THE UNIVERSITY OF CHICAGO

COMPUTATIONAL DISSECTION OF ETIOLOGY OF COMPLEX DISEASE WITH OBSERVATIONAL DATA

A DISSERTATION SUBMITTED TO

THE FACULTY OF THE DIVISION OF THE BIOLOGICAL SCIENCES

AND THE PRITZKER SCHOOL OF MEDICINE

IN CANDIDACY FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

COMMITTEE ON GENETICS, GENOMICS, AND SYSTEMS BIOLOGY

BY

HANXIN ZHANG

CHICAGO, ILLINOIS

JUNE 2021

TABLE OF CONTENTS

LIST OF FIGURES ..... **vii**

LIST OF TABLES ..... **xi**

LIST OF SUPPLEMENTARY FILES ..... **xii**

ACKNOWLEDGMENTS ..... **xv**

ABSTRACT ..... **xvii**

CHAPTER 1. INTRODUCTION **1**

CHAPTER 2. MEASURING THE HEALTH EFFECTS ASSOCIATED WITH THE DAYLIGHT-SAVING TIME SHIFT **6**

2.1 INTRODUCTION 6

2.2 METHODS 7

2.2.1 Data and other materials 7

2.2.2 Statistical analyses 9

2.3 RESULTS 11

2.4 DISCUSSION 16

2.5 LIMITATIONS 20

2.6 COMPLETE DETAILS OF MATERIALS AND METHODS. 22

2.6.1 Materials and methods 22

2.6.1.1 The MarketScan all-patient database 22

2.6.1.2 The MarketScan inpatient database 24

2.6.1.3 The Swedish inpatient registry 25

2.6.1.4 Mappings 25
2.6.1.5 Models 26
2.6.1.6 Alternative models 30
2.6.1.7 The multiple comparisons problem 33
2.6.1.8 Negative controls 34
2.6.1.9 Geographic location and other covariates 36
2.6.1.10 First-time diagnoses 36
2.6.1.11 Methodological limitations 37
2.6.1.12 Data limitations 38
2.6.2 Analyses summary 39
2.6.2.1 The signal selection procedure for presenting the US results 41
2.6.2.2 Methods comparison 46
2.6.2.3 Geographic location comparison 47
2.6.2.4 Results of the first-diagnoses analyses 48
2.6.2.5 Absolute human cost of the daylight saving time shift 49
CHAPTER 3. BAYESIAN GENERATIVE MODELS FOR PSYCHIATRIC DISEASES' SEASONALITY AND TREND 60
3.1 INTRODUCTION 60
3.2 BAYESIAN MODEL SUMMARY 62
3.3 RESULTS 66
3.3.1 Uncorrected seasonality analysis 69
3.3.2 Corrected seasonality analysis 86
3.4 DISCUSSION 95

3.5 COMPLETE DETAILS OF MATERIALS AND METHODS 100
3.5.1 Data and assumptions 100
3.5.2 Methods and techniques 102
CHAPTER 4. GENE-ENVIRONMENT INTERACTIONS IN PSYCHIATRIC DISORDERS 110
4.1 INTRODUCTION 110
4.2 METHODS 114
4.2.1 Data 114
4.2.2 Modeling 115
4.3 RESULTS 117
4.4 DISCUSSION 132
4.5 COMPLETE DETAILS OF MATERIALS AND METHODS 136
4.5.3 Models 136
4.5.3.1 Linear model 0 (LM0) 136
4.5.3.2 Linear model 1 (LM1) 137
4.5.3.3 Interaction Model 1 (IM1) 139
4.5.3.4 Linear Model 2 (LM2) 140
4.5.3.5 Interaction Model 2 (IM2) 142
4.5.4 Bayesian Inference 142
4.5.4.1 Priors 142
4.5.4.2 Sampling 143
CHAPTER 5. CONCLUSION 144

REFERENCES..... 148

LIST OF FIGURES

Figure 2.1: RR was evaluated as a ratio of the observed diagnosis rate and the expected diagnosis rate..... 9

Figure 2.2: Daylight saving time (DST) shifts appear to affect the relative risk (RR) of numerous diseases, spanning several human biological systems. 13

Figure 2.3: The joint distribution estimation of the spring DST's RR versus the negative control's RR in inpatients, and the analysis results in Sweden..... 14

Figure 2.4: The top 30 conditions exhibiting the largest increasing or decreasing risks (effect sizes) for the results of the US all-patient analyses. 18

Figure 2.5 Enrollee variation in the US MarketScan data. 23

Figure 2.6 Number of enrollees in the MarketScan data set (Y axis) by week of study (Y axis). 24

Figure 2.7 Different methods result in similar RR and interval estimates (US all-patient)..... 51

Figure 2.8 Different methods result in similar RR estimates and intervals (US inpatient). 52

Figure 2.9 Selecting conditions with increased risk by comparing spring DST shift RRs to the negative control on pseudo-DST shift dates (US all-patient). 53

Figure 2.10 Selecting conditions with increased RR by comparing spring DST shift RRs to the negative control on pseudo-DST shift dates (US inpatient). 54

Figure 2.11 Selecting conditions with decreased RR by comparing spring DST shift RRs to the negative control on pseudo-DST shift dates (US all-patient). 55

Figure 2.12 Selecting conditions with decreased RR by comparing spring DST shift RRs to the negative control on pseudo-DST shift dates (US inpatient). 56

Figure 2.13 The Bayesian method shrinks the estimates to the prior mean when there is not enough information (US all-patient). 57

Figure 2.14 The Bayesian method shrinks the estimates to the prior mean when there is not enough information (US inpatient). 57
Figure 2.15 The geographic and cultural diversity of the DST shift's effects on health (US all-patient) 58
Figure 2.16 The geographic and cultural diversity of the DST shift's effects on health (US-inpatient) 59
Figure 3.1 US data characteristics and how they influenced our model design. 65
Figure 3.2 The method and procedure to infer seasonality. 67
Figure 3.3 The baseline seasonality of all medical visits in the four northern US states (AK, WA, MT, ND, or AWMN), the whole US, two southern US states (TX and FL), and Sweden (SE) .. 69
Figure 3.4 The uncorrected seasonality plots of the five most-diagnosed psychiatric diseases in the US: depression, anxiety and phobic disorder, adjustment disorder, substance abuse, and attention-deficit/hyperactivity disorder (ADHD). 71
Figure 3.5 The uncorrected seasonality plots of the five most-diagnosed infectious diseases in the US: acute upper respiratory infection, ear infection, acute bronchitis, urinary tract infection, and cellulitis. 73
Figure 3.6 Embedding of uncorrected seasonality curves in a low-dimensional space suggests the homogeneity of the psychiatric diseases' seasonal variation. 74
Figure 3.7 The uncorrected seasonality of skin infection in the US females. 75
Figure 3.8 The uncorrected seasonality of skin infection in US males. 76
Figure 3.9 The uncorrected seasonality of psychiatric diseases in the four high-latitude states: Alaska, Washington, Montana, and North Dakota (AK, WA, MT, and ND)..... 78

Figure 3.10 The uncorrected seasonality of infectious diseases in the four high-latitude states: Alaska, Washington, Montana, and North Dakota (AK, WA, MT, and ND)..... 79
Figure 3.11 The uncorrected seasonality of psychiatric diseases in the two low-latitude states: Texas and Florida (TX and FL). 80
Figure 3.12 The uncorrected seasonality of infectious diseases in the two low-latitude states: Texas and Florida (TX and FL). 81
Figure 3.13 The uncorrected seasonality of schizophrenia-related psychosis in US females 82
Figure 3.14 The uncorrected seasonality of schizophrenia-related psychosis in US males 82
Figure 3.15 The uncorrected seasonality of schizophrenia-related psychosis in Swedish (SE) females 83
Figure 3.16 The uncorrected seasonality of schizophrenia-related psychosis in Swedish (SE) males 84
Figure 3.17 The uncorrected seasonality of migraine in the US females 84
Figure 3.18 The uncorrected seasonality of migraine in US males 85
Figure 3.19 The uncorrected seasonality of migraine in Swedish females..... 85
Figure 3.20 The uncorrected seasonality of migraine in Swedish males..... 86
Figure 3.21 The corrected seasonality plots of the five most-diagnosed psychiatric diseases in the US and Sweden (SE): depression, anxiety and phobic disorder, adjustment disorder, substance abuse, and attention-deficit/hyperactivity disorder (ADHD)..... 89
Figure 3.22 The corrected seasonality plots of the five most-diagnosed infectious diseases in the US and Sweden (SE): acute upper respiratory infection, ear infection, acute bronchitis, urinary tract infection, and cellulitis..... 90

Figure 3.23 The corrected seasonality of psychiatric diseases in the four high-latitude states: Alaska, Washington, Montana, and North Dakota (AK, WA, MT, and ND)..... 91
Figure 3.24 The corrected seasonality of infectious diseases in the four high-latitude states: Alaska, Washington, Montana, and North Dakota (AK, WA, MT, and ND)..... 92
Figure 3.25 The corrected seasonality of psychiatric diseases in the two low-latitude states: Texas and Florida (TX and FL) 93
Figure 3.26 The corrected seasonality of infectious diseases in the two low-latitude states: Texas and Florida (TX and FL)..... 94
Figure 3.27 The corrected seasonality of all psychiatric disorders in eleven to 20 year-olds across four regions. 95
Figure 3.28 Model selection for choosing the number of harmonics. 109
Figure 4.1 Model WAIC estimates and the mean estimates of heritability and environmental statistics..... 119
Figure 4.2 The mean estimates of the geographic random effects for ADHD and PTSD..... 122
Figure 4.3 The posterior distribution of the log-odds (logit) change contributed by one's sex according to each disease's WAIC-best model 124
Figure 4.4 The posterior distribution of the log-odds (logit) change contributed by one's numeric age according to each disease's WAIC-best model..... 125
Figure 4.5 The nonlinear effects estimated for the five EQI domains (air, water, land, sociodemographic, and built environment) given by each disease's WAIC-best model 127
Figure 4.6 Linear slopes of the EQI effect curves 128

LIST OF TABLES

Table 2.1 Summary of DST analyses 40

Table 4.1 Model set-ups and statistics 117

Table 4.2 Mean estimates of the heritability and environmental statistics 129

LIST OF SUPPLEMENTARY FILES

All Supplementary Files are archived in a folder and attached as a compressed file along with the dissertation.

Supplementary File A.1: A mapping from ICD-10 codes to conditions and diseases

Supplementary File A.2: Week-level RR estimates of the US inpatient analysis, via the Bayesian method

Supplementary File A.3: Week-level RR estimates of the US inpatient analysis on pseudo-DST shift dates as a negative control, via the Bayesian method.

Supplementary File A.4: Week-level RR estimates of the US inpatient analysis, via the frequentist method.

Supplementary File A.5: Week-level RR estimates of the US inpatient analysis on pseudo-DST shift dates as a negative control, via the frequentist method.

Supplementary File A.6: Week-level RR estimates of the Swedish inpatient analysis since 1980, via the Bayesian method.

Supplementary File A.7: Week-level RR estimates of the Swedish inpatient analysis before 1980 as a negative control, via the Bayesian method.

Supplementary File A.8: Week-level RR estimates of the Swedish inpatient analysis since 1980, via the frequentist method.

Supplementary File A.9: Week-level RR estimates of the Swedish inpatient analysis before 1980 as a negative control, via the frequentist method.

Supplementary File A.10: Week-level RR estimates of the US all-patient analysis, via the Bayesian method.

Supplementary File A.11: Week-level RR estimates of the US all-patient analysis on pseudo-DST shift dates as a negative control, via the Bayesian method.

Supplementary File A.12: Week-level RR estimates of the US all-patient analysis, via the frequentist method.

Supplementary File A.13: Week-level RR estimates of the US all-patient analysis on pseudo-DST shift dates as a negative control, via the frequentist method.

Supplementary File A.14: A mapping from ICD-9-CM to ICD-8.

Supplementary File A.15: A mapping from the US modification of ICD-8, 9, 10 to conditions.

Supplementary File A.16: A mapping from the Swedish modification of ICD-8, 9, 10 to conditions.

Supplementary File A.17: A count summary of some female-specific diseases in the US data set.

Supplementary File A.18: A count summary of some male-specific diseases in the US data set.

Supplementary File A.19: A summary of conditions with increased RRs (Bayesian, US all-patient).

Supplementary File A.20: A summary of conditions with increased RRs (frequentist, US all-patient).

Supplementary File A.21: A summary of conditions with increased RRs (Bayesian, US inpatient).

Supplementary File A.22: A summary of conditions with increased RRs (frequentist, US inpatient).

Supplementary File A.23: A summary of conditions with decreased RRs (Bayesian, US all-patient).

Supplementary File A.24: A summary of conditions with decreased RRs (frequentist, US all-patient).

Supplementary File A.25: A summary of conditions with decreased RRs (Bayesian, US inpatient).

Supplementary File A.26: A summary of conditions with decreased RRs (frequentist, US inpatient).

Supplementary File A.27: Estimation of cost associated with the DST shift (Bayesian, US all-patient).

Supplementary File A.28: Estimation of cost associated with the DST shift (frequentist, US all-patient).

Supplementary File A.29: Estimation of cost associated with the DST shift (Bayesian, US inpatient).

Supplementary File A.30: Estimation of cost associated with the DST shift (Frequentist, US inpatient).

ACKNOWLEDGMENTS

I received the most generous support through my graduate studies. With gratitude, I would like to start by thanking my advisor, Professor Andrey Rzhetsky, for devoting his enthusiasm and energy to direct me in my research. Andrey is the warmest and most patient friend, who afforded me all the support and independence possible, allowing me to explore the most intriguing problems in computational biomedicine. Andrey's deep understanding of the area has inspired me to work out the puzzles, and his exceptionally high standard in science is something I always want to pursue.

I also want to express my sincere appreciation to my dissertation committee. The committee chair, Professor Hae Kyung Im, taught the statistical genetics class and introduced me to numerous techniques in statistics and genetics. Professor Risi Kondor taught the machine learning class and provided the most thorough overview of the field. Also on the committee, both Professors Robert Grossman and Bana Jabri offered many suggestions and helped me adjust my research direction.

As my first-year academic advisor, Professor Xin He offered me great tutorials and rotation research in his lab when I had just joined the university. Thank you, Xin, for introducing me to the computational field and advising on my graduate studies!

In addition, it has always been a pleasure to study and work in the Genetics, Genomics, and Systems Biology program and the Institute for Genomics and Systems Biology, where we received lavish care from our warm-hearted administrators Sue Levison and Temi Okubadejo. I am incredibly grateful for Sue and Temi's generous support.

My friends, classmates, and coworkers provided me mental and technical support throughout this journey. I want to thank my colleagues and mentors, Gengjie Jia, Atif Khan,

Yanan Long, Ed Sudzilovsky, Chao Zhang, Chengjian Shi, Xikuan Wang, Rachel Melamed, Ryan Mork, and Bohdan Khomtchouk, for their enduring mentoring and insightful advice. I also want to thank our Research Manager, Erin Gannon for reviewing and commenting on my manuscripts and writing. I also want to thank my classmate Yeonwoo Park, whose scientific brilliance and perception motivated me.

Last but not least, my parents Xueying and Shichun, provide the most pleasant harbor for me, where I always feel peaceful and secure. I want to especially thank my wife Yunqi Li for her patience and attentiveness, for understanding my difficulties, and for exploring science with me as a comrade.

ABSTRACT

Decoding the etiology of complex human diseases is one of the central topics in biomedical research. This dissertation investigates how genetic and environmental factors contribute to disease incidence by leveraging the power of computational models and large-scale observational data. Chapter 1 briefly overviews the classical assumptions, models, and methods underlying the problem. We examine how we might build computational models to infer the genetic and environmental effects on disease etiology based on the observable outcome. Chapter 2 introduces a paradigm that studies the health effects of a short-term exterior intervention: the daylight-saving time shift that changes time forward and backward by one hour in spring and autumn. In Chapter 3, we expand our subject to long-term environmental factors and discuss using a Bayesian model to probe psychiatric disease seasonality and trends. In Chapter 4, we consider an even more complicated problem and explore how we might construct models to jointly infer the effects of various environmental qualities, geographic locations, genetic factors, and gene-environment interactions. The study shows explicitly that varying environmental factors can sway the genetic influence on disease etiology (the heterogeneity of genetic effects). Lastly, Chapter 5 sums up our studies and concludes our computational dissection of complex human disease etiology. We discuss what additional knowledge we have contributed towards understanding the cause and complex disease development. Also, in the last chapter, we present some directions for future exploration.

CHAPTER 1. INTRODUCTION

Human traits are the result of the complex interaction between genetic lineages and environmental influences. As one of the most classic assumptions claims, any phenotype can be modeled by adding up the genetic and environmental effects: Phenotype (P) = Genotype (G) + Environment (E). Many biomedical analyses center on understanding how genetic and environmental factors induce or affect disease incidences, given the significance for preventive measures in clinical settings.

If we take a closer look at the $P = G + E$ equation, we will realize that, in most cases, the environmental effect E is a function of time t and geographic location $\mathbf{x}$ (a two-dimensional coordinate vector). Therefore, while the G term does not vary, assuming consistent genetic composition for an individual, the outcome P can still change dynamically on the temporal and spatial scale: $P(t, \mathbf{x}) = G + E(t, \mathbf{x})$. Moreover, the genetic effect may not be homogenous across time and space because the environment may greatly impact how the genotype contributes to the trait. Such interactions between genetic and environmental factors complicate our analysis, and the equation becomes $P(t, \mathbf{x}) = G + E(t, \mathbf{x}) + G \times E(t, \mathbf{x})$, where the crossed term denotes a simplified interaction effect between genes and the environment.

Despite the theoretical simplicity, it is difficult, if not impossible, to accurately study how genetic variants and changing environmental factors contribute to the phenotype due to the lack of suitable, comprehensive health data. To solve the equation $P(t, \mathbf{x}) = G + E(t, \mathbf{x}) + G \times E(t, \mathbf{x})$, we will need to collect all genetic information and track health records, living environments, and other possible exterior interventions for a large enough cohort from their birth. Hence, researchers commonly resort to large-scale population data and transform this

problem to statistical analysis; they fit regression models that associate genetic and environmental effects to the phenotype instead of solving the equation.

To further simplify the problem, it is also typical to use a cross-section of the complete health records to conjecture disease incidences for individuals in a population. We can impute the exterior environment by associating the individual's living time and location to available environmental records, such as the public data of climate, weather, and environmental quality. Mixed-effects models are also popular statistical tools in solving this problem. Suppose the genetic and environmental effects can be modeled as random effects. It is possible to estimate how genetic and environmental factors influence the phenotype without the actual genetic and environmental information. We only need the relationship matrices describing how individuals in the test group are associated genetically and environmentally [1, 2].

In summary, it is hard to consistently keep an eye on the dynamics of environment and every individuals' traits. Nevertheless, the availability of large-scale data still allows us to statistically infer how genetics, variable environments, and their interactions might shape an observable trait, considering enough genetic and environmental variation within the population.

At the same time, the monumental size of new data and the complexity of new models have also challenged us to upgrade our statistical and computational methods. When using large data and intricate models, it is not convenient to apply simple, deterministic methods like the ordinary least squares found in linear regression. Let us consider the likelihood function $\ell(\boldsymbol{\Theta}|\mathbf{Y})$ for the parameter $\boldsymbol{\Theta}$ and observable data $\mathbf{Y}$. There exist multiple frameworks to estimate $\boldsymbol{\Theta}$. One method (called frequentist) treats the parameter as fixed and latent and relies on maximum likelihood (ML) methods to approximate the actual value. Traditionally, statisticians might attempt to find the maximum likelihood estimate with simplification, approximation, and

optimization techniques. In this example [3], Sulc et al. first derived the likelihood function for a gene-environment interaction model and simplified it with some statistical assumptions. The final form of the likelihood function tends to be easier for optimization. This manual work can be arduous for more complicated models. Advances in computational sciences have provided us with more labor-saving approaches that are scalable for complex models. Machine learning frameworks, such as PyTorch [4, 5] and TensorFlow [6], allow us to declare models as computational graphs [7] and optimize likelihood in the form of a loss function. In this procedure, one no longer needs to consider the actual form of the likelihood function or how to simplify and solve it. The back-end will automatically find the estimates through automatic differentiation [8] and stochastic gradient descent [9].

