Published in final edited form as:

Hum Genet. 2014 November ; 133(11): 1369–1382. doi:10.1007/s00439-014-1466-9.

Improving the Power of Genetic Association Tests with Imperfect Phenotype Derived from Electronic Medical Records

Jennifer A. Sinnott,

Department of Biostatistics, Harvard School of Public Health, Boston MA 02115, USA

Wei Dai,

Department of Biostatistics, Harvard School of Public Health, Boston MA 02115, USA

Katherine P. Liao,

Division of Rheumatology, Immunology and Allergy, Brigham and Women's Hospital, Boston, Massachusetts, 02115 USA

Stanley Y. Shaw,

Center for Systems Biology, Massachusetts General Hospital, Boston, Massachusetts, 02114 USA

Ashwin N. Ananthakrishnan,

Division of Gastroenterology, Massachusetts General Hospital, Boston, Massachusetts, 02114 USA

Vivian S. Gainer,

Research Computing, Partners Healthcare, Charlestown, Massachusetts, 02129 USA

Elizabeth W. Karlson,

Division of Rheumatology, Immunology and Allergy, Brigham and Women's Hospital, Boston, Massachusetts, 02115 USA

Susanne Churchill,

I2b2 National Center for Biomedical Computing, Boston, Massachusetts 02115, USA

Isaac Kohane,

I2b2 National Center for Biomedical Computing, Boston, Massachusetts 02115, USA; Center for Biomedical Informatics, Harvard Medical School, Boston, Massachusetts, 02115, USA

Peter Szolovits,

Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts, 02139 USA

Shawn Murphy,

Laboratory of Computer Science, Massachusetts General Hospital, Boston, Massachusetts, 02114 USA

Robert Plenge, and

Merck Research Laboratories, Boston, Massachusetts, 02115, USA

Tianxi Cai

Department of Biostatistics, Harvard School of Public Health, Boston MA 02115, USA

Jennifer A. Sinnott: jsinnott@hsph.harvard.edu

Abstract

To reduce costs and improve clinical relevance of genetic studies, there has been increasing interest in performing such studies in hospital-based cohorts by linking phenotypes extracted from electronic medical records (EMRs) to genotypes assessed in routinely collected medical samples. A fundamental difficulty in implementing such studies is extracting accurate information about disease outcomes and important clinical covariates from large numbers of EMRs. Recently, numerous algorithms have been developed to infer phenotypes by combining information from multiple structured and unstructured variables extracted from EMRs. Although these algorithms are quite accurate, they typically do not provide perfect classification due to the difficulty in inferring meaning from the text. Some algorithms can produce for each patient a *probability* that the patient is a disease case. This probability can be thresholded to define case-control status, and this estimated case-control status has been used to replicate known genetic associations in EMR-based studies. However, using the estimated disease status in place of true disease status results in outcome misclassification, which can diminish test power and bias odds ratio estimates. We propose to instead directly model the algorithm-derived probability of being a case. We demonstrate how our approach improves test power and effect estimation in simulation studies, and we describe its performance in a study of rheumatoid arthritis. Our work provides an easily implemented solution to a major practical challenge that arises in the use of EMR data, which can facilitate the use of EMR infrastructure for more powerful, cost-effective, and diverse genetic studies.

Keywords

Case-Control Studies; Electronic Health Records; Electronic Medical Records; Genetic association studies; Outcome Misclassification

1 Introduction

For numerous pressing goals of modern disease genomics, including quantifying the effects of rare variants, gene-gene interactions, and gene-environment interactions, studies with very large sample sizes are essential. As the technology to measure genetic features continues to improve and become less expensive, the costs and timelines of studies become driven by study infrastructure, acquisition of biosamples, and phenotype characterization. Many large genetic studies are nested in traditional cohort studies with banked blood samples; however, such studies are necessarily of restricted size and historically of limited ethnic diversity (McCarthy et al. 2008). To increase size and better reflect current population demographics, genetic studies are being implemented in health care systems with electronic medical records (EMRs) linked to biorepositories (Kohane 2011). Such studies can be extremely cost-effective because they rely primarily on pre-existing infrastructure developed for routine care: genotyping can be performed on discarded biosamples from medical tests, and phenotypes can be extracted from medical records through a combination of computer algorithms and record review by disease experts. Recent EMR-based genetic studies have successfully replicated associations observed in traditional genetic studies (Ritchie et al. 2010; Kurreeman et al. 2011; Kho et al. 2012). They also offer opportunities to extend the

sorts of outcomes available for study to include, for example, adverse drug reactions or treatment response in the context of current clinical practice (Wilke et al. 2011).

One of the primary impediments to using EMRs for genetic studies is the difficulty in extracting accurate information from them on patients' exposures, diseases, and treatments. There are two main types of EMR data: *codified data*, which are entered in a structured format and may include demographic information, laboratory test results, and billing codes, and *narrative data*, which are extracted from free form text such as radiology reports or physicians' notes. Methods using codified data alone are simpler to implement, but can lead to extensive misclassification of disease status which can severely bias results. Extracting precise information from narrative EMR data usually requires the use of natural language processing techniques and typically requires several iterations of algorithm refinement, in which algorithm results are compared with true disease status as assessed by disease experts undertaking time-consuming chart-review. This process can produce excellent phenotype identification algorithms, which can be evaluated using metrics such as the sensitivity, the proportion of true cases being classified as cases; and the positive predictive value (PPV), the proportion of the individuals classified as cases by the disease algorithm who are true cases. For example, in the Electronic Medical Records and Genomics (eMERGE) Network, the algorithms developed to predict seven different case-control phenotypes showed PPVs between 67.6% and 100%, with the majority having PPVs over 90% (Newton et al. 2013). However, as evidenced by these numbers, the predicted disease status is still typically imperfect due to the difficulty in accurately interpreting the content of the text.

After using an algorithm to identify probable cases and controls from EMRs, biological samples linked to those records are genotyped. Typically, each genotyped single nucleotide polymorphism (SNP) is tested for association with case-control status using logistic regression, and the magnitude of the association is estimated. However, because the EMR-estimated case-control status is imperfect, these results will be biased; in general, we expect reduction of test power and attenuation of effect estimates. Power and sample size calculations for genetic studies with phenotype misclassification are available (Gordon et al. 2002); they have even been extended into the EMR setting for studies seeking to combine gold standard cases and controls with imperfectly phenotyped cases and controls (McDavid et al. 2013). The setting of outcome misclassification has been addressed in the measurement error literature, and methods to reduce estimation bias are available when the rates of outcome misclassification are known (Carroll et al. 2006).

However, none of the existing work has been extended to take advantage of a unique aspect of EMR phenotyping – specifically, that not just estimated disease status, but the *probability of having the disease*, is output from the algorithm. For example, the Informatics for Integrating Biology and the Bedside (i2b2) Center, an NIH-funded National Center for Biomedical Computing based at Partners HealthCare System, has developed algorithms for several phenotypes including rheumatoid arthritis (RA), Crohn's disease, and ulcerative colitis (Liao et al. 2010; Carroll et al. 2012; Ananthakrishnan et al. 2013). In existing EMR-based genetic studies, the probability is thresholded to classify individuals as cases and controls for subsequent analyses (Liao et al. 2010; Kurreeman et al. 2011). However, this

probability captures information about the uncertainty of disease classification which is lost when individuals are simply classified as probable cases or controls.

In this paper, we propose to model the probability of disease directly, instead of relying on thresholded case-control status, and demonstrate that by doing so, we can improve both power and estimation accuracy. In Section 2, we describe the approach and its application in three common EMR-based study designs. In Section 3 we compare the approaches in simulation and in a study of RA. Final comments are in Section 4, and derivations are provided in the Appendix. In the Appendix, we also provide power and sample size calculations for planning future studies with EMR phenotyping, and software is available upon request to implement the proposed approach as well as these power and sample size calculations.

