American Journal of Epidemiology

Vol. 188, No. 5
DOI: 10.1093/aje/kwz031
Advance Access publication:
March 16, 2019

Commentary

Signals Among Signals: Prioritizing Nongenetic Associations in Massive Data Sets

Arjun K. Manrai, John P. A. Ioannidis, and Chirag J. Patel*

* Correspondence to Dr. Chirag J. Patel, Department of Biomedical Informatics, Harvard Medical School, 10 Shattuck Street, Room 302, Boston, MA, 02115 (e-mail: chirag_patel@hms.harvard.edu).

Initially submitted October 30, 2018; accepted for publication January 30, 2019.

Massive data sets are often regarded as a panacea to the underpowered studies of the past. At the same time, it is becoming clear that in many of these data sets in which thousands of variables are measured across hundreds of thousands or millions of individuals, almost any desired relationship can be inferred with a suitable combination of covariates or analytic choices. Inspired by the genome-wide association study analysis paradigm that has transformed human genetics, X-wide association studies or "XWAS" have emerged as a popular approach to systematically analyzing nongenetic data sets and guarding against false positives. However, these studies often yield hundreds or thousands of associations characterized by modest effect sizes and miniscule *P* values. Many of these associations will be spurious and emerge due to confounding and other biases. One way of characterizing confounding in the genomics paradigm is the genomic inflation factor. An analogous "X-wide inflation factor," denoted λ_X, can be defined and applied to published XWAS. Effects that arise in XWAS may be prioritized using replication, triangulation, quantification of measurement error, contextualization of each effect in the distribution of all effect sizes within a field, and pre-registration. Criteria like those of Bradford Hill need to be reconsidered in light of exposure-wide epidemiology to prioritize signals among signals.

big data; inflation factor; machine learning; *P* values; X-wide association study

Abbreviations: GWAS, genome-wide association study; Q-Q, quantile-quantile; SNP, single nucleotide polymorphism; XWAS, X-wide association study.

It is currently common for investigators to amass data from thousands to millions of individuals in observational epidemiologic investigations of large populations, health systems, or entire countries. Consortia harmonize data across dozens of cohorts, and institutional and national biobanks (1) are merging patient records with biorepositories. In principle, these investigations may transform discovery. However, the larger the data set, the greater the chance of it being "overpowered" to detect small associations that are of limited clinical significance or entirely spurious (2).

Agnostic analysis frameworks have emerged as a way to guard against spurious findings in massive parallel investigations by addressing issues of multiplicity, candidate variable selection, and confounding. The classic example is the genome-wide association study (GWAS), an approach in human genetic epidemiology in which several million common single nucleotide polymorphisms (SNPs) are systematically tested for associations with a phenotype. SNPs are deemed genome-wide significant if they meet stringent significance cutoffs (e.g., $P < 5 \times 10^{-8}$), and standard methods exist to correct for confounding due to differences between cases and controls in ancestry (population stratification). Although the study design, predictive capability, and biological relevance of GWAS have been debated (e.g., see Visscher et al. (3) for a summary), one can also argue that the reproducible genetic leads have dramatically improved (4).

Although nongenetic investigations, such as those in nutritional or environmental epidemiology, have traditionally associated one or a handful of variables at a time with an outcome, emerging technologies are measuring more and more of the exposome, the comprehensive set of exposures encountered from birth to death (5). GWAS have inspired a range of X-wide association studies (XWAS) (6) to deal with nongenetic data in a similarly agnostic way. X denotes an entire domain of variables (e.g., environment-wide (7), nutrient-wide (8), medication-wide (9), or sociodemographic-wide (10) variables) executed in a variety of academic and industrial institutions with access to large data sets, such as Stanford University (Stanford, California), Imperial

College London (London, United Kingdom), Harvard University (Cambridge, Massachusetts), Columbia University (New York, NY), Penn State University (State College, PA), and the Marshfield Clinic (Marshfield, WI). In the big data setting, in which power approaches 100%, XWAS often yield numerous results that survive multiple testing and covariate adjustment. A central challenge is identifying which signals among the many hits are causal and therefore relevant and actionable for medicine or public health.

