Profiling phenome-wide associations: a population-based observational study

RECEIVED 2 September 2014
REVISED 23 October 2014
ACCEPTED 2 November 2014
PUBLISHED ONLINE FIRST 5 February 2015

Shabbir Syed-Abdul[1], Max Moldovan[2,3], Phung-Anh Nguyen[1], Ruslan Enikeev[4], Wen-Shan Jian[5], Usman Iqbal[1], Min-Huei Hsu[6], Yu-Chuan Li[7]

AMIA INFORMATICS PROFESSIONALS. LEADING THE WAY. OXFORD UNIVERSITY PRESS

ABSTRACT

Objectives To objectively characterize phenome-wide associations observed in the entire Taiwanese population and represent them in a meaningful, interpretable way.

Study Design In this population-based observational study, we analyzed 782 million outpatient visits and 15 394 unique phenotypes that were observed in the entire Taiwanese population of over 22 million individuals. Our data was obtained from Taiwan's National Health Insurance Research Database.

Results We stratified the population into 20 gender-age groups and generated 28.8 million and 31.8 million pairwise odds ratios from male and female subpopulations, respectively. These associations can be accessed online at http://associations.phr.tmu.edu.tw. To demonstrate the database and validate the association estimates obtained, we used correlation analysis to analyze 100 phenotypes that were observed to have the strongest positive association estimates with respect to essential hypertension. The results indicated that association patterns tended to have a strong positive correlation between adjacent age groups, while correlation estimates tended to decline as groups became more distant in age, and they diverged when assessed across gender groups.

Conclusions The correlation analysis of pairwise disease association patterns across different age and gender groups led to outcomes that were broadly predicted before the analysis, thus confirming the validity of the information contained in the presented database. More diverse individual disease-specific analyses would lead to a better understanding of phenome-wide associations and empower physicians to provide personalized care in terms of predicting, preventing, or initiating an early management of concomitant diseases.

Keywords: phenotype, association, electronic health records, disease complications

BRIEF COMMUNICATION

INTRODUCTION

Over a lifetime, a person can experience multiple diseases. Some diseases are considered to be concomitant diseases or complications of preexisting diseases, whereas others are considered to be independent. Despite rapid advances in the biomedical sciences, empirical phenome-wide associations remain a relatively inadequately studied area.[1,2] Traditional methods for exploring phenome-wide associations rely mostly on case reports and cohort studies that investigate one-on-one disease relationships.[3] The rapid adoption of electronic health records has resulted in the accumulation of an unprecedented amount of patient-level data by health care providers.[4,5] It would be highly valuable to analyze such patient-level clinical data in a way that would aid physicians to predict or detect at an early stage concomitant diseases to initiate preventive and curative measures.[6] The present study took advantage of the fact that in Taiwan's highly accessible and universal health care system, people visit their physicians on an average of 15 times per year for even the slightest of illnesses such as the common cold. Moreover, because the National Health Insurance Bureau crosschecks all diagnostic and medication codes before reimbursement, the codes are of good fidelity. Thus, the Taiwanese health system presents a unique opportunity to observe the dynamics of disease-wide associations. The study aims to objectively characterize phenome-wide associations observed in the entire population and represent them in a meaningful, interpretable way.

METHODS

In this population-based observational study, we analyzed 782 million outpatient visits and 15 394 unique diagnosis codes that were observed in the entire Taiwanese population of over 22 million individuals. We obtained our data from Taiwan's National Health Insurance Research Database (NHIRD). Using a mapping tool, all diagnoses (phenotypes) from the claim records were converted from NHI codes to International Classification of Diseases (ICD), Ninth Revision, Clinical Modification (ICD9-CM) codes of up to 5 digits.[7] This study was based on the assumption that 2 diseases co-occurred if and only if they were recorded for the same individual within the 36-month observation window despite the number of times a phenotype was recorded for the same individual. Any disease occurring outside the observation window was excluded from the analysis. Each subject was attributed to a certain gender-age–specific group (male and female, aged 0 to 9, 10 to 19, . . . , and 90 to 99 years) and followed for 36 months from January 1, 2000 through December 31, 2002. Against each pair of ICD identified disorders, we recorded the number of disease occurrences and pairwise co-occurrences. Knowing the size of a given gender-age–stratified subpopulation segment, we assumed that the