2 Methods

We consider a setting in which the true disease status is not observed for everyone; instead, we assume that we have EMRs for a large number of patients, and that we can construct an algorithm that extracts information from each patient's medical records and produces the *probability* that the patient has the disease. We let p_D denote the probability of disease estimated by the algorithm. Of real interest, however, is the association between a SNP and true disease status. To establish notation, let D be the indicator of true disease status, taking the values $D = 1$ if the patient has the disease and $D = 0$ otherwise. Let Z be the number of risk alleles at the SNP, and $\mathbf{W}$ be a vector of covariates we wish to control for, such as age, gender, and principal components capturing population stratification (Price et al. 2006). We assume that a standard logistic regression model holds:

$$P(D=1|Z, \mathbf{W}) = g(\beta_0 + \beta_1 Z + \beta_2^{\hat{\alpha}\hat{\tau}\hat{n}} \mathbf{W}), \quad (1)$$

for some parameters $\beta = (\beta_0, \beta_1, \beta_2^{\hat{\alpha}\hat{\tau}\hat{n}})^{\hat{\alpha}\hat{\tau}\hat{n}}$, where throughout g will denote the inverse-logit function – i.e.,

$$g(x) = \text{logit}^{-1}(x) = \frac{\exp(x)}{1 + \exp(x)}, \text{ where } \text{logit}(x) = \log\left(\frac{x}{1-x}\right). \quad (2)$$

We may wish to test for an association between the SNP and disease by testing $H_0 : \beta_1 = 0$ in this model. We may also wish to estimate the parameter β_1 , which is the increase in the logodds of being a case associated with each additional risk allele. Our objective is to determine the best way to test H_0 and estimate β_1 in the setting where p_D is observed instead of D .

Throughout, we will assume that conditional on true disease status D , the EMR-based prediction for a given person is independent of that person's genotype Z and the covariates $\mathbf{W}$ that we wish to control for. Mathematically, we assume (A):

$$\hat{p}_D \perp \{Z, \mathbf{W}\} | D. \quad (\text{A})$$

For example, in a setting with no covariates $\mathbf{W}$, this assumption implies that the distribution of the algorithm predictions p_D among true cases only (or among true controls only) does not differ based on the genotype at Z , which is reasonable since the algorithm is built without information on genetics. If we are including covariates $\mathbf{W}$ – for example, if we control for gender in the model – then assumption (A) could potentially be violated if gender is a major contributor to creating the algorithm predictions p_D .

For a chosen threshold value p , we could define an *estimated disease status* $D_p = \mathbf{I}\{p_D > p\}$, where for any event A , $\mathbf{I}\{A\} = 1$ if A happens and 0 otherwise. That is, probable cases are those individuals with probability of disease larger than p , and probable controls have probability of disease smaller than p . One may choose a threshold p_S during algorithm development (which typically involves comparing algorithm predictions to time-consuming chart review) to achieve a desired specificity $S = P(p_D > p_S | D = 0) = P(D_p = 0 | D = 0)$ – i.e., to maintain a low rate of false positives, where $D = D_{p_S}$. Thresholding to maintain a certain specificity then also fixes the sensitivity $\text{SE}(S) = P(p_D > p_S | D = 1) = P(D_p = 1 | D = 1)$, the rate of true positives. After identifying probable cases and controls, one potential analysis approach which has been used in the literature (Kurreeman et al. 2011) is to fit a logistic regression model using estimated disease status D in place of D :

$$P(\tilde{D}=1|Z, \mathbf{W})=g(\gamma_0+\gamma_1 Z+\gamma_2^{\hat{p}_D} \mathbf{W}) \quad (3)$$

where $\gamma=(\gamma_0, \gamma_1, \gamma_2^{\hat{p}_D})$ are parameters.

Unfortunately, the parameter γ_1 does not in general equal the parameter of interest β_1 . Under assumption (A), a nonlinear relationship exists between them:

$g(\gamma_0+\gamma_1 Z+\gamma_2^{\hat{p}_D} \mathbf{W})=(\text{SE}(S)+S-1)g(\beta_0+\beta_1 Z+\beta_2^{\hat{p}_D} \mathbf{W})+(1-S)$ (Magder and Hughes 1997). In the absence of covariates $\mathbf{W}$, γ_1 does in fact equal 0 under the null $H_0: \beta_1 = 0$, and

thus, a test of $H_0: \gamma_1=0$ is a valid test of $H_0: \beta_1 = 0$, but test power may be hampered.

Estimates of the genetic effect using model (3) will tend to be attenuated; the expected amount of bias can be approximated by methods discussed in Appendix 6.4. When the model includes clinical covariates $\mathbf{W}$, both tests and estimation based on model (3) will in general be invalid.

The relationship between γ_1 and β_1 may be used to construct unbiased estimates of β_1 as proposed in the measurement error literature by viewing D as a misclassified outcome for the true outcome D (Carroll et al. 2006). In preliminary simulations, we found that this approach reduced estimation bias but did not improve power (simulations not shown). In our setting, we can reduce bias *and* improve power by instead modeling the probability of disease p_D , which is not available in traditional outcome misclassification settings. The intuition is that a subject with p_D far from the threshold p_S has much more certain disease status than a subject with p_D near the threshold, but this uncertainty is not incorporated when

modeling D . By modeling the probability of disease $p_{\hat{D}}$, we can leverage this uncertainty to gain efficiency.

In what follows, we assume the logistic regression model (1) for D holds, and find a linear transformation of $p_{\hat{D}}$, which we denote Y , whose expectation given Z and $\mathbf{W}$ is

$g(\beta_0 + \beta_1 Z + \beta_2^{\hat{\alpha}\hat{\tau}\hat{n}} \mathbf{W})$. With this unbiased relationship, we can perform better-powered tests of $H_0 : \beta_1 = 0$ and can accurately estimate β_1 using the same estimating equations used for fitting logistic regression models, but with Y in place of the usual case-control outcome.

Specifically, writing $\mathbf{X}_i = (1, Z_i, \mathbf{W}_i^{\hat{\alpha}\hat{\tau}\hat{n}})^{\hat{\alpha}\hat{\tau}\hat{n}}$ throughout for convenience, we solve the estimating equations $\sum_{i=1}^n \mathbf{X}_i (\mathcal{Y}_i - g(\beta^{\hat{\alpha}\hat{\tau}\hat{n}} \mathbf{X}_i)) = 0$ where i indexes the observed values on n subjects, and where Y_i is the appropriate linear transformation of the algorithm probability $p_{\hat{D}}$ calculated for the i^{th} person. The form of the necessary linear transformation is fundamentally the same regardless of study design, but we describe it separately for three common EMR-based study designs that are useful in practical settings because different constants are readily available depending on study design. Explicit derivations are provided in the Appendix. Because existing software for the binomial and quasibinomial models (e.g., `glm` in R) requires that the outcome Y be between 0 and 1, we solve the estimating equation directly using a Newton-Raphson algorithm since our linear transformation of $p_{\hat{D}}$ may take it out of this range. Software for the methods and for power calculations is available upon request.

2.1 Design A

In Design A, we take a random sample of size n from the collection of patients with EMR data, we genotype everyone in this sample, and we apply the algorithm to everyone to calculate $p_{\hat{D}}$. This design might be useful in practice when the outcome of interest is a common disease and the proportion of cases in a random sample is likely to be large, or when multiple disease outcomes in the same population are of interest (Ritchie et al. 2010; Denny et al. 2011; Kho et al. 2012). It may also be useful in so-called phenome-wide association studies or studies of pleiotropy, in which genes are queried for simultaneous associations with more than one disease (Denny et al. 2010, 2013).

As we show in Appendix 6.1, under this design $E[Y_A | \mathbf{X}] = g(\beta^{\top} \mathbf{X})$, where

$$\mathcal{Y}_A = \frac{\hat{p}_D - \zeta_0}{\zeta_1 - \zeta_0}$$

and $\zeta_d = E[p_{\hat{D}} | D = d]$, $d = 0, 1$. The parameters ζ_1 and ζ_0 are the average values of the algorithm predictions $p_{\hat{D}}$ among true cases and controls; these constants may be calculated during algorithm development.

2.2 Design B

In Design B, we begin as in Design A by taking a random sample of size n from the EMR and genotyping everyone. We then observe on everyone the value of a screening variable U

which serves as a *perfect negative predictor*, in that $P(D = 0 | U = 0) = 1$. Thus, individuals with $U = 0$ are definite controls, while case-control status for individuals with $U = 1$ is less clear, so we develop an algorithm for p_D to predict disease status among those individuals with $U = 1$. For example, in a study of RA, the value $U = 1$ could indicate having at least one billing code for RA or a mention in the narrative notes, since individuals without any such RA mention are extremely unlikely to be RA cases. Among those with such a reference to RA in their medical records, there will still be many individuals without RA; for example, individuals tested for a marker of RA but whose test results were negative (Gabriel 1994; Singh et al. 2004; Katz et al. 1997). As in Design A, this study design is useful in situations where the total study population is already determined, such as when multiple phenotypes are of interest or when existing studies are being re-used for a new phenotype; this design is always preferable to Design A when an appropriate screening variable U is available. Identifying and using a screening variable U is especially attractive when the disease is uncommon since the disease prevalence is typically higher among those with $U = 1$, so we can more easily develop an algorithm with high PPV; however, we are of course limited by the number of diseased individuals in the overall sample.

