Hypothesis-Free Search for Connections between Birth Month and Disease Prevalence in Large, Geographically Varied Cohorts

John P. Borsi[1]

[1]**Explorys, an IBM Company, Cleveland, OH**

Abstract

We have sought to replicate and extend the Season-wide Association Study (SeaWAS) of Boland, et al[1] *in identifying birth month-disease associations from electronic health records (EHRs). We used methodology similar to that implemented by Boland on three geographically distinct cohorts, for a total of 11.8 million individuals derived from multiple data sources. We were able to identify eleven out of sixteen literature-supported birth month associations as compared to seven of sixteen for SeaWAS. Of the nine novel cardiovascular birth month associations discovered by SeaWAS, we were able to replicate four. None of the novel non-cardiovascular associations discovered by SeaWAS emerged as significant relations in our study. We identified thirty birth month disease associations not previously reported; of those, only six associations were validated in more than one cohort. These results suggest that differences in cohort composition and location can cause consequential variation in results of hypothesis-free searches.*

Introduction

The human urge to assign importance to birth season is well-documented: astrologists have been attempting to divine human fates based on birth timing for millennia[2]. Such endeavors were put on solid scientific footing in 1929, with the publication of a work on the connection between mental disorders such as schizophrenia and birth month[3]. Little doubt remains that birth month does have a measurable impact on many facets of life; over 250 studies had confirmed the importance of birth season before the year 2000[4]. Links have been established between disease prevalence and birth month for numerous conditions: allergies and rhinitis[5,6], reproductive performance[7,8], attention deficit hyperactivity disorder (ADHD)[9], dermatitis[10], Crohn's disease[11], and otitis media[12], among others. Studies have also suggested connections between birth month and height[13], life expectancy[14], and life events[15]. Potential explanations for the observed links have incorporated diverse causes such as neonatal vitamin D exposure[16], exposure to allergens[17], and the impact of social age[15].

In 2015, Boland *et al*[1] implemented a hypothesis-free, phenome-wide method to systematically identify associations between birth month and disease prevalence. Their work is part of a growing acceptance of using electronic health record (EHR) data to conduct retrospective studies. EHRs have been mined to better understand health care utilization of diabetic patients[18], detect adverse drug events[19], and find health care fraud[20]. "Hypothesis-free systems," algorithms which proceed systematically through a dataset without *a priori* hypotheses, have achieved success in identifying clinically relevant associations. For example, EHRs were used in conjunction with genetic data to validate associations between single nucleotide polymorphisms and specific diseases[21].

As the authors of SeaWAS admitted, EHR observational studies have limitations—the existence of bias in health care data is well-known and well-documented[22]. Comparison of large-scale EHR research with gold-standard, manually curated research has demonstrated that the two can produce inconsistent results[23]. Known examples of systemic biases in EHRs include selection bias[24], coding bias[25], and missing or inaccurate records[26,27]. In addition to EHR biases, retrospective studies in general may overstate effects and be subject to confounding factors[28]. However, EHR research has been shown to provide an approximation to traditional research and has had several successful replications of large studies[24,25].

Boland, *et al*'s hypothesis-free methodology was able to confirm known disease-birth month associations, discover new associations, and find clusters of birth month dependencies among disease types[1]. Their work was significant because it applied sophisticated statistical techniques to a large dataset to derive novel insights. However, their work was limited by the cohort to which they applied their method. The SeaWAS study investigated records of 1.7 million individuals at New York-Presbyterian/Columbia University Medical Center. Observed associations from this population may be local effects that do not generalize to a more general population.

We conducted a retrospective study to apply the SeaWAS approach to larger cohorts from different geographical regions. By simultaneously applying this methodology to three separate cohorts, we were able to discern the effect of geographical, administrative, and population-based cohort differences. The increased sample size of our cohorts

allowed increased power in detecting birth month associations. In addition, we proposed a change to the methodology used by SeaWAS to address statistical concerns first raised by Boland, *et al* and tested the impacts of its implementation.

