dedicated to the study of G–E interaction. This issue has been investigated in this article with a conclusion that the genetic model mis-specification can not only undermine the power of detecting G–E interaction in both case-control and case-only designs but also distort the type I error rate in case-control design. To tackle this problem, a class of robust tests, namely MAX3, have been proposed for both the case-control and case-only designs. The proposed tests can well control the type I error rate and yield satisfactory power even when the genetic model is mis-specified. The asymptotic distribution and the *p*-value formula for MAX3 have also been derived. Comprehensive simulation studies and a real data application on the genome-wide association study (GWAS) have been conducted using these analytical tools and the results demonstrate desirable operating characteristics of the proposed robust tests.

1. Introduction

Rapid development in human genetics and epidemiology has revealed that genetic susceptibility and environmental exposures play a synergistic role in many complex diseases. This understanding has boosted the development of gene–environment (G–E) interaction study in population genetics, which investigates the joint genetic and environmental interactive effect on the risk of developing diseases (Hunter, 2005). The case-control design has been commonly used to detect the G–E interaction, where the interactive effect can be conveniently modeled by a multiplicative term of genotypes and exposure levels based on a prospective logistic regression model. However, in such design, samples are classified by both the genotypes and exposure levels, which may result in a substantial loss of power (Mukherjee et al., 2012; Marigorta

[*] Corresponding author at: Department of Biostatistics, Indiana University, Indianapolis, IN, USA.
E-mail address: zangy@iu.edu (Y. Zang).

[1] 410 West 10th St., Suite 5000, Indianapolis IN 46202.

https://doi.org/10.1016/j.csda.2018.08.014
0167-9473/© 2018 Published by Elsevier B.V.

Table 1
Case-control data with a diallelic marker G and a binary environmental exposure factor E.

	$G=0$		$G=1$		$G=2$		Total
	$E=0$	$E=1$	$E=0$	$E=1$	$E=0$	$E=1$	
$D=0$	r_{000}	r_{001}	r_{010}	r_{011}	r_{020}	r_{021}	m_0
$D=1$	r_{100}	r_{101}	r_{110}	r_{111}	r_{120}	r_{121}	m_1
Total	n_{00}	n_{01}	n_{10}	n_{11}	n_{20}	n_{21}	n

and Gibson, 2014). Alternatively, under the assumption of G–E independence and rare disease, the G–E interaction can be evaluated by simply assessing the G–E association on the cases only. Such case-only design can yield a higher power than a case-control design when these assumptions hold (Piegorsch et al., 1994; Umbach and Weinberg, 1997).

In both case-control and case-only designs, if the genetic model of inheritance can be specified a priori, then a score test can be performed to detect the G–E interaction. The genetic model determines the orders of individuals' risk of having the disease based on the number of risk alleles in the genotype. Generally speaking, for a diallelic marker, three genetic models, namely the recessive (REC), multiplicative (MUL) and dominant (DOM) are commonly used (Sasieni, 1997; Freidlin et al., 2002). For each genetic model, an optimal set of scores should be used to maximize the power of the test. In particular, the value 0, 1/2 and 1 are the optimal scores to code the genotype conferring one risk allele when the genetic model is REC, MUL and DOM, respectively (Zheng et al., 2003). Hence, if the genetic model is correctly specified, the corresponding optimal scores can maximize the power of the score test. However, for many complex diseases, the underlying genetic models are unknown and using an inappropriate genetic model can substantially undermine the power of the tests (Zheng et al., 2003). Therefore, robust tests against genetic model mis-specification are in urgent demand.

Despite intensive studies on robust tests for detecting the main genetic effect (Wang and Sheffield, 2005; Gonzalez et al., 2008; Zheng and Ng, 2008; Yamada and Okada, 2009; Zang et al., 2010a, b), little has been dedicated for the G–E interaction effect. Hence, the purpose of this paper is to fill this research gap. Specifically, we first investigate the impact of genetic model mis-specification on testing the G–E interaction. Interestingly, we find that the genetic model mis-specification highly affects the case-control design by distorting both the type I error rate and power, but only decreases the power for the case-only design. Furthermore, to handle the genetic model uncertainty, we have developed robust tests for both designs. The asymptotic formulas to calculate the p-value of the robust tests together with an user-friendly software are also released in this paper to facilitate the use of the proposed methods in practice. Simulation study demonstrates that the proposed robust tests could control type I error rate under the null hypothesis and yet yield satisfactory power under the alternative hypothesis, even when the genetic model is mis-specified. The proposed method is also applied to a real genome-wide association study (GWAS) dataset for illustrative purpose.

The rest of this paper is organized as follows. We develop the robust tests for the case-control design and case-only design in Sections 2 and 3. In Sections 4 and 5 we extend the proposed tests to handle non-monotonic genetic model and categorical environment factor with possible environmental level mis-classification. In Section 6, we carry out comprehensive simulation studies to investigate the operating characteristics of the proposed tests. In Section 7, we apply the robust tests to analyze a genome-wide association study (GWAS) of bladder cancer (Rothman et al., 2010). We provide a brief discussion and concluding remarks in Section 8.

2. Robust test for case-control design

Assume m_1 cases and m_0 controls being genotyped in a case-control study and let $n = m_0 + m_1$ denoting the total sample size. For ease of presentation, we consider a binary environmental factor E and a diallelic marker G, for which we are interested in testing the impact of the gene–environmental (G–E) interaction effect on the disease risk. Let $G = 0, 1, 2$ denote the three genotypes aa, Aa and AA with A indicating the minor allele conferring high risk of the disease. Let $E = 0(E = 1)$ denote an unexposed (exposed) individual. Let D denote the disease status with $D = 0(D = 1)$ representing an unaffected (affected) individual. The case-control data can be displayed in the form of a 2 × 6 table as presented in Table 1. As expressed in Table 1, we use r_{ijk} to denote the number of individuals with $D = i$, $G = j$ and $E = k$ and define $n_{jk} = r_{0jk} + r_{1jk}$.

Let D_l, G_l and E_l be the phenotype, genotype and environmental factor for the lth sample in case-control study. We define $f_{jk} = \Pr(D_l = 1|G_l = j, E_l = k)$ as the penetrance level conditional on $G = j$ and $E = k$, by which the recessive (REC), multiplicative (MUL) and dominant (DOM) genetic models correspond to $f_{1k} = f_{0k}$, $f_{1k} = \sqrt{f_{0k}f_{2k}}$ and $f_{1k} = f_{2k}$ for $k = 0{,}1$ respectively (Sasieni, 1997).

According to the definition, when the genetic model is specified, the impact of G_l and E_l on the disease status D_l can be conveniently formulated by the following logistic regression model:

$$\log\left(\frac{\Pr(D_l = 1|G_l, E_l)}{\Pr(D_l = 0|G_l, E_l)}\right) = \alpha + \delta E + x(\beta + \lambda E)\mathrm{I}(G = 1) + (\beta + \lambda E)\mathrm{I}(G = 2), \tag{1}$$

where $\mathrm{I}(\cdot)$ is an indicator function, δ is the main environmental effect, β is the main genetic effect, λ is the G–E interaction effect and x is a real number between 0 and 1 representing the underlying genetic model. The interest here is to test the null hypothesis $H_0 : \lambda = 0$. It is straightforward to see that $x = 0$ and 1 correspond to the REC and DOM models, respectively. If

Please cite this article in press as: Zang Y., et al., Robust tests for gene–environment interaction in case-control and case-only designs. Computational Statistics and Data Analysis (2018), https://doi.org/10.1016/j.csda.2018.08.014.