In this setting, we let $P_D^{\sim} = p_D^{\hat{}}$ among those individuals with $U = 1$ and $P_D^{\sim} = 0$ among those with $U = 0$, the definite controls. We assume U is independent of $\mathbf{X}$ given disease status D ; this is similar to Assumption (A) since U is likely derived from medical records. We show in Appendix 6.2 that $E[Y_B | \mathbf{X}] = g(\beta^T \mathbf{X})$, where

$$\mathcal{Y}_B = \frac{\tilde{p}_D - \tilde{\mu}_0}{\mu_1 - \tilde{\mu}_0} \text{ where } \tilde{\mu}_0 = \mu_0 \frac{\pi_U - \rho\pi_U}{1 - \rho\pi_U}$$

Here, $\mu_d = E[p_D^{\hat{}} | U = 1, D = d]$ are the average values of the algorithm predictions $p_D^{\hat{}}$ among true cases and controls in the screen-positive population; $\rho = P(D = 1 | U = 1)$ is the PPV of the filtering variable; and $\pi_U = P(U = 1)$ is the prevalence of positive screening in the study population. These quantities are typically calculated during algorithm development.

2.3 Design C

In Design C, we assume as in Design B that we have a screening variable that is a perfect negative predictor, but here we use the predictor to separate individuals into potential cases and potential controls, and perform sampling within these two pools. For example, in a study of RA, we could define a control-mart ($M = 0$) of individuals without any billing code for RA, and a disease-mart ($M = 1$) of individuals with an RA billing code. As in Design B, we assume $P(D = 0 | M = 0) = 1$ – that is, individuals in the control-mart are definite controls. The EMR-based predictions are developed in the disease-mart, but here we select cases for inclusion into our study only if they have $p_D^{\hat{}} > p_S$, where p_S is a threshold selected to maintain specificity S in the disease-mart. If n_1 subjects are selected as cases, we then select mn_1 controls from the control-mart. The number of controls per case, m , is typically set as 1, 2 or 3 depending on resources. This is essentially a case-control design with uncertainty in the case status (Breslow et al. 1980). This study design is useful when the disease is very uncommon in the general population (Kurreeman et al. 2011; Ananthakrishnan et al. 2013).

Let V indicate that an individual is sampled into our study as either a case or control. In this setting, let $p_{\tilde{D}} = p_{\hat{D}}$ in the disease-mart ($M = 1$) and $p_{\tilde{D}} = 0$ in the control-mart ($M = 0$). Under this design, we show in Appendix 6.3 that $E[Y_C | \mathbf{X}, V = 1] = g(\mathbf{X}^\top \beta^*)$, where

$\beta^* = (\beta_0^*, \beta_1, \beta_2^{\hat{\alpha} \hat{\tau} \hat{\eta}})$ is a parameter vector that differs from β only in the intercept β_0 , and where

$$\mathcal{Y}_C = \frac{\tilde{p}_D - \xi_0(1 - \pi)}{\xi_1 - \xi_0(1 - \pi)}.$$

Here, $\xi_d = E[p_{\hat{D}} | D = d, M = 1, p_{\hat{D}} > p_S]$ are the average values of the algorithm predictions $p_{\hat{D}}$ among *true* cases and controls selected from the disease-mart to *serve* as cases in the

analysis, and $\pi = \frac{m}{m + (1 - \text{PPV}(S))}$, where $\text{PPV}(S)$ is the PPV of the algorithm in the disease-mart – i.e., $\text{PPV}(S) = P(D = 1 | M = 1, p_{\hat{D}} > p_S)$. As before, these quantities are typically calculated during algorithm development.

In practice, this study design is used when the disease is very uncommon, and typically every patient in the disease-mart with $p_{\hat{D}} > p_S$ is included as a case. Thus the effect of the threshold p_S is especially worth investigating. By requiring high specificity S , we maintain a high proportion of true disease cases in our case group. By lowering the threshold, we increase the number of cases in our study while including more misclassified disease-free individuals in the case group. We assess the impact of p_S on power in the simulation studies in Section 3.

2.4 Power and Bias Calculations

To estimate the power to detect an OR of $\exp(\beta_1)$ at a SNP for a given α -level using an EMR-based probability of disease $p_{\hat{D}}$, we can use the asymptotic normality of the estimator

$\hat{\beta}_1$ and calculate the power as $1 - \Phi\left(c_{\alpha/2} - \sqrt{n} \frac{\beta_1}{\sigma_{\hat{p}_D}}\right) + \Phi\left(-c_{\alpha/2} - \sqrt{n} \frac{\beta_1}{\sigma_{\hat{p}_D}}\right)$ where Φ denotes the normal cdf and $c_{\alpha/2}$ satisfies $\Phi(c_{\alpha/2}) = 1 - \alpha/2$; estimation of $\sigma_{\hat{p}_D}$ is described in the Appendix. We also describe in the Appendix how to estimate the power that results from using the thresholded D in the misspecified model. These expressions can be helpful during the planning of a new EMR-based study. They can be used to compare the power from a study using EMR-based phenotyping to a potentially more expensive study with traditional phenotyping. Also, they rely on quantities relating to the accuracy of the algorithm, and thus may be useful either when an algorithm's accuracy properties are known, or when a phenotyping algorithm is in development and target values for these accuracy parameters would be helpful.

3 Results

3.1 Simulations—In simulation, we consider two tasks of interest in relating SNP Z to outcome D assuming model (1): (i) testing the null $H_0 : \beta_1 = 0$ and (ii) estimating the odds ratio (OR) $\exp(\beta_1)$. We compare the performance of the two primary methods we described:

dichotomizing $p_{\hat{D}}$ into $\tilde{D}$ and using the misspecified logistic model, as has been done in the literature, and using the continuous outcome Y , the design-specific linear function of $P_{\tilde{D}}$ that satisfied $E[Y | \mathbf{X}] = g(\beta^\top \mathbf{X})$. For simplicity, we focus on the setting without additional confounders; hence, $\mathbf{X} = (1, Z)^\top$. In this setting, we expect both procedures to be valid tests of H_0 , but only the proposed methods with Y to provide unbiased estimates of β_1 . Thus, in testing, we are most interested in which approach provides better power; in particular, since the method using $\tilde{D}$ has been used in the past and is slightly simpler to apply, it will be of interest to see whether using $p_{\hat{D}}$ provides a substantial power increase. For estimation, it will be of interest to compare the accuracy of the OR estimates of β_1 . Using the program R (R Development Core Team 2009), we ran 2000 simulations in each setting.

We generate the distribution of medical record risk scores R from a mixture distribution, where 30% of the risk scores come from a $N(\mu_{\text{neg}}, 1)$ distribution and 70% come from a $N(\alpha, \beta^2)$ population; we used this mixture distribution to reflect that typically some proportion of medical records are clearly negative for the disease of interest (here, those centered at $\mu_{\text{neg}} = -3$) while the rest belong to a spectrum where disease status is less obvious (here, those from the $N(\alpha, \beta^2)$ population, where parameters α and β are selected to guarantee a specific disease prevalence and algorithm accuracy as measured by the Area Under the Receiver Operating Characteristic Curve (AUC) when using $p_{\hat{D}}$ to predict D). This setup reflects what we have previously observed using algorithms developed in the i2b2 Center. We do this for all individuals in Design A, while for Designs B and C, we generate risk scores R only among those who screen positive (i.e., $U = 1$ or $M = 1$). We calculate the predicted probability of disease $p_{\hat{D}} = g(R)$ and generate the true disease status $D \sim \text{Bernoulli}(p_{\hat{D}})$. We choose a minor allele frequency (MAF) among the controls and an odds ratio (OR) quantifying the effect of each additional risk allele on disease status D , and define η_0 and η_1 by letting $\text{logit}^{-1}(\eta_0)$ be the MAF among the controls and $\exp(\eta_1)$ be the OR. We generate the number of risk alleles $Z \sim \text{Binomial}(n = 2, p = \text{logit}^{-1}(\eta_0 + \eta_1 D))$. In the null setting (OR=1), we found that all tests maintained their nominal Type I error rate, so we only present results in non-null settings. For all designs, we considered algorithms with AUC = 0.92 and 0.95, specificity thresholds $S = 0.95$ and 0.97 , MAF = 0.1 and 0.3, and OR = 1.2 and 1.5.