XWAS has substantial differences from GWAS. First, nongenetic variables are more heterogeneous, densely correlated (11), and time dependent and are often measured with substantial error (12, 13). SNPs, by contrast, are static, locally and predictably dependent (in "haploblocks" along the chromosome), and well-measured. Second, confounding is often difficult to address in XWAS. In GWAS, the primary source of confounding is population stratification (14). Because ancestry is strongly tied to genotype frequency, associations between genotypes and a phenotype of interest are biased if ancestry is not balanced between cases and controls. Methods that can infer ancestry have made adjustment for population stratification routine (14), although they have yet to penetrate all clinical applications (15). As we discuss below, confounding in XWAS presents thornier challenges.

THE GENOMIC INFLATION FACTOR DETECTS CONFOUNDING (AND POLYGENICITY)

Confounding in GWAS can be detected by examining the distribution of test statistics (e.g., χ^2 test statistics or P values) to measure how it deviates from a null distribution (16), often visualized as a quantile-quantile (Q-Q) plot of test statistics. In these plots, test statistics are ranked from lowest to highest and plotted against the corresponding test statistics under the null (17). In GWAS, an assumption often used is that most genotypes are not associated with the outcome. Thus, substantial deviations from the diagonal imply systematic differences between cases and controls or "polygenicity," in which many small genetic effects contribute to the phenotype (18, 19). The degree of deviation is called the genomic inflation factor, denoted by λ_G, and is the ratio of the median observed test statistic divided by the median test statistic under the null. A λ_G close to 1 suggests acceptable control of confounding; a λ_G greater than 1 is often indicative of systematic bias. After appropriate control, for example, by applying principal components analysis (14), GWAS Q-Q plots that initially showed sharp departures from the diagonal are often corrected. However, it has been shown (18) that even if confounding from population stratification and cryptic relatedness is eliminated, λ_G may still deviate from 1. Specifically, it may deviate more when there is strong and common linkage disequilibrium among the tested genetic variants and when there are more causal genetic factors and larger heritability. Finally, λ_G increases with larger sample sizes.

THE X-WIDE INFLATION FACTOR LIKELY IMPLIES MOSTLY MASSIVE CONFOUNDING AND CORRELATED STATISTICAL TESTS

In conducting a Q-Q analysis in a typical nongenetic XWAS, we observe that there is substantial departure from the diagonal, often much more than what is typically seen in GWAS (Figure 1). We define the X-wide inflation factor, denoted λ_X, as the ratio of median test statistics in a nongenetic XWAS to that expected under the null, which can be calculated as:

$$\lambda_X = \text{median}(-\log_{10}(p_{\text{observed}}))/\text{median}(-\log_{10}(p_{\text{null}})).$$

Of note, in a GWAS, λ_G is usually computed as the median of the χ^2 test statistics observed across the SNPs as opposed to the observed P values directly. Because XWAS use a range of statistical models (e.g., Cox proportional hazards regression (20), survey-weighted linear regression (21)), further work should evaluate whether a robust analogue exists across XWAS. In