Correspondence to Dr Yu-Chuan Li, Graduate Institute of Biomedical Informatics, College of Medicine Science and Technology, Chair, Department of Dermatology, Wan Fang Hospital, Taiwan. Taipei Medical University, WuXing Street No. 250 Taipei, 11031 Taiwan; jack@tmu.edu.tw

For numbered affiliations see end of article.

number of pairwise co-occurrences were distributed by a (noncentral) hypergeometric distribution characterized by a single parameter given by an odds ratio— the odds of a person being affected by one disorder given the presence of another disorder.[8] We have recorded the odds ratio estimates, complemented by *P* values (under the null hypothesis of the odds ratio is less than 1) and 95% confidence limits based on hypergeometric testing (see Rivals *et al.*). We did not apply multiple comparison corrections to the resulting *P* values.

Moldovan *et al.*[9] comprehensively described the methodology of the underlying objective characterization, which also revealed some potential limitations behind the resulting estimates. For example, one violation of the hypergeometric distribution assumption is the potential dependence between disease occurrences observed in different subjects, as in the case for infectious diseases. One more limitation behind the suggested objective characterization is the specifics of how disorders are recorded by medical practitioners, who may often assign closely related but essentially different ICD diagnoses to the same medical condition. This would cause a bias toward upward pairwise association estimates between closely related conditions (which are unlikely to be of a major interest anyway) but would not distort pairwise associations with other disorders. While recognizing the limitations of the objective characterization methodology suggested, we maintain that pairwise disease-wide association estimates still bring unique and valuable information if used with proper discretion.[10]

To validate the information contained in the pairwise association estimates and to provide a demonstration of the presented database, we performed the following straightforward correlation analysis. First, we arbitrarily selected essential hypertension (ICD9-CM 401.9) as a reference disorder, which was commonly observed in both sexes of the groups aged 40 years and older. Next, we selected 100 disorders with the greatest odds ratio estimates (strongest positive association) that were common among the groups aged 40 to 49 through 90 to 99 years. We expected to observe a certain monotonic change in association patterns across the groups. After observing the monotonic change and noting that there was no overlap between the subjects in each of the considered gender-age–stratified groups, we were able to count the outcome as an elementary validation for the information contained in the phenome-wide association database we used.

RESULTS

The results of the Spearman rank correlation for the top 100 phenotypes most strongly and positively associated with essential hypertension have been categorized into gender-age groups in table 1.

As expected, association patterns were closely related between adjacent age groups (reflected in stronger positive correlations), while the correlation strength tended to diminish with an increased distance in age between individuals. The monotonic decrease in correlation coefficients persisted across the majority of age group pairs, with the single exception of females aged 80–89 years vs groups aged 90–99 years. This finding would benefit from further investigation.

The monotonic divergence between association patterns with the increasing age gap between groups can be explained through population dynamics; a prevalence of individual disorders were observed across age groups, complemented by a decreasing number of individuals in older age groups. Once again, the presented analysis is largely illustrative, though it can be further extended to different contexts and directions. For example, the recorded information on the statistical uncertainty accompanying the odds ratio estimates obtained from the data can be incorporated into an analysis involving *P* values and confidence limits, leading to an entirely new scope of information and related analysis opportunities.

DISCUSSION

Using a phenome-wide association approach, we developed a comprehensive database that can be accessed online at http://associations.phr.tmu.edu.tw/index.php/. Further, using correlation analysis, we analyzed 100 phenotypes that were observed to have the strongest positive association estimates with respect to essential hypertension. The Spearman rank correlation results revealed that phenome-wide associations significantly differed with age. For example, a woman with obesity in her 20s (ICD9-CM 278.0) had higher odds of experiencing hyperlipidemia, diabetes mellitus, or hypothyroidism; whereas a woman in her 50s was prone to experience Cushing syndrome, corticoadrenal insufficiency, or hypothyroidism; and a woman with obesity in her 70s had higher odds of developing a mental disorder, Cushing syndrome, or enthesopathy of the knee. This observational database can be used for identifying potential associations that warrant further consideration but should not imply causality.