3.1.1 Design A: In this setting, we consider a common disease with disease prevalence 20%. We consider a sample of size $n = 2000$, classify everyone using our algorithm, and include everyone in our analysis. The PPVs of the algorithm at specificity levels of $S = (0.95, 0.97)$ are (0.77, 0.83) when AUC = 0.92 and (0.79, 0.86) when AUC = 0.95. Results are shown in Figure 1 for the different configurations.

The method using the dichotomous $\tilde{D}$ performs differently according to the specificity threshold S . With respect to power, we can see that when AUC is lower, it can be marginally better to choose specificity threshold 95% instead of 97% – i.e., the gain from adding additional cases to the analysis outweighs the contamination of truly disease-free individuals among the cases; when AUC is higher, this gain disappears. If we lower the specificity much further, we expect ultimately a decrease in power due to a case group overly diluted with true controls. With respect to estimation, we can see at times quite significant downward

bias when $\tilde{D}$ is used, and using 95% specificity results in even higher bias than using 97%.

Using $p_{\hat{D}}$ results in better power everywhere. We see the most improvement when the algorithm AUC is low, reflecting the fact that $p_{\hat{D}}$ carries forward into the analysis more information about the uncertainty in the algorithm classification. The method using $p_{\hat{D}}$ also has minimal bias in all settings, and does not depend on any specificity threshold parameter S . For example, when AUC = 0.92, MAF = 0.3 and OR = 1.5, the power of the $\tilde{D}$ -based method is 0.88 for both specificity levels while the proposed $p_{\hat{D}}$ -based method yields a power of 0.94, and the estimated OR from $\tilde{D}$ has downward biases as high as -34% and -29% of the true OR for the two specificity levels, while the bias is only 3% using the $p_{\hat{D}}$ -based approach. It is also important to note the power loss due to the uncertainty in disease status D is nontrivial compared to the setting where D is known, especially for weaker signals. This suggests that the accuracy of the algorithm in predicting D is crucial to ensure adequate power for subsequent genetic studies.

3.1.2 Design B: In this setting, we consider an uncommon disease in a larger population (sample size 5000). We have a preliminary screening variable U which satisfies $P(D = 0 | U = 0) = 1$. We assume that 20% of the EMR patients screen positive, and that among those screening positive, 40% have the disease, for an overall prevalence in the EMR data of 8%. We develop the EMR-based algorithm among those who screen positive. The performance of the methods is compared in Figure 2.

In this setting, the screening variable U assists us in developing an EMR-based classification with high accuracy; thus, both $\tilde{D}$ and $p_{\tilde{D}}$ have high accuracy in predicting D . For example, the PPVs of the algorithms at specificity levels $S = (0.95, 0.97)$ are (0.90, 0.92) when AUC=0.92 and (0.91, 0.94) when AUC=0.95. Consequently, the power lost from using either $\tilde{D}$ or $p_{\tilde{D}}$ compared to the true disease status D is less severe than in Design A, and the bias is also reduced. Nevertheless, as in Design A, we see uniformly better power and less bias using $p_{\hat{D}}$ when compared to the methods based on $\tilde{D}$.

3.1.3 Design C: In Design C, we first partition the EMR into a disease-mart and a control-mart, and include as cases all subjects in the disease-mart with $p_{\hat{D}} > p_S$; for each case, we select $m = 1$ control from the control-mart, and the controls are assumed to be perfectly classified. We assume the disease-mart has 5000 individuals and 20% disease prevalence, so for specificity 95% this would lead to genotyping approximately 854 cases and 854 controls; for specificity 97%, 697 of each. The PPVs of $\tilde{D}$ in this setting are the same as the PPVs of $\tilde{D}$ in Design A. However, because we exclude disease-mart subjects with low $p_{\hat{D}}$ and our controls are perfectly labeled, the overall accuracy of $\tilde{D}$ or $p_{\tilde{D}}$ in predicting D is much higher. Consequently, we expect less power loss due to misclassification in case-control status, when compared with Designs A and B. The performance of the methods is compared in Figure 3.

Indeed, in this design, the methods have similar power, though the $p_{\hat{D}}$ -based method tends to be the best; the improvement is most noticeable when the algorithm AUC is higher and the specificity threshold is lower, because in that setting $p_{\hat{D}}$ contains much more information

than a thresholded $\tilde{D}$ about true disease status. With respect to estimation, these results again demonstrate the significant downward bias from using $\tilde{D}$ as the outcome, especially for larger ORs; the proposed method based on $p_{\hat{D}}$ consistently produces very small bias.

One important point to highlight is that in this setting, unlike Designs A and B, the $p_{\hat{D}}$ -based method is also affected by the specificity threshold S because it is used to exclude individuals from the study. Reducing the specificity threshold increases the total number of cases (and controls) in our study, though at the cost of including more incorrectly classified cases. We see from simulation that no matter which technique is used, using the lower specificity level always yields better power due to the increased sample sizes and potential cases. However, it is also apparent that lowering the specificity threshold S increases the downward bias in the estimated OR when using $\tilde{D}$, since a larger proportion of truly disease-free individuals are misclassified as cases.

3.1.4 Empirical Power Evaluation: To demonstrate how the power of each method varies with sample size in a genome-wide association study context, we calculated the empirical power to detect a genome-wide significant result (i.e., $\alpha = 5e - 8$) as a function of sample size for each Design (Figure 4). We considered a SNP with MAF 0.3 and OR 1.5; we assumed the algorithm AUC is 95%; and we selected thresholds to guarantee 95% specificity. We used the same prevalence settings as in the simulations.

The curves demonstrate the significant loss, especially in Designs A and B, of using either $p_{\hat{D}}$ or $\tilde{D}$ in place of D . In Design A, when D is known, we can detect the SNP with 80% power for $n < 4000$; using $p_{\hat{D}}$ requires $n \approx 6000$ and $\tilde{D}$ requires $n \approx 7000$. In Design B, where the disease is rarer, we need $n \approx 7000$ to detect the SNP with 80% power when D is known, and $n \approx 9000$ when using $p_{\hat{D}}$ and $n > 10000$ when using $\tilde{D}$. Designs A and B can be useful tools when genotyping exists (e.g., when reusing other study subjects) but Design C is clearly the best choice when designing a study from scratch. In that design, 50% of genotyped subjects are (estimated) cases; in Design A, about 20% are estimated cases, and in Design B, only 8%. In Design C, perfect knowledge of D yields more power than using $p_{\hat{D}}$, which again yields more power than using $\tilde{D}$, but all need fewer than 4000 samples to yield 80% power to detect the signal.

3.2 Rheumatoid Arthritis Study—To demonstrate the performance of our approach in real data, we consider a study relating known RA risk alleles to incidence of RA in an EMR-based study following Design C. As described previously (Liao et al. 2010), an algorithm was developed to identify RA cases in the Partners Health EMR, a system used by two large academic hospitals serving the Boston, MA metropolitan area. Specifically, an RA-mart was defined by selecting individuals with at least one International Classification of Diseases 9 (ICD-9) code for RA, or who had been tested for antibodies to cyclic citrullinated peptide (anti-CCP). The RA-mart ultimately contained 29,432 individuals, and all other individuals in the EMR belonged to a control-mart. Individuals with $p_{\hat{D}} > p_S$ for $S = 95\%$ were included as cases when blood samples became available from discarded clinical specimens acquired through routine care, and were matched to controls in the RA-Mart with blood samples similarly obtained; this data was analyzed previously (Kurreeman et al. 2011). The Partners HealthCare Institutional Review Board approved the protocol. In this analysis, we restrict to

cases and controls of European descent, and require that cases are anti-CCP positive, because previous studies have reported associations in this setting; extensions to other populations and anti-CCP negative cases have also been considered (Kurreeman et al. 2011).