Contrary to the frequentists' view, Bayesians regard parameters as random variables. They assign prior distributions to the unknown parameter and mold them to posterior distributions given observed data according to the Bayes' theorem: $\text{Posterior} \propto \ell(\boldsymbol{\Theta}|\mathbf{Y}) \times \text{Prior}$. This major difference in definitions means Bayesians count on completely different methods from frequentists to find the parameter estimate. Instead of optimizing the likelihood function directly via gradient descent or higher-order methods, Bayesians use sampling (simulation-based methods) or variational inference (VI, also known as variational Bayes, VB) to delineate the shape of the posterior distribution. Well-liked sampling methods include Markov chain Monte Carlo (MCMC) [10] and its variant: Hamiltonian Monte Carlo (HMC) [11], the No-U-Turn sampler (NUTS) [12], and sequential Monte Carlo (SMC) [13]. For those who would like to avoid the interminable sampling process, variational inference provides an analytical approach to transform the problem to an optimization problem by imposing strong restrictions on the posterior distribution [14, 15]. For example, one can restrict the posterior distributions in a

particular family, such as the Gaussian distribution, and also assume mean-field approximation (the high-dimensional unknown parameter can be factorized into independent, one-dimensional variables) [14, 15].

In terms of recent implementation trends, Bayesian computational frameworks (e.g., PyMC3 [16] and Stan [17]) share many commonalities with the above-mentioned machine learning modules designed for frequentists – despite the fundamental gap between frequentists' and Bayesians' presumption on the unknown parameter. Specifically, modern Bayesian workers commonly use probabilistic programming (PP) [16-18] to specify the (hierarchical) model, and the inference will be done, automatically and later, by the framework. This is similar to what frequentists do with machine learning frameworks like PyTorch [4, 5] and TensorFlow [6]. In these model-centered implementations, models are represented by computational graphs. The graphs (also known as the "forward" paths) describe how the input might generate observable output, just like the hierarchical, generative model does in Bayesian probabilistic programming. Machine learning frameworks' auto-differentiation engine will also conduct the inference by itself, as in probabilistic programming. To sum up, for modern implementations of Bayesian and frequentist methods, the model itself is at the center of the whole procedure. One no longer needs to direct the program in how to compute the likelihood or sample through the MCMC. All of the inferences will be made automatically, which grants researchers more flexibility to study complex models.

In the first part of this introduction, we reviewed one of the most elemental assumptions in biology, which claims that any phenotypic trait can be modeled as the sum of genetic, environmental, and possibly the interactive effects between genetic and environmental factors. We acknowledge that the availability of large health data provides us with many opportunities to

explore how the dynamics of the environment might influence many disease incidences. We also discussed the fundamental division between frequentists' and Bayesians' view on the nature of unknown parameters: The former group treats them as fixed, inferable values, the latter group insists unknown parameters must be modeled as random variables. Frequentists and Bayesians also rely on different statistical approaches to approximate parameters given their distinct mindsets. Nevertheless, modern computational implementations (e.g., PyTorch [4, 5], TensorFlow [6], PyMC3 [16], and Stan [17]) allow us to focus on the model itself and leave the whole inference problem to the program's back-end, regardless of one's preferred method – either it is Bayesian's MCMC or frequentists' maximum likelihood estimation. The programming paradigm (auto-differentiation [8] and probabilistic programming [18]) makes it painless to study more complicated models. To some extent, it sets biomedical researchers' imaginations free when building models.

This dissertation will introduce three sample studies based on a large health data set, MarketScan [19]. The MarketScan data set records the health history of over 150 million unique patients in the United States. For a subset of the entire population, the data also incorporates their family relationship, demographic, and geographic information. As a comparison group, we will also use the Swedish national register [20] for the health history of almost all Swedes. We will show large data's power when investigating both short-term and long-term environmental interventions. Our models' complexity and diversity will also demonstrate how modern computational tools empower us when we investigate convoluted problems.

CHAPTER 2. MEASURING THE HEALTH EFFECTS ASSOCIATED WITH THE DAYLIGHT-SAVING TIME SHIFT

This chapter is adapted from the manuscript "**Measurable health effects associated with the daylight saving time shift**" authored by Hanxin Zhang, Torsten Dahlén, Atif Khan, Gustaf Edgren, and Andrey Rzhetsky.

2.1 INTRODUCTION

The idea of introducing daylight saving time (DST) was attractive at the time of candles and gas lamps, as it allowed workers to use sunlight a bit longer during working hours, as well as saving employers' energy for lighting. Much has changed since then. Today, only a small fraction of electricity expenditure actually corresponds to producing light after sunset (in the US, it is about six percent in the residential sector and eight percent in the industrial sector) [21]. Yet, over a quarter of the world population is subjected to the DST shift twice a year, which disrupts both human work and rest schedules, and possibly their circadian clock rhythms [22]. DST shifts have been shown to have a measurable effect on electric power consumption, although not necessarily in the intended direction [23, 24]. Previous studies have demonstrated that the spring DST shift causes noticeable alterations in human behavior in terms of waking-up time and self-reported alertness, [25] a significant increase in fatal traffic accidents (up to 30 percent on the day of DST commencement), [26] a short-term rise in workplace injuries (5.7 percent after the spring DST shift as employees sleep 40 minutes less on average), [27] and elevated rates of acute myocardial infarction (up by about 3.9 percent) [28]. The study described in [26] and [29] reported conflicting results regarding whether DST shifts are associated with accident incidence. The study described in [30] found increased mental health- and behavioral health-oriented emergency department visits in certain seasons, but did not obtain conclusive results on whether they could be linked to DST shifts.

Remarkable progress has been made in the past decade towards understanding the neurology of sleep-wake cycles and circadian rhythms, and how they affect our behavior [31]. Despite these advances, significant gaps remain in our knowledge of how changes in the social clock (DST shifts) interact with the body's biological clock and impact human health. Recent studies have urged for investigations into the clinical implications of DST shifts on human health [32, 33].

The DST shift represents a natural exposure experiment which allows us the unique opportunity of linking health outcomes to an external, state-wide event in the US and Sweden. Earlier analyses of DST shift effects typically examined a single medical condition per study, often with conflicting or inconclusive results [26, 29, 30]. In addition, these studies often relied on small, disease-specific data sets with thousands of observations from a single country or a single hospital, making it impossible to run phenome-wide screening. In the present study, we used the electronic health records (EHRs) of hundreds of millions of people across two countries, for the purpose of: (1) examining the temporal disease risk dynamics in relationship to DST shifts, and; (2) identifying those population strata which manifest health changes linked to DST-related schedule disruption.

2.2 METHODS

2.2.1 Data and other materials

Our study accessed EHRs from two countries: In the US, through the IBM Watson Health MarketScan data set, [19] and in Sweden, through the Swedish national inpatient register [34]. The version of the MarketScan data set we used in this study incorporates health information about more than 150 million unique patients, observed during the time interval

between 2003 and mid-2014. Individuals followed asynchronous enrollment and disenrollment on insurance policies, leading to variance in their "visibility" durations and endpoints in the data. The mean follow-up time for the patient in the US MarketScan database was 154 weeks. The Swedish register described more than nine million unique Swedes, nearly all observed continuously from 1968 to 2011, except in cases of death or emigration.

In both data sets, disease diagnoses were represented with codes defined by the World Health Organization (WHO) International Classification of Diseases (ICD) taxonomies, versions 9 and 10 for the US, and versions 8, 9, and 10 for Sweden, along with a day-level temporal label recording the date of diagnosis in the US data or the discharge date in the Swedish data. To evaluate the risk of DST shifts across the whole spectrum of diseases, we grouped ICD codes into 263 condition classes under 31 biological systems using the WHO ICD-10 guidelines (Supplementary File A.1) [35]. The grouping is hierarchical and exhaustive, so neighboring codes fall in the same or similar condition classes, and no ICD code is left uncategorized.

2.2.2 Statistical analyses

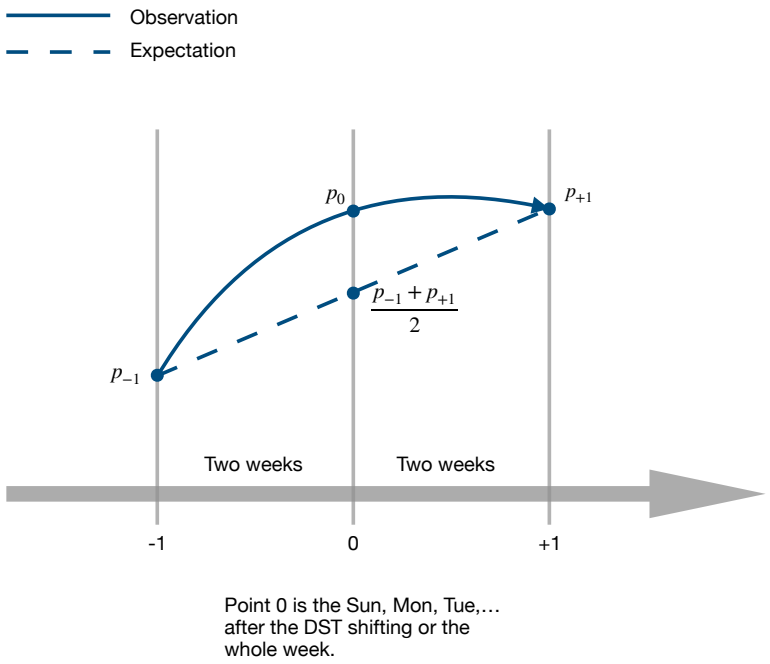

Figure 2.1: RR was evaluated as a ratio of the observed diagnosis rate and the expected diagnosis rate.

Taking advantage of this ICD code grouping, we summarized the daily incidences of each class of conditions for females and males in the following age groups: 0–20 years old (or, alternatively, in the larger US data set, separately 0–10, and 11–20), 21–40, 41–60, and over 60. Building on a previous study's methodology [28], we quantified the relative risk (RR) involved with shifting to and from DST by comparing the diagnosis rate for each day during the week following a DST shift to the linear expectation, which is the average diagnosis rate of the same week day, two weeks before and two weeks after the day of interest (Figure 2.1). The diagnosis rate is the expected proportion of patients diagnosed with a given disease on the day of interest out of all enrollees at a given time point. We estimated week-level RR values by collapsing incidences reported during the whole week following a DST shift and comparing it to the average of corresponding week-level diagnoses rates two weeks before and after the time point.

We corrected the possible effect of holidays (see Section 2.6.1.5 for a full list of holidays considered) and different day lengths (23 hours at the spring DST shift and 25 hours at the autumn). We obtained all the RR estimates in a Bayesian framework and shrunk them towards one. The shrinkage represented our prior belief that we expected the DST shifts to show no health effects for most conditions. The Bayesian procedure pooled information about multiple diseases – with a hierarchical prior distribution imposed over RR estimates. This shrinkage technique resolved the multiple comparisons problem and also yielded statistically efficient estimates [36]. To provide our readers with a frame of reference, we supplemented all of our Bayesian analyses with their frequentist counterparts (see Section 2.6.1.6). The results of our Bayesian and frequentist analyses were very similar after a false coverage rate (FCR) adjustment of frequentist confidence intervals (see Section 2.6.1.7), [37] suggesting a practical equivalence of FCR-corrected frequentist and hierarchical Bayesian (with shrinkage prior) frameworks.

Bayesian and frequentist methods should produce compatible but not identical results. Each method comes with its own advantages and disadvantages. The most obvious difference between them is associated with specification of a prior distribution over parameter values. In our Bayesian analysis, we assumed that most of the estimated RR values would be close to one (shrinkage assumption, implemented as a prior distribution strongly pulling estimates towards the central mean). This assumption forces the results to be conservative (which eliminates weaker signals) and removes the need to correct the results for multiple tests. The frequentist analysis is agnostic in regard to the likely distribution of priors. Some statisticians argue that this type of analysis is less subjective and easier to interpret. The frequentist analyses require explicit corrections for multiple statistical tests and are likely to produce estimates with larger absolute deviations from one.

To control for possible false positive discoveries, we designed a few negative control experiments. Because DST was not adopted in Sweden until 1980, we compared the RRs of time transitions before 1980 (at some "pseudo-DST" shift points, i.e., when a time shift would have happened, see Section 2.6.1.8 for details) and after 1980 at real DST shift points. For the US data, we analyzed all patients residing in states not observing DST as a negative control. Furthermore, we introduced another negative control by repeating the RR estimation procedure at "pseudo-DST" shift dates, which were set to 28 days after each real DST shift in the spring and 28 days before each real DST shift in the autumn. The latter negative control resulted in the most statistically powerful test among the three, because it covered the largest population comparable in size to the groups being tested for association.

Because we ran all our analyses, in parallel, in both Bayesian and frequentist frameworks, we decided to present Bayesian results primarily, highlighting the differences between the two approaches when relevant. The decision to analyze inpatient data separately was driven by the consideration that patients who were hospitalized ("inpatient") may have been subject to fewer of the social and environmental confounders that drive spurious associations. Inpatient admissions are typically associated with a set of health problems distinct from outpatient visits, with more severe conditions, such as acute heart attacks, most commonly treated in hospital inpatient settings.

2.3 RESULTS

In the US inpatient cohort, we detected a significant risk elevation in a number of disease and condition groups (see Section 2.6.2.1 and the summary table); for example, complications related to pregnancy, childbirth, and puerperium (PCP), as well as injuries, symptoms, and signs across various systems, and circulatory diseases (Figure 2.2A). We observed stand-out RR

increases for some injuries, immune disorders, heart diseases, and possibly in related conditions such as renal failure (urinary system-related, shown in Supplementary File A.2 but not Figure 2.2A as it is not among the top 30 for the effect size) and circulatory/cognition symptoms and signs (Figure 2.2A). The RR change signals presented in Figure 2.2 were automatically selected according to their effect size and for their significance as shown in a comparison between the experimental and negative control tests (see Section 2.6.2.1). For the whole spectrum of conditions considered in this study, we present the RR changes with DST shifts in Supplementary File A.2. The results of experiments with the negative controls are shown in Supplementary File A.3. We conducted similar analyses with the frequentist approach and found results consistent with those using the Bayesian framework. In some cases, we even noticed larger estimated effect sizes after adjusting for FCR (Supplementary File A.4 and Supplementary File A.5).

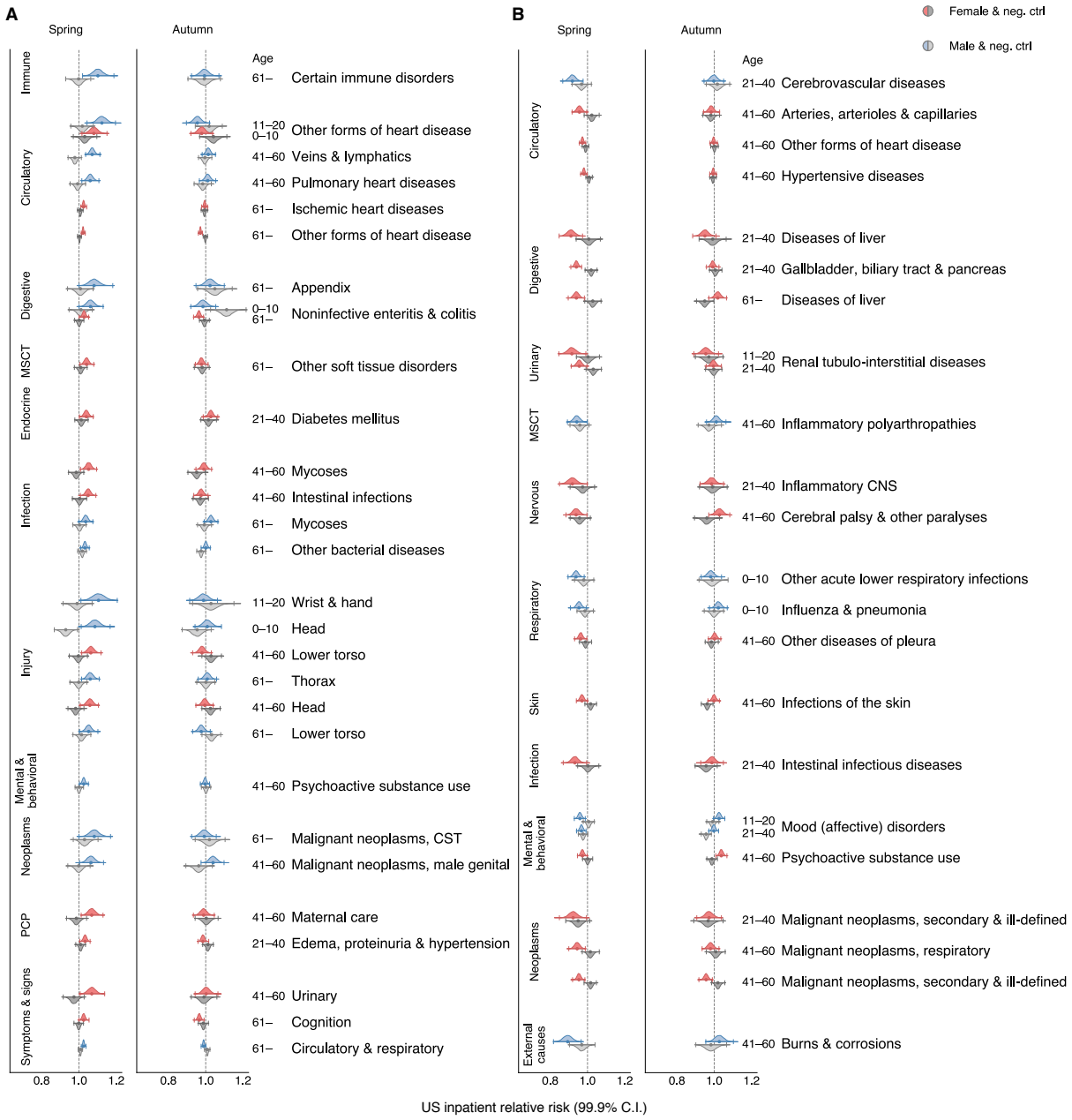

Figure 2.2: Daylight saving time (DST) shifts appear to affect the relative risk (RR) of numerous diseases, spanning several human biological systems.

Figure 2.2 Continued: Color-coded violin plots and error bars represent RR estimates' posterior density distributions and credible interval (CI) boundaries, respectively. We adjusted the credible intervals (CI) for multiple tests with a Bayesian shrinkage procedure that ensured 99.9 and 99 percent significance levels for US and Swedish data, respectively. Gray-colored violin plots and error bars indicate analogous RR distribution results computed for the negative control (pseudo-DST shift dates, the same populations as for the real DST shift dates). The pseudo-DST shift date was selected as 28 days after the real DST shift in spring and 28 days before the real one in autumn. For the Swedish tests, we performed negative controls on data before 1980 when DST was not observed in Sweden. We selected all depicted signals automatically (see Section 2.6.2.1), as the significant RR largest in effect size that were: (1) significantly greater than one in the spring DST shift analyses (colored); (2) not significantly greater than one in the negative control (gray) analyses, or; (3) *vice versa* for decreased signals. We also excluded too broadly defined, ambiguous clinical and laboratory findings, examinations, and health services from the figures (they are retained in the Supplementary Files). **(A)** The top 30 conditions exhibiting the largest increasing RRs (effect sizes) for the results of the US inpatient analyses. The results suggest risk expansion in diseases involving the immune, circulatory, and digestive systems, the musculoskeletal system and connective tissue (MSCT), the endocrine systems, some infections and injuries, mental and behavioral disorders, neoplasms, problems with pregnancy, childbirth, and the puerperium (PCP), and symptoms and signs across various systems. **(B)** All disease conditions with significantly decreased RR in inpatient data after the spring DST shift (minus ambiguous procedures).

