Bioinformatics, 2016, 1–3
doi: 10.1093/bioinformatics/btw391
Advance Access Publication Date: 4 July 2016
Applications Note

OXFORD

Genetics and population analysis

TwoPhaseInd: an R package for estimating gene–treatment interactions and discovering predictive markers in randomized clinical trials

Xiaoyu Wang[1] and James Y. Dai[1,2,*]

[1]Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA, USA and [2]Department of Biostatistics, University of Washington, Seattle, WA, USA

*To whom correspondence should be addressed.

Associate Editor: Oliver Stegle

Received on March 17, 2016; revised on May 24, 2016; accepted on June 16, 2016

Abstract

Summary: In randomized clinical trials, identifying baseline genetic or genomic markers for predicting subgroup treatment effects is of rising interest. Outcome-dependent sampling is often employed for measuring markers. The R package TwoPhaseInd implements a number of efficient statistical methods we developed for estimating subgroup treatment effects and gene–treatment interactions, exploiting the gene–treatment independence dictated by randomization, including the case-only estimator, the maximum estimated likelihood estimator and the semiparametric maximum likelihood estimator for parameters in a logistic model. For rare failure events subject to censoring, we have proposed efficient augmented case-only designs, a variation of the case–cohort design, to estimate genetic associations and subgroup treatment effects in a Cox regression model. The R package is computationally scalable to genome-wide studies, as illustrated by an example from Women's Health Initiative.

Availability and Implementation: The R package TwoPhaseInd is available from http://cran.r-project.org/web/packages.

Contact: jdai@fredhutch.org

1 Introduction

Depending on genetic background and clinical characteristics, individuals respond differently to treatment or prevention. In randomized clinical trials, there is a rising interest in identifying baseline biomarkers that predict subgroup or individual treatment effects, for example pharmacogenenetics or pharmacogenomics studies, that may lead to personalized or precision medicine. A common design of such studies is to measure predictive biomarkers of treatment efficacy from achieved specimens after trial completion (Simon *et al.*, 2009). The genetic or genomic markers are often expensive to measure, particularly for genome-wide high-throughput assays such as whole-genome sequencing. The outcome-dependent sampling, such as case–control or case–cohort sampling are widely used for cost efficiency in this type of retrospective studies.

Efficient statistical methods have been developed for estimating gene–treatment interaction and subgroup treatment effects for a range of outcome-dependent sampling schemes, exploiting the gene–treatment independence dictated by randomization (Dai *et al.*, 2009, 2012, 2014, 2016). The efficiency gain can be as much as 50% when compared to approaches not using the gene–treatment independence. For rare events often studied in prevention trials, the case-only estimator has been advocated: biomarkers are measured only in the cases, yet subgroup treatment effects and gene–treatment interactions are estimated with the same efficiency as the full cohort approach where all participants are measured. For not-so-rare dichotomized trial endpoints (for example cancer therapeutic trials) and case–control sampling for biomarkers (Dai *et al.*, 2012, 2014), semiparametric maximum likelihood estimators and maximum estimated likelihood estimators have been developed with efficient algorithms (Dai *et al.*, 2009). We recently proposed augmented case–

only designs for trials with rare events subject to censoring, in which genetic/biomarker main effects are also of interest (Dai *et al.*, 2016).

The R package TwoPhaseInd assembles a number of functions to compute the estimates and provide standard error and *P*-values for subgroup treatment effects and gene–treatment interactions in various aforementioned study settings. The 'TwoPhase' part in the name of the package refers to the retrospective sampling for biomarker measurement after the completion of the prospective trial, and the 'Ind' part refers to the unifying theme in implemented methods that we exploit gene–treatment independence to improve the estimation efficiency. Data examples are included in the package and presented in an R package vignette.

