

Birth month affects lifetime disease risk: a phenome-wide method

RECEIVED 7 January 2015
 REVISED 23 March 2015
 ACCEPTED 18 April 2015
 PUBLISHED ONLINE FIRST 3 June 2015

Mary Regina Boland^{1,2}, Zachary Shahn³, David Madigan^{2,3}, George Hripcsak^{1,2},
 Nicholas P Tatonetti^{1,2,4,5,*}



ABSTRACT

Objective An individual's birth month has a significant impact on the diseases they develop during their lifetime. Previous studies reveal relationships between birth month and several diseases including atherothrombosis, asthma, attention deficit hyperactivity disorder, and myopia, leaving most diseases completely unexplored. This retrospective population study systematically explores the relationship between seasonal affects at birth and lifetime disease risk for 1688 conditions.

Methods We developed a hypothesis-free method that minimizes publication and disease selection biases by systematically investigating disease-birth month patterns across all conditions. Our dataset includes 1 749 400 individuals with records at New York-Presbyterian/Columbia University Medical Center born between 1900 and 2000 inclusive. We modeled associations between birth month and 1688 diseases using logistic regression. Significance was tested using a chi-squared test with multiplicity correction.

Results We found 55 diseases that were significantly dependent on birth month. Of these 19 were previously reported in the literature ($P < .001$), 20 were for conditions with close relationships to those reported, and 16 were previously unreported. We found distinct incidence patterns across disease categories.

Conclusions Lifetime disease risk is affected by birth month. Seasonally dependent early developmental mechanisms may play a role in increasing lifetime risk of disease.

Keywords: electronic health records, personalized medicine, seasons, cardiovascular diseases, embryonic and fetal development, prenatal nutritional physiological phenomena, pregnancy, maternal exposure.

INTRODUCTION

Hippocrates described a connection between seasonality and disease nearly 2500 years ago, "for knowing the changes of the seasons . . . how each of them takes place, he [the clinician] will be able to know beforehand what sort of a year is going to ensue . . . for with the seasons the digestive organs of men undergo a change."¹ Following in footsteps laid more than 2 millennia ago, recent studies have linked birth month with neurological,^{2–4} reproductive,^{5–9} endocrine¹⁰ and immune/inflammatory disorders,¹¹ and overall lifespan.¹²

Many disease-dependent mechanisms exist relating disease-risk to birth month. For example, evidence linking a subtype of asthma to birth month was presented in 1983.¹³ They found that individuals born in seasons with more abundant home dust mites had a 40% increased risk of developing asthma complicated by dust mite allergies. Their finding was corroborated later when it was found that sensitization to allergens during infancy increases lifetime risk of developing allergies.¹⁴ In addition, some neurological conditions may be associated with birth month because of seasonal variations in vitamin D and thymic output.¹⁵ Understanding disease birth month dependencies is challenging because of the diversity of seasonal affects and connections to disease-risk.

The recent adoption of electronic health records (EHRs) allows meaningful use¹⁶ of data recorded during the clinical encounter for high-throughput exploratory analyses.^{17,18} Using EHR data requires

overcoming problems with definition discrepancies,¹⁹ data sparseness, data quality,²⁰ bias,²¹ healthcare process effects,²² and privacy issues.²³ Informatics methods overcome these challenges, e.g., standardized ontologies minimize definition discrepancies,²⁴ concordance measured across integrated datasets allows for data sparseness and quality assessment,²⁰ and statistical methods can minimize bias and healthcare process effects.^{25–27} Using informatics approaches, EHR discovery methods²⁸ were developed with successful applications in diverse areas including: dentistry,²⁹ genetics,^{30–32} and pharmacovigilance.^{33,34} Novel disease association patterns^{35,36} and seasonal dependencies^{37–39} have also been established using EHRs.

Advances in health informatics coupled with the availability of large clinical databases enable systematic investigation of birth month-disease dependencies. All previous disease-birth month association studies were hypothesis-driven and focused on popular diseases leaving rare diseases unstudied (selection bias). Also, in the literature there is a propensity to publish studies that find an association over those that fail to find a relationship, illustrating publication bias.^{26,27,40,41} In contrast, we developed a high-throughput, hypothesis-free algorithm that mines for disease-birth month associations across millions of records. We call our approach: Season-Wide Association Study (SeaWAS) as it finds all conditions associated with birth month. We show that SeaWAS detects diseases with seasonal components related to early development.

* Correspondence to Nicholas Tatonetti, PhD, Department of Biomedical Informatics, Department of Systems Biology, Department of Medicine, Columbia University, 622 West 168th Street, VC-5, New York, NY 10032, USA; nick.tatonetti@columbia.edu

© The Author 2015. Published by Oxford University Press on behalf of the American Medical Informatics Association.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com For numbered affiliations see end of article.

METHODS

Population

We used the Columbia University Medical Center (CUMC)’s health record data, previously converted to the standardized Common Data Model (CDM) developed by the Observational Medical Outcomes Partnership (now the Observational Health Data Sciences and Informatics).⁴² CUMC data was initially recorded using *International Classification of Diseases, version 9* (ICD-9) codes. These ICD-9 codes were mapped to *Systemized Nomenclature for Medicine-Clinical Terms* (SNOMED-CT) codes according to the CDM v.4.⁴² We selected SNOMED-CT because it captures more clinical content than ICD-9 codes,⁴³ making SNOMED-CT ideal for phenotype classification. Additionally, using this standardized CDM increases the portability of our method across institutions enhancing data sharing.⁴⁴

We extracted all individuals born between 1900 and 2000 inclusive (N = 1749400 individuals) who were treated at CUMC (between 1985 and 2013), demographics given in [Table 1](#). The median age of our population was 38 years (interquartile range, IQR: 22–58). We performed a Fisher-exact test between the birth month distributions for each sex vs the average birth month distribution. Likewise the birth month distributions by birth decade (e.g., 1900–1909, 1990–1999) were compared to the overall average birth month distribution. No statistically significant differences were found ($P = 1$ for all comparisons). Therefore, yearly and sex-based variation in the birth month distribution is minimal and should not affect our analyses ([SI Appendix Figure S1](#) and [S2](#)).

We verified that our monthly birth rate data was consistent with known New York City (NYC) births using data from the Centers of Disease and Control (CDC) for 1990–2000 inclusive.⁴⁵ CUMC data were highly correlated with CDC birth rates from the Bronx ($r = 0.833$, $P = .001$), New York ($r = 0.796$, $P = .002$), and Queens ($r = 0.791$, $P = .002$) counties ([SI Appendix Figure S3](#)). We performed this verification check because confirming the place of birth for individuals can be complex,⁴⁶ and was not possible for our CUMC dataset. Subsequently, for the 1990–2000 period we were able to obtain data regarding the number of babies admitted to CUMC on the day of their birth for the 1990–2000 period and found that the proportion (no. of patients admitted to CUMC on their day of birth/no. of patients included in SeaWAS) ranged from 17.97% to 31.28% by birth year with the average proportion being 22.98%. CUMC’s Institutional Review Board approved this study.

Methods

We investigated associations with birth month across all recorded conditions. A condition is defined as any SNOMED-CT code mapped using the CDM.⁴² For controls, we randomly sampled individuals from the same EHR population without the disease ensuring that our control sample size was ten times the size of the case population. We then modeled the association between birth month (as an integer) and each condition as a logistic regression model with significance assessed using chi-square (R v.3.1.0). Therefore, the monthly birth rate was compared between the case and control populations for each condition adjusting for monthly birth month variation effects. For multiplicity correction, we only selected conditions passing the Benjamini-Hochberg adjustment that controls for the false discovery rate (FDR).⁴⁷ To ensure sufficient sample size across all 12 months, we only investigated conditions having at least 1000 individuals born between 1900 and 2000 inclusive (this amounted to 1688 conditions).