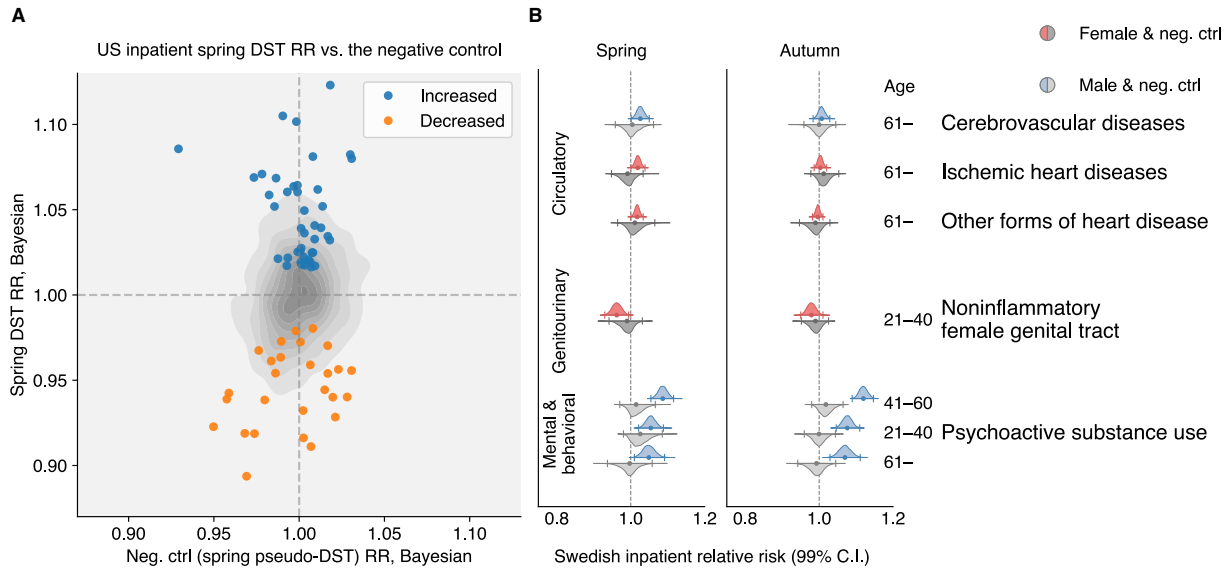

Figure 2.3: The joint distribution estimation of the spring DST's RR versus the negative control's RR in inpatients, and the analysis results in Sweden.

Figure 2.3 Continued: (A) Spring RR estimates versus the negative control results in the US inpatient population. The gray contour represents the empirical estimation of the joint distribution for all RR estimates. The blue and orange markers accent the increased and decreased signals selected by an impartial procedure based on effect size and significance (see Section 2.6.2.1) **(B)** All conditions showing significant change around the DST shift analyses in the Swedish data after 1980. None of their corresponding negative controls were significantly different from one. As in the US data, we observed an increased RR in ischemic and other forms of heart disease in the senior population and mental and behavioral disorders due to psychoactive substance use in middle-age males. The RR for cerebrovascular diseases in senior inpatients increased in Sweden in the week following the spring DST shift, confirming the increase (with no statistical significance) in US inpatients. By contrast, in the US all-patient population, cerebrovascular diseases actually decreased significantly (Supplementary File A.10 and Supplementary File A.12).

To the best of our knowledge, we are the first to report the DST-related RRs of disorders involving the digestive system (such as noninfective enteritis and colitis), which rose three percent after the spring DST shift in females over 60 and six percent in males under ten.

We also observed, for the first time, that the RR for a number of diseases appears to reduce after the spring DST shift (see Figure 2.2B). Such diseases include a set of infectious and inflammatory diseases. In the Bayesian analysis, the effect sizes of the negative RR changes tend to be smaller than those for positive RR changes. The gray contour in Figure 2.3A shows the joint distribution estimation of the spring DST's RR versus the negative control's RR in inpatients, spotlighting those increased and decreased RR change signals (see Section 2.6.2.1 for more details).

In the smaller Swedish data set, as expected, we found fewer significant RR change signals. Under a less conservative significance level (99 percent versus 99.9 percent in the US analyses), we were able to reproduce the RR change signals for a subset of cardiovascular diseases in the elderly population (Figure 2.3B, Supplementary File A.6, and , Supplementary File A.7). The RRs of some heart and cerebrovascular diseases go up in the spring, when the day

length shrinks by one hour, but not in the autumn. As would be expected for a real effect, the RR for circulatory diseases increased after 1980 in the spring DST shift but not in the autumn one. Corresponding frequentist results are shown in Supplementary File A.8 and Supplementary File A.9. Interestingly, the RR of psychoactive substance use increases as much as nine percent with the spring DST shift and 12 percent with the autumn DST shift but only among males age 20 or above population (Figure 2.3B and Supplementary File A.6 and Supplementary File A.7).

2.4 DISCUSSION

Our analyses reproduced the major past literature's findings, such as an elevation in ischemic heart disease rates in males and females older than 60, [28, 38-40] and a rise in accidents [26, 41] and injuries [27]. We also discovered novel significant RR change signals, such as a DST-shift-associated increase in mental and behavioral disorders due to the aforementioned elevated psychoactive substance use in the male adult population. The strongest effect size was observed among males between the ages 41-60 and the signal was consistent in both the US and Sweden. A large body of studies have shown that circadian disruption increases the risk of substance abuse [42-45], with some studies providing in-depth mechanistic details [46]. Because psychoactive substance users generally have very disrupted diurnal rhythms [47], it seems plausible that further acute disruption due to DST shifts may lead to abnormal clock function, resulting in increased vulnerability for substance abuse.

The findings in the US "all-patients" data set (Figure 2.4, Supplementary Files A.10-A.13) resolves the inconclusive results of a previous study focusing only on emergency admissions [30]. To the best of our knowledge, never reported before, immune-related disorders tend to become more common than expected in the first week following each spring DST shift. We observed the largest effect sizes for the following conditions: an approximate ten percent

increase in the RR for some cardiovascular and heart diseases in inpatients under 20, injuries at various locations and ages (in the frequentist framework, RR estimates increased by 30 percent), and some immune disorders in senior males. The absolute risk posed by the DST shift is discussed in Section 2.6.2.5. The comparison of Bayesian and frequentist analyses indicates that there is not much data in support of the estimates computed for a subset of diseases, and that the prior has a strong influence on the estimate.

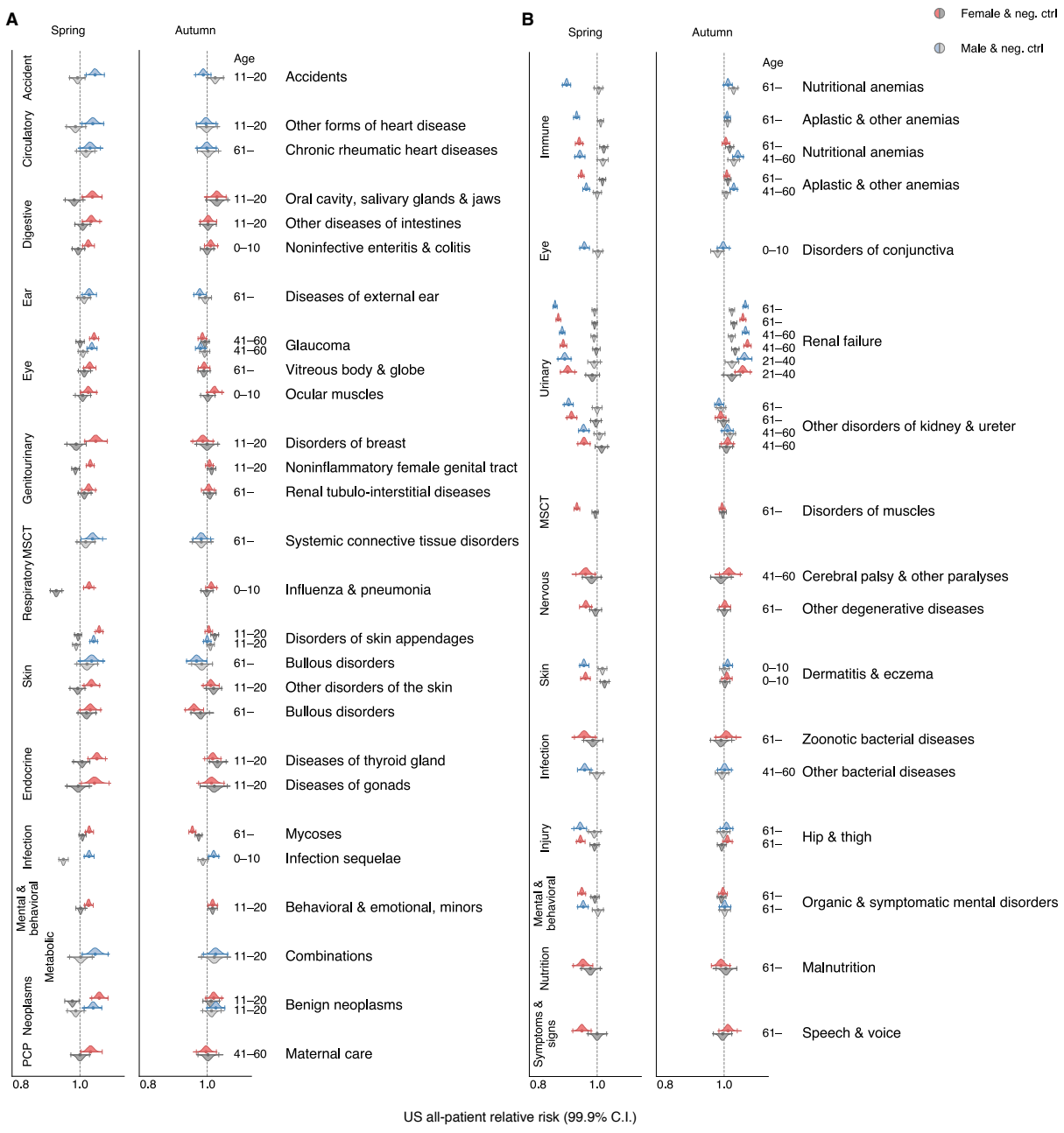

Figure 2.4: The top 30 conditions exhibiting the largest increasing or decreasing risks (effect sizes) for the results of the US all-patient analyses.

(A) Increasing signals shown in the US all-patient analysis. **(B)** Decreasing signals shown in the US all-patient analysis.

Our analysis identified several population strata that appeared responsive to the time and schedule disruption in terms of changes in disease RR: (1) very young (0–10, or 11–20) and older (41–60 and over 60) patients with (probably preexisting) chronic diseases, for example, acute myocardial infarction, behavioral and emotional disorders, and stress-related immune disorders such as inflammatory bowel diseases and noninfective enteritis; (2) children (0–10) and teenagers/young adults (11–20) who were prone to accidents resulting in injuries of the head, wrist, and hand; and (3) older adults (41–60 and older than 60) who were more likely to injure the lower torso or thorax. Note that while MarketScan represents the United States fairly evenly, geographically, it excludes the uninsured population.

In a DST shift effect analysis, one should keep in mind that there are two distinct phenomena in play: (1) the natural variation of day length, the amount of light over the year, and the daily sunlight intensity cycle affects the human circadian clock, behavior, and numerous diseases [48-64], and; (2) the disruption of individuals' daily schedule due to the DST shift. Our analyses were designed to target the effects associated with the latter, adjusting for the former with negative controls. Furthermore, it is not always possible to distinguish between increased diseased incidence (for a truly new diagnosis) and increased care (for an established ongoing diagnosis). For a subset of patients under observation in the US data set since their birth, we were able to carefully identify the first disease diagnosis and validate the RR change signal for noninfective enteritis and colitis after the spring DST shift (see Sections 2.6.1.10 and 2.6.2.4). We performed an analogous “first-diagnosis” analysis for older US patients in the all-patients data set, which resulted in only one statistically significant result: an increased rate of “other forms of heart disease” in females over 60 (see Sections 2.6.1.10 and 2.6.2.4).

We have considered using a more complicated model, such as a mixed-effect regression model, which would allow us to include some chronic diseases as confounding factors to obtain better estimates of disease relative risks. We did not attempt to control for any particular disease for two reasons: (1) introducing control diseases would cause a combinatorial explosion in the model space, and; (2) controlling for specific diseases would exponentially increase the number of necessary statistical tests. Both issues would increase analysis complexity, make interpretation more difficult, and decrease our ability to detect RR change signals. Therefore, we decided against using this approach.

What mechanism could plausibly explain the reduced RR observed for some of the infectious and immune diseases during the spring DST shift? The spring DST shift might act as a short-term stressor, as one hour is subtracted from a normal night's schedule. A transient, mild stress was shown to enhance the immune system (as opposed to long-term stress, which has the opposite, suppressive action), possibly accounting for the reduced RR for urinary, skin, and other infections. As F.S. Dhabhar put it, a "short-term stress can enhance the acquisition and/or expression of immune-protective (wound healing, vaccination, anti-infectious agent, anti-tumor) or immuno-pathological (pro-inflammatory, autoimmune) responses." [63, 64] Of course, this is but one possible explanation and possible alternative explanations could be suggested.

2.5 LIMITATIONS

This study suggests that even a one-hour change of the clock may impact population health significantly, but a number of caveats accompany this assertion.

First, it is important to keep in mind that diseases are not truly independent of each other; one illness can facilitate the development of another, and environmental insults may also lead to the exacerbation of a (pre-existing) chronic disease. Therefore, our analysis cannot distinguish

between “driver” and “passenger” diseases. Note that the low p-values and narrow credible/confidence intervals are driven by the large data set. This does not necessarily mean that the data set is free of bias, and any bias will propagate and can very well drive associations.

Second, while the US data set records the actual diagnoses dates (rather than “billing dates”), chronic diseases, such as hypertension and diabetes, are likely to have been developing for a long time before the actual diagnoses were made and recorded. It is possible that environmental insult (mild stress) acts as a trigger to worsen already pre-existing conditions, so that patients are forced to see a physician (which results in diagnoses entered in records). The Swedish national registry records the “discharge dates” instead of the actual dates of diagnosis. In most cases, neither the date of diagnosis as in the US data nor the discharge date is expected to be the same as the actual time of disease attack. But these dates still more accurately reflect the time of attack than the “billing dates” do. We tried to smooth out the discrepancy between the dates by adding up all the codes of interest in the following week of the DST shift and estimated the week-level RR based on it.

Third, it is possible that disease coding errors could influence our RR estimates. A simple way to spot miscoding is to look for sex-specific codes assigned to the wrong genders. For instance, the data shows males having pregnancy, ovarian cancer, and females having prostate cancer. There are two scenarios to explain the origins of such errors: Either the sex was recorded wrongly or the code itself is inaccurate. The former scenario would not affect our analyses substantially because we anticipated for symmetric coding errors in both sexes that offset against each other. On the other hand, if a diagnosis itself is miscoded, it may influence the RR estimation and tests. We estimated that the miscoding rate at the MarketScan data is around 0.52 percent, (see Section 2.6.1.12. Data Limitations, for details of this estimation). The error rate is

positive but is small in comparison to the observed DST shift effect sizes. Importantly, simple disease coding errors are unlikely to be in any way related to DST shifts, and would only bias RR estimates towards the null model.

Fourth, the major difficulties in our analysis were associated with changing coding standards (especially in Sweden, which went from ICD8 to ICD9, then to ICD10) and insufficient data sample despite the fact that data sets spanned whole countries. In particular, we did not have sufficient data to analyze every condition in every age/sex group. We alleviated the coding problem using two synergistic measures: (1) careful ICD version mapping, and; (2) considering relatively large constellations of diseases instead of following very specific conditions (for example, we analyzed "ischemic heart diseases," including acute myocardial infarction, instead of specifically "acute myocardial infarction"). We felt that these categories of larger disease groups were robust enough to withstand changes in medical practice and diagnosis criteria. The larger disease group we used in our analysis also helped with increasing disease-specific patient sample size, as the collective incidence of a group of diseases is the sum of incidences of individual diseases in the group.

2.6 COMPLETE DETAILS OF MATERIALS AND METHODS.

2.6.1 Materials and methods

2.6.1.1 The MarketScan all-patient database

The IBM Watson Health MarketScan compiles data from over a hundred large, US-based insurance companies. We used a 2016 snapshot of this database which contained 5,197,121,918 diagnosis records for 151,104,811 unique patients in the US, enrolled from 2003 to mid-2014. For each patient, we knew when and where they entered and left our database. Note that not all

patients were enrolled in our database from the start date to the end date. Patients might have been visible for a few weeks, a few months, or several years (Figure 2.5 and Figure 2.6). For each diagnosis entry, the database documented the date, the patient's age, and an International Classification of Diseases (IDC) 9th Version, Clinical Modification (ICD-9-CM) code. Because inpatient and outpatient hospitalizations were not discriminated in this database, we call it the "all-patient" database. We have also grouped patients by the state of enrollment; therefore, the experimental group includes patients from all states that observe DST, and the negative control group includes patients from states where DST has not been consistently observed (Arizona, Hawaii, and Indiana before 2006). We did not find patients registered in insular territories or minor outlying possessions in the data.

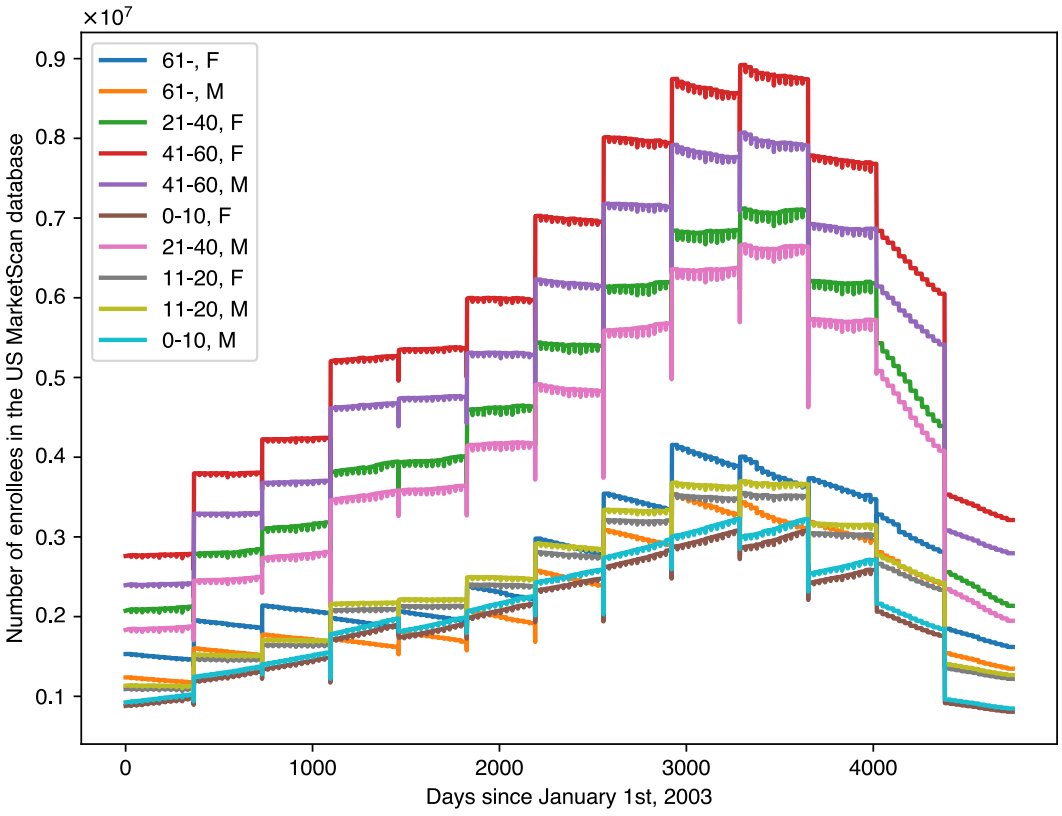

Figure 2.5 Enrollee variation in the US MarketScan data.