2 Methods

Consider an ancillary biomarker study for a two-arm randomized clinical trials with treatment assignment Z ($Z=1$ if investigational treatment, $Z=0$ if control treatment), disease endpoint Y and baseline biomarker G. For ease of exposition, additional covariates are omitted. Suppose an outcome-dependent design was employed for measuring G in archived samples. The R package TwoPhaseInd provides functions for each of the designs below:

2.1 Case-only estimator

The case-only design can be used to estimate the gene–treatment interaction and subgroup treatment effects in trials when the disease endpoint is rare (for example the prevalence of the event ≤ 0.05). The following logistic regression model is fitted to biomarker data in cases only by the function caseonly: $\text{Logit}\{\mathbb{E}(Z|G, Y=1)\} = \log\{\frac{\pi}{1-\pi}\} + \beta_0 + \beta_1 G$, where π is the randomization fraction to the investigational treatment arm. The resulting estimates for β_0 and β_1 are the subgroup effect when $G=0$ and the gene–treatment interaction, respectively.

2.2 Estimators for case–control sampling

For case–control studies to assess the effect modification of baseline biomarkers, where the event rate is not necessarily rare, semiparametric models can be used to estimate subgroup treatment effects and gene–treatment interactions (Dai *et al.*, 2009). The data structure has essentially a two-phase sampling form: the first-phase data contain the treatment assignment and the disease outcome known for every subject; the second-phase data contain the biomarker measured for a case–control subsample, and possibly a collection of adjusting covariates (for example eigen vectors from principal component analysis of genome-wide genetic data). We denote X to be the collection of covariates measured in the case–control samples, including the biomarker G. The semiparametric likelihoods using the independence $X \perp Z$ can be written as $L^{\perp}(\beta, G) = \prod_{j\in V} f_\beta(y_i|x_i, z_i) g_{x_i} \prod_{j\in\bar{V}} \left(\sum_{x_l\in\mathcal{X}} g_{x_l} f_\beta(y_j|x_l, z_j)\right)$, where $f_\beta(y|x, z)$ is the parametric regression model with parameters β, which often takes the form of a generalized linear model; $g(x)$ is the density function for x irrespective of z. We assume sampling takes place so that only a subset of subjects have X measured in phase two, and $\mathcal{X}$ denotes the set of observed X regardless of Z. Because of the orthogonality of Z and X, the density function of X does not have to condition on Z, thereby improving efficiency. Two functions are available to estimate the β parameter of the regression model. The first function spmle applies a profile likelihood based on the Newton–Raphson algorithm to compute semi-parametric maximum likelihood estimate, and the second one mele computes maximal estimated likelihood estimator. The details can be found in Dai et al. (2009).

2.3 Estimators for case–cohort sampling

For usual failure time endpoints subject to censoring, the case-only estimator still applies when the event is rare. However the genetic main effect (as a prognostic marker) cannot be assessed by case-only design. We proposed augmented case-only designs to supplement controls to achieve estimation for all parameters in a Cox model (Dai *et al.*, 2016). The R package TwoPhaseInd has a function acoarm to implement a multi-step method that employs classical case–cohort estimation methods, but incorporating the case-only estimators. Specifically, consider a proportional hazards model with gene–treatment interaction $\lambda(t; Z, G, V) = \lambda_0(t)\exp(\beta_1 G + \hat{\beta}_{2co} Z + \hat{\beta}_{3co} GZ + \beta_4 V)$, where $\lambda_0(t)$ is a baseline hazard function, t denotes time, G denotes the baseline biomarker of interest, V denotes a set of pre-treatment variables to be adjusted in risk association, $\hat{\beta}_{2co}$ and $\hat{\beta}_{3co}$ are case-only estimators. The controls sampled in augmented case–cohort designs are used to obtain estimates of β_1 and β_4, as well as the baseline hazard functions. It considers two scenarios of adding controls to the case-only design: classical case–cohort design with controls drawn from both the treatment arm and the placebo arm, and augmented case-only design with controls draw from one of the two arms (Dai *et al.*, 2016).