To evaluate SeaWAS, we extracted all articles from PubMed with the term “birth month” and an additional article referenced by a located article (n = 156). We manually reviewed all abstracts and

Table 1: Demographics of Patients Included in SeaWAS Study (N = 1749400)

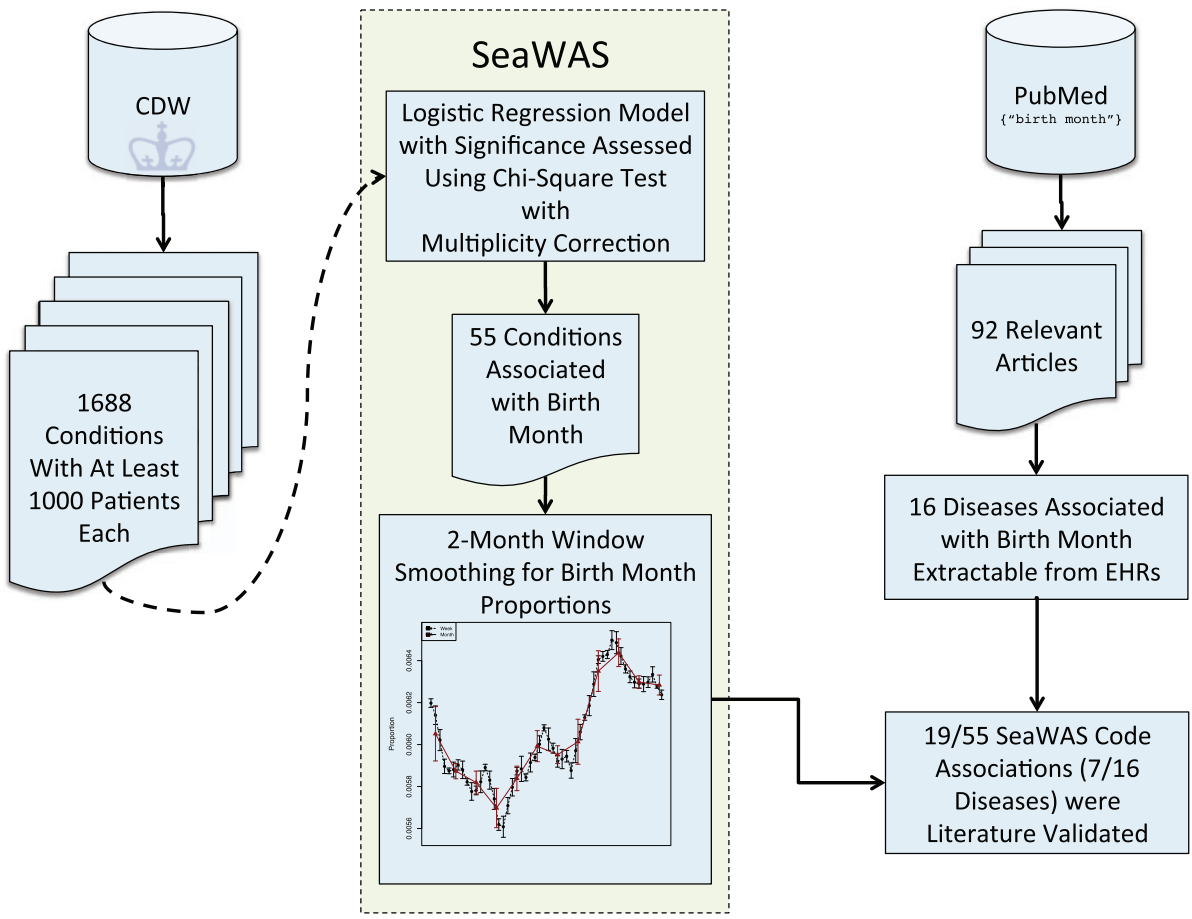
| Demographic | N (%) |
|--------------------------------------|-------------------------------|
| Sex ^a | |
| Female | 956 465 (54.67) |
| Male | 791 534 (45.25) |
| Other/unidentified | 1401 (0.08) |
| Race | |
| White | 665 366 (38.03) |
| Other ^a | 456 185 (26.08) |
| Unidentified | 386 533 (22.10) |
| Black | 189 123 (10.81) |
| Declined | 29 747 (1.70) |
| Asian | 20 746 (1.19) |
| Native American/Indian | 1511 (0.09) |
| Pacific Islander | 189 (0.01) |
| Ethnicity | |
| Non-Hispanic | 590 386 (33.75) |
| Unidentified | 458 071 (26.18) |
| Hispanic | 361 123 (20.64) |
| Declined | 339 820 (19.42) |
| Other attributes | Median (first–third quartile) |
| Total SNOMED-CT codes per patient | 6 (1–32) |
| Distinct SNOMED-CT codes per patient | 3 (1–8) |
| Age (year of service–year of birth) | 38 (22–58) |
| Years of Follow-up | 1 (1–3) |

^aOther (includes Hispanics not otherwise identified)

removed articles related to nonhumans (n = 8), breeding (n = 7), sports (n = 10), or where birth month was used for another purpose, e.g., for matching controls (n = 34), perspective/meta-analysis papers (n = 2), papers not available in English (n = 2), and one paper with a statistical error noted in PubMed. This process identified 92 relevant articles. We then manually classified each paper by the disease studied, and whether they found or failed to find an association. Some conditions associated with birth month in the literature, e.g., height, were not extractable from our EHR (36 diseases were not extractable). In total, 19 diseases reported in the literature could be mapped to EHR conditions. Of those diseases, 16 were positively associated (>50% of literature supported an association) and 3 were not associated (≤50% of literature failed to find an association). We extracted all relevant EHR codes for each of the 16 positive associations (n = 172 codes). These literature associations were used for quality assessment of SeaWAS results.

We used an internal evaluation technique to evaluate novel associations discovered by SeaWAS. We ran the SeaWAS algorithm on a restricted sample comprising 80% of the original sample, randomly chosen. We then corrected for multiplicity using the Benjamini-

Figure 1: Overview of the SeaWAS algorithm. The algorithm takes all 1688 conditions as initial input, finds significant associations over all months, then it models each birth month's association with the condition by smoothing the birth month proportions using a 2-month window. We then extracted all relevant birth month articles from PubMed ($n = 92$) and mapped the results to extractable codes from electronic health records. SeaWAS found 7 of the 16 diseases reported as associated with birth month in the literature corresponding to 19/55 associated codes.



RESEARCH AND APPLICATIONS

Hochberg adjustment that controls the FDR. We took all novel associations (i.e., not reported in the literature) revealed in the restricted sample, and then validated them using the validation set (containing 20% of the original population). Twelve of the 16 discovered associations were validated in this manner.

Permutation analysis was also used for empirical evaluation of SeaWAS. We randomly selected 55 diagnosis codes from the set of 1688 codes included in our study. We then set all codes in this randomly derived set as “positive” associations. Next, the number of positive literature results in each random sample was measured. This was done for 1000 random samples. The overall distribution of these random samples was compared to our SeaWAS results. This allowed us to assess the true positive rate, false positive rate, positive predictive value, and the total number of confirmed literature associations obtained from SeaWAS.

For all significant associations, we calculated the proportion of individuals having the condition using their birth month and day out of all individuals with the same birth month and day. This generated a set of proportions for every day in the year (366 days). We then used a 2-month window¹⁰ to smooth the daily proportion rate (1 month before

the date and 1 month after the date). The weekly and monthly averages were then computed. An overview of the algorithm is shown in Figure 1.