Methods

Data Preparation

The individuals of interest were derived from the Explorys platform[31]. Patients were separated into three cohorts based on ZIP3 codes corresponding to regions in three different states. The first cohort (C1) consisted of patients in a southern US state at approximately 31° N latitude. The second cohort (C2) was constructed of patients from a midwestern US state around 40° N. The third cohort (C3) included patients from a western US state at approximately 38° N. Each of these cohorts represents an aggregation of multiple clinical and claims data sources, grouped by patient. In order to match patients across data sources, demographic records were matched on date of birth, gender, ZIP3, and the New York State Identification and Intelligence System (NYSIIS)[32] representation of the patient's name. All records were de-identified prior to analysis and all records derived from Centers for Medicare & Medicaid Services (CMS) data were excluded from the study.

Demographic information about both the original and replication cohorts is summarized in Table 1.

Table 1. Demographic information about original and replication cohorts.

	SeaWAS (NY)	**Replication (C1)**	**Replication (C2)**	**Replication (C3)**
Total Individuals [Count]	1,749,400	4,588,300	4,840,500	2,331,000
Sex [Count (%)]				
Female	956,465 (54.67)	2,598,592 (56.64)	2,597,027 (53.65)	1,275,957 (54.73)
Male	791,534 (45.25)	1,988,451 (43.34)	2,240,720 (46.29)	1,054,662 (45.25)
Other/unidentified	1,401 (0.08)	1,289 (0.02)	2,757 (0.06)	362 (0.02)
Race [Count (%)]				
White	665,366 (38.03)	2943136 (64.14)	3,087,731 (63.79)	1,563,300 (67.07)
Other/unidentified	842718 (48.18)	951485 (20.74)	983,021 (20.31)	623,719 (26.76)
Black	189,123 (10.81)	553,657 (12.07)	752,837 (15.55)	56,987 (2.44)
Declined	29,747 (1.70)	350,33 (0.76)	48,903 (1.01)	38,336 (1.64)
Asian	20,746 (1.19)	69,909 (1.52)	48,523 (1.02)	38,162 (1.64)
Native American/Indian	1,511 (0.09)	15,704 (0.34)	9,147 (0.19)	10,486 (0.45)
Pacific Islander	189 (0.01)	19,376 (0.42)	971 (0.02)	0 (0.00)
Ethnicity [Count (%)]				
Non-Hispanic	590,386 (33.75)	2,831,416 (61.71)	2,559,409 (52.87)	1,282,259 (55.01)
Unidentified	458,071 (26.18)	1,177,856 (25.67)	2,066,994 (42.70)	693,478 (29.75)
Hispanic	361,123 (20.64)	560,298 (12.21)	83,991 (1.74)	277,226 (11.89)
Declined	339,820 (19.42)	18,730 (0.41)	1,301,06 (2.69)	78,037 (3.35)
Other [Median (IQR†)]				
Age	38 (22-58)	45 (27-62)	49 (29-66)	46 (29-64)
Years of follow-up	1 (1-3)	1 (0-3)	3 (0-7)	1 (0-3)

† Interquartile range

A list of diagnoses was derived from each patient's medical history documents, problem lists, billing records, and other clinical findings. All patient records that contained an *International Classification of Diseases*, version 9 or version 10 (ICD-9, ICD-10) code were aggregated. Using a custom map based on the Common Data Model (CDM) mapping[33] and Intelligent Medical Objects (IMO) data[34], ICD concepts were converted to *Systemized Nomenclature for Medicine-Clinical Terms* (SNOMED) codes. The ICD to SNOMED map used in this study is not one-to-one; that is, an ICD code may map to more than one SNOMED code. In this case, all relevant SNOMED codes were recorded and included in analysis.

Statistical Methodology

For each SNOMED code with more than 1,000 distinct patients, a Pearson's chi-squared test of independence[35] was performed comparing the birth month distribution of patients with the condition to the birth month distribution of all

patients with records in the given system. In each cohort, a different number of SNOMED codes met the sample size restriction. In the original SeaWAS study, 1,688 conditions met the sample size cutoff. In the C1, C2, and C3 replication cohorts, 4,218, 6,379, and 2,973 SNOMED codes were evaluated, respectively.

