Hindawi
Journal of Healthcare Engineering
Volume 2017, Article ID 7653071, 5 pages
https://doi.org/10.1155/2017/7653071

Research Article

The Impact of Diagnostic Code Misclassification on Optimizing the Experimental Design of Genetic Association Studies

Steven J. Schrodi[1,2]

[1]*Center for Human Genetics, Marshfield Clinic Research Institute, Marshfield, WI, USA*
[2]*Computation and Informatics in Biology and Medicine, University of Wisconsin-Madison, Madison, WI, USA*

Correspondence should be addressed to Steven J. Schrodi; schrodi.steven@mcrf.mfldclin.edu

Received 17 May 2017; Accepted 13 September 2017; Published 18 October 2017

Academic Editor: Richard Segall

Diagnostic codes within electronic health record systems can vary widely in accuracy. It has been noted that the number of instances of a particular diagnostic code monotonically increases with the accuracy of disease phenotype classification. As a growing number of health system databases become linked with genomic data, it is critically important to understand the effect of this misclassification on the power of genetic association studies. Here, I investigate the impact of this diagnostic code misclassification on the power of genetic association studies with the aim to better inform experimental designs using health informatics data. The trade-off between (i) reduced misclassification rates from utilizing additional instances of a diagnostic code per individual and (ii) the resulting smaller sample size is explored, and general rules are presented to improve experimental designs.

1. Introduction

Clearly, a wealth of important clinical information is contained within large electronic health record (EHR) systems. Such information can be an invaluable resource for measuring disease prevalence [1] and disease comorbidity [2], the association between birth month and disease susceptibility [3], the prediction of outcomes [4], the measurement of economic impact of health care [5], and the discovery of etiological factors [6]. A key feature of these data is in the diagnostic codes given by medical professionals to patient records. However, the accuracy of inferring disease phenotypes from electronic diagnostic codes can vary widely across diseases and is often subject to high degrees of error [7–10]. These studies have noted the substantial misclassification effects from the use of electronic diagnostic code data, sufficient to undermine experiments utilizing cases and controls defined by the International Classification of Diseases (ICD) codes alone. The ICD coding system is instituted by the World Health Organization and has been adopted in the United States by the National Center for Health Statistics. More sophisticated approaches to disease classification, such as those using a variety of EHR data and machine learning methods, are difficult to generalize across all diseases and implement in a high-throughput manner. That said, I anticipate that machine learning methods applied to problems of phenotype prediction using EHR variables as features in the predictive modeling will eventually supplant the sole use of ICD code data. Until that time, the use of ICD data may still have utility in initial screens, to be subsequently validated through methods with higher positive and negative predictive values.

2. Related Work

In a general setting, the effect of phenotypic misclassification on statistical power of genetic association studies has been previously explored [11–14]. Edwards and colleagues characterized the noncentrality parameter in asymptotic power distributions given the presence of phenotypic misclassification [11]. The authors use cost functions to capture the effect of misclassification and show that the cost of misclassifying a

control as a case becomes large and the cost of misclassifying a case individual as a control becomes small as the disease prevalence becomes small. Similarly, Ji et al. also investigated the calculation of a noncentrality parameter capturing phenotype errors for subsequent use in a likelihood ratio test for genetic association studies [12]. Later, Gordon and colleagues showed how to incorporate misclassification error rates into a trend test for genetic association in case/control studies [13]. More recently, Manchia and colleagues investigated the impact of heterogeneity within a clinical phenotype on genetic association [14].

Considering ICD data with misclassification, the type I and type II error rates for genomic association studies were recently thoroughly explored by Duan et al. [15]. The Duan et al. study found little inflation in false-positive rates, but not in considerable false-negative rates under certain allele frequency, effect size, and disease prevalence parameters. In the context of initial screens of ICD codes in EHR systems, several studies have investigated the relationship between the number of instances of particular ICD codes and the measures of diagnostic utility [1, 16–18]. In general, the accuracy of diagnoses improves with the number of instances of the code; however, this is at the expense of smaller sample sizes/increasing false negatives. Hence, there is a trade-off between type I and type II error rates with the number of ICD code instances used to define a disease. In this work, I investigate this trade-off and provide a framework for determining highly powered EHR-based experimental designs using diseases defined by different numbers of instances of ICD codes.