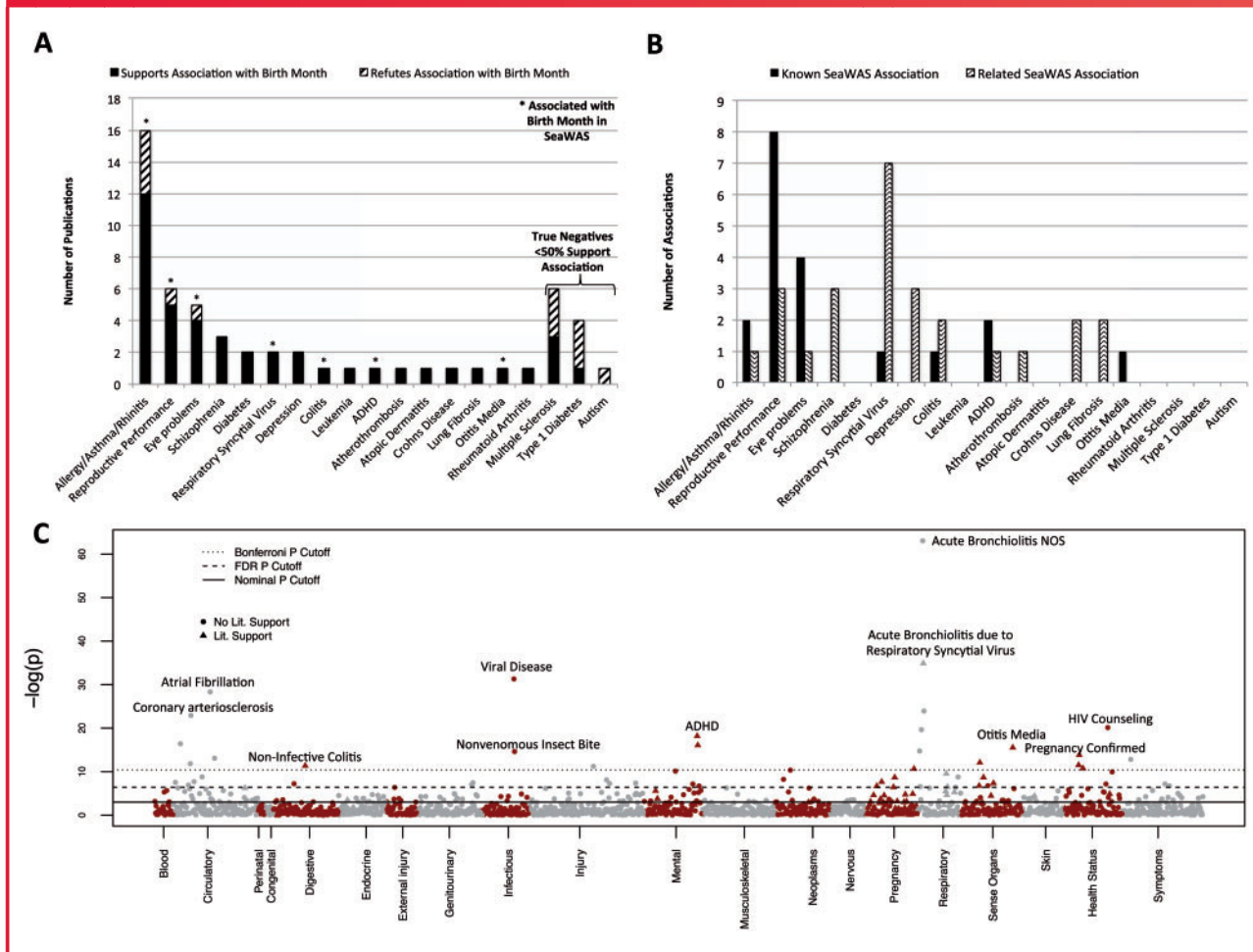
All SeaWAS results were compared to the literature in a binary manner to ascertain if the association was previously reported. Afterwards, we analyzed the disease-birth month risk plots from the literature. We used three criteria to select studies, namely: 1) published raw data; 2) raw data includes some adjustment for natural variation in birth month depending on study region; and 3) disease-birth month data were at a similar granularity level to allow for effective comparisons (e.g., this criterion would exclude studies that grouped multiple diseases together or removed certain disease subtypes). We sought to include pattern data for at least one study per disease category to compare with SeaWAS.

RESULTS

EHR Mining of 1688 Conditions Reveals 55 Conditions Dependent on Birth Month

We used SeaWAS to mine birth month associations for 1688 SNOMED-CT conditions with at least 1000 individuals recorded at

Figure 2: SeaWAS Results Show Enrichments for Literature Associations. (A) shows the breakdown of SeaWAS results by number of publications demonstrating a relationship. (B) shows the number of SeaWAS associations known to be related to disease from the literature (solid black), and those that are closely related to known diseases (curvy lines). (C) Depicts all birth month–disease associations in a Manhattan plot organized by their respective ICD-9 disease categories (x axis). A significant SeaWAS association is a disease–birth month association remaining significant after FDR adjustment.



CUMC. After multiplicity correction using FDR ($\alpha = 0.05$, $n = 1688$ conditions), 55 conditions were found associated with birth month. All reported *P*-values are FDR adjusted (*q*-values).

Literature Validation of SeaWAS Results

Using our curated reference set of 16 conditions (that mapped to 172 SNOMED-CT codes), we found 19 SeaWAS results (7 distinct diseases) were supported by the literature (SI Appendix Table S1), representing a significant enrichment with $OR = 3.4$ (95% CI: 1.9–6.0, $P < .0001$, Figure 2a). SeaWAS successfully ruled-out associations between birth month and disease risk for all “true negatives” in our reference set (Figure 2a). We compared SeaWAS results for known and closely related diseases (Figure 2b) to help elucidate gaps in the literature. We found that some diseases, e.g., reproductive performance, featured prominently in both the literature and SeaWAS results, whereas, other diseases featured heavily in the literature but not as strongly in our results, e.g., asthma/allergy and rhinitis. A potential literature gap exists for respiratory syncytial virus (2 publications Figure 2a), which had many SeaWAS known or highly related associations (8 total associations, Figure 2b). A Manhattan plot visualizes our results

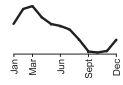
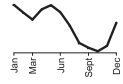
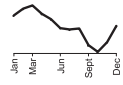
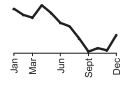
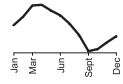
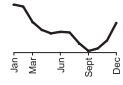
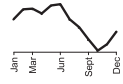
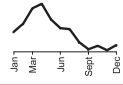
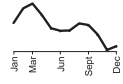
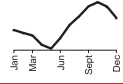
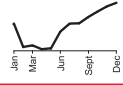
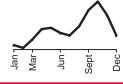
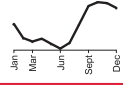
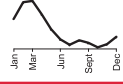
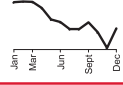
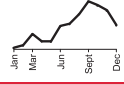
by disease category (Figure 2c) showing that some categories including, circulatory, and respiratory diseases appear prominently in our results.

We found 20 conditions associated with birth month that were similar to those in our reference set (SI Appendix Table S2) and 16 that were completely novel (Table 2). Nine of these 16 associations were cardiovascular conditions including: atrial fibrillation ($P < .001$), essential hypertension ($P < .001$), congestive cardiac failure ($P < .001$), angina ($P = .001$), cardiac complications of care ($P = .027$), mitral valve disorder ($P = .024$), pre-infarction syndrome ($P = .036$), cardiomyopathy ($P = .009$), and chronic myocardial ischemia ($P = .022$). Seven discovered associations were non-cardiovascular: primary malignant neoplasm of prostate, malignant neoplasm of overlapping lesion of bronchus and lung, acute upper respiratory infection, nonvenomous insect bite, venereal disease screening, bruising, and vomiting.

Internal Evaluation of Discovered Associations

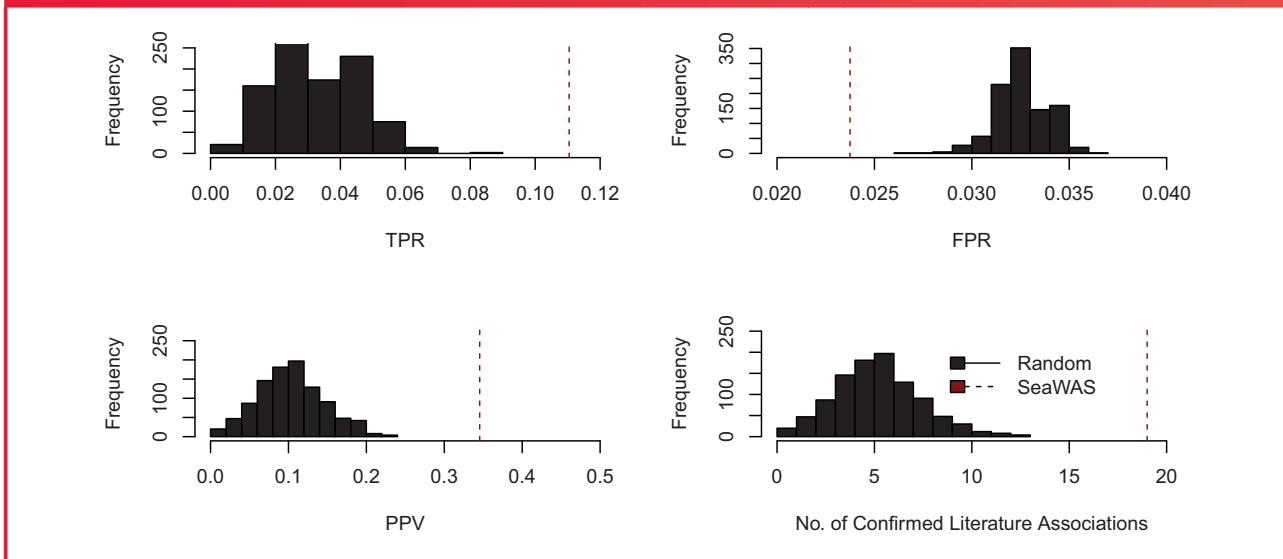
We internally evaluated all novel associations found using SeaWAS. We ran SeaWAS on an 80% restricted sample and then validated the novel