During algorithm development, it was determined that $p_{\hat{D}}$ should be thresholded at $p_{95} = 0.53$ to maintain 95% specificity. Using two data sets considered during algorithm development and validation, we estimate that $\xi_0 = 0.75$ is the average value of $p_{\hat{D}}$ among truly disease-free individuals satisfying $p_{\hat{D}} > p_{95}$, and $\xi_1 = 0.87$ is the average value among the true RA cases with $p_{\hat{D}} > p_{95}$. The PPV of the algorithm was estimated as 84%, and with 811 cases and 1225 controls available for analysis, $m = 1.5$. These parameter estimates may be slightly inaccurate because ancestry and anti-CCP status was not known for the RA cases used in algorithm development, but the cases in our analysis are anti-CCP positive and of European descent. Regardless, we use these parameter values to fit the model described in Section 2.3, and we compare to results from using thresholded disease status $\tilde{D}$ and to estimates from a recent meta-analysis (Stahl et al. 2010). For each SNP, we also calculated the bias we would expect from using $\tilde{D}$ as the case-control outcome assuming the true OR is the OR estimated using the unbiased $p_{\hat{D}}$ -based method; the expected bias calculations are based on asymptotic results as discussed in the Appendix.

Results comparing using $\tilde{D}$ and using $p_{\hat{D}}$, shown in Figure 5, are consistent with what we expect to see based on simulation. In Design C, where we have already restricted focus to cases with $p_{\hat{D}} > p_S$, further gradations of $p_{\hat{D}}$ are likely to make less of an impact on results than in other settings where the range of $p_{\hat{D}}$ is larger. Based on simulation (Figure 3) we expect similar power across the two methods, while expecting the $\tilde{D}$ -based OR estimates to be attenuated, especially for larger ORs and those based on $p_{\hat{D}}$ to be unbiased in general. In the example, we see that the $p_{\hat{D}}$ -based estimates are typically further away from the null than the $\tilde{D}$ -based estimates, with larger differences for larger ORs, while also having slightly wider confidence intervals. The differences are similar to the expected bias calculated using asymptotic results. For example, for the *HLA* SNP rs6457620, the OR estimate using $\tilde{D}$ is 1.96 (95% CI: 1.72, 2.24), while the OR estimate using $p_{\hat{D}}$ is 2.28 (95% CI: 1.92, 2.70); if the true OR is 2.28, the expected bias from using $\tilde{D}$ is -0.44, while the observed bias is -0.32.

While the relationship between the methods using $\tilde{D}$ and using $p_{\hat{D}}$ is what we expect based on simulation, the estimates from $\tilde{D}$ and $p_{\hat{D}}$ are not always in line with what we expect from the literature. For several of the SNPs, most notably the *HLA* SNP rs6457620, we see that the $\tilde{D}$ -based estimate is closer to the null than the literature-based estimate, and using $p_{\hat{D}}$ instead helps pull that estimate closer to what we expect based on the literature. However, for several other SNPs, such as the *PTPN22* SNP rs6679677 and the *TNFAIP3* SNP rs10499194, the estimate using $\tilde{D}$ is more extreme than the literature-based estimate, and the $p_{\hat{D}}$ -based estimate is even more extreme. While the difference between $\tilde{D}$ and $p_{\hat{D}}$ is approximately what we expect from the bias calculations, it is slightly concerning that the estimates are different from other studies. A possible explanation for these discrepancies is that the individuals in our study – both cases and controls – are different than those in previous RA studies, many of which are conducted within cohort studies. For example, the RA cases in our study are likely to have more severe disease than a random sample of RA

cases. Our RA disease-mart is drawn from a patient population at an academic medical center which may attract patients seeking treatment for more severe disease. Moreover, the patients who enter our genetic study have both a high probability of disease based on the information in their EMR and available blood samples available from discarded clinical specimens. Thus, they are likely to have extensive documentation of their disease as well as available blood from, for example, monitoring of drug therapy. Thus, if some of these SNPs predict not only RA incidence but also RA severity, the magnitude of the association in our study may differ from that estimated in cohort-based studies. While some promising genetic predictors of RA severity (Weyand et al. 1992; Brinkman et al. 1997; Gonzalez-Gay et al. 2002; Kastbom et al. 2008) have been suggested, evidence is not substantial enough to make a meaningful comparison here, but these associations are worth following up with subsequent studies. Furthermore, the controls in our study may be quite different from the “healthy controls” frequently used in cohort-based studies, who are often selected among individuals without other morbidities. The controls in our study are going to the hospital and providing biospecimens for some reason – thus, they are more likely to have other diseases.

4 Discussion

Linking EMRs to discarded biospecimens can provide an amazing resource for studying the relationships between phenotypes and genotypes provided that methods exist to effectively extract the phenotype information from the EMRs. Algorithms combining codified EMR data with narrative EMR data have proven their ability to predict disease status for many different diseases with good accuracy, but the predictions are still imperfect, and more complex diseases provide an especially significant challenge. Phenotype misclassification can negatively impact power in genetic association tests, so we proposed here a simple method to improve power to detect phenotype-genotype association by using the predicted probability of disease from the algorithm.

This approach has several benefits over the more standard approach of thresholding this probability and using a dichotomous estimated disease status $\tilde{D}$. It uniformly improves test power; the gains are sometimes modest, but noticeable especially when the algorithm AUC is low or the true OR is high. The difference between using $\tilde{D}$ and $p_{\hat{D}}$ is least dramatic in Design C, in which individuals are selected into the study based on thresholding $p_{\hat{D}}$; in this setting, the variability in $p_{\hat{D}}$ can be quite small, so the additional information to be gained is less than in Designs A and B. While the power improvements are small in some settings considered, even modest power improvements would be welcome in settings where the number of tests is quite large; this is evident from the power curves presented for genome-wide significance levels, where we see gains in power from using $p_{\hat{D}}$ instead of $\tilde{D}$ in a genome-wide context. It bears repeating that using our approach with $p_{\hat{D}}$ also always provides a valid test; testing with $\tilde{D}$ with misspecified link is valid (though less powerful) when there are no control covariates, but when we want to control for clinical covariates or population stratification, we are no longer assured that tests will maintain the nominal Type I error. Another benefit of using $p_{\hat{D}}$ is that in two of the three designs discussed, it also obviates the need to select a threshold.

While the gains in power from using $p_{\hat{D}}$ are modest, the reduction in bias is dramatic. Using the truncated $\tilde{D}$ to estimate ORs can produce estimates that are severely biased towards the null, especially when the true OR is large. Modeling $p_{\hat{D}}$ eliminates this bias. Some EMR-based studies use an algorithm which simply classifies individuals as cases or controls, or excludes them, with no probability output. Our method does not immediately apply to these algorithms, but in fact, we see that the benefits of using the probability over estimated disease status suggest that it is better to work with algorithms that produce probabilities rather than dichotomous predictions.

In Designs B and C, we assume the screening variable M or U is a perfect negative predictor, but in practice this may not be the case, so in fact the controls may not be perfectly selected as we assume here. Our method can be easily adapted to this and more complicated settings, as long as pertinent parameters such as the sensitivity, specificity and PPV are available.

For large scale implementation of EMR-based genetic association testing, we may want to aggregate information across multiple sites with EMRs. EMR implementation practices vary by institution, and disease prevalence varies too, either due to population differences or due to hospital characteristics (e.g., cancer prevalence at a cancer research hospital). For example, in one study, RA disease-marts defined by billing code had different prevalences of true RA cases across EMRs – 49%, 26%, and 19% (Carroll et al. 2012). Thus, care must be taken to transport a disease classification algorithm to a different institution, and the probability of disease (which should have expectation equal to the disease prevalence) and any threshold choices must be recalibrated. Extending our method to include ranges of possible sensitivities and specificities is an area of future research.

Ultimately, improvements in extracting information from EMRs will improve the discriminatory capability of EMR-based phenotyping, and for some phenotypes that are easy to detect from EMRs, the difference between using a thresholded $p_{\hat{D}}$ and $p_{\hat{D}}$ itself will not differ substantially; however, for particularly complex phenotypes such as psychiatric disorders or phenotypes that are otherwise difficult to diagnose, making better use of imperfect algorithm-derived phenotype information can bring about more powerful genetic discovery research (Perlis et al. 2011). Our simple method provides one way of improving power and estimation when case-control phenotypes are defined by an algorithm, and we recommend its usage as one component of a powerful, well-implemented, EMR-based study for discovery genetic research.

Acknowledgments

JAS was supported by the National Institutes of Health (NIH) grants T32 GM074897 and T32 CA09001 and the A. David Mazzone Career Development Award. TC was supported by NIH grants R01 GM079330, U01 GM092691 and U54 LM008748.