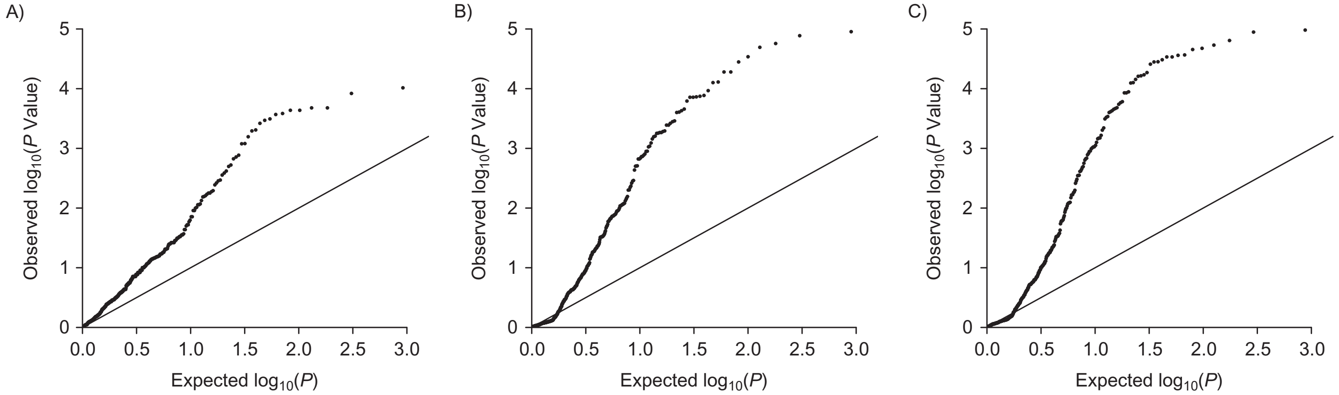

Figure 1. Quantile-quantile (Q-Q) plots for X-wide association studies (XWAS). All XWAS Q-Q plots show substantial deviation from the diagonal (solid black line) expected under the null. Equivalently, the Q-Q plots have X-wide inflation factors, defined as the $\text{median}(-\log_{10}(p_{\text{observed}}))/\text{median}(-\log_{10}(p_{\text{null}}))$, that are substantially greater than 1. A) Q-Q plot for XWAS correlating exposures, behaviors, and clinical variables with telomere length in Patel et al. (20) (inflation factor = 1.59; sample size = 20,997; number of associations tested = 461). B) Q-Q plot for medication-wide association study for breast cancer from Patel et al. (21) (inflation factor = 2.18; sample size = 9,014,975; number of associations tested = 536). C) Q-Q plot from medication-wide association study for prostate cancer in Patel et al. (20) (inflation factor = 2.39; sample size = 9,014,975; number of associations tested = 528).

existing XWAS, we have observed in practice that λ_X tends to be much greater than 1. Figure 1 shows Q-Q plots of test statistics from previous XWAS that associate a set of exposures, behaviors, and clinical variables with telomere length (21) (Figure 1A) and a prescription-wide association study in breast cancer (Figure 1B) and in prostate cancer (20) (Figure 1C). The λ_X values are 1.59 (sample size $N = 20{,}997$; number of tested associations = 461), 2.18 ($N = 9{,}014{,}975$; number of tested associations = 536), and 2.39 ($N = 9{,}014{,}975$; number of tested associations = 528), respectively.

This gross deviation likely implies several features that characterize nongenetic XWAS: 1) model misspecification and residual confounding (much stronger in XWAS than in GWAS), 2) densely correlated factors (exposures tend to be much more widely correlated than genetic factors that are correlated through linkage disequilibrium), and/or 3) proportionally more nongenetic factors that are causally associated with outcomes and/or a larger proportion of the variance that is explained by nongenetic factors than by genetic factors. The latter explanation needs to be examined on a case-to-case basis in each XWAS, but very often the impact of confounding and dense correlation may be far more influential than the impact of genuine causality (2, 11, 22, 23).

Even if causally related factors can be reliably identified, there is an extra challenge to determine which of those are sizeable enough to have clinical or public health impact, how easy it is to modify them in real-life, and how we can verify that their modification yields the desired improvements. Regardless, unless we can manage to select an initial set of factors that are enriched in truly causal ones, these additional considerations remain moot from the perspective of intervention. However, noncausal factors may still be of some use in prognostic/predictive models.

THE X-WIDE INFLATION FACTOR WILL GROW WITH LARGE BIOBANK STUDIES

Q-Q plots seen in XWAS to date are heralds of those that will emerge from future nongenetic association studies from much larger biobanks and other cohort data that are becoming commonplace around the world. Deviation for nongenetic association studies (e.g., nutritional, clinical, and environmental exposures) will increase in larger sample size regimens. Instead of typical observational studies of exposures with several thousands of subjects, current and future biobanks and observational studies will include hundreds of thousands to millions of individuals, with thousands of measurements per individual. We will be fully powered to detect tiny effects; however, most will be spurious, and among those that are not spurious, only a subset of yet unpredictable volume will be clinically important, actionable, or even just biologically insightful.