Figure 2.5 Continued: Integrating the models' real-time enrollee number is an easy step to account for the confounding enrollment variation. Around DST shifts, the total enrollee number does not change much (around one percent on average, see the plot below), but the yearly difference is significant. We observe that the number of enrollees changes several folds (yearly) from 2003 to 2014. The majority of these changes occur at the beginning of each year when new enrollees either joined or left the insurance policy. However, these yearly changes would not significantly affect the RR estimation, as we always compared the incidences from the few weeks around DST shifts.

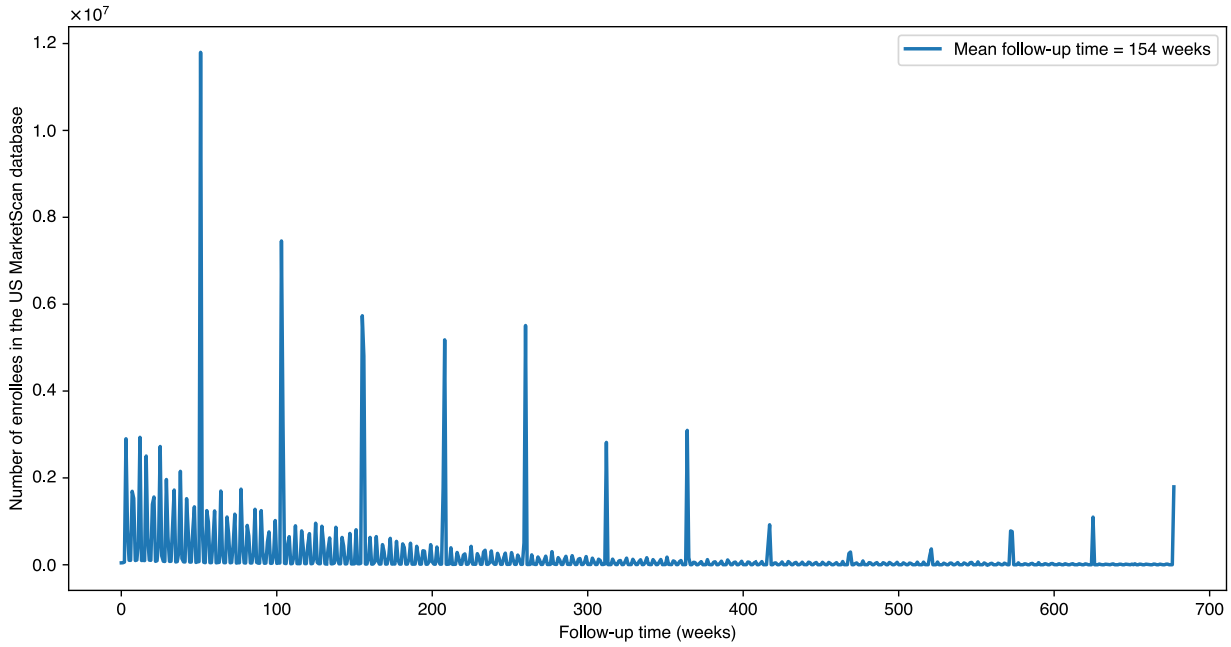

Figure 2.6 Number of enrollees in the MarketScan data set (Y axis) by week of study (Y axis).

Week 0 in this representation starts on January 1, 2003, and the weeks are numbered sequentially thereafter.

2.6.1.2 The MarketScan inpatient database

For our analysis, we used an inpatient version of the MarketScan database which documents 496,885,296 inpatient diagnoses records in ICD-9-CM for 175,208,465 unique people from 2003 to 2015. Most patients are duplicates of the enrollees of the all-patient database, but only their inpatient diagnoses were taken into account in this "inpatient database."

2.6.1.3 The Swedish inpatient registry

The Swedish inpatient database we used incorporates 94,669,631 inpatient diagnoses of 9,419,692 Swedish people from 1968 to 2010. Diagnoses are coded in Sweden's modification of the ICD, 8^{th} , 9^{th} or 10^{th} versions (ICD-8-SE, ICD-9-SE, or ICD-10-SE). The data are collected from all Swedish hospitals as discharge codes. In the Swedish data, enrollment is nearly static within the weeks surrounding either DST shift. Theoretically, this is because all Swedish inpatient hospitalizations are documented in this database. Unlike the insurance-company-curated MarketScan database, Swedish patients were dis-enrolled only if they died or left Sweden.

2.6.1.4 Mappings

The data sets we employed consisted of ICD codes (versions 8, 9, and 10 in the US) and their Swedish modifications. In order to fully utilize these codes, we first created mappings between the different ICD versions. We referred to the CDC General Equivalence Mappings for the translation between ICD-9-CM and ICD-10-CM [65]. We curated the mapping between ICD-9-CM and ICD-8 by ourselves (Supplementary File A.14). Additionally, we grouped the ICD-10 diagnosis codes, based on the first three digits, into 263 conditions under 31 systems using the WHO ICD-10 reference (Supplementary File A.1) [35]. Neighboring codes tended to fall in the same or related conditions. All ICD-10 codes are categorized in this grouping, so it is also exhaustive. Using all the mappings mentioned above, we produced a UniICD (`ICD_VERSION:ICD_CODE`)-to-condition mapping, and tuned it for both US and Swedish health records in compliance with the difference between US and Swedish ICD modifications in trailing digits (Supplementary File A.15 and Supplementary File A.16).

2.6.1.5 Models

We borrowed the idea for the general methodology, and the correction of holidays and day length from a previous study [28]. Thus, we calculated the base diagnosis rate using the expected proportion of patients out of all the enrollees in a certain age group and sex who were recorded as having a specific condition at a specific time point. A time point of interest is a day or week in this study, and it's RR can be quantified as follows:

$$\widehat{\mathrm{RR}} = \frac{p_0}{\frac{1}{2}(p_{-1} + p_{+1})}, \tag{2-1}$$

where p_{-1}, p_0, and p_{+1} are diagnoses rates for a particular test group (for example, for a condition in a certain age group and sex, such as diabetes for males aged over 60) at 0, the time point of interest, -1, two weeks before the time point of interest, or +1, two weeks after the time point 0 (Figure 2.1). If some influential holidays or celebrations fell into the week following the time point -1, or +1, the -1 and +1 points are then adapted to three weeks or one week before and after point 0. In the US data, we considered the following holidays and celebrations: President's Day in February, Western Easter, St. Patrick's Day, Memorial Day, Thanksgiving, Veterans Day, Columbus Day, and Labor Day. In the Swedish data, we considered only Western Easter. In the US, Easter and Thanksgiving showed the largest effect on disease reporting, as we examine in our studies.

The time intervals (called "time points" below) over which we counted disease code incidence, were either day- or week-long. We estimated both day- and week-level RRs for every test group in a Bayesian framework with a hierarchical model. We chose a set of flat, non-informative priors [66] for the diagnosis rates two weeks before and after the day of interest:

$$p_{-1} \sim \text{Beta}(1,1), \tag{2-2a}$$

$$p_{+1} \sim \text{Beta}(1,1). \tag{2-2b}$$

In addition, we drew RRs across all test groups from a Gamma prior with the mean μ and the standard deviation σ:

$$\text{RR} \sim \text{Gamma}(\text{mean} = \mu, \text{sd} = \sigma). \tag{2-3}$$

This prior distribution shrank all RRs towards the mean and let information flow across all conditions and test groups. As Gelman *et al.* suggested [36], the multiple comparisons problem is alleviated this way.

Note that all RR estimates (for all disease groups) were sampled simultaneously within the same inference framework; individual estimates constrained each other within one hierarchical model.

Hyper-priors of μ and σ were assumed to be nearly flat:

$$\mu \sim \text{HalfCauchy}(5), \tag{2-4a}$$

$$\sigma \sim \text{HalfCauchy}(5). \tag{2-4b}$$

We computed the diagnosis rate on the time point of interest as:

$$p_0 = \text{RR} \times \frac{p_{-1} + p_{+1}}{2}. \tag{2-5}$$

We assumed that the observed incidence values followed binomial distributions:

$$x_{-1} \sim \text{Binomial}(n_{-1}, p_{-1}), \tag{2-6a}$$

$$x_0 \sim \text{Binomial}(n_0, p_0), \tag{2-6b}$$

$$x_{+1} \sim \text{Binomial}(n_{+1}, p_{+1}), \tag{2-6c}$$

in which x_* and n_* ($*$ is -1, 0, or +1) were observable incidences and the total numbers of enrollees accumulated through years at spring or autumn DST shifts, respectively:

$$x_* \sim \sum_{y\text{: year of data}} x_{*,y}\,, \tag{2-7a}$$

$$n_* \sim \sum_{y\text{: year of data}} n_{*,y}\,. \tag{2-7b}$$

We adjusted for the varied day lengths at a DST shift by multiplying the observed day-level incidences by a factor of either 23/24 or 25/24, for the spring and autumn, respectively.

Finally, we estimated the posterior RR distribution via a Markov chain Monte Carlo (MCMC) sampler (PyMC3) [16] and computed the highest posterior density (HPD) interval as the credible interval. We used the highly efficient No-U-Turn sampler (NUTS) [12], initialized by a variational inference (ADVI, Automatic Differentiation Variational Inference [67]), which generated four independent Markov traces. Each trace was composed of 2,000 tuning iterations and an additional 2,000 drawing steps. Other arguments of NUTS and ADVI such as target acceptance rate, max tree depth, and step scale were PyMC3 3.6's default choices. We computed the Gelman-Rubin convergence diagnostic [68, 69] for all RR estimates by comparing the difference between the four traces. The final diagnostic results indicated that all samplings were well-mixed and RR estimates converged rapidly. We have supplied all the Gelman-Rubin statistics we used against the RR estimates in the Supplementary Files. We computed the final RR estimate distributions based on 2000×4 drawing steps. Again, please note that because the RRs were constrained by an across-the-board, hierarchical prior, we did not need to make formal corrections for multiple tests after sampling.

However, after applying this model to assess the effect of changing to and from DST, we found that, for most test groups, the risk is a little bit smaller than one and the mean of RR is smaller than one. This conflicted with our prior belief that, if DST shifts do not influence health, the relative risk should be approximately one. The less-than-one phenomenon was due to the fact

that the disease trend tends to be convex (bent downwards) at the DST shift time points. To compensate for this bias, we corrected the RR using the following equation:

$$\widehat{\mathrm{RR}}_{\mathrm{corrected}} = \frac{\widehat{\mathrm{RR}}}{\mathrm{E}\left[\widehat{\mathrm{RR}}\right]}. \tag{2-8}$$

In the above, we estimated the RR's expectation $\mathrm{E}\left[\widehat{\mathrm{RR}}\right]$ by estimating the Bayesian model's corresponding parameter, μ. This correction ensured that $\widehat{\mathrm{RR}}_{\mathrm{corrected}}$'s expectation was one, and for most test groups, the RR was inclined to one. An observed $\widehat{\mathrm{RR}}_{\mathrm{corrected}}$ that was significantly greater than one would mean the upward curvature at the DST shift had exceeded the average natural bent across all test groups. Such a correction procedure is equivalent to initializing μ in Expression (2-3) to 1.

Because we do not have the enrollment information for Swedish data, that corresponding model was slightly different. The Swedish data was characterized by high coverage and low mobility. Almost all Swedes were visible in the data set throughout their entire lives. Therefore, we determined it was safe to presume that enrollments did not change from two weeks before to two weeks after DST shifts. We quantified the RR as:

$$\widehat{\mathrm{RR}} = \frac{l_0}{\frac{1}{2}(l_{-1} + l_{+1})}, \tag{2-9}$$

where the diagnosis rate l_* here is not the proportion but the exact number of incidences expected to be documented for a certain condition at a time point. We still assumed RR to follow a Gamma distribution with an across-all-condition mean μ and we assumed σ as the standard deviation. We set priors for l_{-1}, l_{+1}, μ, and σ to follow a flat, half-Cauchy distribution with a large-scale parameter. Again, we assumed the observed incidences to follow Poisson distributions:

$$x_{-1} \sim \text{Poisson}(l_{-1}), \tag{2-10a}$$

$$x_0 \sim \text{Poisson}(l_0), \tag{2-10b}$$

$$x_{+1} \sim \text{Poisson}(l_{+1}). \tag{2-10c}$$

Finally, we accounted for the convex tendency by using Expression (2-8).

2.6.1.6 Alternative models

As an alternative to the Bayesian method, we also tested frequentist models based on random variables' asymptotic properties. Both types of analysis, Bayesian and frequentist, led to nearly identical conclusions.

Frequentist method

For large x and n , we used a normal distribution to approximate the error. The normal approximation of the diagnosis rate is as follows:

$$p \dot{\sim} \text{Normal}\left(\hat{p}, \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right), \tag{2-11}$$

where $\hat{p} = \frac{x}{n}$ is a realization of random variable $p = \frac{X}{N}$. Throughout this text, we use notation $\text{Normal}(\mu, \sigma)$ to denote a normal distribution with mean μ and variance σ^2.

Using this approximation, we can compute the normally-approximated p_{-1}, p_0, and p_{+1}, corresponding to the diagnosis rate two weeks before, on, and after the day or week of interest. The expected diagnosis rate on the time point of interest, $\bar{p}_0 = \frac{1}{2}(p_{-1} + p_{+1})$, is also normal:

$$\bar{p}_0 \dot{\sim} \text{Normal}\left(\frac{1}{2}(\hat{p}_{-1} + \hat{p}_{+1}), \sqrt{\frac{\hat{p}_{-1}(1-\hat{p}_{-1})}{4n_{-1}} + \frac{\hat{p}_{+1}(1-\hat{p}_{+1})}{4n_{+1}}}\right). \tag{2-12}$$

Here, we assume p_{-1} and p_{+1} are conditionally independent given a consistent, disease-specific incidence rate. This assumption is generally true if we admit that every incidence is unrelated but only depends on the disease's intrinsic attributes. The RR is $\frac{p_0}{\bar{p}_0}$, which is the ratio of two normal, random variables. This ratio's distribution was discussed by Hinkley in 1969 [70]. Specifically, for two independent, normally-distributed, random variables $X_1 \sim \text{Normal}(\theta_1, \sigma_1)$ and $X_2 \sim \text{Normal}(\theta_2, \sigma_2)$, the cumulative distribution function $F(w)$ of their ratio $W = \frac{X_1}{X_2}$ can be approximated by

$$F(w) \to \Phi\left(\frac{\theta_2 w - \theta_1}{\sigma_1 \sigma_2 a(w)}\right) \quad \text{as} \quad \frac{\theta_2}{\sigma_2} \to \infty, \tag{2-13}$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution, and

$$a(w) = \sqrt{\frac{w^2}{\sigma_1^2} + \frac{1}{\sigma_2^2}}. \tag{2-14}$$

Let

$$\theta_1 = p_0 = \frac{x_0}{n_0},$$

$$\theta_2 = \bar{p}_0 = \frac{1}{2}(p_{-1} + p_{+1}) = \frac{1}{2}\left(\frac{x_{-1}}{n_{-1}} + \frac{x_{+1}}{n_{+1}}\right),$$

$$\sigma_1 = \sqrt{\frac{\hat{p}_0(1-\hat{p}_0)}{n_0}},$$

$$\sigma_2 = \sqrt{\frac{\hat{p}_{-1}(1-\hat{p}_{-1})}{4n_{-1}} + \frac{\hat{p}_{+1}(1-\hat{p}_{+1})}{4n_{+1}}}.$$

The $(1-q) \times 100\%$ confidence interval of $\text{RR} = \frac{p_0}{\bar{p}_0} = \frac{\theta_1}{\theta_2}$ can be found by solving the equation

$$\frac{\theta_2 w - \theta_1}{\sigma_1 \sigma_2 a(w)} = z. \tag{2-15}$$

in which z is the $1 - \frac{q}{2}$ or $\frac{q}{2}$ quantile of the standard normal distribution.

Half-Bayesian method

For a binomial proportion $p = \frac{x}{n}$, we also used a Bayesian method with a Jeffrey's prior, a beta distribution with parameters $\alpha = \beta = \frac{1}{2}$, to approximate its distribution. The posterior distribution is also a beta with parameters:

$$\beta = \frac{1}{2} + n - x.$$

For large α and β, we may use a normal distribution with mean and variance

$$\mu = \frac{\alpha}{\alpha + \beta},$$

$$\alpha = \frac{1}{2} + x,$$

$$\sigma^2 = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)},$$

as an approximation. Subsequently, the process of estimating the RR and its confidence interval would be identical to what we have discussed below, in the frequentist method section.

Frequentist method for Swedish data

Due to the consistency of enrollment, the frequentist model is simpler for the Swedish data. Again, we assumed Poisson distributions to observe diagnoses as done in Expression (2-10a). Notice that the sum of two independent Poisson random variables is still Poisson:

$$x_{-1} + x_{+1} = x_s \sim \text{Poisson}(l_s = l_{-1} + l_{+1}). \tag{2-16}$$

We considered the confidence interval for the halved RR:

$$\frac{1}{2} \cdot \mathrm{RR} = \frac{l_0}{l_s}. \tag{2-17}$$

Ederer and Mantel have shown that the Poisson ratio's confidence interval in the above form can be deduced from the confidence interval of the binomial parameter, $\theta = \frac{l_0}{l_0 + l_s}$. This is because x_0, given $x_0 + x_s$, is a conditional binomial distribution [71].

We used the Wilson score interval [72] for θ. The lower and upper endpoints of the confidence interval of RR are as follows:

$$R_L = \frac{2\theta_L}{1 - \theta_L}, \tag{2-18a}$$

$$R_U = \frac{2\theta_U}{1 - \theta_U}, \tag{2-18b}$$

where θ_L and θ_U are the lower and upper limits of the Wilson score interval of θ.

2.6.1.7 The multiple comparisons problem

The multiplicity involved in constructing thousands of credible or confidence intervals simultaneously in the present study may have given rise to erroneous inferences. We controlled this problem in different – but appropriate – ways for the Bayesian and frequentist (and half-Bayesian) methods.

For the Bayesian method, we controlled multiplicity by an across-the-board prior (Expression (2-3)). We parameterized the prior by using the mean and standard deviation of Gamma but not the commonly-used shape α and rate β. Nevertheless, for the RRs in our study, the mean μ is usually close to one and for $\sigma \ll \mu$, the shape and rate $\alpha = \frac{\mu^2}{\sigma^2}, \beta = \frac{\mu}{\sigma^2}$ are large, leading to an approximately normal distribution. Sampling for such a prior would shrink all the RRs towards the mean and restrain the significance of posterior intervals. We determined how

closely we controlled the multiplicity by using the σ scale. A smaller σ would induce more conservative (wider) posterior credible intervals. In our study, $\sigma$was naturally inferred from the whole data pool with a nearly-flat prior.

We adjusted the simultaneous confidence intervals constructed by the frequentist analyses by controlling the false coverage rate (or false coverage-statement rate, FCR) [37]. Benjamini and Yekutieli [37] proposed a very simple procedure to control FCR $\leq q$, through adjusting the significance level q:

$$q \to q \cdot \frac{s}{n}, \tag{2-19}$$

where s is the number of selected estimates among all candidates, and n is the number of candidates. The selection procedure in our study consisted of simply choosing those RR estimates significantly less or greater than one, meaning the **unadjusted** confidence intervals before any correction procedure covered one. The number of candidates is the total number of RRs we were trying to estimate. It is worth noting that we applied these correction procedures only to the selected estimates, whose unadjusted confidence intervals do not cover zero, as suggested by Benjamini and Yekutieli in the original article [37]. Other insignificant confidence intervals remain unchanged. This type of correction procedure ensures that the FCR, i.e., the proportion of the true parameter's failure coverage rate, is less than or equal to q, among all selected intervals.