Boland, *et al* applied the Benjamini-Hochberg[36] (BH) multiplicity correction to the p-values resulting from the chi-squared test. The BH multiplicity correction is a sequential hypothesis rejection procedure designed to control the False Discovery Rate (FDR) of *independent* test statistics. However, as Boland states, "Study limitations include the lack of condition independence […] potentially affecting multiplicity correction" (1051). We applied a more conservative multiplicity correction to our p-values, the sequential Holm multiplicity correction[37]. To evaluate the impact of the change in multiplicity correction, we calculated the results using both multiplicity corrections and compared the output of the different methodologies.

Results

In all replication cohorts, we identified several literature-supported birth month-disease associations. We used the curated reference set of literature-supported associations from Boland, *et al* as a baseline to compare the results from the four cohorts. SeaWAS identified seven out of sixteen associations. Using the same multiplicity correction as the original study, the replication cohorts identified ten, eleven, and three associations at the adjusted $p<.05$ significance level. When considering associations only identified using the more conservative Holm correction, the replication cohorts identified nine, seven, and one of the literature-supported associations. The literature-supported associations identified in each cohort are given in Table 2.

Table 2. Recall of literature-supported birth month-disease association in original and replication cohorts. "X" represents an exact match and "(BH)" represents a match only significant with the Benjamini-Hochberg multiplicity correction.

Literature-Supported Association	**SeaWAS (NY)**	**Replication (C1)**	**Replication (C2)**	**Replication (C3)**
Allergy/Asthma/Rhinitis	X	X	X	(BH)
Reproductive Performance	X	X	(BH)	
Eye Problems	X	X	X	
Schizophrenia		(BH)	(BH)	
Diabetes		X	X	
Respiratory Syncytial Virus	X	X	X	X
Depression		X		
Colitis	X			
Leukemia				
ADHD	X	X	X	(BH)
Atherothrombosis			(BH)	
Atopic Dermatitis			X	
Crohn's Disease				
Lung Fibrosis				
Otitis Media	X	X	X	
Rheumatoid Arthritis		X	(BH)	
Multiple Sclerosis				
Type 1 Diabetes				
Autism				

In addition to identifying correlations from the literature, we evaluated the previously unidentified associations discovered in each cohort. Out of the sixteen new associations from SeaWAS, four of them were replicated in at least one of the replication cohorts. All of these replicated findings were cardiovascular conditions. None of the non-cardiovascular associations discovered by SeaWAS were replicated. Out of the thirty-five unique conditions identified in an Explorys cohort, eight were replicated in another cohort. A summary of these findings, broken down by multiplicity correction and cohort, is presented in Table 3. The list of conditions identified is in Appendix A.

Table 3. Comparison of results for Holm and Benjamini-Hochberg (BH) multiplicity corrections. SeaWAS results are given in columns labeled "NY"; replication results given in "C1," "C2," and "C3" columns.

	Total Number of Associations Identified			
	NY	C1	C2	C3
Holm	-	39	16	5
BH	55	81	46	9
	Number of Literature Supported Associations Identified [# (%Recall)]			
	NY	C1	C2	C3
Holm	-	9 (56)	7 (44)	1 (6)
BH	7 (44)	10 (63)	11 (69)	3 (19)

	Number of Novel Associations Discovered			
	NY	C1	C2	C3
Holm	-	30	9	4
BH	16	71	35	6
	Number of Novel Associations Validated in Other Cohort [# (%Precision)]			
	NY	C1	C2	C3
Holm	-	6 (20)	6 (67)	3 (75)
BH	4 (25)	8 (11)	14 (40)	4 (67)

Discussion

In this replication attempt, we considered fifty data sources in three widely separated regions. The results differed greatly between different geographic regions. The most striking difference between cohorts in different geographical areas is the relatively few associations detected in cohort C3. Cohort C1 produced almost eight times more associations than did the C3. It has been shown that the strength of a birth month-disease association depends on the latitude[38], but birth month effects are typically stronger in higher latitudes. It is unclear why cohort C3, which is at a higher latitude than C1, produced fewer associations. It may be due to different health care processes used in different locations, or it may be related to differences in the characteristics of the individuals in the cohort. The variation in total number of associations observed and the fact that only six out of thirty newly observed associations were statistically significant in more than one cohort suggests that differences between cohorts may be a driver of substantial variation in the results of hypothesis-free searches.