Table 2: Birth Month-Disease Associations Discovered Using SeaWAS (n = 16)

| EHR Condition in SeaWAS | N | Passed Internal Validation? | Adjusted P^1 | Seasonal Pattern | Birth Month Risk | |
|---|---------|-----------------------------|----------------|---|------------------|-----------|
| | | | | | High | Low |
| Cardiovascular (n = 9) | | | | | | |
| Atrial fibrillation | 48 961 | Yes | <0.001 |  | March | October |
| Essential hypertension | 269 913 | Yes | <0.001 |  | January | October |
| Congestive cardiac failure | 61 448 | Yes | <0.001 |  | March | October |
| Angina | 20 741 | Yes | <0.001 |  | April | September |
| Cardiac complications of care | 13 653 | Yes | 0.027 |  | April | September |
| Cardiomyopathy | 17 873 | Yes | 0.009 |  | January | September |
| Pre-infarction syndrome | 25 028 | No | 0.036 |  | June | October |
| Chronic myocardial ischemia | 10 010 | No | 0.022 |  | April | November |
| Mitral valve disorder | 22 966 | No | 0.024 |  | March | November |
| Other (n = 7) | | | | | | |
| Acute upper respiratory infection | 112 487 | Yes | <0.001 |  | October | May |
| Bruising | 8904 | Yes | 0.015 |  | December | April |
| Nonvenomous insect bite | 7435 | Yes | 0.001 |  | October | February |
| Venereal disease screening | 69 764 | Yes | 0.003 |  | October | June |
| Primary malignant neoplasm of prostate | 20 353 | Yes | 0.002 |  | March | October |
| Malignant neoplasm of overlapping lesion of bronchus and lung | 2714 | Yes | 0.014 |  | February | November |
| Vomiting | 30 495 | No | 0.029 |  | September | January |

¹P-values adjusted using Benjamini-Hochberg method (see Methods)

Figure 3: SeaWAS vs random reveals higher true positive rate, lower false positive rate, higher positive predictive value, and more confirmed literature associations. We used 1000 randomly generated samples. For each sample, 55 random codes were pulled (from the set of 1688), and then the number of confirmed literature associations was measured. SeaWAS consistently performed better than random across all measures.



associations in the validation set (20% original sample size). 12 of the 16 novel associations were validated including 6 out of 9 novel cardiovascular conditions. Table 2 denotes the discovered conditions that passed the internal validation. Four conditions were not significant after correction in the restricted sample including: mitral valve disorder, pre-infarction syndrome, chronic myocardial ischemia, and vomiting.

Evaluation Using Permutation Analysis

We used permutation analysis to assess the concordance we found between our SeaWAS results and what was reported in the literature. We randomly selected 55 codes from the set of 1688 codes included in our study and set them as “positives.” We then measured the number of positive literature results in our random samples and compared to SeaWAS. We did this for 1000 random samples. Results are shown in Figure 3. SeaWAS consistently and significantly ($P < .001$) outperformed random for TPR, FPR, and PPV at finding more literature validated associations.

SeaWAS Replicates Established Birth Month Trends: Asthma, Reproductive Performance, and ADHD

We calculated smoothed birth month proportions for all 55 SeaWAS birth month associations. We then compared conditions with known associations to birth month and their published trends. The smoothed weekly and monthly proportions are shown in Figure 4 for 3 established associations: asthma, Attention Deficit Hyperactivity Disorder (ADHD) and reproductive performance and three discovered associations: atrial fibrillation, mitral valve disorder, and chronic myocardial ischemia. Relative risk plots for the associations are given in SI Appendix Figure S4. To compare our results with the published proportions from other studies, we used an asthma study from Denmark,¹³ a reproductive performance study from Austria,⁸ and an ADHD study from Sweden.³

Comparing our results with Denmark’s asthma study¹³ showed highly similar seasonal patterns. They found two large peaks in May

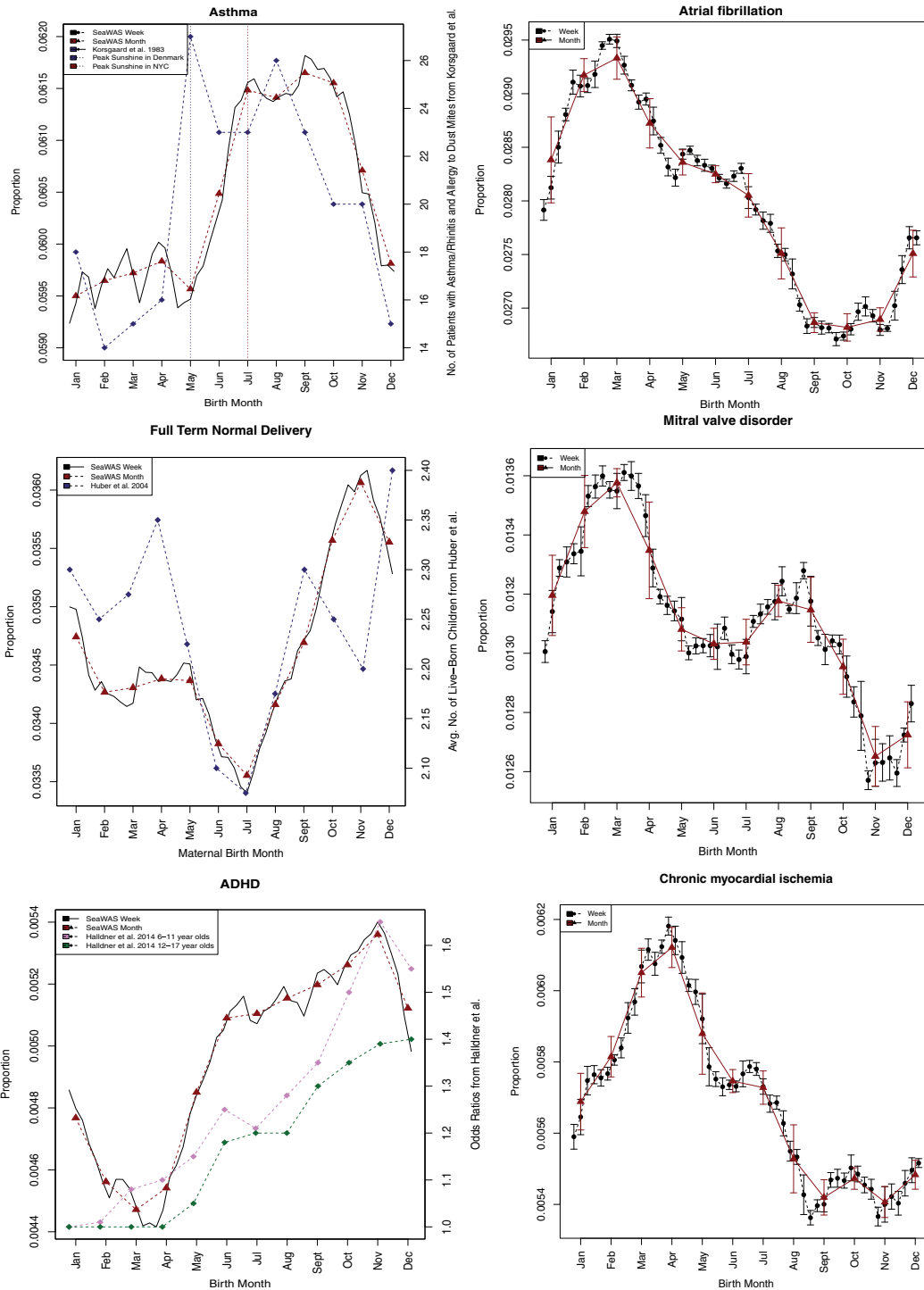
and August, with 2 smaller peaks in June and July.¹³ Our results were shifted by 2 months with large peaks in July and October and smaller peaks in August and September. We extracted data on the average monthly sunshine exposure for NYC and Denmark^{48,49} for comparison (Figure 4). For reproductive performance, we compared our results to an Austrian study⁸ (Figure 4). We validated a dip in births among females born in May through September as this was also found in the Austrian study. We compared our ADHD smoothed proportions to odds ratios reported by a Swedish study and found a similar upward trend towards the later part of the year peaking in November³ (Figure 4).