As seen in GWAS, sample size influences the value and therefore the interpretation of λ_G. For example, in a case-control study (24) of $N/2$ cases and $N/2$ controls that have a genetic distance f (measured by the fixation index, or F_{ST}) from one another that, uncorrected, would confound the association study, λ_G is given by $1 + N \times f$. In other words, a λ_G of 1.3 with $N = 300$ is more concerning than the same λ_G with $N = 30{,}000$ because f is 0.3/300 = 0.001 in the former but 0.3/30,000 = 0.00001 in the latter. Methods like linkage disequilibrium score regression (19) have been developed to detect signals of polygenicity in large genetic studies, and analogous work for nongenetic XWAS would likely be fruitful. The same principle should hold with λ_X. In settings in which very large cohorts are involved, such as the Swedish Cancer Register Q-Q plot shown in Figure 1B and 1C, it is expected that the inflation factor will be increased even given the same degree of confounding and other things being equal.

Of note, test statistics in GWAS are routinely adjusted for the stratification and genomic inflation. We argue that the corresponding correction for the X-wide inflation factor should be used also in XWAS. Given the large values of these inflation factors, these corrections may be very consequential.

These observations call for both caution in the interpretation of associations that stem from future large-scale XWAS and the development of new methods that can separate the "signals from signals" and identify which of the thousands of statistically significant associations are worth prioritizing.

PRIORITIZING FINDINGS GOING FORWARD

Prioritization of XWAS signals is a related but distinct challenge from improving the overall reproducibility or replicability of the scientific literature. For the latter, it is important, for

Table 1. Approaches to Prioritizing Signals Among the Many Significant Findings From Massive Data Sets: Replication, Triangulation, Contextualization, Error Quantification, and Pre-Registration

Approach	Strengths	Limitations
Replication: statistical consistency across independent data sets	Often easy to do some version of replication even using the original data set (e.g., cross-validation, held out data)	Can contain the same biases (e.g., confounding, measurement error) if using the same data; may not address residual confounding
Triangulation: assimilate findings across methodologies and data	Leverages differing approaches and data, each with their own assumptions; moves towards causality	May be expensive or difficult to achieve/coordinate; subjective how to weigh complementary but distinct approaches
Contextualization: assess associations across a field of study	Allows for a relationship to be framed among an entire field of similar investigations; enables meta-analysis	Data hungry; may require investigators to "spend" statistical power (in correcting for hypotheses that may not be important)
Error quantification: estimate misclassification of exposures and outcomes	Potential to remove some of the bias that is epidemic in some fields (e.g., nutritional epidemiology)	Often difficult to execute or remeasure a previously studied population; may be difficult to get funding
Pre-registration: prespecify hypotheses	Helps mitigate publication and other selective reporting biases	May be perceived of as reducing scientists' creativity or independence

example, to devise ways to enhance data and code sharing, as well as collaborative team science (23). Nevertheless, in large biobanks and observational studies with thousands of significant associations, we may have widely shared data and code but still lack a systematic approach with which we can identify robust and clinically meaningful relationships.