Despite the lack of associations produced in C3, we have confidence that the replication cohorts were well-suited to test the hypothesis-free search. The C1 and C2 cohorts identified more of the literature-supported associations than did SeaWAS. This indicates that the sample size and extent of data was sufficient to detect legitimate birth month associations present in the cohorts.

In all of the cohorts tested, we failed to replicate the non-cardiovascular novel associations discovered by SeaWAS. These conditions include common ailments such as bruising, nonvenomous insect bite, vomiting, and venereal disease screening. Because these conditions are not typically very serious, their inclusion in an EHR may be dependent on the completeness of documentation and may suffer from seasonal bias.

Four new associations were discovered in the replication cohorts that were statistically significant in more than one geographical: coronary arteriosclerosis, tobacco use, hypoxemia, and reversible ischemic neurologic disease (RIND). The first association, coronary arteriosclerosis, may share a mechanism with the other cardiovascular conditions identified in SeaWAS. Although tobacco use has no obvious biological connection to birth month, the connection may be cultural, related to the impact of relative social age (as in Halldner[9] or Skirbekk[15]). The association of hypoxemia and birth month is mostly seen in individuals younger than five, suggesting that the observed effect may be transient. We are not aware of any explanation for the observed association of RIND with birth month.

The results discussed above strongly suggest that the Holm correction is an appropriate multiplicity correction to use for this hypothesis-free search. Although the recall of the search was higher with the Benjamini-Hochberg correction, there were many more non-replicable potential false positives identified using that correction. In addition to the non-replications of SeaWAS associations, the BH correction produced non-replicated associations in the new cohorts such as constipation, diaper rash, tinnitus, and headache. There is naturally a trade-off in choosing any threshold for significance; we suggest that the Holm multiplicity correction more accurately represents the uncertainties involved in this context.

This study had several limitations that could be overcome with future work. All the cohorts considered were in the northern hemisphere and the United States. Results from different regions of the globe may reveal different cultural

or environmental impacts on birth month associations. In addition, there were several confounding factors that we were unable to control in this retrospective study; two of which, age and coding differences between data sources, are known to impact studies[39,25]. Future studies should more closely examine the impact of those confounding factors on the results of hypothesis-free searches.

Conclusion

This study is the largest of its kind with over 11 million unique patients, and is the first to include results from geographically distinct data sources. The size and breadth of this study allowed it to identify subtle trends and associations over large populations. We identified new birth month-disease associations and showed that results from a previous hypothesis-free search may not be generalizable. We have shown that the Benjamini-Hochberg multiplicity correction may not be well suited to a hypothesis-free search with dependent test statistics and shown that the Holm multiplicity correction is a better choice. It is clear that this approach can be a powerful hypothesis generator and tool for investigating the role of seasonally dependent early developmental mechanisms in general health, but obtaining generalizable results requires evaluating different sets of data and accounting for potential biases.

Acknowledgements

The feedback and assistance of the Innovations Team of Explorys, an IBM company was invaluable in preparing this work. I especially wish to thank Matthew Pohlman for his support throughout the research process and Amanda Yoho and Yifan Xu for their comments on the manuscript.