6 Appendix

6.1 Design A

In Design A, we consider a random sample of size n from the entire EMR data, and calculate $p_D^{\hat{}}$ for everyone in the sample. Using Assumption (A), we see $P(p_D^{\hat{}} > c \mid \mathbf{X}) = P(p_D^{\hat{}} > c \mid D = 1)g(\beta^T \mathbf{X}) + P(p_D^{\hat{}} > c \mid D = 0)(1 - g(\beta^T \mathbf{X}))$. Then, since for any positive random variable T ,

$$\begin{aligned} E[\hat{p}_D \mid \mathbf{X}] &= \int_0^1 P(\hat{p}_D > c \mid \mathbf{X}) dc \\ &= g(\beta^T \mathbf{X}) \int_0^1 P(\hat{p}_D > c \mid D = 1) dc \\ &\quad + (1 - g(\beta^T \mathbf{X})) \int_0^1 P(\hat{p}_D > c \mid D = 0) dc \end{aligned}$$

$$E[T] = \int_0^\infty P(T > c) dc, \text{ we have: } \quad = \zeta_0 + (\zeta_1 - \zeta_0)g(\beta^T \mathbf{X}) \quad \text{where } \zeta_d = E[p_D^{\hat{}} \mid D = d].$$

Thus, letting $\mathcal{Y}_A = \frac{\hat{p}_D - \zeta_0}{\zeta_1 - \zeta_0}$, we have $E[Y_A \mid \mathbf{X}] = g(\beta^T \mathbf{X})$.

6.2 Design B

In Design B, we also genotype a random sample of size n , but observe on everyone a perfect negative predictor U satisfying $P(D = 0 \mid U = 0) = 1$. The EMR algorithm is developed among those individuals with $U = 1$. In addition to assumption (A), we assume that U is independent of $\mathbf{X}$ conditional on true disease status D . We let $P_D^{\sim}(U) = p_D^{\hat{}} U$. Defining $\mu_d = E[p_D^{\hat{}} \mid U = 1, D = d]$ for $d = 0, 1$, $\rho = P(D = 1 \mid U = 1)$, $\pi_U = P(U = 1)$, and

$\tilde{\mu}_0 = \mu_0 \frac{\pi_U - \rho\pi_U}{1 - \rho\pi_U}$, we may calculate $E[P_D^{\sim} \mid \mathbf{X}] = \sum_{d \in \{0,1\}} \mu_d P(U = 1, D = d \mid \mathbf{X}) = \sum_{d \in \{0,1\}} \mu_d P(U = 1 \mid D = d) P(D = d \mid \mathbf{X}) = \mu_0 + (\mu_1 - \mu_0)g(\beta^T \mathbf{X})$ since $P(U = 1 \mid D = 1) = 1$ and

$P(U = 1 \mid D = 0) = \frac{\pi_U - \rho\pi_U}{1 - \rho\pi_U}$ by an application of Bayes rule. Thus, letting $\mathcal{Y}_B = \frac{\tilde{p}_D - \tilde{\mu}_0}{\mu_1 - \tilde{\mu}_0}$, we have $E[Y_B \mid \mathbf{X}] = g(\beta^T \mathbf{X})$.

6.3 Design C

In Design C, we first partition the full EMR into a disease-mart ($M = 1$) that includes all disease cases and a control-mart ($M = 0$) of disease-free individuals. We develop and apply our algorithm to calculate $p_D^{\hat{}}$ only among individuals with $M = 1$. Let $PPV(S) = P(D = 1 \mid M = 1, p_D^{\hat{}} > p_S)$, and we assume a design with m controls per case sampled from the control-mart. Let V be the indicator that an individual is sampled in our study, and let $P_D^{\sim} = P_D^{\hat{}}(M) = p_D^{\hat{}} M$.

We may calculate $E[P_D^{\sim} \mid \mathbf{X}, D, V = 1] = E[P_D^{\sim} \mid D, V = 1] = DE[p_D^{\hat{}} \mid D = 1, M = 1, p_D^{\hat{}} > p_S]P(M = 1 \mid D = 1, V = 1) + (1 - D)E[p_D^{\hat{}} \mid D = 0, M = 1, p_D^{\hat{}} > p_S]P(M = 1 \mid D = 0, V = 1) = D\xi_1 + (1 - D)\xi_0(1 - \pi)$ where $\xi_d = E[p_D^{\hat{}} \mid D = d, M = 1, p_D^{\hat{}} > p_S]$ and $\pi = P(M = 0 \mid D = 0, V = 1)$. In this calculation, we have used that $p_D^{\hat{}} = 0$ when $M = 0$, and that $P(M = 1 \mid D = 1, V = 1) = 1$ because the initial partition has perfect sensitivity. We further calculate:

$$\begin{aligned}
\pi &= \frac{P(M=0, D=0|V=1)}{P(D=0|V=1)} \\
&= \frac{P(M=0|V=1)}{P(M=0|V=1) + P(D=0|M=1, V=1)P(M=1|V=1)} \\
&= \frac{\frac{m}{m+1}}{\frac{m}{m+1} + (1 - \text{PPV}(S))\frac{1}{m+1}} \\
&= \frac{m}{m + (1 - \text{PPV}(S))}. \quad \text{Then, letting} \\
\mathcal{Y}_C &= \frac{\tilde{p}_D - \xi_0(1 - \pi)}{\xi_1 - \xi_0(1 - \pi)}, \text{ we have that } E[Y_C | \mathbf{X}, V = 1] = E[D | \mathbf{X}, V = 1].
\end{aligned}$$

Finally, by using Bayes rule, we see that

$$\begin{aligned}
E[D | \mathbf{X}, V=1] &= \frac{P(V=1|D=1)g(\beta^{\hat{\mathbf{a}}\tilde{\mathbf{T}}\tilde{\mathbf{n}}}\mathbf{X})}{P(V=1|D=1)g(\beta^{\hat{\mathbf{a}}\tilde{\mathbf{T}}\tilde{\mathbf{n}}}\mathbf{X}) + P(V=1|D=0)(1 - g(\beta^{\hat{\mathbf{a}}\tilde{\mathbf{T}}\tilde{\mathbf{n}}}\mathbf{X}))} = \frac{\exp(\beta^{\hat{\mathbf{a}}\tilde{\mathbf{T}}\tilde{\mathbf{n}}}\mathbf{X})}{\lambda + \exp(\beta^{\hat{\mathbf{a}}\tilde{\mathbf{T}}\tilde{\mathbf{n}}}\mathbf{X})} \\
\text{where } \lambda &= \frac{P(V=1|D=0)}{P(V=1|D=1)}. \text{ Letting } \beta_0^* = \beta_0 - \log\{\lambda\}, \text{ we have} \\
E[D | \mathbf{X}, V=1] &= \frac{\exp(\beta_0^* + \beta_1 Z + \beta_2^{\hat{\mathbf{a}}\tilde{\mathbf{T}}\tilde{\mathbf{n}}}\mathbf{W})}{1 + \exp(\beta_0^* + \beta_1 Z + \beta_2^{\hat{\mathbf{a}}\tilde{\mathbf{T}}\tilde{\mathbf{n}}}\mathbf{W})}. \text{ Thus, } E[Y_C | \mathbf{X}, V = 1] = g(\beta^{*\top}\mathbf{X}), \text{ where} \\
\beta^* &= (\beta_0^*, \beta_1, \beta_2^{\hat{\mathbf{a}}\tilde{\mathbf{T}}\tilde{\mathbf{n}}})^{\hat{\mathbf{a}}\tilde{\mathbf{T}}\tilde{\mathbf{n}}}.
\end{aligned}$$

