A Climate-Wide Journey to Explore Mechanisms Underlying Birth Month-Disease Risk Associations: A Call for Collaboration

Mary Regina Boland[1-4], MA, George Hripcsak[1-2], MD, MS,
Patrick Ryan[1-2], PhD, Nicholas P Tatonetti[1-4], PhD
[1]Department of Biomedical Informatics, Columbia University; [2]Observational Health Data Sciences and Informatics, Columbia University; [3]Department of Systems Biology,
[4]Department of Medicine, Columbia University, New York

Abstract

Season of birth can affect many biological processes. Factors that vary with the season (e.g., sunlight exposure, allergens, diet, exercise) can alter developmental mechanisms leading to complications later in life. These exposures may occur during the prenatal or perinatal stages and become observable as disease risk-birth month associations. Previously, a Season-Wide Association Study (SeaWAS) was performed using data from New York City to systematically uncover birth month-disease relationships. To tie these relationships back to the underlying environmental factors that are responsible for the increase in disease risk, we require data from diverse climates and locales. We describe differences in climate metrics among members of the Observational Health Data Sciences and Informatics (OHDSI) consortium and some effects that this could have on data reproducibility for collaboration members. We are calling for additional collaborators to perform this large-scale climate-wide study.

Introduction

The relationship between seasonality, climate, and disease has been described and studied for millennia. Modern researchers study relationships between developmental seasonality (using birth month as a proxy), disease risk, and overall lifespan (1). Widespread adoption of Electronic Health Records (EHRs) has enabled researchers to reuse these data for diverse high-throughput exploratory analyses (2, 3). Previously, Boland et al. developed an algorithm, called SeaWAS for Season-Wide Association Study, to systematically investigate dependencies between birth month and disease risk across all diseases recorded in an EHR having sufficient prevalence (2).

The initial SeaWAS study was conducted using data from just one locale - New York City (NYC) climate. Therefore, additional studies are warranted using data from similar climates (to replicate the original study) and different climates (to probe the contribution of climate and environmental factors on disease risk). By comparing the findings from Boland et al. with SeaWAS results from similar climates, we can seek to understand what findings are due to differences in institutional culture, healthcare process biases (that can differ from region to region) (4), and also demographic differences across regions. Performing this type of 'deep' analysis will also further research reproducibility efforts (5, 6). Reproducibility is a vitally important issue among members of the Observational Health Data Sciences and Informatics (OHDSI) consortium that seek to employ algorithms across many sites. Many of our contributions will be useful to others within the OHDSI community who seek to similarly reproduce their work. We also envision using this forum as a platform to enable further collaboration among members of the community to help contribute birth month-disease risk seasonality results from diverse sites for a complete understanding of the environmental and climatic drivers that contribute to developmental seasonality factors (birth month) and their contribution to disease risk.

Methods and Vision

The scripts necessary to run SeaWAS remotely across sites within the OHDSI community have been made available via GitHub and can be accessed at: https://github.com/maryreginaboland/SeaWAS. Further details for joining this study and links to all relevant sources can be found on the OHDSI wiki site at: http://www.ohdsi.org/web/wiki/doku.php?id=research:seawas. The first script for researchers to run is to collect some aggregate demographics data including ethnicity, race, age and sex distributions. Additionally, it computes the number of total diagnosis codes per patient and also the number of distinct diagnosis codes per patient and then reports the median and intra-quartile ranges. This can enable researchers to compare the diagnosis code densities at different sites to identify different coverage depths. This data is useful for many within the OHDSI research community, because it allows a researcher to compare their data to others.

Another important aspect of our study, and data reproducibility in general, involves how similar your replication site is to your initial study site (i.e., the parent site). Many measures of dataset 'similarity' exist including measures of data quality (7). However, another very important aspect of similarity includes whether the hospital is a rural or urban center – which can influence disease patterns (8) -- and also the overall climate at that particular site.