In the last two decades, extensive research has been conducted on genome-wide associations,[1,11,12] protein interactions,[13–16] and gene expression profiles[17,18] to understand the pathophysiological mechanisms of selected diseases such as hereditary diseases,[19–21] cancers,[22,23] and neurological disorders.[24–26] However, there is no comprehensive understanding of the mechanisms underlying most diseases and their associations.[1,2] The presented phenome-wide information resource enables the development and application of functional methodologies that can lead to a better understanding of previously identified biological mechanisms as well as biological phenomena unnoticed before.[27]

CONCLUSIONS

Better understanding of phenome-wide associations will empower physicians to provide personalized care in terms of predicting, preventing, or initiating an early management of concomitant diseases. This study adopted a reverse translational approach[28] to analyze the large phenotypic data source from the NHIRDB. We argue that this type of approach to understanding phenome-wide associations is a critical milestone toward the comprehensive understanding of complex associations among diseases. Phenome-wide associations will provide an array of opportunities and challenges with regard to classifying diseases and developing new treatment approaches.[10] We made this database open for researchers who are interested in exploring phenome-wide associations.

CONTRIBUTORS

SS-A, P-AN, UI, W-SJ, and Y-CL contributed to the conception and design of the work. SS-A, MM, RE, P-AN, and UI were involved with the acquisition, analysis, or interpretation of the data. SS-A, UI, W-SJ, and Y-CL drafted the work or revised it critically for important intellectual content. SS-A, W-SJ, and Y-CL approved the final version. SS-A, P-AN, UI, MM, RE, W-SJ, M-HH, and Y-CL agreed to be accountable for all aspects of the work, ensuring that questions related to the accuracy or integrity of any part of the work would be appropriately investigated and resolved.

FUNDING

SS-A, P-AN, UI, W-SJ, and Y-CL have been sponsored in part by the National Science Council (NSC), Taiwan, under grants NSC100-2622-E-038-001-CC2(1/2), NSC99-2511-S-038-005-MY3, NSC 100-2320-B-038-034, and NSC100-2325-B-038-006; by the Department of Health, Executive Yuan, Taiwan, under grant DOH101-TD -C-111-008; and by Taipei Medical

Table 1: The Spearman Rank Correlation for the Top 100 Phenotypes Associated with Essential Hypertension

Female and Male Groups by Age, y						
Age, y	40–49	50–59	60–69	70–79	80–89	90–99
Female						
40–49	1.00					
50–59	0.956**	1.00				
60–69	0.860**	0.936**	1.00			
70–79	0.599**	0.731**	0.885**	1.00		
80–89	0.273**	0.382**	0.546**	0.763**	1.00	
90–99	0.060	0.153	0.266**	0.353**	0.265**	1.00
Male						
40–49	1.00					
50–59	0.965**	1.00				
60–69	0.825**	0.909**	1.00			
70–79	0.595**	0.707**	0.863**	1.00		
80–89	0.342**	0.427*	0.611**	0.815**	1.00	
90–99	0.158	0.202*	0.234*	0.254*	0.246*	1.00
Females vs Males						
Female and Male Groups by Age, y	Spearman Rank Correlation (ρ)					
40–49	0.863***					
50–59	0.868***					
60–69	0.733***					
70–79	0.727***					
80–89	0.597***					
90–99	0.408***					

*Correlation is significant at the 0.05 level (2-tailed).
**Correlation is significant at the 0.01 level (2-tailed).
***Correlation is significant at the 0.001 level (2-tailed).

University, under grant A0051-4100. MM and RE received no formal funding toward the project.

COMPETING INTERESTS

None.

REFERENCES

1. Baker M. Genomics: the search for association. *Nat.* 2010;467(7319): 1135–1138.
2. Li Y, Huang J, Amos CI. Genetic association analysis of complex diseases incorporating intermediate phenotype information. *PLoS One.* 2012;7(10): e46612.
3. Siri-Tarino PW, Sun Q, Hu FB, Krauss RM. Meta-analysis of prospective cohort studies evaluating the association of saturated fat with cardiovascular disease. *Am J Clin Nutr.* 2010;91(3):535–546.
4. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. *Nat Rev Genet.* 2012;13 (6): 395–405.
5. Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. *J Am Med Inform Assoc.* 2013;20(1):117–121.
6. Roque FS, Jensen PB, Schmock H, *et al.* Using electronic patient records to discover disease correlations and stratify patient cohorts. *PLoS Comput Biol.* 2011;7(8):e1002141.
7. Hebbring SJ. The challenges, advantages and future of phenome-wide association studies. *Immunol.* 2014;141(2):157–165.
8. Rivals I, Personnaz L, Taing L, Potier MC. Enrichment or depletion of a GO category within a class of genes: which test? *Bioinformatics. 2007*;23(4): 401–407.
9. Moldovan M, Enikeev R, Syed-Abdul S, Nguyen P-A, Chang Y-C, Li Y-C. Disease universe: visualisation of population-wide disease-wide associations. *Adv Sys Sci Appl.* 2013;14:144–157.
10. Neuraz A, Chouchana L, Malamut G, *et al.* Phenome-wide association studies on a quantitative trait: application to TPMT enzyme activity and thiopurine therapy in pharmacogenomics. *PLoS Comput Biol.* 2013;9(12):e1003405.
11. Hirschhorn JN, Daly MJ. Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet.* 2005;6(2):95–108.