2.6.1.8 Negative controls

We implemented negative controls in a few ways. For the Swedish data, we analyzed records before 1980, when the DST shift was not obeyed, as a negative control. First, we

calculated the pseudo-DST shifts' RRs from 1968 to 1979. We determined the pseudo-DST shift using the following equation:

$$\begin{aligned}&\text{Pseudo-date of DST shift in year } y\\&= \text{Date of DST shift in year } (y+12) - 12 \times 365 - 2\end{aligned} \tag{2-20}$$

where $1968 \leq y \leq 1979$. We deducted two days in this equation to calibrate the pseudo-DST dates to Sundays. We compared these pre-DST-shift results to the results of the RR estimation for data after 1980. Because we applied the intervention to the entire population of Sweden (comparing patients to themselves before and after intervention), we treated Swedish data as a natural experiment and deduced a stronger conclusion.

All the US health data were collected after the DST shift policy was implemented, so we were unable to contrast results for actual DST shifts with no-treatment control observations (observing DST shift dates, but no DST shift, as we were able to do with Swedish data). We therefore designed negative controls in two different ways:

1. We used health statistics from US states that do not observe the DST shift (Arizona, Hawaii, and Indiana before 2006) as a negative control. However, we found the population too small to inform any statistically reliable conclusions. Compared to the experimental group, which included hundreds of millions of patients who observe the DST shift, this control group only consisted of a few million people. For many diseases, we only have less than ten incidences a day for a specific age-sex stratification.

2. We applied our pipeline to dates other than those with DST shifts. For spring, we repeated the whole analysis for 28 days after the DST shift (day and week). For autumn, the control date was selected at 28 days before the DST. Again, we adjusted for holidays: We avoided President's Day in February, Western Easter, St. Patrick's Day, Memorial Day,

Thanksgiving, Veterans Day, Columbus Day, and Labor Day. Similar to the previous Swedish experiment in 1980, we addressed this negative control as "pseudo-DST" analyses on other dates.

2.6.1.9 Geographic location and other covariates

Geographic location will drive variance in daytime length and therefore may affect the DST shift's influence on health. Thus, we divided the applicable states that observe the DST shift into two groups (northern and southern) parts and performed separate analyses. The northern part includes Oregon, Idaho, Wyoming, Nebraska, Iowa, Illinois, Indiana, Ohio, Pennsylvania, New Jersey, and states on the northern side of the above-mentioned states' southern border. The southern part includes states on the south side of the northern states' southern border.

We also considered another covariate: the difference in culture and work-life balance between western and (north) eastern states. The western part includes Washington, Oregon, California, and Nevada. The eastern part includes Maine, New Hampshire, Vermont, Massachusetts, Connecticut, Rhode Island, New York, Pennsylvania, New Jersey, Maryland, Delaware, Virginia, and the District of Columbia. We performed analyses separately for these two areas.

2.6.1.10 First-time diagnoses

The potential dissimilarity between the effects of DST shifts on a condition's first incidence and a recurrent follow-up diagnoses is also noteworthy. Because of most patients' incomplete enrollment in the MarketScan data, we were not able to extract a condition's first diagnoses. We had the entire medical history from birth of only a small proportion of zero to ten-

year old children (around six million), and, in those cases, we were able to extract first diagnoses. We performed analyses on these patients combining all the DST shift states, along with a negative control experiment on "pseudo-DST" shift dates. A negative control on the non-DST states (Arizona, Hawaii, and Indiana before 2006) was unfeasible due to lack of data.

For completeness and comparison purposes, we still performed another analysis on a disease's first-time diagnosis in each patient's insurance claim sequence in the MarketScan data (US all patients). This first observable diagnosis is not necessarily a disease's true first incidence, but could be useful for comparison.

As the Swedish register has the complete hospitalization profile for all patients, it allows us to distinguish the first-time diagnoses of chronic disorders from follow-up visit diagnoses. Thus, we also performed tests for all Swedish inpatients. However, because the reduced data provided a much smaller sample size (lifetime first-time diagnoses only, for everyone), this prevented our signals from reaching statistical significance using either the Bayesian or the frequentist method. Note that the Swedish data set only contains inpatient diagnoses, so we are not able to know if there were any identical but non-hospitalized diagnoses occurring before the first inpatient records of a particular condition.

2.6.1.11 Methodological limitations

The fundamental assumption of the model, illustrated in Figure 2.1, is that a disease incidence's short-term trend is approximately linear for the weeks surrounding a DST shift. If this is true, we would be able to detect irregular disease incident variation by comparing the observed versus the expected (average) diagnosis rates. This assumption might be violated for some highly seasonal and fluctuating conditions. We implemented the negative controls and validated the results in the two countries to alleviate this issue. We are comfortable interpreting

only DST shift signals supported by the results of statistical tests of various types across two countries. The set of US-only signals has to be interpreted more carefully: While results are "real" in the statistical sense, very large data sets capture a plethora of various signals (social, ethnic, economic, cultural, and climate- and weather-related) that do not lend themselves to easy deconvolution.

2.6.1.12 Data limitations

Because we focused solely on diagnostic codes used around DST shift dates, the data we used in this study were not large enough for some rare diseases. In addition, because many people's insurance enrollment intervals do not overlap with the DST shift dates, the US insurance claim data only covered around one-tenth of the US population (35 million) during and around those dates. For some rare conditions, we observed only a few instances in both data sets (the US and Swedish), so the statistical power for detecting putative DST shift signals was insufficient.

It is possible that disease coding errors could influence our RR estimates. A simple way to spot miscoding is to look for sex-specific codes assigned to the wrong gender. For instance, the data shows males diagnosed with pregnancy, ovarian cancer, and females diagnosed with prostate cancer. There are two scenarios to explain the origins of such errors: Either the sex was recorded wrongly or the code itself is inaccurate. The former scenario would not affect our analyses substantially because we anticipated for symmetric coding errors in both sexes that offset against each other. On the other hand, if the diagnosis itself is miscoded, it may influence the RR estimation and tests. We summarized data for some female- and male-specific diseases as anchor points to estimate the coding error rate (Supplementary File A.17 and Supplementary File A.18). We approximated the error rate using the following equation:

$$(\mathrm{FP}+\mathrm{FN})\ \textsc{Error Rate} = 100\% \times \frac{2 \times \left(F_m + M_f\right)}{F_f + F_m + M_m + M_f} = 0.52\% \tag{2-21}$$

where FP stands for "false positive," FN stands for "false negative," F_m is the number of male-assigned, female-specific diagnoses, F_f is the number of female-assigned, female-specific diagnoses, M_m is the number of male-assigned, male-specific diagnoses assigned to females, and M_f is the number of female-specific diagnoses assigned to males.

With our coding error estimates, the error rate is positive, but is small in comparison to the observed DST shift effect sizes.

Importantly, simple disease coding errors are unlikely to be in any way related to DST shifts, and would only bias RR estimates towards the null model.

2.6.2 Analyses summary

All week- and day-level results can be found on the project's web site https://github.com/hanxinzhang/dst

We performed analyses for all 263 disease groups for females and males separately, partitioned into five age groups (0 to 10, 11 to 20, 21 to 40, 41 to 60, and greater than or equal to 61). We started from $263 \times 10 = 2630$ test groups in total and filtered out those with lower than ten incidences on any day of study. This quality-control step ensured that all the analyzed conditions were statistically meaningful. For sub-common diseases, none of our Bayesian, frequentist, or half-Bayesian methods could give dependable results. Notice that, for the Swedish data, we arranged the age groups slightly differently: 0 to 20, 21 to 40, 41 to 60, and over 60. This difference in analyses groupings does not affect our discussion or conclusions, as we did not find any significant signals in the younger populations in Sweden.

The time periods covered by the study are: (1) the seven days of the week, two weeks before a DST shift; (2) the seven days just after a DST shift, and; (3) the seven days of the week, two weeks after a DST shift.

For negative controls performed on other dates, including the 28 days before or after a DST shift (or pseudo-DST dates before 1980 in Sweden), the days of study were taken around the pseudo-DST dates in a similar way. If any day of study coincided with a holiday or celebration (see Section 2.6.1.5 Models), we used either one or three weeks before and after the DST or pseudo-DST shift.

The number of test groups and the number of spring week-level RRs that were significantly greater or less than one are summarized in Table 2.1.

Table 2.1 Summary of DST analyses

Experiment	No. tests	No. sig. (Bayesian)	No. sig. (frequentist)	No. sig. (half-Bayesian)
US all-patient, all DST states	2025	290	265	265
US all-patient, northern DST states	1691	180	287	287
US all-patient, southern DST states	1802	181	191	191
US all-patient, eastern DST states	1471	118	137	137
US all-patient, western DST states	1258	63	81	81
US all-patient, neg. ctrl on other states	546	4	17	17
US all-patient, neg. ctrl on other dates	2041	229	331	331

Table 2.1 Continued

Experiment	No. tests	No. sig. (Bayesian)	No. sig. (frequentist)	No. sig. (half-Bayesian)
US inpatient, all DST states	1635	70	64	64
US inpatient, northern DST states	1117	19	28	28
US inpatient, southern DST states	1291	35	35	35
US inpatient, eastern DST states	854	16	8	8
US inpatient, western DST states	633	11	16	15
US inpatient, neg. ctrl on other states	218	3	1	1
US inpatient, neg. ctrl on other dates	1639	28	31	31
Swedish inpatient since 1980	836	7	4	
Swedish inpatient before 1980	242	0	0	

2.6.2.1 The signal selection procedure for presenting the US results

Conditions with increased risk during spring DST shifts

We designed multiple controls and compared them to spring DST shift tests to corroborate the risk estimation associated with these shifts.

To perform the US analyses, we chose negative-control dates (pseudo-DST) near the actual DST shift dates; in addition, we used US states that do not observe DST ("non-DST

states") as negative controls (Section 2.6.1.8). Our analyses of Swedish data were somewhat simpler because Sweden did not observe DST before 1980, providing a natural negative control. We present all significant spring signals for the Swedish analyses in the Section 2.3 (Figure 2.3B) because all of their natural negative controls before 1980 are not significantly different from one. We used similar selection criteria – by comparing the negative controls to the responses of actual DST shifts – to identify potential conditions associated with spring DST shifts in the US analyses (see the descriptions in the following paragraphs).

Autumn DST shift dates could be viewed as controls for spring's disease RR changes. We did not expect much risk to be associated with the period surrounding autumn DST shifts. A previous study on DST shift association with acute myocardial infarction used solely the autumn tests as control [28]. In contrast to that study's design, we decided to focus on the pseudo-DST shift's negative control near the actual DST shift dates for several reasons.

First, the negative controls we used on the non-DST states did not render a sufficient number of observations; the data was too sparse to allow for a fair comparison to the experimental tests on actual DST shifts. We filtered out many conditions during our quality control step due to their low incidence in the non-DST state data, and it was not even possible to make comparisons for these diseases.

Second, for some diseases, the one-hour disruption – in either direction – seemed to lead to an RR increase. For example, behavioral and emotional disorders in young adults increase after the DST shift in both spring and autumn (Figure 2.4 panel A). We might conclude that the autumn DST shift may also have negative effects if this signal were not compared to the negative control tests, which actually revealed a dubious effect in autumn because the negative control test showed a close result.

Third, we accounted for seasonal variation patterns and disease incidence curvature by comparing actual DST time points with pseudo-DST time points, both chosen close enough (a few weeks) to actual DST points. The goal in such an analysis is to account for seasonal confounding factors. Panel A in Figure 2.4 shows the example mentioned, in which autumn mental and behavioral disorders seem to increase for both actual DST and pseudo-DST shifts with no differentiation. This indicates such inflation is more likely to be caused by the trend's upward curvature or disease seasonality as opposed to DST time shifts.

We selected all diseases with an increased risk that could be putatively associated with the spring DST shift using the following criteria: (1) The estimated spring DST shift's RR should be significantly greater than one after Bayesian shrinkage or frequentist FCR correction, and; (2) The RR associated with the pseudo-DST shift near the actual spring DST shift dates should not be significantly greater than one after correction for multiple comparisons. The results are shown in Supplementary File A.19 (US all-patient, Bayesian), Supplementary File A.20 (US all-patient, frequentist), Supplementary File A.21 (US inpatient, Bayesian), and Supplementary File A.22 (US inpatient, frequentist). The half-Bayesian estimates closely followed the frequentist (see Figure 2.7 and Figure 2.8), so we focused on comparing the more divergent Bayesian and frequentist results. Figure 2.9 (US all-patient) and Figure 2.10 (US inpatient) plot spring DST shifts' RRs versus the spring negative control's RR, with selected conditions based on the above-mentioned criteria showing increased risk in blue. The top five conditions with the largest absolute effect sizes ($\widehat{\mathrm{RR}} - 1$) are text-labeled.

Using the above-mentioned selection procedure in conjunction with the Bayesian method, we found 82 increased signals for the US all-patient analysis (Supplementary File A.19 and Figure 2.9 Panel A). We found 69 when using the frequentist method (Supplementary File

A.20 and Figure 2.9 Panel B). Inspecting these results, we noticed a number of ill-defined clinical and laboratory findings, examinations, and health services, for which the increased risk could be attributed to various diseases. In addition, some infections, and possibly infection-related eye, ear, genitourinary, and respiratory diseases showed increased risk. The results also suggest possible inflated risk in some circulatory, digestive, metabolic, endocrine, nutritional, musculoskeletal, skin, neoplasm, and mental/behavioral/nervous system diseases, childbirth problems, and injuries in various body sites (Supplementary File A.19 and Supplementary File A.20). Some anemias also stand out, though only in the frequentist results (and not in the more conservative Bayesian results, Supplementary File A.20).

For the US inpatient analyses, we selected 42 conditions using the same two selection criteria (formulated above) with a Bayesian analysis (Supplementary File A.21 and Figure 2.10 Panel A), and 39 with frequentist estimates (Supplementary File A.22 and Figure 2.10 Panel B). Again, we saw infections, genitourinary, and respiratory diseases' risks enlarge. We also saw increased risks in immune, circulatory, digestive, endocrine, metabolic, musculoskeletal, neoplastic, mental/behavioral/nervous system diseases, childbirth problems, and injuries in various body sites (Supplementary File A.21 and Supplementary File A.22).

Conditions with decreased risk during spring DST shifts

We performed all our analyses bi-directionally (looking for both increases and decreases in disease risk). *A priori,* we expected to find as many decreased signals as increased ones (assuming that signals are distributed randomly with mean zero). This is because, if we assume that there is no effect from changing time, the RR distribution estimates should then be approximately zero-mean normal. Thus, we focused here on significantly decreased RR signals during the spring DST shift period.

Using a similar positive signal selection, but a reverse procedure, we selected conditions with a spring RR of significantly less than one and a negative control RR not significantly less than one. We summarized these selected conditions in Supplementary File A.23 (US all-patient, Bayesian), Supplementary File A.24 (US all-patient, frequentist), Supplementary File A.25 (US inpatient, Bayesian), and Supplementary File A.26 (US inpatient, frequentist). Figure 2.11 (US all-patient) and Figure 2.12 (US inpatient) show the spring DST RRs versus the spring negative-control RRs and selected conditions with decreased risk shown in orange. The top five largest-effect conditions, along with their absolute effect sizes ($1 - \widehat{\mathrm{RR}}$), are text-labeled.

Our tests revealed a number of decreased RR signals. The Bayesian all-patient results showed 115 diseases with decreased RR immediately after the actual DST shift, but no significant decrease after the pseudo-DST shift dates (Supplementary File A.23 and Figure 2.11, Panel A). When we used the frequentist method, we were able to identify 80 diseases with such behavior (Supplementary File A.24 and Figure 2.11, Panel B). We observed that these protective signals were distributed differently with regards to human biological systems – rather than with regards to diseases possessing increased risk. The diseases with significantly decreased RRs are associated with infection, genitourinary/urinary systems, skin, musculoskeletal functions, nervous system, neoplasms, blood diseases (anemia), and some injuries.

We also found many decreased RR signals in mental and behavioral disorders across various age groups and both sexes in the US all-patient analysis (Supplementary File A.23, Supplementary File A.24, and Figure 2.4, Panel B). We observed these signals in disease groups that were very different from those with increased risks.

For instance, organic mental disorders showed decreased RR signals in the senior population, while neurotic, stress-related disorders, and youth behavioral and emotional

disorders (including attention-deficit/hyperactivity disorder) showed increased RR signals in the US all-patient data set. Decreased RRs for circulatory and digestive conditions were also dissimilar from their corresponding, increased RR conditions. Remarkably, cerebrovascular diseases in the middle-aged and senior populations showed decreased RR in the US analysis.

We did not, however, observe any childbirth and pregnancy-related conditions with decreased RRs in the US all-patient analyses. This observation gives more credibility to one of this study's largest-effect signals – those related to the spring DST shift's increase in disease RR related to maternal care for women of advanced ages.

The inpatient results were generally consistent with the all-patient results (Supplementary File A.25, Supplementary File A.26, and Figure 2.12). However, some diseases' RR signals reversed signs across age groups. For example, for disease codes associated with "other forms of heart disease," spring DST shift RRs decreased in females aged 41 to 60 (Supplementary File A.25 and Supplementary File A.26), but increased in some young and senior age groups (Supplementary File A.21 and Supplementary File A.22).

2.6.2.2 Methods comparison

We evaluated DST shift's effects on health based on the results of a Bayesian analysis (which had a tendency to be more conservative in terms of the number of signals detected) – though the other methods, Bayesian, frequentist, or half-Bayesian approaches, would have led us to very similar conclusions. Figure 2.7 and Figure 2.8 show comparisons of the methods we used to estimate disease RR. Each triad – consisting of the blue, red, and purple bars – shows a particular test group's confidence or credible intervals (CI) and RR estimates. All three methods gave us close RR estimates with comparable interval widths, especially for estimates close to one. In those more extreme cases, in terms of effect size, the Bayesian method did provide a

smaller RR estimate due to its shrinkage property. The Bayesian method tended to "pull" RRs towards the across-the-board mean if the information from the observation did not surpass the prior. Figure 2.13 (US all-patient) and Figure 2.14 (US inpatient) clarify this claim. Figure 2.13, Panel A and Figure 2.14, Panel A show a scatter-plot of frequentist versus Bayesian estimates. One can see there are two types of estimates on the plot forming two lines of dots – one diagonal and the other off-diagonal. The diagonal line shows conditions with enough observed incidences that all three methods returned very close RR estimates. By contrast, the off-diagonal line shows that the frequentist method returns more extreme estimates (in terms of absolute effect size), while both Bayesian methods shrink the estimates to the prior, no-effect assumption. The RR estimate distribution, shown in Figure 2.13, Panel B and Figure 2.14, Panel B, partially explains why there are as many decreased signals as increased signals. The RR estimates follow a normal distribution with symmetric tails. Again, the Bayesian method shows a conservatively shrunk estimation.

2.6.2.3 Geographic location comparison

The DST shift's health effects may differ between the southern, northern, eastern, and western areas of the US (Figure 2.15 and Figure 2.16, Bayesian estimates). For example, the population appears to suffer more from heart diseases in the southern and western populations (Figure 2.16), but the northern and eastern populations may be at higher risk for injuries and neurotic and stress-related mental disorders (Figure 2.15). Nevertheless, we opt not to make any inference from these comparisons because too many other covariates should be considered, and a larger population is required to draw any meaningful conclusion.