Several steps might help going forward (Table 1). First, replication in both held out and independent data offers the opportunity to test the stability of associations across different cohorts and settings. We need more empirical data on how heterogeneous epidemiologic signals are in XWAS settings. Second, triangulating a candidate association using multiple sources of other evidence (25), each of which has its own strengths, weaknesses, and biases, may help move beyond associations from an initial large-scale analysis toward cause-and-effect relationships. Third, considering individual association effect sizes and where they lie in the distribution of all possible effects in addition to statistical significance allows for richer analysis. The same effect size may have different importance in a field in which most effect sizes are as large compared with a field in which equally sizeable effects are rare. Moreover, it is traditionally taught that large effects are more reliable than small ones, but this needs to be revisited in the XWAS setting, where (like in GWAS) very large effects may point mostly to errors (26). Fourth, large measurement error for many nongenetic studies (e.g., nutrition recall studies) may dwarf other considerations. In these situations, careful consideration of whether the measurement error is likely to be independent nondifferential or differential may help understand the impact of misclassification on a large scale. Massive data and complex analyses are unlikely to salvage error-laden data, much like a Lamborghini engine may not help if placed on a wooden cart. Finally, even when data and code are shared, it is often unclear what investigations took place and remain unreported, including what models (e.g., which covariate combinations (27)) were explored and discarded (e.g., as "negative"). Efforts to pre-register fully specified hypotheses might help reduce publication and selective analysis bias. When a sizeable set of convincingly null associations is known (from strong mechanistic reasoning and prior large-scale data), these sets of null associations can be used as prespecified falsification endpoints (28) and new proposed discoveries can be calibrated against them. A Q-Q plot could be generated for the falsification endpoints to be used as guidance for other tested associations. Other causality criteria of Bradford Hill may also be considered, but they need to be probed for their validity and perhaps recast in the XWAS setting (26). For example, one Bradford Hill criterion includes strength or size of association, where Hill writes ". . . we must not be too ready to dismiss a cause and effect hypothesis merely on the grounds that the observed observation appears to be slight . . ." (29, p. 296). In XWAS, all associations will be slight; therefore, associations could be recast into comparing those slight sizes with prior reported associations that leverage different methodologies or study designs (30).

Massive data sets offer the promise of achieving a more comprehensive understanding of human health and disease through simultaneous systematic analyses across thousands of variables. XWAS can move beyond piecemeal candidate studies on diverse exposures and outcomes, but now the challenge shifts to prioritizing select signals among the multitude of those deemed significant.

ACKNOWLEDGMENTS

Author affiliations: Computational Health Informatics Program, Boston Children's Hospital, Boston Massachusetts (Arjun K. Manrai); Department of Biomedical Informatics, Harvard Medical School, Boston, Massachusetts (Arjun K. Manrai, Chirag J. Patel); Department of Pediatrics, Harvard Medical School, Boston, Massachusetts (Arjun K. Manrai); Stanford Prevention Research Center, Department of Medicine, Stanford University, Stanford, California (John P. A. Ioannidis); Department of Health Research and Policy, Stanford University, Stanford, California (John P. A. Ioannidis); Department of Biomedical Data Science, Stanford University, Stanford, California (John P. A. Ioannidis); and Department of Statistics, Stanford University, Stanford, California (John P. A. Ioannidis).

This work was supported by the Laura and John Arnold Foundation, the National Institute of Environmental Health Sciences (grants R00 R00ES023504 and R21 R21ES025052), the National Institute of Allergy and Infectious Diseases (grant R01127250), the National Science Foundation (grant 1636870), and the National Heart, Lung, and Blood Institute (grant K01HL138259).

The funding agencies had no role in the study design and opinions here.

Conflict of interest: none declared.