We sought to include at least one seasonality comparison for each disease category ($n=7$) of known associations to those found by SeaWAS (SI Appendix Table S1). This includes: allergy/asthma/rhinitis, reproductive performance, ADHD, eye conditions/problems, respiratory syncytial virus, otitis media, and colitis. Literature studies on eye conditions/problems failed our 3 criteria for inclusion as data was presented at different disease granularity levels (e.g., mild myopia was excluded) preventing effective comparisons. We found data for conditions in the three remaining categories, otitis media, colitis, and respiratory syncytial virus (SI Appendix Figures S5–S7). We found many similarities among these data, but the exact mechanistic relationship between these conditions and birth seasonality remains obscure.

Discovered Associations: Cardiovascular Conditions and Birth Month

We found 16 associations with no prior literature, we highlight 3 of these in Figure 4, including: atrial fibrillation, mitral valve disorder, and chronic myocardial ischemia. For illustration purposes, we selected cardiovascular conditions whose pattern of association between birth month and disease risk differs. Mitral valve disorder demonstrates a clear bimodal seasonal pattern with a major disease risk peak among those born in March and a second smaller disease risk peak for those born in August. Whereas, risk for atrial fibrillation is unimodal and

Figure 4: Birth month distribution plots for 3 literature validated SeaWAS results and 3 discovered SeaWAS associations. We selected 3 well-known literature associations: asthma, ADHD, and reproductive performance to compare with SeaWAS birth month trends. We compared our results to findings published in articles for each of these diseases: 1) for asthma we used a Denmark study by Korsgaard *et al.*¹³; 2) for reproductive performance we used an Austrian study by Huber *et al.*,⁸ which we compared to full-term normal delivery (i.e., general birth code); and 3) for ADHD we used a Swedish study by Halldner *et al.*³ To facilitate comparison between asthma studies from different locales, we used data on the average monthly sunshine exposure for New York, USA and Skagen, Denmark obtained from World Weather and Climate Information.^{48,49} We also found 3 interesting new associations: atrial fibrillation, mitral valve disorder, and chronic myocardial ischemia.



RESEARCH AND APPLICATIONS

peaks among those born in March with a trough between September and November.

Patterns of Birth-month Dependencies Cluster by Disease Type

Of nine discovered cardiovascular associations, six had high-risk birth months in March or April suggesting that high-risk birth months may cluster by disease category. We examined the disease category–birth month relationship and found that individuals born in March were at increased risk for cardiovascular diseases (Figure 5), but they had greater protection against respiratory illnesses and neurological conditions. Contrastingly, individuals born in October were at increased risk for respiratory conditions with increased protection against developing cardiovascular conditions. Overall, we found that some months, namely May and July, had zero at risk diseases (Figure 5, top). The complete list of protective and at risk diseases by birth month is given in SI Appendix Table S3 with all 55 conditions and their patterns given in SI Appendix Table S4.

Cardiovascular Disease Risk–Birth Month and Lifespan–Birth Month

We compared our cardiovascular disease findings ($n=10$) from SeaWAS to published data relating overall lifespan and birth month,¹² see Figure 6. Months with lower cardiovascular disease risk corresponded with months having longer life expectancies from Doblhammer et al.'s previous study.¹² Six of the 10 cardiovascular conditions were significantly anti-correlated with life-expectancy data. The strongest anti-correlation was cardiac complications of care (Denmark: $r=-0.815$, $P=.001$; Austria: $r=-0.863$, $P<.001$); followed by chronic myocardial ischemia (Denmark: $r=-0.810$, $P=.001$; Austria: $r=-0.826$, $P<.001$); pre-infarction syndrome (Denmark: $r=-0.712$, $P=.009$; Austria: $r=-0.918$, $P<.001$); coronary atherosclerosis (Denmark: $r=-0.617$, $P=.030$; Austria: $r=-0.773$, $P=.003$); atrial fibrillation (Denmark: $r=-0.615$, $P=.033$; Austria: $r=-0.763$, $P=.004$); and angina (Denmark: $r=-0.611$, $P=.035$; Austria: $r=-0.771$, $P=.003$).

DISCUSSION

Many diseases demonstrate birth month dependencies with known mechanistic etiologies, including: asthma,¹³ ADHD,³ reproductive performance,⁸ and myopia.⁵⁰ In these studies birth month was used as a proxy for seasonal variations in physiological state or changes in environmental exposures. Understanding dependencies between diseases and these variations is an important and challenging research task. Large clinical databases, such as EHRs, represent a novel resource for systematically investigating diseases.^{17,18} We present a novel method, SeaWAS, for investigating birth-month dependencies across all diseases in a large EHR. Prior studies analyzed a single disease, or a disease spectrum (e.g., Immune-mediated Diseases) at a time. These hypothesis-driven methods suffer from publication bias, whereby papers demonstrating an association between a disease of interest and birth month are more likely to be published than papers that fail to find an association.^{26,27,40} Prior methods also suffer from disease selection bias whereby diseases of popular interest are studied more frequently potentially overlooking other important disease–birth month associations. SeaWAS overcomes these challenges using a hypothesis-free method that does not relying on a priori hypotheses.

SeaWAS Confirms Known Disease–Birth Month Associations

SeaWAS confirmed a literature-validated association between asthma (hyper-reactive airway disease) and birth month reported by studies from Denmark¹³ and Sweden.⁵¹ When we compared our findings to the Denmark study,¹³ we found a 2-month shift in the birth month–

asthma pattern that corresponds with a shift in the peak sunshine (a factor in asthma complicated by dust mite allergies) between Denmark and NYC^{48,49} (Figure 4).

Likewise, comparing our reproductive performance results to an Austrian study⁸ revealed that the dip in births among females born in May through September was observed in both studies.⁸ Importantly, the female reproductive system, unlike males, is established early with females being born with their lifetime maximum number of oocytes.^{52,53} Oocyte count is thought to be linked to fertility.⁵⁴ Many studies show a link between maternal birth month and number of offspring supporting the belief that prenatal and early developmental effects can alter a female's lifetime fertility.^{5–9} SeaWAS findings bolster this body of literature.

We compared our ADHD smoothed proportions to odds ratios reported by a Swedish study and found a similar upward trend towards the later part of the year peaking in November.³ A rationale for their findings (and ours) is that relative immaturity (born later in the year) may result in increased ADHD detection.³ This occurs because more immature children (i.e., younger in age) face higher demands early on in their school years making them more susceptible to ADHD diagnosis. The age cutoff for schools in Sweden is 31 December, which is the same for NYC public schools. Alternatively, the relationship between Vitamin D and ADHD and learning patterns has been established in rats^{55,56} and Vitamin D deficiency in early development (in utero or shortly after birth) could be related to ADHD.

Discovered Cardiac Condition–Birth Month Relationship

SeaWAS revealed nine cardiovascular conditions associated with birth month. Importantly, children born to survivors of the H1N1 1918 subtype were associated with a >20% excess risk of cardiovascular disease,⁵⁷ suggesting a relationship between maternal infection and cardiovascular disease risk that is independent of maternal malnutrition.⁵⁷ Therefore, maternal infection during the winter months (January–March) could contribute to the increased cardiovascular disease risk among children born in those months.

Looking at all 10 (9 novel) cardiovascular conditions revealed that individuals born in the autumn (September–December) were protected against cardiovascular conditions while those born in the winter (January–March) and spring (April–June) were associated with increased cardiovascular disease risk (Figure 5). Interestingly, one study found that people born in the autumn (October–December) lived longer than those born in the spring (April–June).¹² Furthermore the relationship between cardiovascular disease risk and lifespan is established.⁵⁸ We compared our results to the Doblhammer et al. study investigating lifespan's dependency on birth month and found 6 cardiovascular diseases were significantly anti-correlated. This indicates that birth months with low risk for 6 cardiovascular diseases in our study were also associated with longer lifespan in Doblhammer's study¹² (Figure 6). Our findings suggest that the relationship between lifespan and birth month¹² could be explained by increased cardiovascular disease risk.