References

1. Boland MR, Shahn Z, Madigan D, Hripcsak G, Tatonetti NP. Birth month affects lifetime disease risk: a phenome-wide method. J Am Med Inform Assoc. 2015;22:1042-1053.
2. Tester, SJ. A history of western astrology. Boydell & Brewer, 1987.
3. Tramer, M. Uber die biologische Bedeutung des Geburtsmonates insbesondere fur die Psychoseerkrankung. / The biological significance of the birth month, with special reference to psychosis. Schweizer Archiv fur Neurologie und Psychiatrie. 1929;24:17-24.
4. Torrey EF, Miller J, Rawlings R, Yolken RH. Seasonality of births in schizophrenia and bipolar disorder: a review of the literature. Schizophrenia Research. 1999;28(1):1-38.
5. Aberg N. Birth season variation in asthma and allergic rhinitis. Clin Exp Allergy. 1989;19(6):643-648.
6. Pearson DJ, Freed DLJ, Taylor G. Respiratory allergy and month of birth. Clin Exp Allergy. 1977;7(1):29-33.
7. Huber S, Fieder M. Strong association between birth month and reproductive performance of Vietnamese women. Am J Hum Biol. 2009;21(1):25-35.
8. Lummaa V, Tremblay M. Month of birth predicted reproductive success and fitness in pre-modern Canadian women. Proc Royal Soc Biol Sci. 2003;270(1531):2355-2361.
9. Halldner L, Tillander A, Lundholm C, *et al*. Relative immaturity and ADHD: findings from nationwide registers, parent- and self-reports. J Child Psychol Psychiatry. 2014;55(8):897-904.
10. Nilsson L, Bjorksten B, Hattevig G, Kjellman B, Sigurs N, Kjellman NM. Season of birth as predictor of atopic manifestations. Arch Dis Child. 1997;76:341-344.
11. Joossens M, Joossens S, Van Steen K, *et al*. Crohn's disease and month of birth. Inflammatory Bowel Diseases. 2005;11(6):597-599.
12. Biles RW, Buffler PA, O'Donell AA. Epidemiology of otitis media. Am J Pub Health. 1980;70(6):593-598.
13. Weber GW, Prossinger H, Seidler H. Height depends on month of birth. Nature. 1998;391:754-755.
14. Doblhammer G, Vaupel JW. Lifespan depends on month of birth. Proc Natl Acad Sci. 2001;98(5):2934-2939.
15. Skirbekk V, Kohler HP, Prskawetz A. Birth month, school graduation, and the timing of births and marriages. Demography. 2004:41(3):547-568.
16. McGrath JJ, Burne TH, Eyles DW. Developmental vitamin d deficiency and risk of schizophrenia: a ten-year update. Schizophrenia Bulletin. 2011;37:56.
17. Businco L, Catani A, Farinella F, Businco E. Month of birth and grass-pollen or mite sensitization in children with respiratory allergy - a significant relationship. Clinical Allergy. 1988;18(3):269-274.
18. Lee N, Laine AF, Hu J, Wang F, Sun J, Ebadollahi S. Mining electronic medical records to explore the linkage between healthcare resource utilization and disease severity in diabetic patients. 2011 IEEE International Conference on Healthcare Informatics.
19. Trifiro G, Pariente A, Coloma PM, *et al*. Data mining on electronic health record databases for signal detection in pharmacovigilance: which events to monitor? Pharmacoepidem Dr S. 2009;18(12):1176-1184.