2.6.2.4 Results of the first-diagnoses analyses

US children (aged zero to ten) with a complete medical history

The results of these analyses can be found on the project's web site https://github.com/hanxinzhang/dst/tree/master/us_allpatient/results_AllStatesWithDst_trueFirstDiag0-10 . The first-diagnoses analyses show a lack of statistical power. Most conditions were filtered out during the quality control stage. We would not trust much in the results, as many conditions were not even tested, and those analyzed were not constrained adequately because of the small test number. The sporadic signals we found in the all-patient, first-diagnoses experiment are more likely to be due to seasonal convexity or concavity – dermatitis, eczema, respiratory, and various communicable diseases. The only remarkable signal is non-infective enteritis and colitis for female children, going up about 3.4 percent in RR in the spring all-patient data. This is also one of the most important discoveries we observed in other analyses, and it is possibly associated with other, stress-related mental health issues and immune disorders.

US all patients, condition's first incidence in the insurance claim sequence

Again, the first-diagnoses analyses show a lack of statistical power due to data limitation. The only significant signal we replicated is the notable circulatory condition in senior females (other forms of heart diseases in female patients over 60) using the frequentist method. The results can be found on https://github.com/hanxinzhang/dst/tree/master/us_allpatient/results_AllStatesWithDst_firstDiag .

Swedish inpatients, condition's first hospitalized diagnosis

Most of the conditions were filtered out due to their low incidences. There is no significant signal in the results. The results can be found on https://github.com/hanxinzhang/dst/tree/master/se_inpatient/first%20diag .

2.6.2.5 Absolute human cost of the daylight saving time shift

Due to asynchronous enrollments (not every patient joined from the first day of the MarketScan database and left on the last day), our data only covered around one-tenth of the US population (35 million) during and around DST shift dates. We estimated the cost of spring DST shifts in terms of incident elevation (Supplementary Files A.27-30). For selected, increasing conditions (see Section 2.6.2.1 for the selection procedure), we estimated the incident increase by $(\widehat{\mathrm{RR}}-1) \times$ # of expected incidences averaging points $+1$ and -1. To compute the total number of incidences possibly associated with spring DST shifts, we combined conditions with significantly increased RRs in the spring that were not ruled out by the negative control experiment based on our selection criteria, discussed in Section 2.6.2.1.

The RR elevation translates to the following incidences in the first week of a DST shift on average, out of 35 million people:

- 600 more inpatient incidents of other forms of heart disease
- 300 more inpatient incidents of ischemic heart diseases in people over 60
- 500 more behavioral and emotional disorders for eleven to 20-year olds
- 200 more diagnoses of non-infective enteritis and colitis in 21 to 40-year olds

These numbers are all based on our Bayesian estimates (shown in Supplementary File A.27 for US all-patient and Supplementary File A.29 for US inpatient). We also approximated

costs using the frequentist results where the estimation is larger (Supplementary File A.28 for US all-patient and Supplementary Files A.30 for US inpatient).

In all, we found that around 15,000 incidences of all kinds on average per year in the first week of a DST shift could be linked to the time change. Considering the coverage rate of our data, 0.15 million disease incidents and conditions could emerge due to DST shifts in the US every year. Globally, there could be 0.88 million more disease incidents during the week after the spring DST shift, every year.

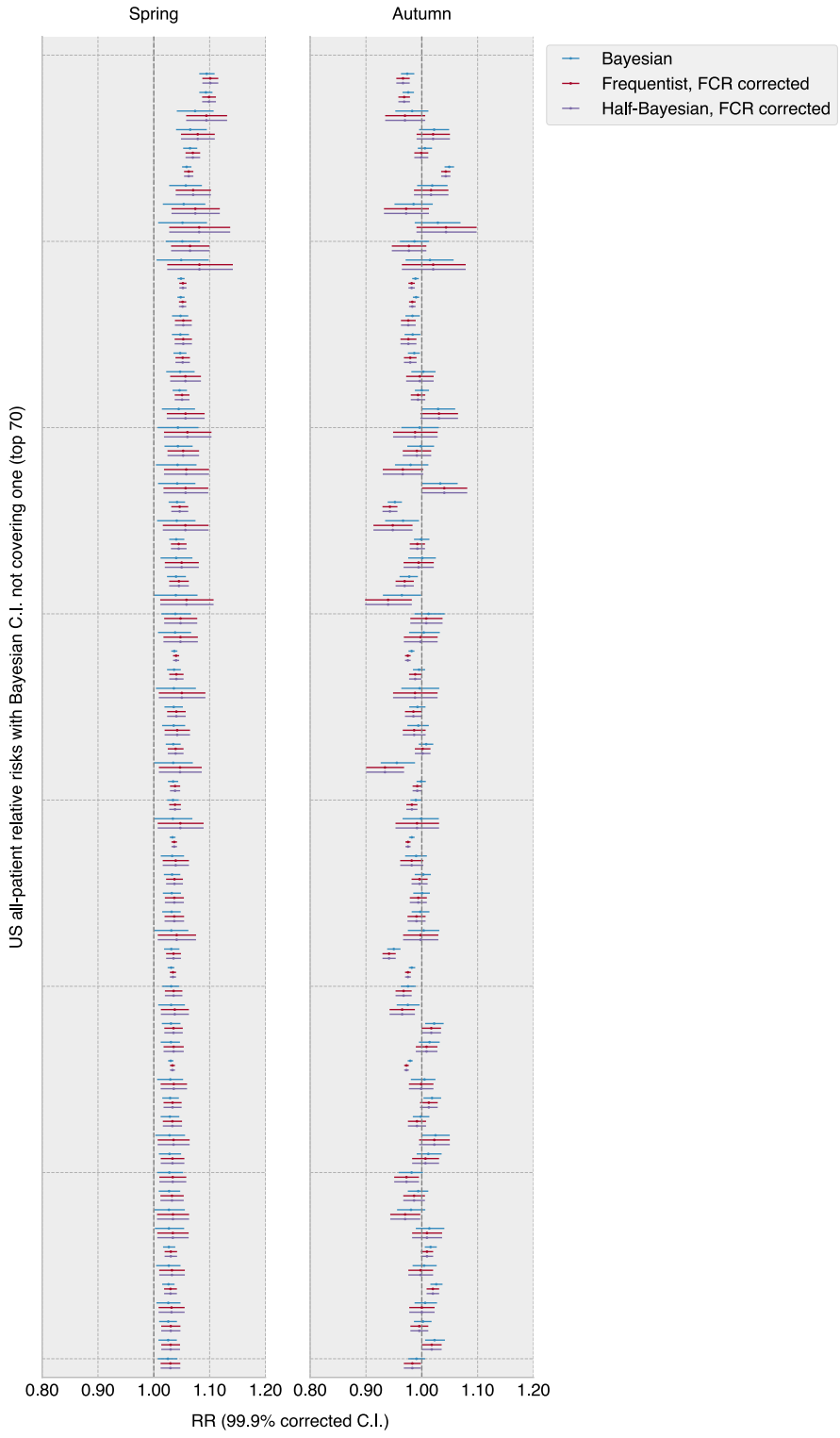

Figure 2.7 Different methods result in similar RR and interval estimates (US all-patient).

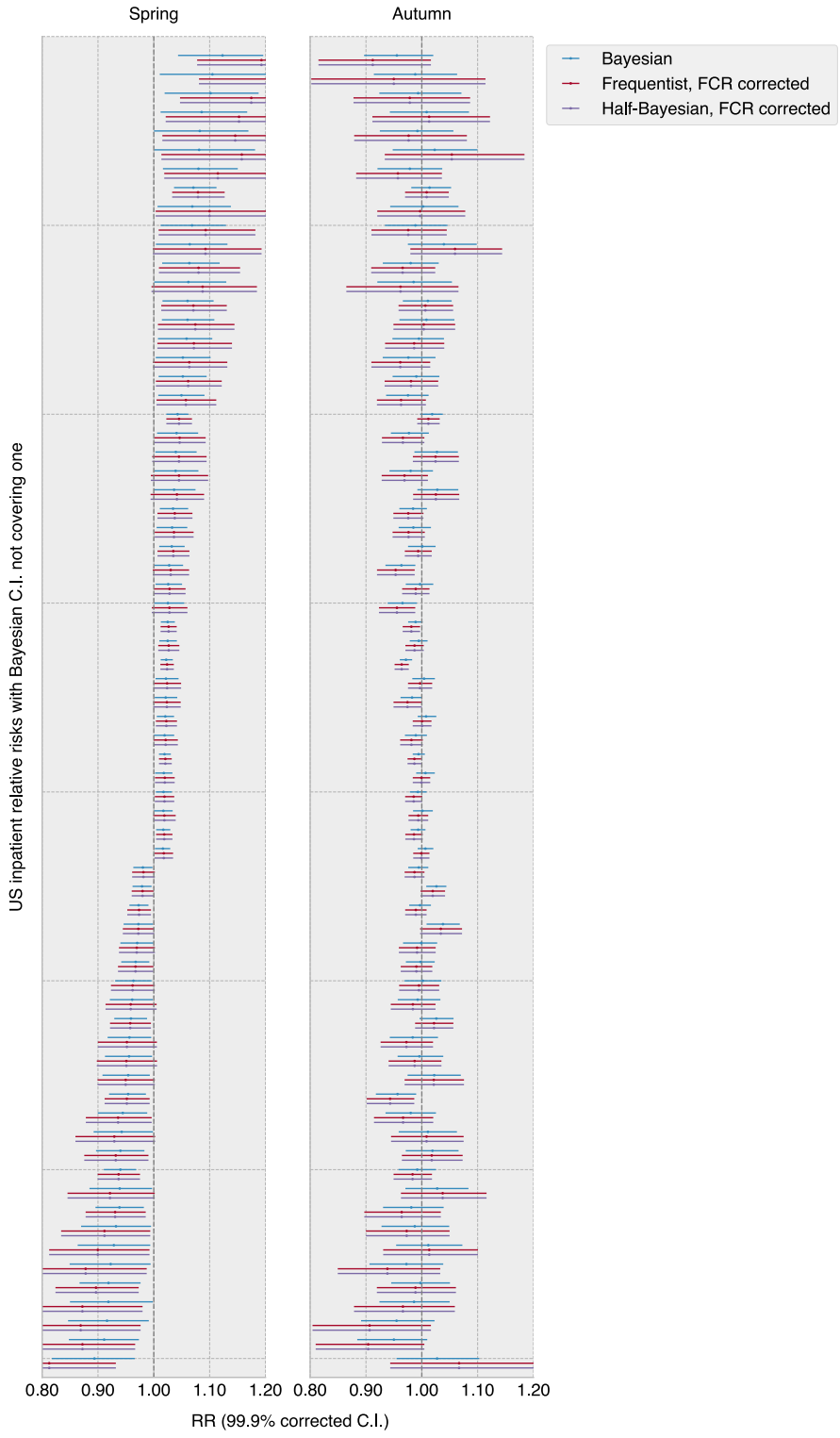

Figure 2.8 Different methods result in similar RR estimates and intervals (US inpatient).

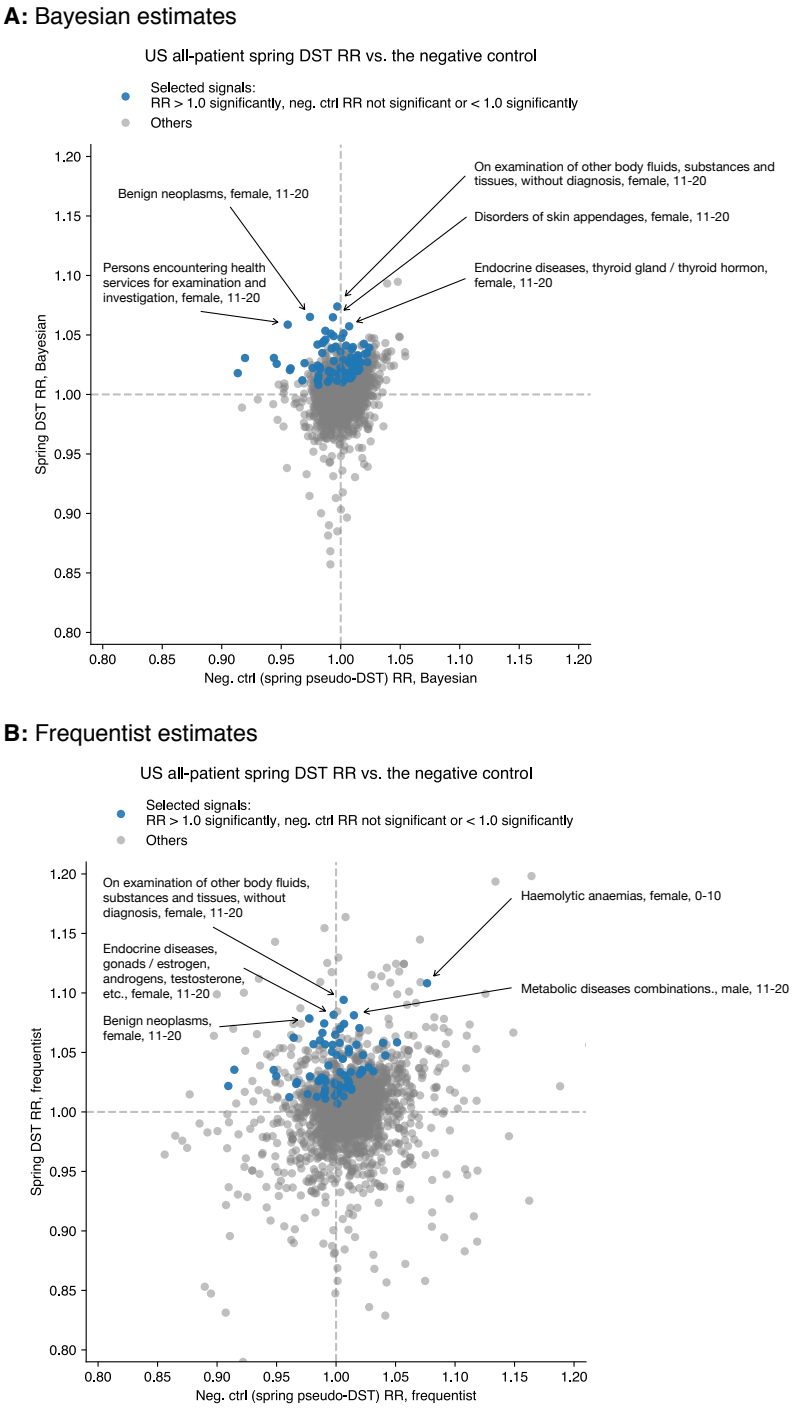

Figure 2.9 Selecting conditions with increased risk by comparing spring DST shift RRs to the negative control on pseudo-DST shift dates (US all-patient).

The top five in effect size are annotated. **(A)** RR estimates generated by the Bayesian method. **(B)** RR estimates generated by the frequentist method.

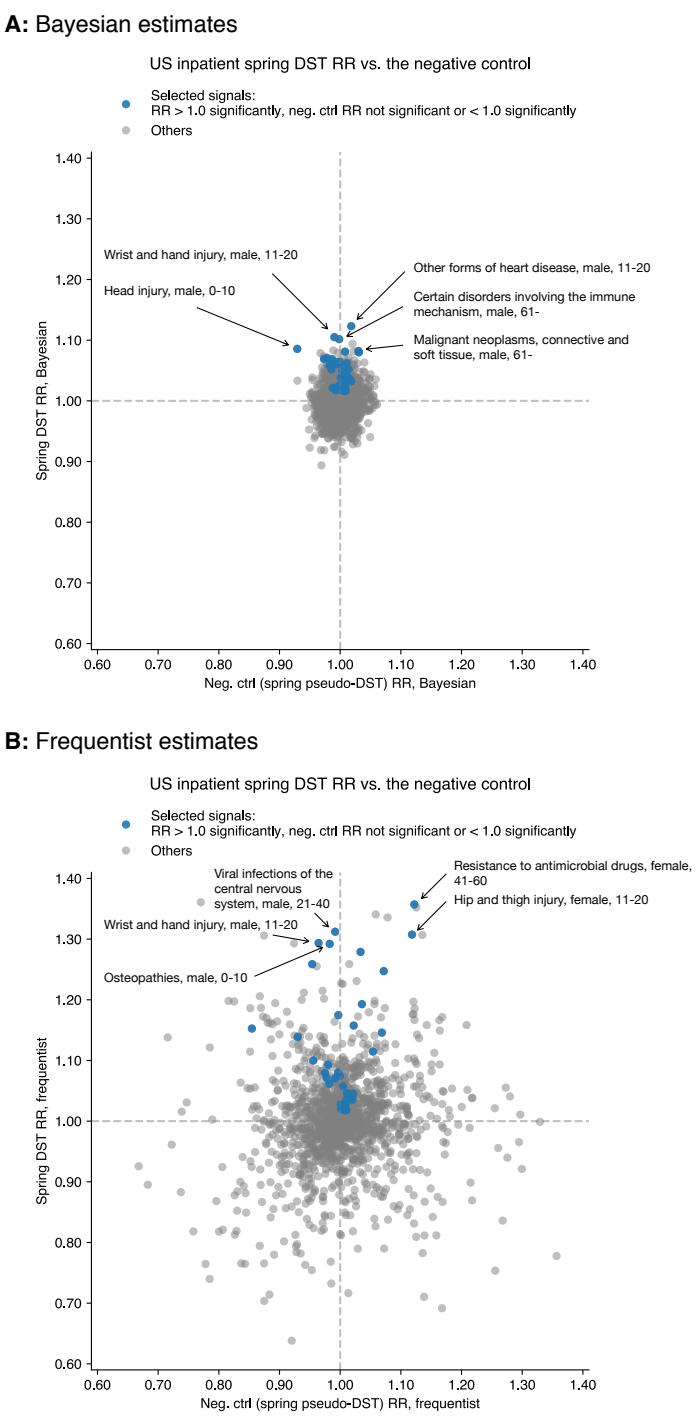

Figure 2.10 Selecting conditions with increased RR by comparing spring DST shift RRs to the negative control on pseudo-DST shift dates (US inpatient).

The top five in effect size are annotated. **(A)** RR estimates generated by the Bayesian method. **(B)** RR estimates generated by the frequentist method.

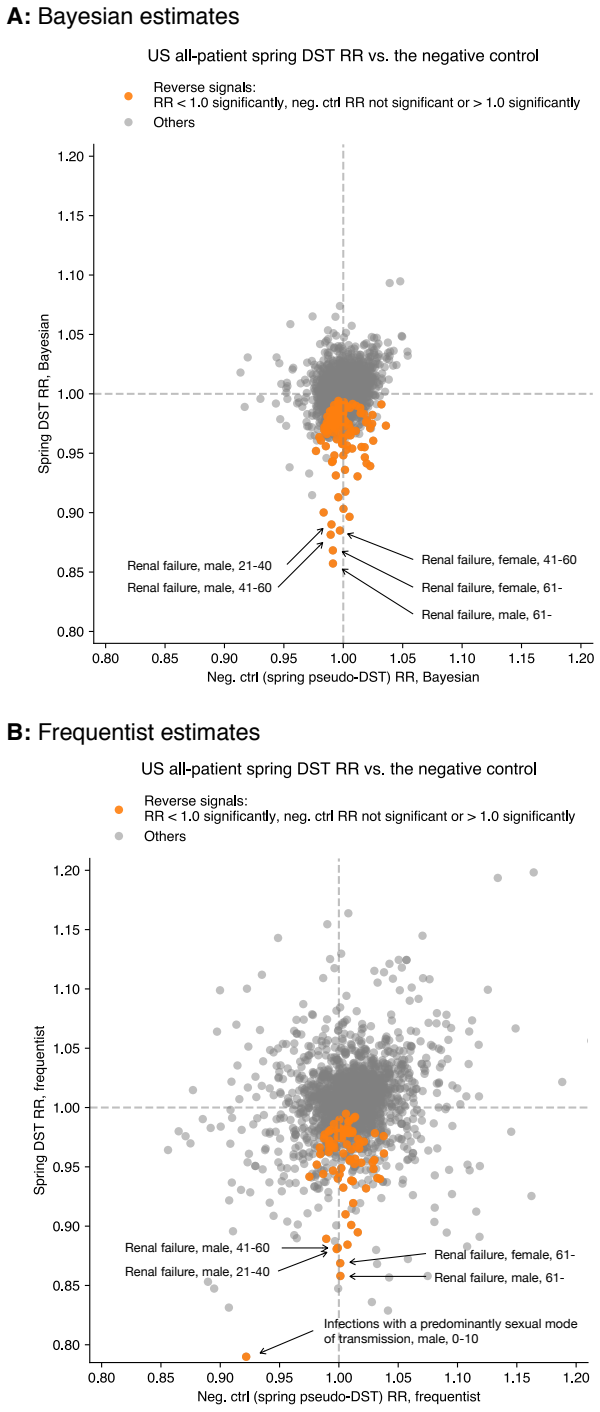

Figure 2.11 Selecting conditions with decreased RR by comparing spring DST shift RRs to the negative control on pseudo-DST shift dates (US all-patient).