6.4 Power and Bias Calculations

For simplicity we derive expressions under Design A. When using $p_{\hat{D}}$, the estimator $\hat{\beta}$ solves

$$\begin{aligned}
U(\beta) &= \frac{1}{n} \sum_{i=1}^n \psi_i(\beta) = \frac{1}{n} \sum_{i=1}^n \mathbf{X}_i (\mathcal{Y}_i - g(\beta^{\hat{\mathbf{a}}\tilde{\mathbf{T}}\tilde{\mathbf{n}}}\mathbf{X}_i)) = 0, \text{ so } \sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{\mathcal{D}} N(0, V(\beta)) \text{ where} \\
V(\beta) &= B(\beta)^{-1} A(\beta) (B(\beta)^{-1})^\top \text{ where } B(\beta) = E[g(\beta^\top \mathbf{X})(g(\beta^\top \mathbf{X}) - 1)\mathbf{X}\mathbf{X}^\top] \text{ and } A(\beta) = E[(Y - g(\beta^\top \mathbf{X}))^2 \mathbf{X}\mathbf{X}^\top] \\
&= E[\{\text{Var}(Y|D) + E[Y|D]^2\} E[\mathbf{X}\mathbf{X}^\top | D]] - 2E[E[Y|D] E[g(\beta^\top \mathbf{X})\mathbf{X}\mathbf{X}^\top | D]] + E[g(\beta^\top \mathbf{X})^2 \mathbf{X}\mathbf{X}^\top], \text{ using assumption (A). We can further expand this since } \mathbf{X} = (1, Z)^\top, \text{ for SNP } Z; \text{ in particular, for any function } f,
\end{aligned}$$

$$\begin{aligned}
E[f(Z)|D] &= \frac{D}{P(D=1)} E[f(Z)g(\beta^{\hat{\mathbf{a}}\tilde{\mathbf{T}}\tilde{\mathbf{n}}}\mathbf{X})] + \frac{1-D}{P(D=0)} E[f(Z)(1 - g(\beta^{\hat{\mathbf{a}}\tilde{\mathbf{T}}\tilde{\mathbf{n}}}\mathbf{X}))]. \text{ Letting } \mu_d = E[Y | D = d] \text{ and } \xi_d = \text{Var}(Y | D = d), \text{ we can rewrite } A(\beta) \text{ as:}
\end{aligned}$$

$$\begin{aligned}
A(\beta) &= \{\xi_0 + \mu_0^2\} E[\mathbf{X}\mathbf{X}^{\hat{\mathbf{a}}\tilde{\mathbf{T}}\tilde{\mathbf{n}}}] + \{\xi_1 + \mu_1^2 - \xi_0 - \mu_0^2 - 2\mu_0\} E[g(\beta^{\hat{\mathbf{a}}\tilde{\mathbf{T}}\tilde{\mathbf{n}}}\mathbf{X})\mathbf{X}\mathbf{X}^{\hat{\mathbf{a}}\tilde{\mathbf{T}}\tilde{\mathbf{n}}}] + (2(\mu_0 - \mu_1) + 1) E[g(\beta^{\hat{\mathbf{a}}\tilde{\mathbf{T}}\tilde{\mathbf{n}}}\mathbf{X})^2 \mathbf{X}\mathbf{X}^{\hat{\mathbf{a}}\tilde{\mathbf{T}}\tilde{\mathbf{n}}}] \\
&\quad .
\end{aligned}$$

To compare power to results using $\tilde{D}$ in the misspecified model, we now consider the

distribution of $\hat{\gamma}$ which solves $U(\gamma) = \frac{1}{n} \sum_{i=1}^n \psi_i(\gamma) = \frac{1}{n} \sum_{i=1}^n \mathbf{X}_i (\tilde{D}_i - g(\gamma^{\hat{\mathbf{a}}\tilde{\mathbf{T}}\tilde{\mathbf{n}}}\mathbf{X}_i)) = 0$. Then $n(\hat{\gamma} - \gamma^*(\beta)) \rightarrow N(0, V^*(\beta))$, where $\gamma^*(\beta)$ is a constant. To estimate γ^* , we can proceed as in Neuhaus (1999) and use results from work on misspecified models to see that estimates from the false model $P_F(D = 1 | Z) = g(\gamma_0 + \gamma_1 Z)$ converge to values (γ_0^*, γ_1^*) that minimize the Kullback-Leibler divergence between the false model and the true model $P_T(D = 1 | Z) = (1 - S) + (SE(S) + S - 1)g(\beta_0 + \beta_1 Z)$ (Neuhaus 1999; Kullback 1959). The Kullback-Leibler divergence between these two models is

$E_Z \left[\log \left\{ \frac{P_T(\tilde{D}=1|Z)}{P_F(\tilde{D}=1|Z)} \right\} P_T(\tilde{D}=1|Z) + \log \left\{ \frac{P_T(\tilde{D}=0|Z)}{P_F(\tilde{D}=0|Z)} \right\} P_T(\tilde{D}=0|Z) \right]$. By taking derivatives with respect to γ_0 and γ_1 and setting them to 0, we find two equations:

$0 = \{\alpha_0 + \alpha_1 g(\beta_0) - g(\gamma_0)\}(1 - p_Z)^2 + \{\alpha_0 + \alpha_1 g(\beta_0 + \beta_1) - g(\gamma_0 + \gamma_1)\} 2p_Z(1 - p_Z) + \{\alpha_0 + \alpha_1 g(\beta_0 + 2\beta_1) - g(\gamma_0 + 2\gamma_1)\} p_Z^2$
 and $0 = \{\alpha_0 + \alpha_1 g(\beta_0 + \beta_1) - g(\gamma_0 + \gamma_1)\} 2p_Z(1 - p_Z) + 2\{\alpha_0 + \alpha_1 g(\beta_0 + 2\beta_1) - g(\gamma_0 + 2\gamma_1)\} p_Z^2$,
 where $\alpha_0 = 1 - S$, $\alpha_1 = SE(S) + S - 1$. Here, we assume that the SNP $Z \sim \text{Bin}(2, p_Z)$, where p_Z is the MAF. Simultaneously solving these two equations for (γ_0, γ_1) yields the desired (γ_0^*, γ_1^*) . The calculation of $V^*(\gamma^*)$ proceeds similarly to the calculation of $V(\beta)$. $V^*(\gamma^*) = B^*(\gamma^*)^{-1} A^*(\gamma^*) (B^*(\gamma^*)^{-1})^\top$, where $B^*(\gamma) = E[g(\gamma^\top \mathbf{X})(g(\gamma^\top \mathbf{X}) - 1)\mathbf{X}\mathbf{X}^\top]$ as before. Here, though, $A^*(\gamma) = (1 - S)E[\mathbf{X}\mathbf{X}^\top] + (SE(S) - 3(1 - S))E[g(\gamma^\top \mathbf{X})\mathbf{X}\mathbf{X}^\top] + (2(1 - S - SE(S)) + 1)E[g(\gamma^\top \mathbf{X})^2\mathbf{X}\mathbf{X}^\top]$.

References

- Ananthakrishnan AN, Cai T, Savova G, Cheng SC, Chen P, Perez RG, Gainer VS, Murphy SN, Szolovits P, Xia Z, et al. Improving case definition of crohn's disease and ulcerative colitis in electronic medical records using natural language processing: a novel informatics approach. *Inflammatory bowel diseases*. 2013; 19(7):1411–1420. [PubMed: 23567779]
- Breslow, NE.; Day, NE., et al. Statistical methods in cancer research Vol 1 The analysis of case-control studies. Vol. 1. Distributed for IARC by WHO; Geneva, Switzerland: 1980.
- Brinkman B, Huizinga T, Kurban S, Van der Velde E, Schreuder G, Hazes J, Breedveld F, Verweij C. Tumour necrosis factor alpha gene polymorphisms in rheumatoid arthritis: association with susceptibility to, or severity of, disease? *Rheumatology*. 1997; 36(5):516–521.
- Carroll R, Ruppert D, Stefanski L, Crainiceanu C. Measurement error in nonlinear models: A modern perspective number 105 in monographs on statistics and applied probability. 2006
- Carroll RJ, Thompson WK, Eyler AE, Mandelin AM, Cai T, Zink RM, Pacheco JA, Boomersshine CS, Lasko TA, Xu H, et al. Portability of an algorithm to identify rheumatoid arthritis in electronic health records. *Journal of the American Medical Informatics Association*. 2012; 19(e1):e162–e169. [PubMed: 22374935]
- Denny J, Ritchie M, Basford M, Pulley J, Bastarache L, Brown-Gentry K, Wang D, Masys D, Roden D, Crawford D. Phewas: demonstrating the feasibility of a phenome-wide scan to discover gene–disease associations. *Bioinformatics*. 2010; 26(9):1205–1210. [PubMed: 20335276]
- Denny JC, Crawford DC, Ritchie MD, Bielinski SJ, Basford MA, Bradford Y, Chai HS, Bastarache L, Zuvich R, Peissig P, et al. Variants near foxe1 are associated with hypothyroidism and other thyroid conditions: Using electronic medical records for genome-and phenome-wide studies. *The American Journal of Human Genetics*. 2011; 89(4):529–542.
- Denny JC, Bastarache L, Ritchie MD, Carroll RJ, Zink R, Mosley JD, Field JR, Pulley JM, Ramirez AH, Bowton E, et al. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nature biotechnology*. 2013
- Gabriel SE. The sensitivity and specificity of computerized databases for the diagnosis of rheumatoid arthritis. *Arthritis & Rheumatism*. 1994; 37(6):821–823. [PubMed: 8003054]
- Gonzalez-Gay, MA.; Garcia-Porrúa, C.; Hajeer, AH. *Seminars in arthritis and rheumatism*. Vol. 31. Elsevier; 2002. Influence of human leukocyte antigen-drb1 on the susceptibility and severity of rheumatoid arthritis; p. 355–360.
- Gordon D, Finch SJ, Nothnagel M, Ott Jr. Power and sample size calculations for case-control genetic association tests when errors are present: application to single nucleotide polymorphisms. *Human heredity*. 2002; 54(1):22–33. [PubMed: 12446984]
- Kastbom A, Verma D, Eriksson P, Skogh T, Wingren G, Söderkvist P. Genetic variation in proteins of the cryopyrin inflammasome influences susceptibility and severity of rheumatoid arthritis (the swedish tira project). *Rheumatology*. 2008; 47(4):415–417. [PubMed: 18263599]