3. Materials and Methods

For a large genetic association scan of using ICD data, define a simple disease classification scheme such that cases are those individuals with x instances of a particular ICD code. Consider a design where individuals with ambiguous numbers of instances (i) of the code (i.e., $0 < i < x$) are excluded from the analysis. Further consider a comparison of well-defined cases (i.e., those with at least x instances) against a large, fixed set of controls. With regard to the genetics, restrict the methods to biallelic markers with minor alleles segregating in the population at a frequency of at least 1% single-nucleotide polymorphisms (SNPs). Define the alleles at a SNP contributing to the susceptibility of the disease as A_1 and A_2. Let the relative risk of the minor allele, A_2, be R, such that $R = P(A_2|\text{cases})[P(A_2|\text{controls})]^{-1}$. Let the frequency of A_2 in the general population be q. Accordingly, $1 - q$ is the frequency of A_1. Define n_x as the number of cases obtained from the definition of having at least x instances of the ICD code being evaluated. Set the number of controls as m, such that $m \gg n_x$. Assume that the A_2 frequency in controls is approximately q. Model the decrease in the misclassification proportion within cases as x increases with a monotonic function $f(x)$, such that the expected number of truly positive cases is $n_x[1 - f(x)]$. The form of $f(x)$ may vary considerably for different ICD codes. Lastly, let α be the statistical threshold for determining a positive finding in analyses where p value $< \alpha$. The statistical test of genetic association considered is the binomial test of proportions which evaluates the null hypothesis of no correlation between the frequency of A_2 and the disease status.

Statistical power will be used to evaluate the impact of increasing x and the resulting experimental design. Under the model specified above, the power to detect association at an autosomal SNP, $1 - \beta$, is calculated by the approximation as follows:

$$\Phi\left[\sqrt{\frac{N(q-s)^2}{q(1-q)+s(1-s)}} - z_{1-\alpha/2}\right], \tag{1}$$

where Φ is the standard Gaussian cumulative distribution function, z is the inverse standard Gaussian score, $N = 4n_x m/n_x + m$, and q and s are the A_2 frequencies in controls and cases, respectively. Using Bayes' theorem, the expected frequency of A_2 within cases under the misclassification model is given by

$$s = f(x)q + Rq[1 - f(x)][1 + (R-1)q]^{-1}. \tag{2}$$

To model the decrease in the misclassification rate with increasing numbers of ICD code instances, consider the simple decay function for $f(x)$:

$$f(x) = (1-\delta)^x, \tag{3}$$

where δ is the parameter that can be estimated for each ICD code. Similarly consider the following form for n_x as a function of $n_{x=1}$ to model the reduction in the number of cases defined by using increasing numbers of instances of an ICD code:

$$n_x = n_{x=1}(1+\varepsilon)^{-x}, \tag{4}$$

where ε is the parameter that captures the rate of decline in case numbers as the definition for case status becomes more stringent with the use of larger numbers of ICD code instances and can also be estimated for each ICD code. The machinery is now in place for the calculation of statistical power to detect disease association at a genetic marker using data from linked ICD coding systems.

4. Results and Discussion

The above model is used to conduct an exploration of the impact of ICD code definitions on power. To obtain a value of x which maximizes power to detect genetic association, one can numerically solve the following differential equation for x:

$$\frac{\partial}{\partial x}\left[\frac{N(q-s)^2}{q(1-q)+s(1-s)}\right] = 0. \tag{5}$$

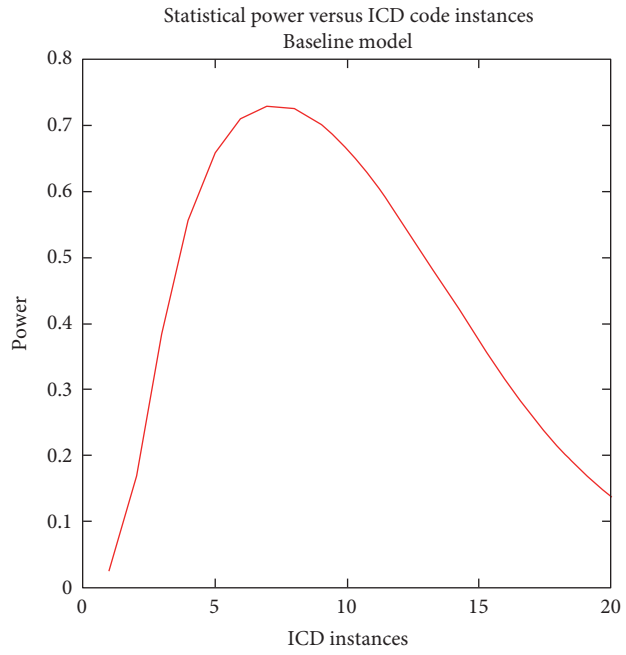

Figure 1: Statistical power versus ICD code instances, baseline model. From the mathematical model specified, power was calculated using the set of parameters from the baseline model. The results show the trade-off between the sample size, misclassification rates, and statistical power to detect genetic association. For the baseline model, the peak of power occurs when the number of instances is 7.