Many climate measurement criteria exist to compare regions. We chose the Koppen-Geiger climate classification system (9, 10) as a high-level tool for comparing locations to each other because of its widespread use by the World Health Organization. Using the Koppen-Geiger climate classification systems each location is assigned a designation for three aspects: main climate, precipitation, and temperature (**Table 1**). The number of categories for each aspect varies with 5 main climate designations and 8 different aspects of temperature. Each of these 3 factors is combined to produce the overall climate designation. For example, NYC climate is designated Cfa meaning that its climate is warm temperate, fully humid with a hot summer.

Table 1. Koppen-Geiger Three Tiered Climate Classification System

Main Climate	Precipitation	Temperature
A: equatorial	W: desert	h: hot arid
B: arid	S: steppe	k: cold arid
C: warm temperate	f: fully humid	a: hot summer
D: snow	s: summer dry	b: warm summer
E: polar	w: winter dry	c: cool summer
	m: monsoonal	d: extremely continental
		F: polar frost
		T: polar tundra

Additionally NYC climate does not change and is 100% Cfa, this is not the case for all areas within the OHDSI consortium. Seven sites are a mixture of two climates with 60%-40% or 80%-20% ratios. For example, Pittsburgh (University of Pittsburgh data) is a mixture of Cfa (like NYC) for 59.8% and Dfb for 40.2%. This makes it the most comparable with Franklin, Ohio (and Ohio State data) having 57.1% Cfa and 42.9% Dfb. Interestingly, four OHDSI sites are a mixture of three different climates. We also will provide the climate designations and their proportions for each of the OHDSI sites within the United States and six international locations.

Another reproducibility issue is that results obtained from international locations may differ from locations within the United States (even among those with the same climates) due to differences in healthcare practice that can be affected by institutional culture. Understanding and studying these differences is important for ensuring the reproducibility of research and to extract healthcare process effects that may be culture-dependent.

Conclusion

In conclusion, we will describe climatic differences among OHDSI sites, which will be helpful to all community members. This can be used by members to select sites for replication purposes that are either similar in climate or different in climate depending on their particular use case. In addition, we will seek collaborators from within the OHDSI community that are interested in an in-depth study of environmental factors at birth that contribute to lifetime disease risk. We hope that this project can garner widespread support from within the OHDSI community to contribute to the impact of our proposed research.

Acknowledgments

M.R.B. was supported by the National Library of Medicine training grant T15 LM00707, G.H. supported by LM006910, N.P.T. supported by R01 GM107145.

References

1 Doblhammer G, Vaupel JW. Lifespan depends on month of birth. Proceedings of the National Academy of Sciences of the United States of America. 2001 Feb 27;**98**(5):2934-9.

2 Boland MR, Shahn Z, Madigan D, Hripcsak G, Tatonetti NP. Birth Month Affects Lifetime Disease Risk: A Phenome-Wide Method. Journal of the American Medical Informatics Association : JAMIA. 2015 Jun 2.

3 Denny JC, Ritchie MD, Basford MA, et al. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. Bioinformatics. 2010 May 1;**26**(9):1205-10.

4 Hripcsak G, Albers DJ. Correlating electronic health record concepts with healthcare process events. Journal of the American Medical Informatics Association : JAMIA. 2013 Dec;**20**(e2):e311-8.

5 Moonesinghe R, Khoury MJ, Janssens AC. Most published research findings are false-but a little replication goes a long way. PLoS medicine. 2007 Feb;**4**(2):e28.

6 Ioannidis JP. Why most published research findings are false. PLoS medicine. 2005 Aug;**2**(8):e124.

7 Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. Journal of the American Medical Informatics Association : JAMIA. 2013 Jan 1;**20**(1):144-51.

8 Majkowska-Wojciechowska B, Pelka J, Korzon L, et al. Prevalence of allergy, patterns of allergic sensitization and allergy risk factors in rural and urban children. Allergy. 2007 Sep;**62**(9):1044-50.

9 Kottek M, Grieser J, Beck C, Rudolf B, Rubel F. World map of the Köppen-Geiger climate classification updated. Meteorologische Zeitschrift. 2006;**15**(3):259-63.

10 Köppen W. The thermal zones of the Earth according to the duration of hot, moderate and cold periods and of the impact of heat on the organic world. Meteorol Z. 1884;**20**:351-60.