3 Data example and illustration

The TwoPhaseInd package can be used in genome-wide studies for gene–treatment interactions. The current version of the package can be used to interrogate gene–treatment interaction for genotypes one at a time. As an example, we illustrated the usages of case-only estimator, the semiparametric maximum likelihood estimator (SPMLE) and the maximum estimated likelihood estimator (MELE) in Women's Health Initiative (WHI) hormone trial. The goal is to estimate the interaction between biomarkers (SNPs) and hormone therapy (estrogen plus progestin) on type II diabetes. A case–control study was conducted to measure genome-wide SNPs. The dataset we illustrated below consists of 1020 diabetes cases and 2127 matched controls, all of which have genome-wide SNPs measured. Figure 1 shows the case-only, the SPMLE and MELE results for interactions of the hormone treatment and 78081 SNPs on chromosome 1. The quantile-quantile plots in the upper panels (Fig. 1A–C) compare the distribution of observed p-values with that of a uniform-distributed p-values. Although there is no significant

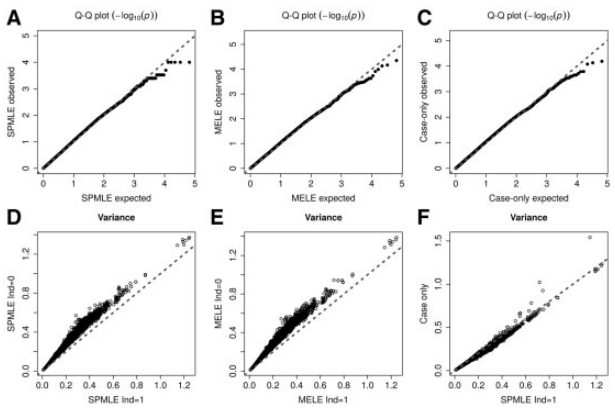

Fig. 1. The results of the three estimators in the WHI study

P-value, the *q*–*q* line is right in the diagonal direction, suggesting the algorithm works well in estimation for all three methods. The first two graphics in the lower panels of Figure 1D,E shows the estimated variances of SNP-treatment interaction, using or without the independence between treatment and the SNP, suggesting that using independence yields a much more precise estimates of interaction. The last graph in the lower panel (Fig. 1F) shows the comparison of the case-only estimator and the SPMLE estimator, suggesting the two agrees well in efficiency of estimation since type II diabetes is relative rare in the WHI hormone trial.

Acknowledgements

The authors thank the members of CRAN team for testing and distributing the package. The authors thank the WHI investigators and staff for their dedication, and the study participants for making the program possible.

Funding

The WHI program is funded by the National Heart, Lung and Blood Institute, National Institutes of Health, U.S. Department of Health and Human Services through contracts HHSN268201100046C, HHSN268201600003C, HHSN268201600002C, HHSN268201600004C, HHSN268201600001C and HHSN271201100004C. This work is supported by the NIH grants P01 CA53996, R01 HL114901.

Conflict of Interest: none declared.

References

Dai,J.Y. *et al.* (2009) Semiparametric estimation exploiting covariate independence in two-phase randomized trials. *Biometrics*, **65**, 178–187.

Dai,J.Y. *et al.* (2012) Two-stage testing procedures with independent filtering for genome-wide gene-environment interaction. *Biometrika*, **99**, 929–944.

Dai,J.Y. *et al.* (2014) Case-only methods for competing risks models with application to assessing differential vaccine efficacy by viral and host genetics. *Biostatistics*, **15**, 196–203.

Dai,J.Y. *et al.* (2016) Augmented case-only designs for randomized clinical trials with failure time endpoints. *Biometrics*, **72**, 30–38.

Simon,R.M. *et al.* (2009) Use of archived specimens in evaluation of prognostic and predictive biomarkers. *J. Natl. Cancer Inst.*, **101**, 1446–1452.