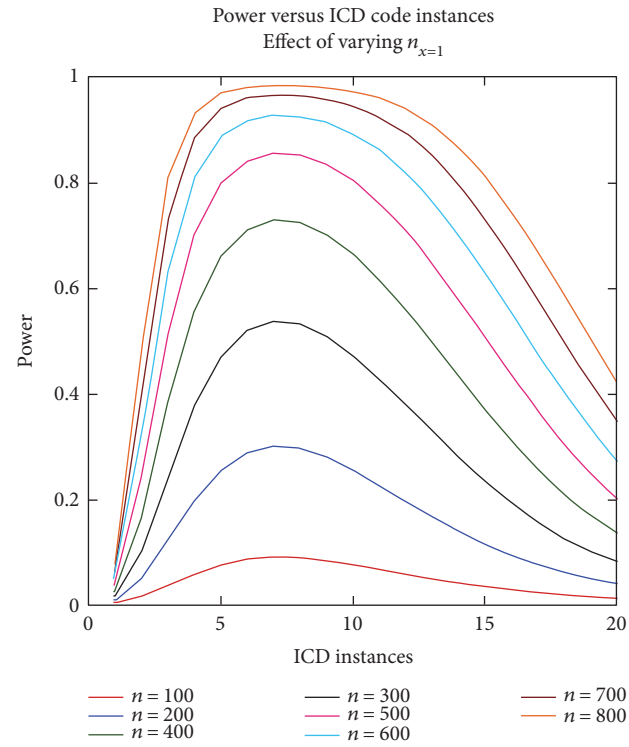

Figure 2: Power versus ICD code instances, effect of varying $n_{x=1}$. The baseline level was used to generate this figure with the exception of $n_{x=1}$, which varied from 100 to 800, and the resulting power was calculated for each number of ICD instances.

The solution to (5) can be solved through standard numerical methods applied to solving

$$n_{x=1}\left[(1-\delta)^2-1\right](q-1)(R-1)^2[F_1F_2\ln(1-\delta) + F_3(F_4-F_5)\ln(1+\varepsilon)] = 0, \tag{6}$$

where

$$\begin{aligned} F_1 &= -(1-\delta)^x[(1+\varepsilon)^x m + n_{x=1}][q(R-1)+1], \\ F_2 &= (1-\delta)^x + 2q(R-1)[(1-\delta)^x+1] + R - R(1-\delta)^x + 3, \\ F_3 &= m(1+\varepsilon)^x[(1-\delta)^x-1], \\ F_4 &= (1-\delta)^x + 1 + \left[(1-\delta)^{2x}+1\right]q^2(R-1)^2 + R - R(1-\delta)^x, \\ F_5 &= q(R-1)\{R(1-\delta)^x[(1-\delta)^x-1] - (1-\delta)^x \\ &\quad - (1-\delta)^{2x} - 2\}. \end{aligned} \tag{7}$$

The closest integer value to the value of x that solves this continuous equation can be used to optimize the power for a given set of parameters. To exemplify the use of this approach, let $m = 10,000$, $n_{x=1} = 400$, $R = 2$, $q = 0.20$, $\delta = 0.15$, and $\varepsilon = 0.15$. Call this set of parameters the baseline model. $x = 7.2265$ solves the differential equation. Therefore, using seven instances of an ICD code will yield the optimal design weighing the trade-off between the case sample size and the misclassification. For that set of parameters, Figure 1 shows the power curve for this set of parameters.

To investigate the power curves, varying the baseline number of cases ($n_{x=1}$), the calculations were performed as the $n_{x=1}$ varied from 100 to 800. Visual inspection shows the peak of power at approximately 7 instances. Figure 2 shows the results.

Next, to determine the role of the δ and ε parameters on the power curves, the calculations were performed fixing the other parameters. Figures 3 and 4 display these results.