REFERENCES

1. Hsing AW, Ioannidis JP. Nationwide population science: lessons from the Taiwan National Health Insurance Research Database. *JAMA Intern Med*. 2015;175(9):1527–1529.
2. Khoury MJ, Ioannidis JP. Medicine. Big data meets public health. *Science*. 2014;346(6213):1054–1055.
3. Visscher PM, Brown MA, McCarthy MI, et al. Five years of GWAS discovery. *Am J Hum Genet*. 2012;90(1):7–24.
4. Visscher PM, Wray NR, Zhang Q, et al. 10 years of GWAS discovery: biology, function, and translation. *Am J Hum Genet*. 2017;101(1):5–22.
5. Wild CP. The exposome: from concept to utility. *Int J Epidemiol*. 2012;41(1):24–32.
6. Fallin MD, Kao WHL. Is "X"-WAS the future for all of epidemiology? *Epidemiology*. 2011;22(4):457–459.
7. Patel CJ, Bhattacharya J, Butte AJ. An environment-wide association study (EWAS) on type 2 diabetes mellitus. *PLoS One*. 2010;5(5):e10746.
8. Tzoulaki I, Patel CJ, Okamura T, et al. A nutrient-wide association study on blood pressure. *Circulation*. 2012; 126(21):2456–2464.
9. Ryan PB, Madigan D, Stang PE, et al. Medication-wide association studies. *CPT Pharmacometrics Syst Pharmacol*. 2013;2(9):1–12.
10. Patel CJ, Bhattacharya J, Ioannidis JP, et al. Systematic identification of correlates of HIV infection: an X-wide association study. *AIDS*. 2018;32(7):933–943.
11. Patel CJ, Manrai AK. Development of exposome correlation globes to map out environment-wide associations. *Pac Symp Biocomput*. 2015;231–242.
12. Ioannidis JP, Loy EY, Poulton R, et al. Researching genetic versus nongenetic determinants of disease: a comparison and proposed unification. *Sci Transl Med*. 2009;1(7):7ps8.

13. Manrai AK, Cui Y, Bushel PR, et al. Informatics and data analytics to support exposome-based discovery for public health. *Annu Rev Public Health*. 2017;38:279–294.
14. Price AL, Patterson NJ, Plenge RM, et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet*. 2006;38(8):904–909.
15. Manrai AK, Funke BH, Rehm HL, et al. Genetic misdiagnoses and the potential for health disparities. *N Engl J Med*. 2016; 375(7):655–665.
16. Pearson TA, Manolio TA. How to interpret a genome-wide association study. *JAMA*. 2008;299(11):1335–1344.
17. Duncanson A, Barrett JC, Burton PR, et al. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*. 2007;447(7145):661–678.
18. Yang J, Weedon MN, Purcell S, et al. Genomic inflation factors under polygenic inheritance. *Eur J Hum Genet*. 2011;19(7): 807–812.
19. Bulik-Sullivan BK, Loh PR, Finucane HK, et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet*. 2015;47(3):291–295.
20. Patel CJ, Ji J, Sundquist J, et al. Systematic assessment of pharmaceutical prescriptions in association with cancer risk: a method to conduct a population-wide medication-wide longitudinal study. *Sci Rep*. 2016;6:31308.
21. Patel CJ, Manrai AK, Corona E, et al. Systematic correlation of environmental exposure and physiological and self-reported behaviour factors with leukocyte telomere length. *Int J Epidemiol*. 2017;46(1):44–56.
22. Patel CJ, Ioannidis JP. Studying the elusive environment in large scale. *JAMA*. 2014;311(21):2173–2174.
23. Ioannidis JP. How to make more published research true. *PLoS Med*. 2014;11(10):e1001747.
24. Price AL, Helgason A, Palsson S, et al. The impact of divergence time on the nature of population structure: an example from Iceland. *PLoS Genet*. 2009;5(6):e1000505.
25. Munafò MR, Davey Smith G. Robust research needs many lines of evidence. *Nature*. 2018;553(7689):399–401.
26. Ioannidis JP. Exposure-wide epidemiology: revisiting Bradford Hill. *Stat Med*. 2016;35(11):1749–1762.
27. Patel CJ, Burford B, Ioannidis JP. Assessment of vibration of effects due to model specification can demonstrate the instability of observational associations. *J Clin Epidemiol*. 2015;68(9):1046–1058.
28. Prasad V, Jena AB. Prespecified falsification end points: can they validate true observational associations. *JAMA*. 2013; 309(3):241–242.
29. Hill AB. The environment and disease: association or causation? *Proc R Soc Med*. 1965;58:295–300.
30. Fedak KM, Bernal A, Capshaw ZA, et al. Applying the Bradford Hill criteria in the 21st century: how data integration has changed causal inference in molecular epidemiology. *Emerg Themes Epidemiol*. 2015;12:14.