The relationship between cardiovascular disease and birth month could be mediated through a developmental Vitamin D-related pathway. Serum 25-hydroxyvitamin D levels are lower and parathyroid hormone levels are higher during the winter when no supplementation is given.⁵⁹ Even with maternal supplementation, seasonally dependent Vitamin D deficiency has been observed among breastfed infants⁶⁰ and newborns.⁶¹ This is important because levels of parathyroid hormone and Vitamin D are associated with cardiovascular disease.^{62,63} Specifically, elevated parathyroid hormone is correlated with increased heart failure in elderly males.⁶⁴ Studies focusing on adolescents found

Figure 5: Disease risk status breakdown by birth month illustrates disease category dependency. Some months, e.g., May, June, August, January, and December, provide no overall advantage or disadvantage to those born in that particular month (Figure 5, top). Other months, e.g., November, are more likely to be associated with *increased disease risk* while others, e.g., February, tend to be associated with *decreased disease risk*. The relationship between birth month and disease risk depends on disease category, and this is shown in the 4 lower subplots. Light gray lines represent risk curves for diseases belonging to a particular category. For example, individuals born in October are at increased risk for respiratory conditions and at the same time are at decreased risk for cardiovascular conditions.

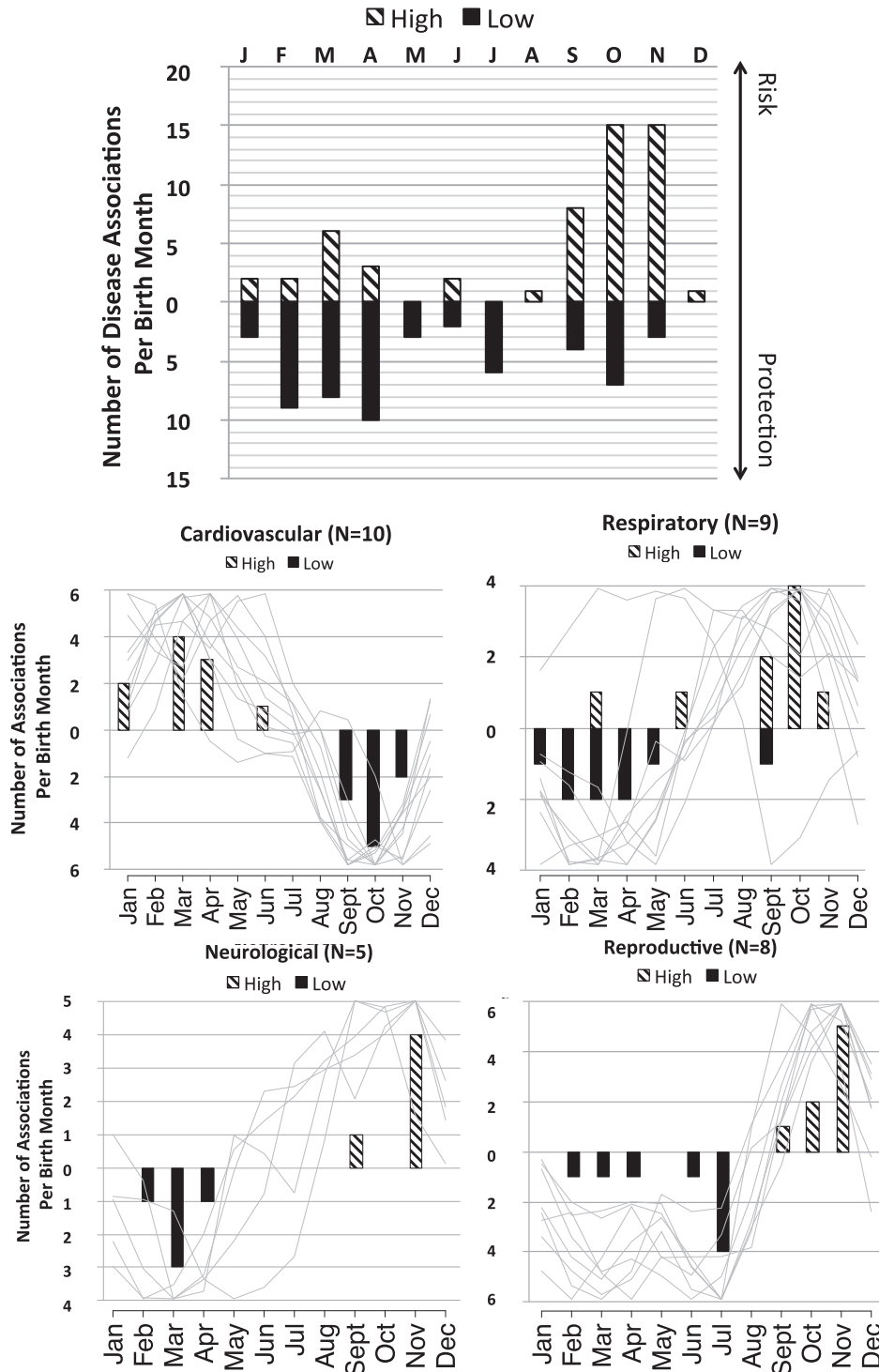
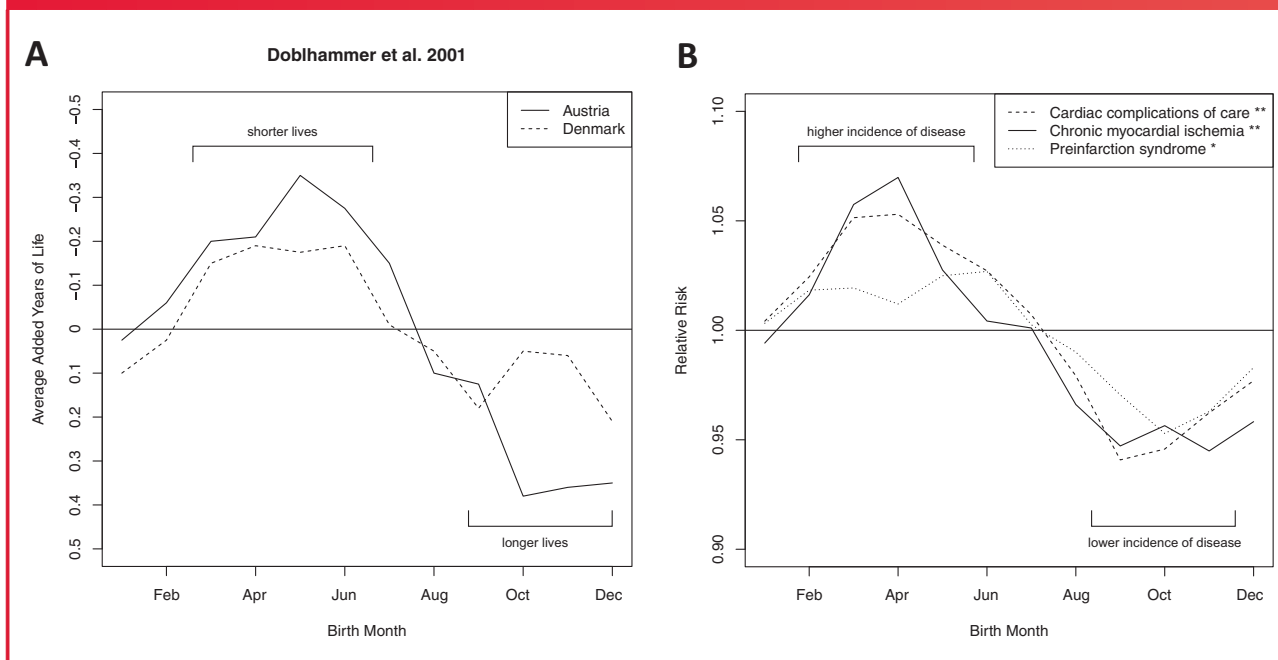


Figure 6: SeaWAS cardiovascular condition–birth month proportions correlate with published lifespan–birth month results from Doblhammer et al. 2001. All 10 (9 novel) cardiovascular disease–birth month associations found by SeaWAS were compared to Doblhammer et al.’s lifespan–birth month dependencies for Denmark and Austria.¹² The lifespan–birth month associations are shown in Figure 6a. Six of the 10 were anti-correlated (i.e., months with low cardiovascular disease risk were also months with longer life expectancies from Doblhammer et al.’s study.¹² The top 3 anti-correlated cardiovascular diseases are shown in Figure 6b, cardiac complications of care (Denmark: $r = -0.815$, $P = .001$; Austria: $r = -0.863$, $P < .001$); chronic myocardial ischemia (Denmark: $r = -0.810$, $P = .001$; Austria: $r = -0.826$, $P < .001$); and preinfarction syndrome (Denmark: $r = -0.712$, $P = .009$; Austria: $r = -0.918$, $P < .001$). In Figure 6b, **denotes $P \leq 0.001$ and *denotes $P < .01$ for both comparisons (Austria and Denmark).



that Vitamin D deficiency resulted in an increased likelihood of hypertension (a SeaWAS discovered association)^{65,66} and high-density lipoprotein cholesterol,⁶⁶ both risk factors for cardiovascular disease.