5. Conclusions

Genetic data linked to longitudinal electronic health records can serve as a very useful tool in modern disease genetics. However, misclassification present in ICD coding systems can severely hamper large-scale screens using those codes for the purpose of genetic association studies. This work has described a simple approach to better understand the impact of misclassification present in EHR systems for the purpose of optimizing experimental designs that screen numerous ICD codes in genetic association studies. Under the mathematical models considered, the methods offer an approach to select the number of instances of an ICD code for the purpose of defining cases and obtaining an optimal experimental design for the identification of genetic markers. Additional work is needed in this area to improve disease

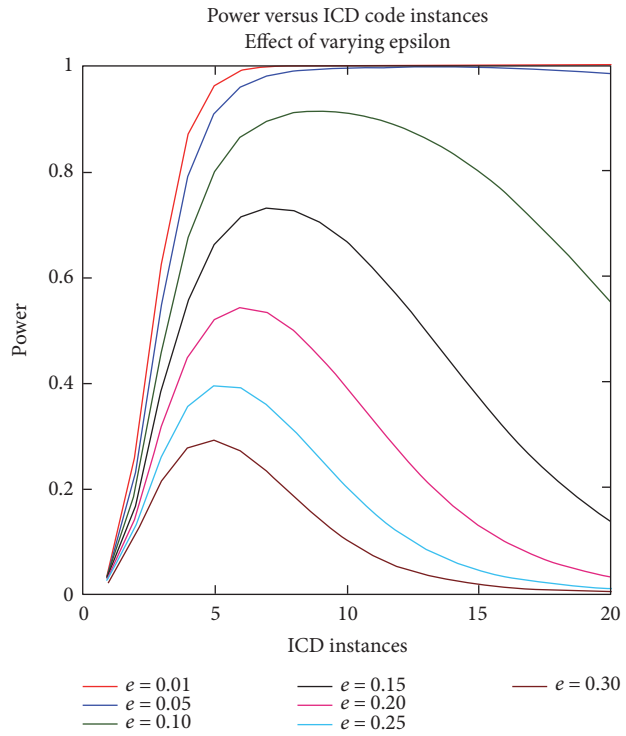

Figure 3: Power versus ICD code instances, effect of varying epsilon. The epsilon parameter varied in the baseline model from 0.01 to 0.30, and the power to detect was subsequently calculated.

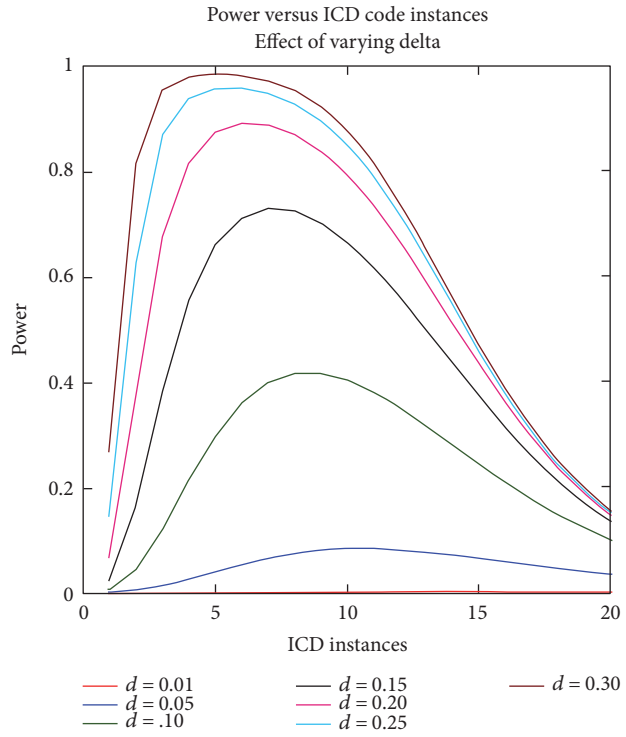

Figure 4: Power versus ICD code instances, effect of varying delta. To explore the effect of the delta parameter on the power calculations, the baseline model was modified to include values of delta from 0.01 to 0.30. The power to detect genetic association was calculated across these delta parameter values.

classification schemes for genetic association studies as well as for other investigations.

Disclosure

The content is solely the responsibility of the author and does not necessarily represent the official views of the National Institutes of Health.

Conflicts of Interest

The author declares that he has no conflicts of interest.

Acknowledgments

The author would like to thank his colleagues Mehdi Maadooliat, Zhan Ye, Scott Hebbring, Ahmad Pahlavan Tafti, and Peggy Peissig for the very useful conversations pertinent to this investigation. The research reported in this publication was supported by generous donors to the Marshfield Clinic, NIMH of the National Institutes of Health Award (4RO1MH097464-04), and the Institute for Clinical and Translational Research supported by the Clinical and Translational Science Award (CTSA) program and the National Center for Advancing Translational Sciences (NCATS) (Grant UL1TR000427).

References

[1] C. A. McCarty, B. N. Mukesh, P. F. Giampietro, and R. A. Wilke, "Healthy people 2010 disease prevalence in the Marshfield Clinic Personalized Medicine Research Project cohort: opportunities for public health genomic research," *Personalized Medicine*, vol. 4, no. 2, pp. 183–190, 2007.