- Katz J, Barrett J, Liang M, Bacon A, Kaplan H, Kieval R, Lindsey S, Roberts W, Sheff D, Spencer R, et al. Sensitivity and positive predictive value of medicare part b physician claims for rheumatologic diagnoses and procedures. *Arthritis & Rheumatism*. 1997; 40(9):1594–1600. [PubMed: 9324013]
- Kho A, Hayes M, Rasmussen-Torvik L, Pacheco J, Thompson W, Armstrong L, Denny J, Peissig P, Miller A, Wei W, et al. Use of diverse electronic medical record systems to identify genetic risk for type 2 diabetes within a genome-wide association study. *Journal of the American Medical Informatics Association*. 2012; 19(2):212–218. [PubMed: 22101970]
- Kohane I. Using electronic health records to drive discovery in disease genomics. *Nature Reviews Genetics*. 2011; 12(6):417–428.
- Kullback S. *Information theory and statistics*. 1959
- Kurreeman F, Liao K, Chibnik L, Hickey B, Stahl E, Gainer V, Li G, Bry L, Mahan S, Ardlie K, et al. Genetic basis of autoantibody positive and negative rheumatoid arthritis risk in a multi-ethnic cohort derived from electronic health records. *The American Journal of Human Genetics*. 2011; 88(1):57–69.
- Liao K, Cai T, Gainer V, Goryachev S, Zeng-treitler Q, Raychaudhuri S, Szolovits P, Churchill S, Murphy S, Kohane I, et al. Electronic medical records for discovery research in rheumatoid arthritis. *Arthritis care & research*. 2010; 62(8):1120–1127. [PubMed: 20235204]
- Magder LS, Hughes JP. Logistic regression when the outcome is measured with uncertainty. *American Journal of Epidemiology*. 1997; 146(2):195–203. [PubMed: 9230782]
- McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, Hirschhorn JN. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews Genetics*. 2008; 9(5):356–369.
- McDavid A, Crane PK, Newton KM, Crosslin DR, McCormick W, Weston N, Ehrlich K, Hart E, Harrison R, Kull WA, et al. Enhancing the power of genetic association studies through the use of silver standard cases derived from electronic medical records. *PloS one*. 2013; 8(6):e63481. [PubMed: 23762230]
- Neuhaus JM. Bias and efficiency loss due to misclassified responses in binary regression. *Biometrika*. 1999; 86(4):843–855.
- Newton KM, Peissig PL, Kho AN, Bielinski SJ, Berg RL, Choudhary V, Basford M, Chute CG, Kullo IJ, Li R, et al. Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the emerge network. *Journal of the American Medical Informatics Association*. 2013
- Perlis R, Iosifescu D, Castro V, Murphy S, Gainer V, Minnier J, Cai T, Goryachev S, Zeng Q, Gallagher P, et al. Using electronic medical records to enable large-scale studies in psychiatry: treatment resistant depression as a model. *Psychological Medicine*. 2011; 1(1):1–10.
- Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*. 2006; 38(8):904–909. [PubMed: 16862161]
- R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing; Vienna, Austria: 2009. URL <http://www.R-project.org>
- Ritchie M, Denny J, Crawford D, Ramirez A, Weiner J, Pulley J, Basford M, Brown-Gentry K, Balser J, Masys D, et al. Robust replication of genotype-phenotype associations across multiple diseases in an electronic medical record. *The American Journal of Human Genetics*. 2010; 86(4):560–572.
- Singh J, Holmgren A, Noorbaloochi S. Accuracy of veterans administration databases for a diagnosis of rheumatoid arthritis. *Arthritis Care & Research*. 2004; 51(6):952–957. [PubMed: 15593102]
- Stahl EA, Raychaudhuri S, Remmers EF, Xie G, Eyre S, Thomson BP, Li Y, Kurreeman FA, Zhernakova A, Hinks A, et al. Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. *Nature genetics*. 2010; 42(6):508–514. [PubMed: 20453842]
- Weyand CM, Hicok KC, Conn DL, Goronzy JJ. The influence of hla-drbl genes on disease severity in rheumatoid arthritis. *Annals of internal medicine*. 1992; 117(10):801–806. [PubMed: 1416553]
- Wilke R, Xu H, Denny J, Roden D, Krauss R, McCarty C, Davis R, Skaar T, Lamba J, Savova G. The emerging role of electronic medical records in pharmacogenomics. *Clinical Pharmacology & Therapeutics*. 2011; 89(3):379–386. [PubMed: 21248726]

Fig. 1.

Presented are power and bias estimates from simulation in Design A, for $n = 2000$, disease prevalence 20%, $\alpha = 0.05$, MAF=0.1 and 0.3, OR=1.2 and 1.5, and algorithm AUC=0.92 and 0.95. In each setting, we compare the results we would get if we could actually observe true disease status (true D) to methods discussed for using disease status estimated from EMR data: $\hat{D}_{95}$ and $\hat{D}_{97}$ use estimated disease status thresholded to guarantee specificity = 95% and 97% as an outcome in the logistic model; and $\hat{p}_D$ uses the predicted probability of disease directly with the correct link function.

Fig. 2.

Presented are power and bias estimates from simulation in Design B, for $n = 5000$, disease prevalence 8% $\alpha = 0.05$, MAF=0.1 and 0.3, OR=1.2 and 1.5, and algorithm AUC=0.92 and 0.95. In each setting, we compare the results we would get if we could actually observe true disease status (true D) to methods discussed for using disease status estimated from EMR data: $\hat{D}_{95}$ and $\hat{D}_{97}$ use estimated disease status thresholded to guarantee specificity = 95% and 97% as an outcome in the logistic model; and $p_{\hat{D}}$ uses the predicted probability of disease directly with the correct link function.

Fig. 3.

Presented are power and bias estimates from simulation in Design C, for a disease-mart of size 5000 with prevalence 20%, $\alpha = 0.05$, MAF=0.1 and 0.3, OR=1.2 and 1.4, algorithm AUC=0.92 and 0.95, and specificity threshold=0.95 and 0.97. In each setting, we compare the results we would get if we could actually observe true disease status (true D) to the methods discussed for using disease status estimated from EMR data: $\tilde{D}$ uses estimated disease status as an outcome in the logistic model; and $\hat{P}_D$ uses the predicted probability of disease directly with the correct link function. Note that in this setting, the threshold p_S affects the performance of $\hat{P}_D$ and $\tilde{D}$ as well as D because it dictates which individuals are included in the analysis.

Fig. 4.

Presented for each design is the empirical power to detect a genome-wide significant result ($\alpha = 5e - 8$) as a function of sample size for: the model where true disease status D is known; the model fit with dichotomized $\hat{D}$; and the model fit with our proposed method using $\hat{p}_D$. We assume a SNP with MAF 0.3 and OR 1.5; we assume the algorithm AUC is 95%; and we select a threshold guaranteeing 95% specificity. In Design A, the overall disease prevalence is 20%; in Design B, 20% of individuals screen positive and 40% of those have the disease; and in Design C, the disease prevalence is 20% in the disease-mart.

Fig. 5.

Presented are the estimated odds ratios and confidence intervals for the RA study described in Section 3.2. SNP IDs are listed with candidate genes in the region. For each SNP, presented are estimates from a meta-analysis (Stahl et al. 2010) (meta-analysis); estimates from using the EMR cohort with the estimated disease status D as the outcome in a logistic regression (D); and estimates from our proposed method modeling p_D directly (p_D). Also shown is the amount of bias expected when using D , assuming that the OR estimated using p_D is the true OR.