SeaWAS vs PheWAS: Looking Towards the Future

We present SeaWAS a Phenome-Wide approach that systematically investigates birth month–disease dependencies using EHRs. Our method uses birth month as a proxy for prenatal or perinatal exposure/effects of seasonality on development, and the disease–risk conferred by these perturbations. Denny et al.’s³⁰ Phenome-Wide Association Study (PheWAS) investigates the relationship between diseases recorded in EHRs and genomic markers in a similar high-throughput manner. Recently, an obesity risk factor gene was found to be associated with year of birth⁶⁷ suggesting the importance of combined genetic–environmental etiologies in complex phenotypes. In the near future it may be possible to harness SeaWAS and PheWAS methods for high-throughput identification of diseases tied to prenatal environmental factors (SeaWAS) and then reveal the genetic drivers (PheWAS) underlying the prenatal seasonality effects from EHRs.

Limitations and Future Work

Study limitations include the lack of condition independence (conditions rarely occur in isolation) potentially affecting multiplicity correction. Also, we cannot rule out indirect mechanisms (e.g., depression affects fertility, and learning ability) behind associations between disease risk and birth month. Some conditions associated with birth

month may be associated because the infant was born in a high-risk period, e.g., acute bronchiolitis–autumn births. These associations differ from lifetime disease effects; however, we do not distinguish between them in our analyses because both are presented in the literature as birth month–disease associations. Another limitation is our exclusive use of EHR data, which is affected by the healthcare process^{22,68} and can introduce bias,²¹ e.g., sick patients tend to be over-represented in EHR populations.⁶⁹ Importantly, we showed that our birth month by year data correlated with CDC data (SI Appendix Figure S3) indicating that our EHR population adequately represents the “true” NYC-born population (which includes healthy people) with respect to birth month. Hence, we do not expect this bias to affect our findings.

Additionally, our study uses one institution’s data only; therefore, all birth month–disease risk findings are based on the NYC climate. Because our data is from one locale and climate, the effects we observe are likely due to the climate effects of the NYC region, and is most comparable to Northern European climates. Future work, involves applying our SeaWAS methodology to other institutions and adjusting for climatic differences, which is important when including data from diverse locales and climates.⁷⁰

CONCLUSION

We present a high-throughput algorithm called SeaWAS that uncovers conditions associated with birth month without relying on a priori hypotheses. SeaWAS confirms many known connections between birth

month and disease including: reproductive performance, ADHD, asthma, colitis, eye conditions, otitis media (ear infection), and respiratory syncytial virus. We discovered 16 associations with birth month that have never been explicitly studied previously. Nine of these associations were related to cardiovascular conditions strengthening the link between cardiac conditions, early development, and Vitamin D. Seasonally-dependent early developmental mechanisms might play a role in increasing lifetime disease risk.

CONTRIBUTIONS

M.R.B.: Ms. Boland designed methodology, conducted all analyses, and wrote the manuscript.

Z.S.: Mr Shahn helped with statistical analyses, reviewed, and noted points of revision for the manuscript.

D.M.: Dr Madigan engaged in the design of the statistical methods, reviewed, and noted points of revision for the manuscript.

G.H.: Dr Hripscak helped design aspects of the methodology particularly as they pertained to appropriate use of Electronic Health Records, provided guidance on the interpretation of the analyses, and reviewed the manuscript.

N.P.T.: Dr Tatonetti was involved in all stages of the study design and implementation. He contributed resources, helped refine aspects of the methodology, provided critical insights into validation of methods, and critically reviewed and edited the manuscript.

FINANCIAL DISCLOSURE

The authors have no financial disclosures relevant to this article.

CONFLICT OF INTEREST

The authors have no conflicts of interest.

CLINICAL TRIAL REGISTRATION

Not Applicable.

SUPPLEMENTARY MATERIAL

Supplementary material is available online at <http://jamia.oxfordjournals.org/>.

ACKNOWLEDGEMENTS

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

- Hippocrates, Adams Ft. *On Airs, Waters, and Places*. <http://classics.mit.edu/Hippocrates/airwatpl.mb.txt>. 460 BCE. Accessed August 7, 2014.
- McGrath JJ, Eyles DW, Pedersen CB, et al. Neonatal vitamin d status and risk of schizophrenia: A population-based case-control study. *Arch General Psychiatr*. 2010;67(9):889–894.
- Halldner L, Tillander A, Lundholm C, et al. Relative immaturity and ADHD: findings from nationwide registers, parent- and self-reports. *J Child Psychol Psychiatr*. 2014;55(8):897–904.
- Willer CJ, Dymant DA, Sadovnick AD, Rothwell PM, Murray TJ, Ebers GC. Timing of birth and risk of multiple sclerosis: population based study. *BMJ*. 2005;330(7483):120.
- Huber S, Didham R, Fieder M. Month of birth and offspring count of women: data from the Southern hemisphere. *Hum Reprod*. 2008;23(5):1187–1192.
- Huber S, Fieder M. Strong association between birth month and reproductive performance of Vietnamese women. *Am J Hum Biol*. 2009;21(1):25–35.
- Huber S, Fieder M. Perinatal winter conditions affect later reproductive performance in Romanian women: intra and intergenerational effects. *Am J Hum Biol*. 2011;23(4):546–552.

- Huber S, Fieder M, Wallner B, Moser G, Arnold W. Brief communication: birth month influences reproductive performance in contemporary women. *Hum Reprod*. 2004;19(5):1081–1082.
- Kemkes A. The impact of maternal birth month on reproductive performance: controlling for socio-demographic confounders. *J Biosoc Sci*. 2010;42(2):177–194.
- Kahn HS, Morgan TM, Case LD, et al. Association of type 1 diabetes with month of birth among US youth the SEARCH for Diabetes in Youth Study. *Diabetes Care*. 2009;32(11):2010–2015.
- Disanto G, Chaplin G, Morahan JM, et al. Month of birth, vitamin D and risk of immune mediated disease: a case control study. *BMC Med*. 2012;10(1):69.
- Doblhammer G, Vaupel JW. Lifespan depends on month of birth. *Proc Natl Acad Sci*. 2001;98(5):2934–2939.
- Korsgaard J, Dahl R. Sensitivity to house dust mite and grass pollen in adults. *Influence of the month of birth*. *Clin Allergy*. 1983;13(6):529–535.
- Wahn U, Lau S, Bergmann R, et al. Indoor allergen exposure is a risk factor for sensitization during the first three years of life. *J Allergy Clin Immunol*. 1997;99(6, Part 1):763–769.
- Disanto G, Watson CT, Meier UC, Ebers GC, Giovannoni G, Ramagopalan SV. Month of birth and thymic output. *JAMA Neurol*. 2013;70(4):527–528.
- Jha AK. Meaningful use of electronic health records: the road ahead. *JAMA*. 2010;304(15):1709–1710.
- 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.
- 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.
- Boland MR, Hripscak G, Shen Y, Chung WK, Weng C. Defining a comprehensive verotype using electronic health records for personalized medicine. *JAMIA*. 2013;20(e2):e232–e238.
- Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *JAMIA*. 2013;20(1):144–151.
- Hripscak G, Knirsch C, Zhou L, Wilcox A, Melton G. Bias associated with mining electronic health records. *J Biomed Discov Collab*. 2011;6:48–52.
- Hripscak G, Albers DJ. Correlating electronic health record concepts with healthcare process events. *JAMIA*. 2013;20(e2):e311–e318.
- Loukides G, Gkoulalas-Divanis A, Malin B. Anonymization of electronic medical records for validating genome-wide association studies. *Proc Natl Acad Sci*. 2010;107(17):7898–7903.
- Elkin PL, Brown SH, Husser CS, et al. Evaluation of the Content Coverage of SNOMED CT: ability of SNOMED Clinical Terms to Represent Clinical Problem Lists. *Mayo Clinic Proc*. 2006;81(6):741–748.
- Hripscak G, Knirsch C, Zhou L, Wilcox A, Melton GB. Using discordance to improve classification in narrative clinical databases: an application to community-acquired pneumonia. *Comput Biol Med*. 2007;37(3):296–304.
- Stern JM, Simes RJ. Publication bias: evidence of delayed publication in a cohort study of clinical research projects. 1997;315:640.
- Dickersin K. The existence of publication bias and risk factors for its occurrence. *JAMA*. 1990;263(10):1385–1389.
- 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.
- Boland MR, Hripscak G, Albers DJ, et al. Discovering medical conditions associated with periodontitis using linked electronic health records. *J Clin Periodontol*. 2013;40(5):474–482.
- 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.
- Kohane IS. Using electronic health records to drive discovery in disease genomics. *Nat Rev Genet*. 2011;12(6):417–428.
- Crawford DC, Crosslin DR, Tromp G, et al. eMERGEing progress in genomics—the first seven years. *Front Genet*. 2014;5:184.