[2] Y. Ko, M. Cho, J.-S. Lee, and J. Kim, "Identification of disease comorbidity through hidden molecular mechanisms," *Scientific Reports*, vol. 6, article 39433, 2016.

[3] M. R. Boland, Z. Shahn, D. Madigan, G. Hripcsak, and N. P. Tatonetti, "Birth month affects lifetime disease risk: a phenome-wide method," *Journal of the American Medical Informatics Association*, vol. 22, no. 5, pp. 1042–1053, 2015.

[4] L. G. Glance, T. M. Osler, D. B. Mukamel, W. Meredith, J. Wagner, and A. W. Dick, "TMPM-ICD9: a trauma mortality prediction model based on ICD-9-CM codes," *Annals of Surgery*, vol. 249, no. 6, pp. 1032–1039, 2009.

[5] J. M. Kinge, K. Saelensminde, J. Dieleman, S. E. Vollset, and O. F. Norheim, "Economic losses and burden of disease by medical conditions in Norway," *Heath Policy*, vol. 121, 2017.

[6] S. E. O'Brien, S. J. Schrodi, Z. Ye, M. H. Brilliant, S. S. Virani, and A. Brautbar, "Differential lipid response to statins is associated with variants in the BUD13-APOA5 gene region," *Journal of Cardiovascular Pharmacology*, vol. 66, no. 2, pp. 183–188, 2015.

[7] M. Icen, C. S. Crowson, M. T. McEvoy, S. E. Gabriel, and H. Maradit Kremers, "Potential misclassification of patients with psoriasis in electronic databases," *Journal of the American Academy of Dermatology*, vol. 59, no. 6, pp. 981–985, 2008.

[8] J. M. Evans and T. M. MacDonald, "Misclassification and selection bias in case-control studies using an automated database," *Pharmacoepidemiology and Drug Safety*, vol. 6, no. 5, pp. 313–318, 1997.

[9] J. A. Singh, A. R. Holmgren, and S. Noorbaloochi, "Accuracy of veterans administration databases for a diagnosis of rheumatoid arthritis," *Arthritis and Rheumatism*, vol. 51, pp. 952–957, 2004.

[10] W. Birman-Deych, A. D. Waterman, Y. Yan, D. S. Nilasena, M. J. Radford, and B. F. Gage, "Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors," *Medical Care*, vol. 43, no. 5, pp. 480–485, 2005.

[11] B. J. Edwards, C. Haynes, M. A. Levenstien, S. J. Finch, and D. Gordon, "Power and sample size calculations in the presence of phenotype errors for case/control genetic association studies," *BMC Genetics*, vol. 6, p. 18, 2005.

[12] F. Ji, Y. Yang, C. Haynes, S. J. Finch, and D. Gordon, "Computing asymptotic power and sample size for case-control genetic association studies in the presence of phenotype and/or genotype misclassification errors," *Statistical Applications in Genetics and Molecular Biology*, vol. 4, article 37, 2005.

[13] D. Gordon, C. Haynes, Y. Yang, P. L. Kramer, and S. J. Finch, "Linear trend tests for case–control genetic association that incorporate random phenotype and genotype misclassification error," *Genetic Epidemiology*, vol. 31, no. 8, pp. 853–870, 2007.

[14] M. Manchia, J. Cullis, G. Turecki, G. A. Rouleau, R. Uher, and M. Alda, "The impact of phenotypic and genetic heterogeneity on results of genome wide association studies of complex diseases," *PLoS One*, vol. 8, no. 10, article e76295, 2013.

[15] R. Duan, M. Cao, Y. Wu et al., "An empirical study for impacts of measurement errors on EHR based association studies," *American Medical Informatics Association Annual Symposium Proceedings*, vol. 2016, pp. 1764–1773, 2017.

[16] J. J. Bazarian, P. Veazie, S. Mookerjee, and E. B. Lerner, "Accuracy of mild traumatic brain injury case ascertainment using ICD-9 codes," *Academic Emergency Medicine*, vol. 13, no. 1, pp. 31–38, 2006.

[17] S. J. Hebbring, S. J. Schrodi, Z. Ye, Z. Zhou, D. Page, and M. H. Brilliant, "A PheWAS approach in studying *HLA-DRB1*1501*," *Genes and Immunity*, vol. 14, no. 3, pp. 187–191, 2013.

[18] J. B. Leader, S. A. Pendergrass, A. Verma et al., "Contrasting association results between existing PheWAS phenotype definition methods and five validated electronic phenotypes," *American Medical Informatics Association Annual Symposium Proceedings*, vol. 2015, pp. 824–832, 2015.