20. Joudaki H, Rashidian A, Minaei-Bidgoli B, *et al*. Improving fraud and abuse detection in general physician claims: a data mining study. International Journal of Health Policy and Management. 2016;5(3):165-172.
21. 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.
22. Hogan WR, Wagner MM. Accuracy of data in computer-based patient records. J Am Med Inform Assoc. 1997;4:342-355.
23. Hripcsak G, Knirsch C, Zhou L, Wilcox A, Melton GB. Bias associated with mining electronic health records. J Biom Disc Collab 2011:6,48-52.
24. Weiskopf NG, Rusanov A, Weng C. Sick patients have more data: the non-random completeness of electronic health records. AMIA Annual Symposium Proc. 2013;1472-1477.
25. Romano PS, Mark DH. Bias in coding of hospital discharge data and its implications for quality assessment. Medical Care. 1994;32(1):81-90.
26. Wells BJ, Nowacki AS, Chagin K, Kattan MW. Strategies for handling missing data in electronic health record derived data. eGEMs (Generating Evidence \& Methods). 2013;1(3):7.
27. Hogan WR, Wagner MM. Accuracy of data in computer-based patient records. J Am Med Inform Assoc. 1997;5:342-355.
28. Concato J, Shah N, Horwitz RI. Randomized, controlled trials, observational studies, and the hierarchy of research designs. N Engl J Med. 2000;342:1887-1892.
29. Kaelber DC, Foster W, Gilder J, Love TE, Jain AK. Patient characteristics associated with venous thromboembolic events: a cohort study using pooled electronic health record data. J Am Inform Assoc. 2012;19(6):965-972.
30. Pfefferle KJ, Gil KM, Fening SD, Dilisio MF. Validation study of a pooled electronic healthcare database: the effect of obesity on the revision rate of total knee arthroplasty. Eur J Orthop Surg Traumatol. 2014;24:1625-1628.
31. Explorys, an IBM company. Cleveland, Ohio. https://www.explorys.com/ Last accessed Mar 1, 2016.
32. Black PE. Dictionary of Algorithms and Data Structures. www.nist.gov/dads/HTML/nysiis.html. Last accessed Mar 1, 2016.
33. Overhage JM, Ryan PB, Reich CG, Hartzema AG, Stang PE. Validation of a common data model for active safety surveillance research. J Am Med Inform Assoc. 2012;19(1):54-60.
34. Intelligent Medical Objects, Inc. "IMO announces enhanced ICD-9 encoded problem list vocabulary with mapping to SNOMED CT." www.e-imo.com. 2004.
35. Pearson K. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. Philos Mag. 1900;5(50):157-175.
36. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Series B. 1995;57:289-300.
37. Holm S. A simple sequentially rejective multiple test procedure. Scand J Stat 1979;6(2):65-70.
38. Davies G, Welham J, Chant D, Torrey EF, McGrath J. A systematic review and meta-analysis of northern hemisphere season of birth studies in schizophrenia. Schizophrenia Bull. 2003;29(3):587-593.
39. Fiddes B, Wason J, Kemppinen A, Ban M, Compston A, Sawcer S. Confounding underlies the apparent month of birth effect in multiple sclerosis. Ann Neurol. 2013;73(6):714-720.

Appendix A: List of conditions with statistically significant associations with birth month

SeaWAS			Replication			
Description	p-value (BH)	Replicated?	SNOMED ID	Description	p-value (Holm)	Validated?
Atrial fibrillation	<.001	Yes	55822004	Hyperlipidemia	<.001	No
Essential hypertension	<.001	Yes	387712008	Jaundice	<.001	No
Congestive cardiac failure	<.001	Yes	53741008	Coronary arteriosclerosis	<.001	Yes
Angina	<.001	No	399269003	Arthropathy	<.001	No
Cardiac complications of care	0.027	No	90708001	Kidney disease	<.001	No
Cardiomyopathy	0.009	No	92065004	Neoplasm of colon	<.001	No
Pre-infarction syndrome	0.036	No	89765005	Tobacco use	<.001	Yes
Chronic myocardial ischemia	0.022	No	40930008	Hypothyroidism	<.001	No
Mitral valve disorder	0.024	Yes	73430006	Sleep apnea	<.001	No
Acute upper respiratory infection	<.001	No	43339004	Hypokalemia	<.001	No
Bruising	0.015	No	93796005	Malignant neoplasm of female breast	<.001	No
Nonvenomous insect bite	0.001	No	198036002	Impotence	<.001	No
Venereal disease screening	0.003	No	201101007	Keratosis	<.001	No
Primary malignant neoplasm of prostate	0.002	No	22325002	Abnormal gait	<.001	No
Malignant neoplasm of overlapping lesion of bronchus and lung	0.014	No	414916001	Obesity	0.003	No
Vomiting	0.029	No	193462001	Insomia	0.001	No
			2169001	Radiculitis	0.001	No
			363746003	Pharyngitis	0.001	No
			389087006	Hypoxemia	0.002	Yes
			48694002	Anxiety	0.002	No
			62315008	Diarrhoea	0.007	No
			42345000	Polyneuropathy	0.013	No
			2776000	Delirium	0.016	No
			8186001	Cardiomegaly	0.024	No
			11381005	Acne	0.027	No
			57406009	Carpal tunnel syndrome	0.028	No
			36179005	RIND syndrome	0.028	Yes
			23056005	Sciatica	0.032	No
			52767006	Hypoglycemia	<.001	No