33. Wang X, Hripcsak G, Markatou M, Friedman C. Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: a feasibility study. *JAMA*. 2009;16(3):328–337.
34. Haerian K, Varn D, Vaidya S, Ena L, Chase H, Friedman C. Detection of pharmacovigilance-related adverse events using electronic health records and automated methods. *Clin Pharmacol Therap*. 2012;92(2):228–234.
35. Holmes AB, Hawson A, Liu F, Friedman C, Khiabani H, Rabadan R. Discovering disease associations by integrating electronic clinical data and medical literature. *PLoS ONE*. 2011;6(6):e21132.
36. Doshi-Velez F, Ge Y, Kohane I. Comorbidity clusters in autism spectrum disorders: an electronic health record time-series analysis. *Pediatrics*. 2014;133(1):e54–e63.
37. Melamed RD, Khiabani H, Rabadan R. Data-driven discovery of seasonally linked diseases from an Electronic Health Records system. *BMC Bioinformatics*. 2014;15 (Suppl 6):S3.
38. Cohen HA, Blau H, Hoshen M, Batat E, Balicer RD. Seasonality of asthma: a retrospective population study. *Pediatrics*. 2014;133(4):e923–e932.
39. Randolph C. Seasonality of asthma: a retrospective population study. *Pediatrics*. 2014;134 (Suppl 3):S165–S166.
40. Easterbrook PJ, Gopalan R, Berlin JA, Matthews DR. Publication bias in clinical research. *The Lancet*. 1991;337(8746):867–872.
41. Vawdrey DK, Hripcsak G. Publication bias in clinical trials of electronic health records. *J Biomed Inform*. 2013;46(1):139–141.
42. Overhage JM, Ryan PB, Reich CG, Hartzema AG, Stang PE. Validation of a common data model for active safety surveillance research. *JAMA*. 2012;19(1):54–60.
43. Campbell JR, Payne TH. A comparison of four schemes for codification of problem lists. *Proceedings / the Annual Symposium on Computer Application [sic] in Medical Care Symposium on Computer Applications in Medical Care*. 1994:201–205.
44. Margolis R, Derr L, Dunn M, et al. The National Institutes of Health's Big Data to Knowledge (BD2K) initiative: capitalizing on biomedical big data. *JAMA*. 2014;21(6):957–958.
45. CDC. Vital Stats Beyond 20/20. National Vital Statistics System US Department of Health and Human Services. 2014. <http://205.207.175.93/Vitalstats/Common/Login/Login.aspx>. Accessed July 1, 2014.
46. Duncan J, Narus SP, Clyde S, Eilbeck K, Thornton S, Staes C. Birth of identity: understanding changes to birth certificates and their value for identity resolution. *JAMA*. 2015;22:e120–e129.
47. 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.
48. Average sunshine in Skagen, Denmark. World Weather and Climate Information. 2014. <http://www.weather-and-climate.com/average-monthly-hours-Sunshine,skagen,Denmark>. Accessed August 8, 2014.
49. Average sunshine in New York, United States of America. World Weather and Climate Information. 2014. <http://www.weather-and-climate.com/average-monthly-hours-Sunshine,New-York,United-States-of-America>. Accessed August 8, 2014.
50. Mandel Y, Grotto I, El-Yaniv R, et al. Season of birth, natural light, and myopia. *Ophthalmology*. 2008;115(4):686–692.
51. Åberg N. Birth season variation in asthma and allergic rhinitis. *Clin Exp Allergy*. 1989;19(6):643–648.
52. Morita Y, Tilly JL. Oocyte apoptosis: like sand through an hourglass. *Dev Biol*. 1999;213(1):1–17.
53. Baker T. A quantitative and cytological study of germ cells in human ovaries. *Proc R Soc Lond*. 1963;158(972):417–433.
54. Tilly JL, Niikura Y, Rueda BR. The current status of evidence for and against postnatal oogenesis in mammals: a case of ovarian optimism versus pessimism? *Biol Reprod*. 2009;80(1):2–12.
55. Burne THJ, Féron F, Brown J, Eyles DW, McGrath JJ, Mackay-Sim A. Combined prenatal and chronic postnatal vitamin D deficiency in rats impairs prepulse inhibition of acoustic startle. *Physiol Behav*. 2004;81(4):651–655.
56. Becker A, Eyles DW, McGrath JJ, Grecksch G. Transient prenatal vitamin D deficiency is associated with subtle alterations in learning and memory functions in adult rats. *Behav Brain Res*. 2005;161(2):306–312.
57. Mazumder B, Almond D, Park K, Crimmins EM, Finch CE. Lingering prenatal effects of the 1918 influenza pandemic on cardiovascular disease. *J Dev Origins Health Dis*. 2010;1(1):26–34.
58. Stamler J, Stamler R, Neaton JD, et al. Low risk-factor profile and long-term cardiovascular and noncardiovascular mortality and life expectancy: findings for 5 large cohorts of young adult and middle-aged men and women. *JAMA*. 1999;282(21):2012–2018.
59. Dawson-Hughes B, Dallal GE, Krall EA, Harris S, Sokoll LJ, Falconer G. Effect of vitamin D supplementation on wintertime and overall bone loss in healthy postmenopausal women. *Ann Int Med*. 1991;115(7):505–512.
60. Halicioğlu O, Sutcuoğlu S, Koc F, Yildiz O, Akman SA, Aksit S. Vitamin D status of exclusively breastfed 4-month-old infants supplemented during different seasons. *Pediatrics*. 2012;130(4):e921–e927.
61. Lee JM, Smith JR, Philipp BL, Chen TC, Mathieu J, Holick MF. Vitamin D deficiency in a healthy group of mothers and newborn infants. *Clin Pediatr*. 2007;46(1):42–44.
62. Lee JH, O'Keefe JH, Bell D, Hensrud DD, Holick MF. Vitamin D deficiency: an important, common, and easily treatable cardiovascular risk factor? *J Am College Cardiol*. 2008;52(24):1949–1956.
63. Wang TJ, Pencina MJ, Booth SL, et al. Vitamin D deficiency and risk of cardiovascular disease. *Circulation*. 2008;117(4):503–511.
64. Wannamethee SG, Welsh P, Papacosta O, Lennon L, Whincup PH, Sattar N. Elevated parathyroid hormone, but not vitamin D deficiency, is associated with increased risk of heart failure in older men with and without cardiovascular disease. *Circ Heart Fail*. 2014;7(5):732–739.
65. Reis JP, von Muhlen D, Miller ER 3rd, Michos ED, Appel LJ. Vitamin D status and cardiometabolic risk factors in the United States adolescent population. *Pediatrics*. 2009;124(3):e371–e379.
66. Kumar J, Muntner P, Kaskel FJ, Hailpern SM, Melamed ML. Prevalence and associations of 25-hydroxyvitamin D deficiency in US children: NHANES 2001–2004. *Pediatrics*. 2009;124(3):e362–e370.
67. Rosenquist JN, Lehrer SF, O'Malley AJ, Zaslavsky AM, Smoller JW, Christakis NA. Cohort of birth modifies the association between FTO genotype and BMI. *Proc Natl Acad Sci*. 2015;112(2):354–359.
68. Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. 2013;20(1):117–121.
69. Weiskopf NG, Rusanov A, Weng C. Sick patients have more data: the non-random completeness of electronic health records. *AMIA Annual Symposium Proceedings; 2013 American Medical Informatics Association*; 2013: 1472–1477.
70. Flamand C, Fabregue M, Bringay S, et al. Mining local climate data to assess spatiotemporal dengue fever epidemic patterns in French Guiana. *JAMA*. 2014;21(e2):e232–e240.

AUTHOR AFFILIATIONS

¹Department of Biomedical Informatics

²Observational Health Data Sciences and Informatics (OHDSI)

³Department of Statistics

⁴Department of Systems Biology

⁵Department of Medicine, Columbia University, New York, NY, USA