12. Vidal M, Cusick ME, Barabási AL. Interactome networks and human disease. *Cell*. 2011;144(6):986–998.
13. Lingappa UF, Wu X, Macieik A, *et al*. Host-rabies virus protein-protein interactions as druggable antiviral targets. *Proc Natl Acad Sci USA*. 2013;110(10):E861-E868.
14. Lee Y, Li H, Li J, *et al*. Network models of genome-wide association studies uncover the topological centrality of protein interactions in complex diseases. *J Am Med Inform Assoc*. 2013;(4):619–629.
15. Gonzalez MW, Kann MG. Chapter 4: protein interactions and disease. *PLoS Comput Biol*. 2012;8 (12):e1002819.
16. Vidal M, Chan DW, Gerstein M, *et al*. The human proteome - a scientific opportunity for transforming diagnostics, therapeutics, and healthcare. *Clin Proteomics*. 2012;9(1):6.
17. Li L, Shiga M, Ching WK, Mamitsuka H. Annotating gene functions with integrative spectral clustering on microarray expressions and sequences. *Genome Inform*. 2010;22:95–120.
18. Ray SS, Bandyopadhyay S, Pal SK. Gene ordering in partitive clustering using microarray expressions. *J Biosci*. 2007;32(5):1019–1025.
19. Zhi D, Chen R. Statistical guidance for experimental design and data analysis of mutation detection in rare monogenic Mendelian diseases by exome sequencing. *PLoS One*. 2012;7(2):e31358.
20. Li MX, Gui HS, Kwan JS, Bao SY, Sham PC. A comprehensive framework for prioritizing variants in exome sequencing studies of Mendelian diseases. *Nucleic Acids Res*. 2012;40(7):e53.
21. Jin W, Qin P, Lou H, Jin L, Xu S. A systematic characterization of genes underlying both complex and Mendelian diseases. *Hum Mol Genet*. 2012; 21(7):1611–1624.
22. Dunlop MG, Dobbins SE, Farrington SM, *et al*. Common variation near CDKN1A, POLD3 and SHROOM2 influences colorectal cancer risk. *Nat Genet*. 2012;44(7):770–776.
23. Kim MS, Lee J, Oh T, *et al*. Genome-wide identification of OTP gene as a novel methylation marker of breast cancer. *Oncol Rep*. 2012;27(5): 1681–1688.
24. Yeo GS, Heisler LK. Unraveling the brain regulation of appetite: lessons from genetics. *Nat Neurosci*. 2012;15(10):1343–1349.
25. Sturiale CL, Puca A, Sebastiani P, *et al*. Single nucleotide polymorphisms associated with sporadic brain arteriovenous malformations: where do we stand? *Brain. 2013*;136(Pt 2):665–681.
26. Malmeström C, Gillett A, Jernås M, *et al*. Serum levels of LIGHT in MS. *Mult Scler*. 2012.
27. Denny JC, Ritchie MD, Basford MA, *et al*. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics*. 2010;26(9):1205–1210.
28. Flintoft L. Disease genetics: phenome-wide association studies go large. *Nat Rev Genet*. 2014;15(1):2.

AUTHOR AFFILIATIONS

[1]Graduate Institute of Biomedical Informatics, College of Medical Science and Technology, Taipei Medical University, Taipei, Taiwan

[2]Centre for Clinical Governance Research, Australian Institute of Health Innovation, Faculty of Medicine, University of New South Wales, Sydney, Australia

[3]School of Population Health, Sansom Institute for Health Research, University of South Australia, South Australian Health & Medical Research Institute (SAHMRI)

[4]The APAC Sale Group, Singapore

[5]School of Health Care Administration, Taipei Medical University, Taipei, Taiwan

[6]Bureau of International Cooperation, Department of Health, Taipei, Taiwan

[7]Graduate Institute of Biomedical Informatics, College of Medicine Science and Technology; Department of Dermatology, Wan Fang Hospital, Taiwan. Taipei Medical University, Taipei, Taiwan