The top five in effect size are annotated. **(A)** RR estimates generated by the Bayesian method. **(B)** RR estimates generated by the frequentist method.

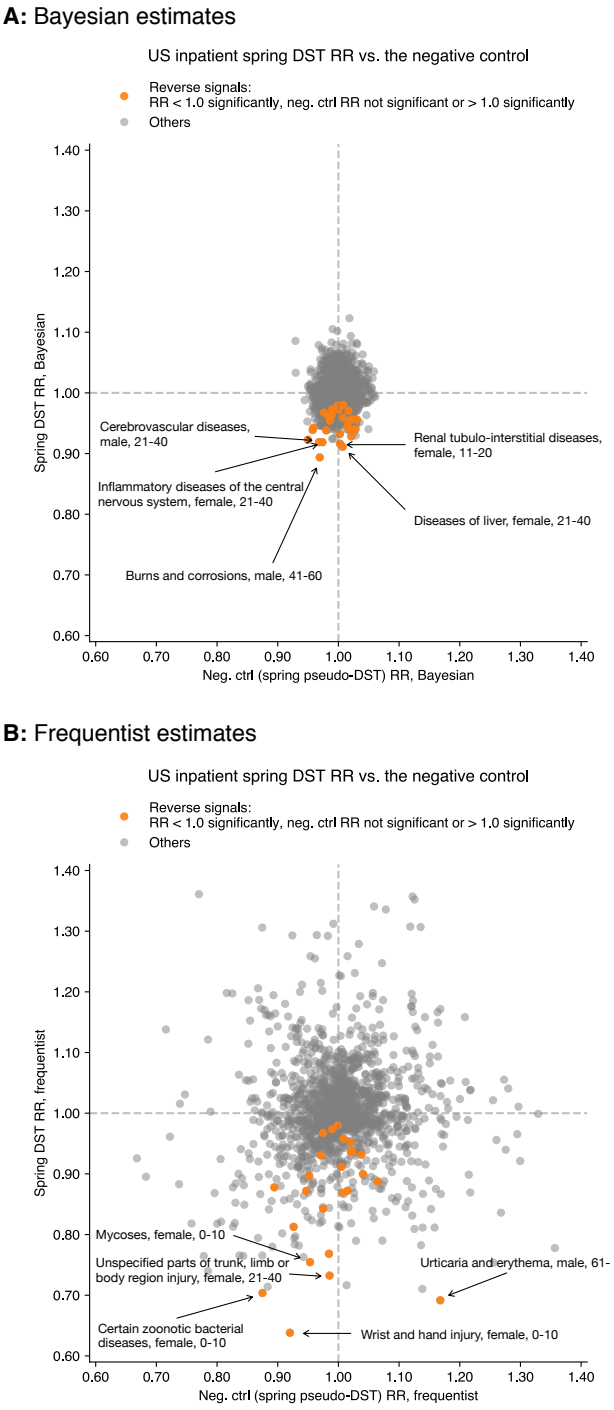

Figure 2.12 Selecting conditions with decreased RR by comparing spring DST shift RRs to the negative control on pseudo-DST shift dates (US inpatient).

The top five in effect size are annotated. **(A)** RR estimates generated by the Bayesian method. **(B)** RR estimates generated by the frequentist method.

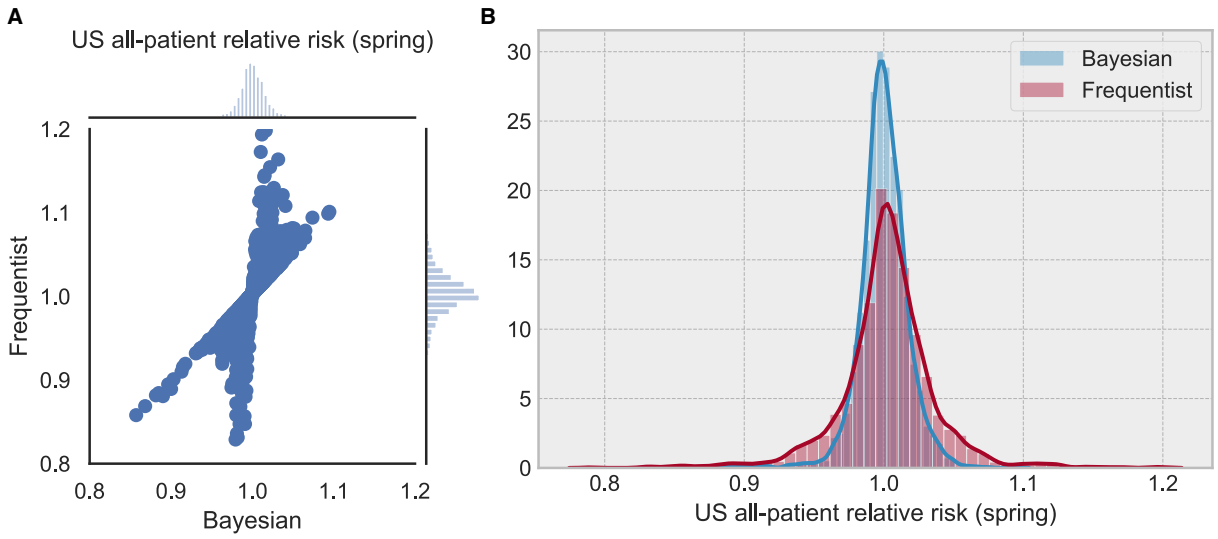

Figure 2.13 The Bayesian method shrinks the estimates to the prior mean when there is not enough information (US all-patient).

(A) A scatter-plot of frequentist versus Bayesian estimates. **(B)** Distributions of RR estimates.

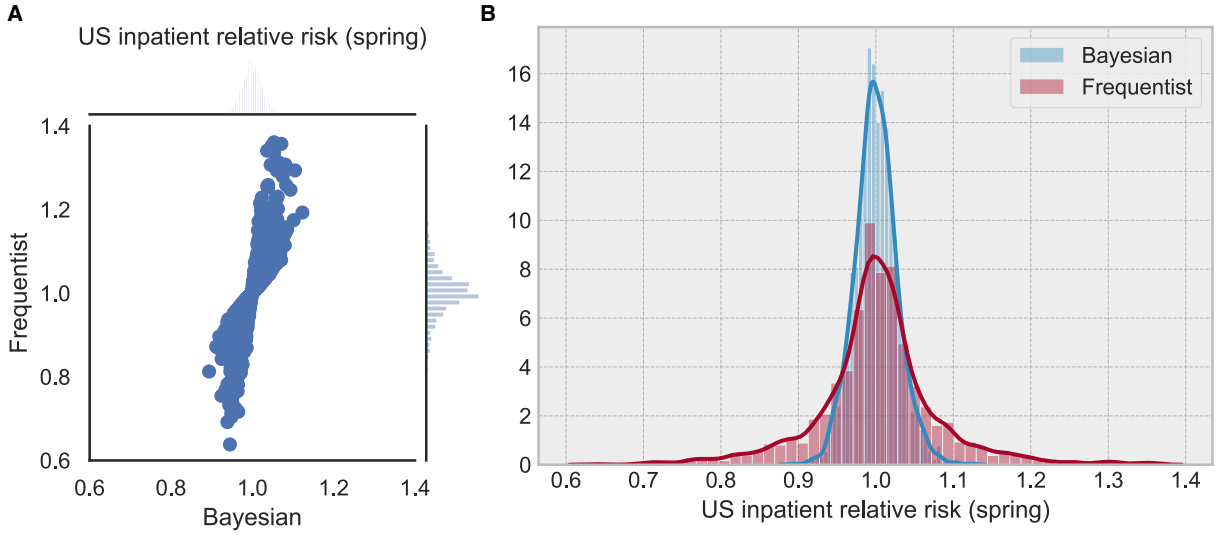

Figure 2.14 The Bayesian method shrinks the estimates to the prior mean when there is not enough information (US inpatient).

(A) A scatter-plot of frequentist versus Bayesian estimates. **(B)** Distributions of RR estimates.

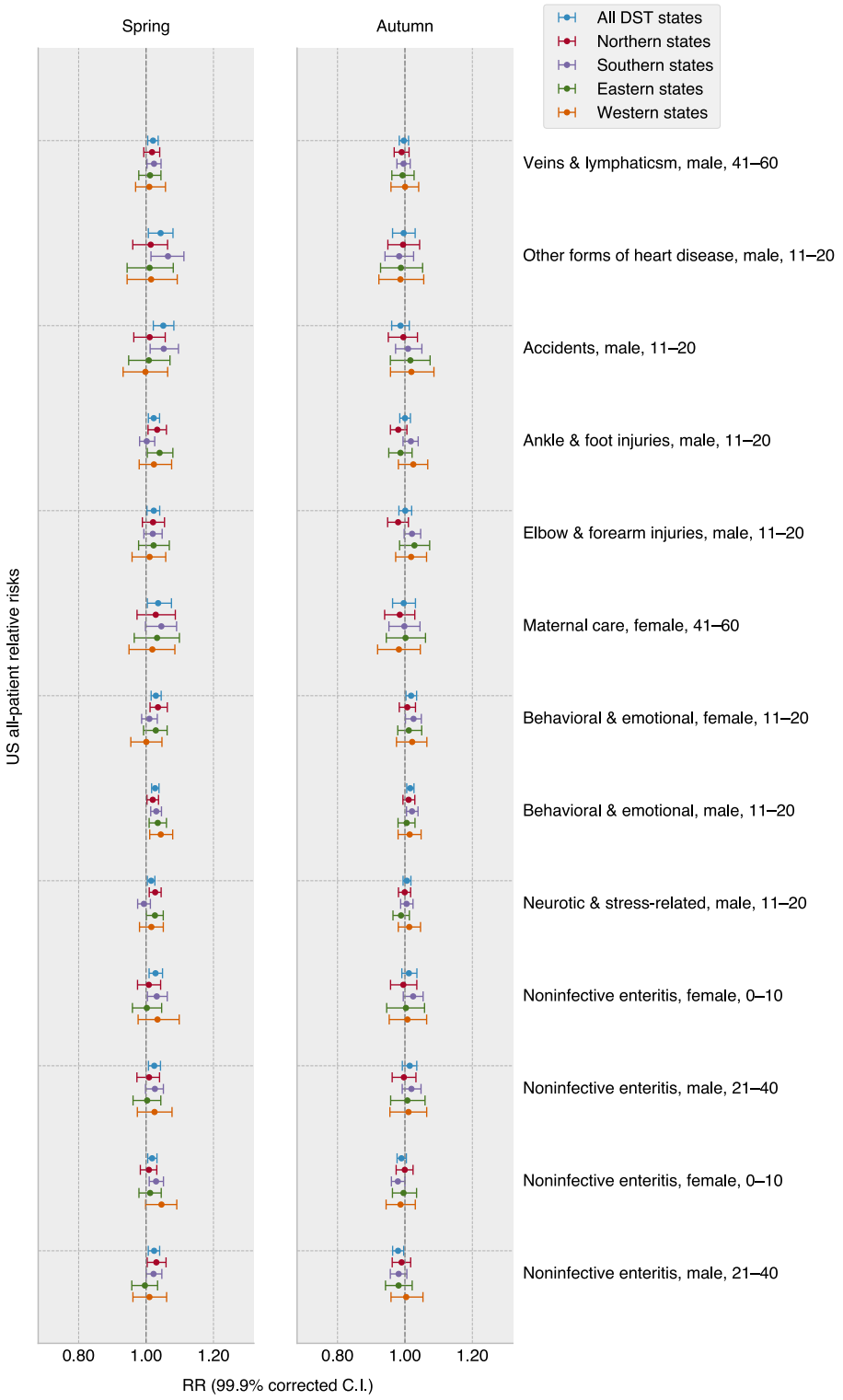

Figure 2.15 The geographic and cultural diversity of the DST shift's effects on health (US all-patient)

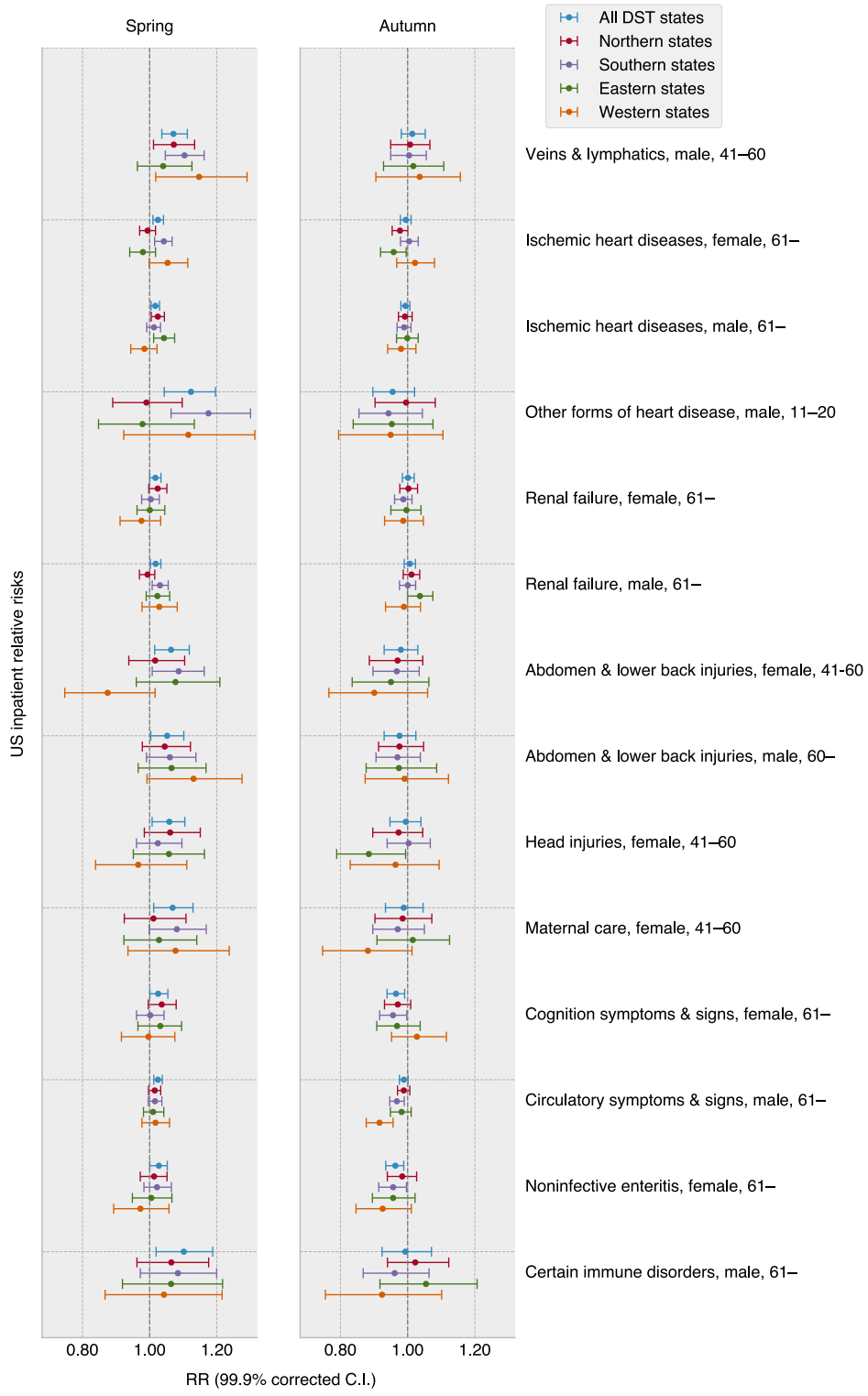

Figure 2.16 The geographic and cultural diversity of the DST shift's effects on health (US-inpatient)

CHAPTER 3. BAYESIAN GENERATIVE MODELS FOR PSYCHIATRIC DISEASES' SEASONALITY AND TREND

This chapter is adapted from the manuscript "**Probing annual disease incidence cycles in US and Sweden**" authored by Hanxin Zhang, Atif Khan, Qi Chen, Henrik Larsson, and Andrey Rzhetsky.

3.1 INTRODUCTION

Psychiatric illness induces profound suffering and profoundly affects the lives of patients and their loved ones. Psychiatric disorders are unique in the realm of complex diseases in that their diagnoses almost exclusively rely on outwardly subjective symptoms, presented by the patient, and interpreted by a psychiatrist. On the other hand, infectious and Mendelian diseases occupy space in the diagnostic continuum's highest-certainty extreme, which typically can be ascertained definitively via specialized experimental tests. Supporting this view of etiologic entanglement, psychiatric disorders appear to share extensive genetic and environmental predispositions. For example, whole-genome association data [73-76] analysis indicates that psychiatric disorders are highly genetically correlated, while large-scale, family-based studies, supporting these high genetic correlations across psychiatric maladies, also suggest that these disorders possess shared environmental risk factors [77, 78]. The estimated shared proportion of environmental risk factors between non-psychiatric complex diseases, and within psychiatric and non-psychiatric disease pairs, tend to be much lower [77, 78].

If psychiatric disorders share many environmental risk factors, it should be possible to identify common environmental stimuli affecting many [79] – or even all – of them. One of the potential environmental drivers of selected psychiatric conditions, such as seasonal affective disorder [80-83] and depression [84-89], is the annual and daily sunlight cycle, which drives the

circadian clock. Both seasonal affective disorder and depression tend to worsen during darker seasons. It is unclear whether this seasonal pattern is shared by other psychiatric disorders and whether this disease seasonality is solely limited to particular geographic areas. This study's main hypothesis is that the bulk of psychiatric maladies share this annual light-dependency cycle. Here, we systematically examine psychiatric conditions *vis-à-vis* their annual cycle of disorder-specific patient visits, as represented in clinical records, across very different geographic zones and two distinct continents, Europe and North America. For reference, we compare annual psychiatric disorders' reporting cycles with those for infectious disease, across American and Swedish populations.

The ideal data input required to answer our questions about disease seasonality would include records of direct, physician-led patient health state evaluations, following patients over many years, directly detecting their health state improvement or deterioration. Unfortunately, such data are yet to be generated. Instead, we used very large collections of electronic medical records, documenting patient's visits to medical practitioners, along with diagnoses, procedures, and prescribed medications. The latter type of data is subject to biases, such as weather events (think of blizzards), holidays, and vacations, all of which affect the behavior of both doctors and patients. To account for both these biases and for possible noise in data, we developed a family of statistical models, estimating the annual disease diagnosis rate's (DR) most likely seasonal oscillation pattern, while striving to account for data biases. We then tested these models against data to determine the best model that did not overfit observations.

Our study used the IBM Watson Health MarketScan data set [19] containing insurance claim records of over 150 millions of unique Americans, and the Swedish National Health Register [34] detailing the health dynamics of virtually all Swedes, with over eleven million

unique people visible in the data. The US data covers the time interval between 2003 and 2014, while the Swedish data encompasses an interval between 1980 and 2013. Although the IBM MarketScan database is one of the largest and most comprehensive collections of US insurance claims, it was built by merging asynchronous subsets collected by multiple private health insurers. As a result, the data has layers of idiosyncratic properties that complicated our analysis (Figure 3.1 US data characteristics and how they influenced our model design.Figure 3.1). To account for systematic biases and noise in data, we designed a multilevel Bayesian model describing generation of the observed disease-specific patient visit counts (see Figure 3.2 and the Complete details of materials and methods section). We present results from two distinct analysis approaches, the first with "uncorrected" counts of diseases-specific visits, and the second with "corrected" seasonality that we adjusted for seasonal changes in all-cause medical visits.

3.2 BAYESIAN MODEL SUMMARY

Figure 3.1B illustrates a typical disease's overall trend and seasonality, summarizing divergent linear trends of population strata (patient cohorts with the same entry and exit points within our database, represented by the grey lines). We calculated the real observation curve by dividing the total diagnoses by the total enrollees (diagnosis rate, DR) at each time point (in this study, by week). The holiday-smooth function uses the average DRs around known holidays to calculate and offset the sharp decrease shown around holidays and other celebrations have on the DR.

The grey lines (Figure 3.1) show the population strata's linear fit trends enrolled in the data asynchronously. For example, the olive curve represents a group of patients enrolled from week one to week 195. Different population strata do not show a uniform trend -- some grey lines go upward, some are flat, and some go downward. Figure 3.1B demonstrates that, due to

heterogeneous insurance enrolling practices, groups of people joining an insurer together do not resemble a random sample from the general US population. In addition to "asynchronous enrollment," we also found that many diseases' diagnosis rates suddenly shifted at the beginning of every year for many population strata of consistent composition. The sample population stratum could give us an idea of such shifts (see the olive curve on the right panel of Figure 3.1B).

We designed a multilevel Bayesian model to describe the generation of the observed DR counts, given several sources of systematic bias and noise (Figure 3.2). First, we grouped patients based on their ages and enrollment dates and defined "population strata," which are cohorts containing patients of the same age group and enrollment date in our data in the same time interval. We then modeled each population stratum's trend and seasonality separately, but not independently. We shared the information across age groups and population stratum because they were sampled from the same priors and hyper priors. For example, for the linear trend intercepts $\alpha_{i,j}$ for population stratum i, we sampled them from a skew-normal distribution with an age-specific center μ_j^{α}, scale σ_j^{α}, and shape h_j^{α} (Figure 3.2 Step 1). These age specific hyperparameters were also sampled from shared Gaussian process hyperpriors that chained them together across age groups so that close ages would have close center μ_j^{α}, scale σ_j^{α}, and shape h_j^{α} (see Section 3.5 Complete details of materials and methods for more information).

We estimated all parameters simultaneously using a Markov chain Monte Carlo (MCMC) sampler [90]. After obtaining all the estimates for every population stratum, we can merge strata and find the age or sex-specific trend and seasonality, as shown in the top panel of Figure 3.2 Step 2. We highlighted a yearly seasonality sample in the bottom panel of Figure 3.2 Step 2. Figure 3.2 Step 2's lower left plot gives the relative seasonal fluctuation of a sample

disease, computed by dividing the raw seasonality estimate by the time-average of observed DR (see Expressions (17) and (18) of Section 3.5 Complete details of materials and methods).

In an attempt to account for season estimates' possible non-biology-driven fluctuations (vacations, bad weather, holidays), we attempted normalizing the raw DR counts using the DR of all-medical visits (shown in the lower-center panel of Figure 3.2 Step 2 and Figure 3.3). The resulting corrected seasonality then represented the count excess/deficit with respect to the baseline medical diagnoses fluctuation (the lower right panel of Figure 3.2 Step 2). In the present work, we refer to the uncorrected seasonality relative to the time-average DR as "uncorrected" seasonality or "$s(t)$". We refer to the seasonality corrected by the all-medical visits baseline as "corrected" seasonality or "$s'(t)$".

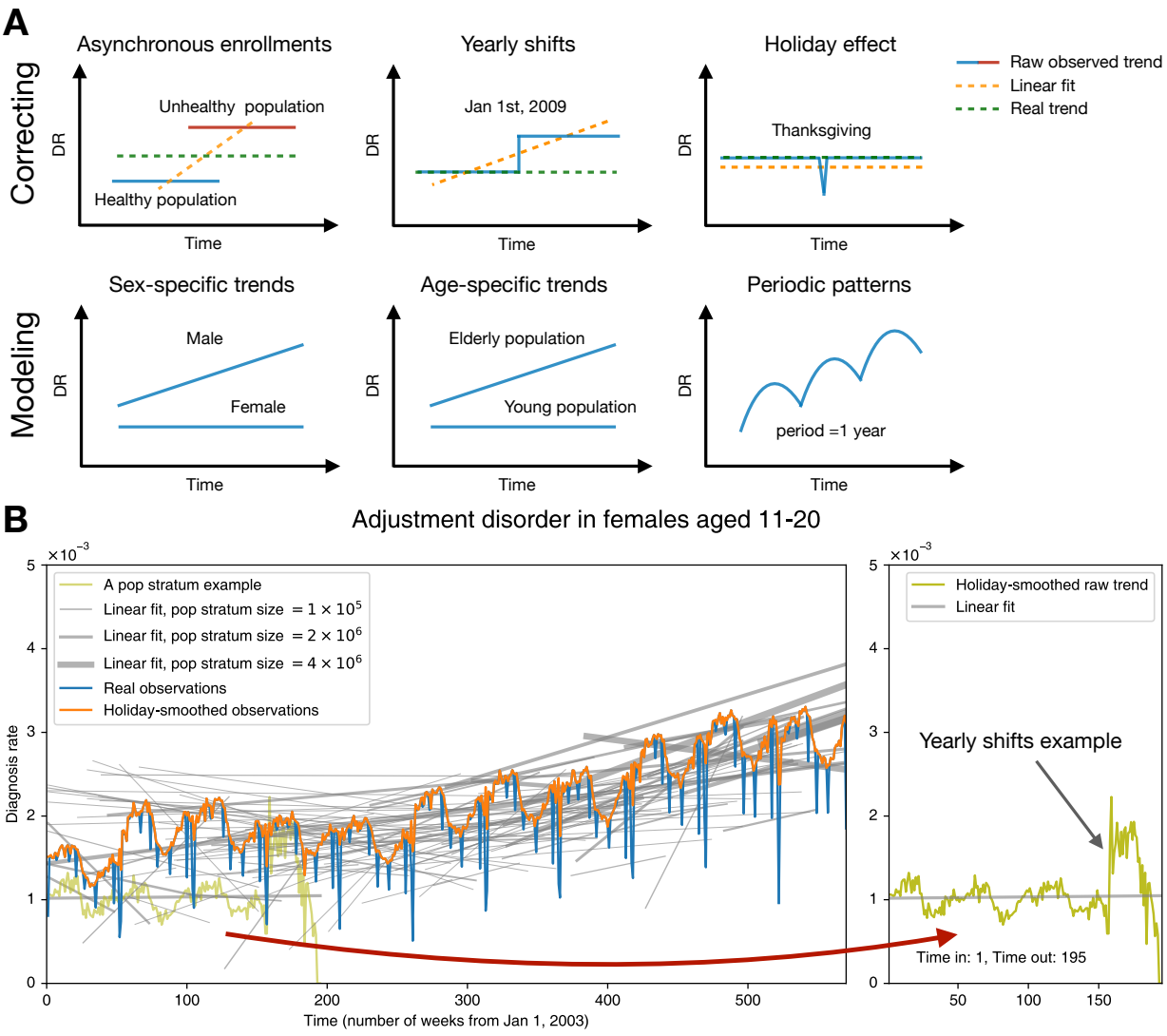

Figure 3.1 US data characteristics and how they influenced our model design.

(A) Our modeling aims to correct biases and noise in the MarketScan® database and to infer a latent disease diagnosis rate (DR) trend and seasonality for specific age-sex groups. **(B)** This subplot shows the overall trend and seasonality for a sample disease. The holiday-smooth function offsets the effect of holidays and celebrations that decrease the DR sharply (the orange curve vs the blue curve). Bear in mind that patients joined and left our US data asynchronously. The gray lines illustrate varying linear fit trends of population strata, defined according to their enrollment dates (see the Methods and techniques part of the Complete details of materials and models section). Some population strata include more people, while others are smaller in size, as marked by different widths of gray lines. A sample population stratum enrolled from week one to week 195 is highlighted in the right panel. Notice that the sudden shift still exists – even for a population with a consistent composition, meaning that the shifts do not result from enrollment changes.

3.3 RESULTS

We applied our statistical models to probe the annual seasonality of 33 psychiatric and 47 infectious diseases in two sexes and multiple age groups. For simplicity of visualization and discussion, we used the meteorological season conventions, defined as follows: winter starts on December 1st and ends on February 28th or 29th, spring starts on March 1st and ends on May 31st, summer starts on June 1st and ends on August 31st, and autumn is the rest of the year. In this description, we focus on the results for the five most prevalent psychiatric disorders and the five most common infectious diseases, but the results for all the diseases studied, using both corrected and uncorrected seasonality analyses, are available on the project repository at https://github.com/hanxinzhang/seasonality.

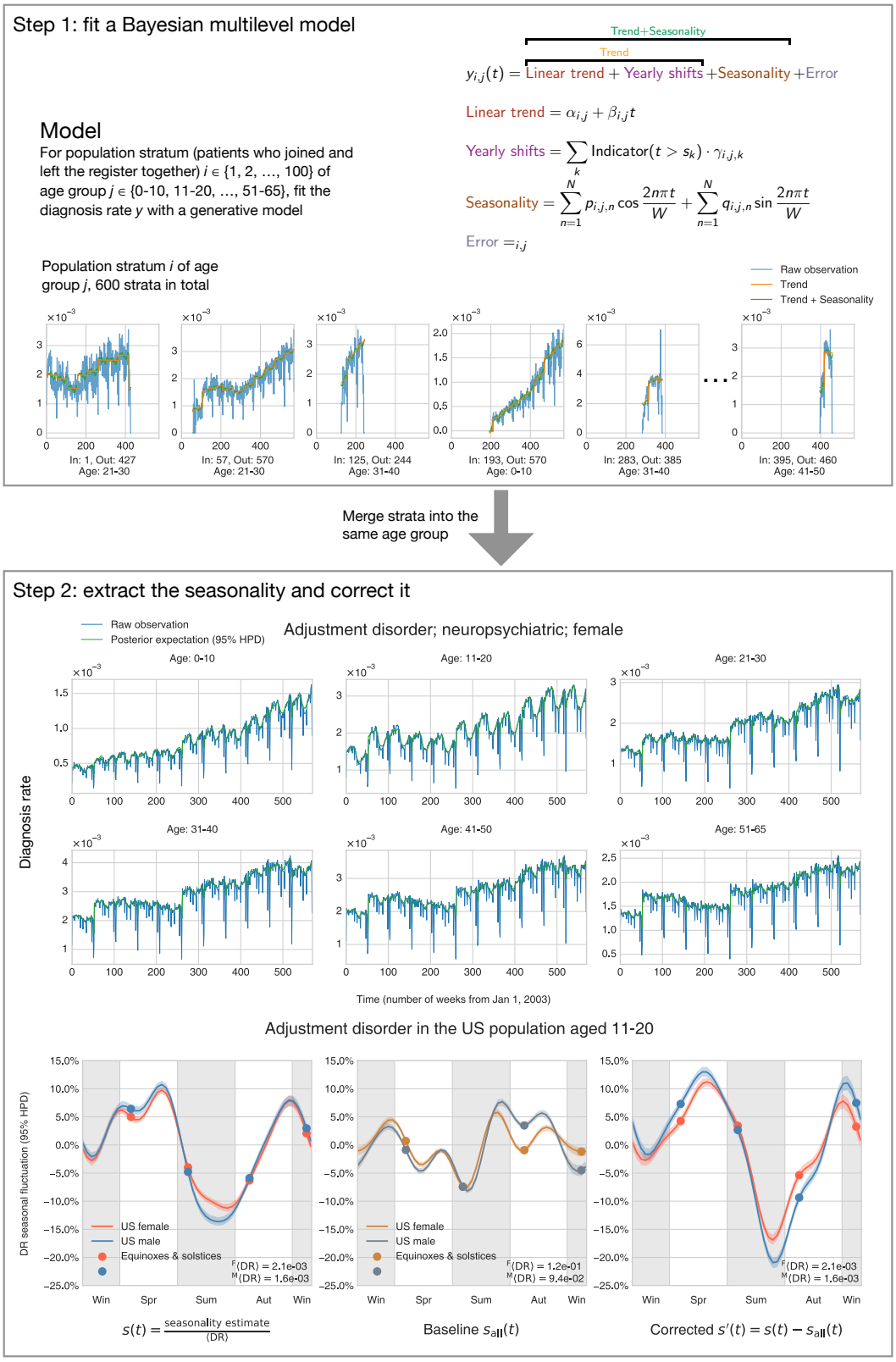

Figure 3.2 The method and procedure to infer seasonality.

Figure 3.2 Continued: Step 1 (Upper Frame): We modeled the diagnosis rate by decomposing it into several parts: the linear trend, yearly shifts, seasonality, and an error term. We assigned people into hundreds of population strata according to their enrollment dates. Six populations of specific age and enrollment dates are shown here. The model fits each population strata separately (but NOT independently), with shared priors and hyperpriors so that information can be shared across populations. **Step 2 (Lower Frame):** After obtaining the estimates of all model parameters, we extracted the seasonality to make inferences. The upper plot shows that the posterior expectation (mean) reproduces our raw observation very well, which partly validates that our model is mixing well. Note that for this particular condition, the 95 percent highest posterior density interval is very small, so it may be difficult to discern in the plot (light green shade). The left subplot at the bottom exemplifies how we can find the relative seasonal fluctuation (uncorrected) $s(t)$ by dividing the seasonality estimates by the time-average DR ($\langle \mathrm{DR} \rangle$, Expression (3-17)). We can possibly correct for the baseline fluctuation of all medical visits by deducting the $s_{\mathrm{all}}(t)$ (representing the uncorrected seasonality of all medical visits) from $s(t)$ and obtain the corrected seasonality $s'(t)$ (right subplot at the bottom).

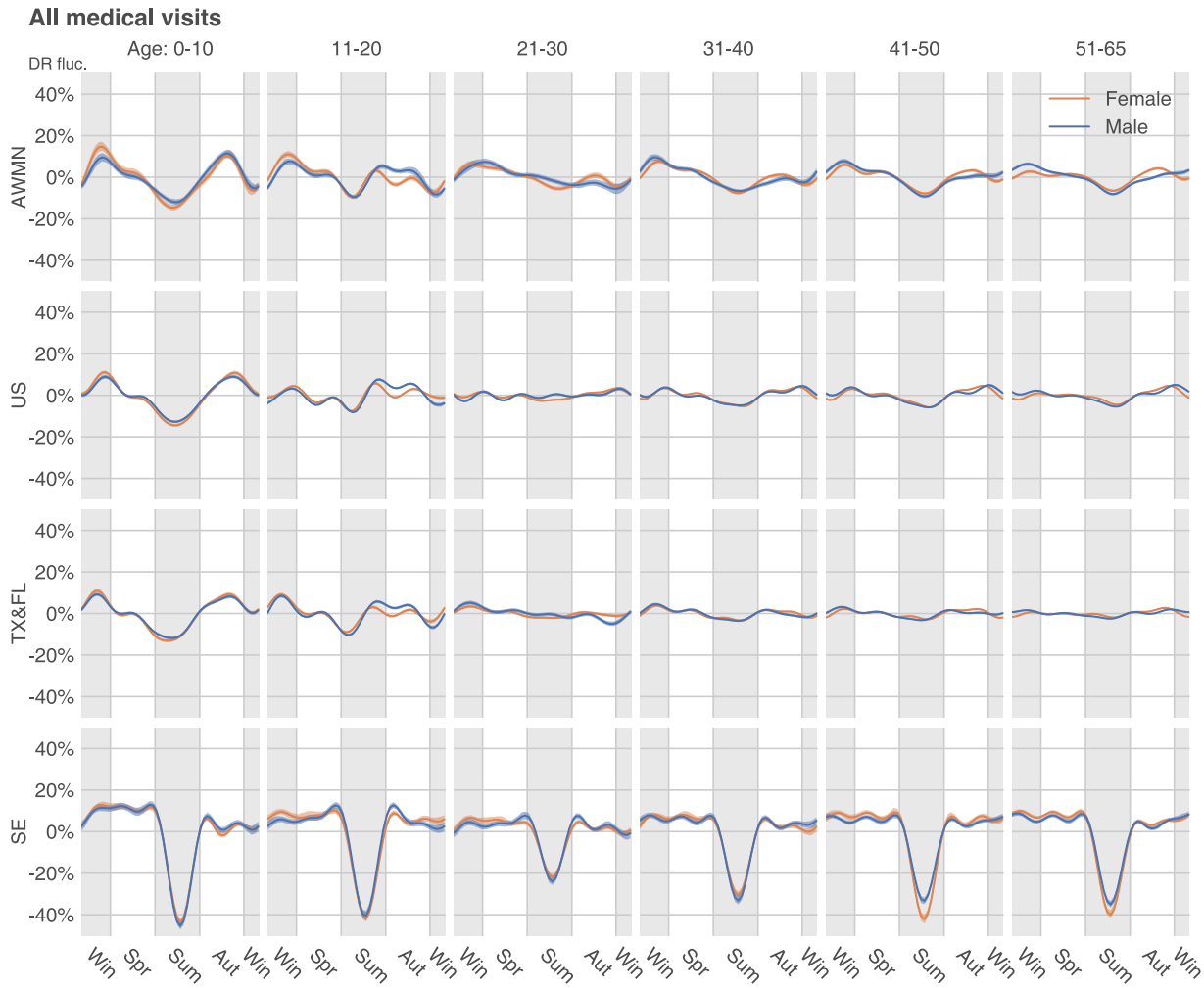

Figure 3.3 The baseline seasonality of all medical visits in the four northern US states (AK, WA, MT, ND, or AWMN), the whole US, two southern US states (TX and FL), and Sweden (SE)

3.3.1 Uncorrected seasonality analysis

We first analyzed diseases' uncorrected seasonality without considering the underlying baseline fluctuation of all medical visits. Psychiatric disorders appear to follow a nearly-identical yearly cycle of care-access patterns; on average, they spike in the darker periods and recede during warmer and brighter times (see Figure 3.4 for uncorrected seasonality), although all

patterns were exceedingly more complicated than a unimodal curve. Figure 3.4 shows disorder-, sex-, and age-group-specific seasonalities for the five most diagnosed psychiatric conditions in the US: depression, anxiety/phobic disorder, adjustment disorder, substance abuse, and attention deficit/hyperactivity disorder (ADHD). We show matching results for Sweden, time-aligned with their US counterparts, but scaled by 0.3 in magnitude for ease of comparison. Clearly, despite significant differences in social, economic, cultural, and healthcare management conditions in the two countries, the curves are surprisingly similar across both countries for the same condition and highly consistent across disorders. The plots are designed to show deviation from the yearly mean value in percent of disorder-specific visits at given time points. A uniform pattern of visits through the year would result in a flat line at zero percent. In the plots, we see around ten to 20 percent fluctuation relative to the yearly mean in the US. Seasonal fluctuations in Sweden are even larger, reaching a 70 percent decrease in patient visits related to, for example, ADHD. For all five diseases, especially in the US, people younger than 20 seem to experience larger-scale seasonal variation psychiatric visit frequency. In Sweden, however, the difference in seasonal variation across age groups is minor. In terms of the discrepancy between the two sexes, females and males bear analogous seasonality in both countries. It is worth mentioning that ADHD in people older than 20 demonstrates an out-of-the-ordinary seasonality that rises gradually from autumn to winter.

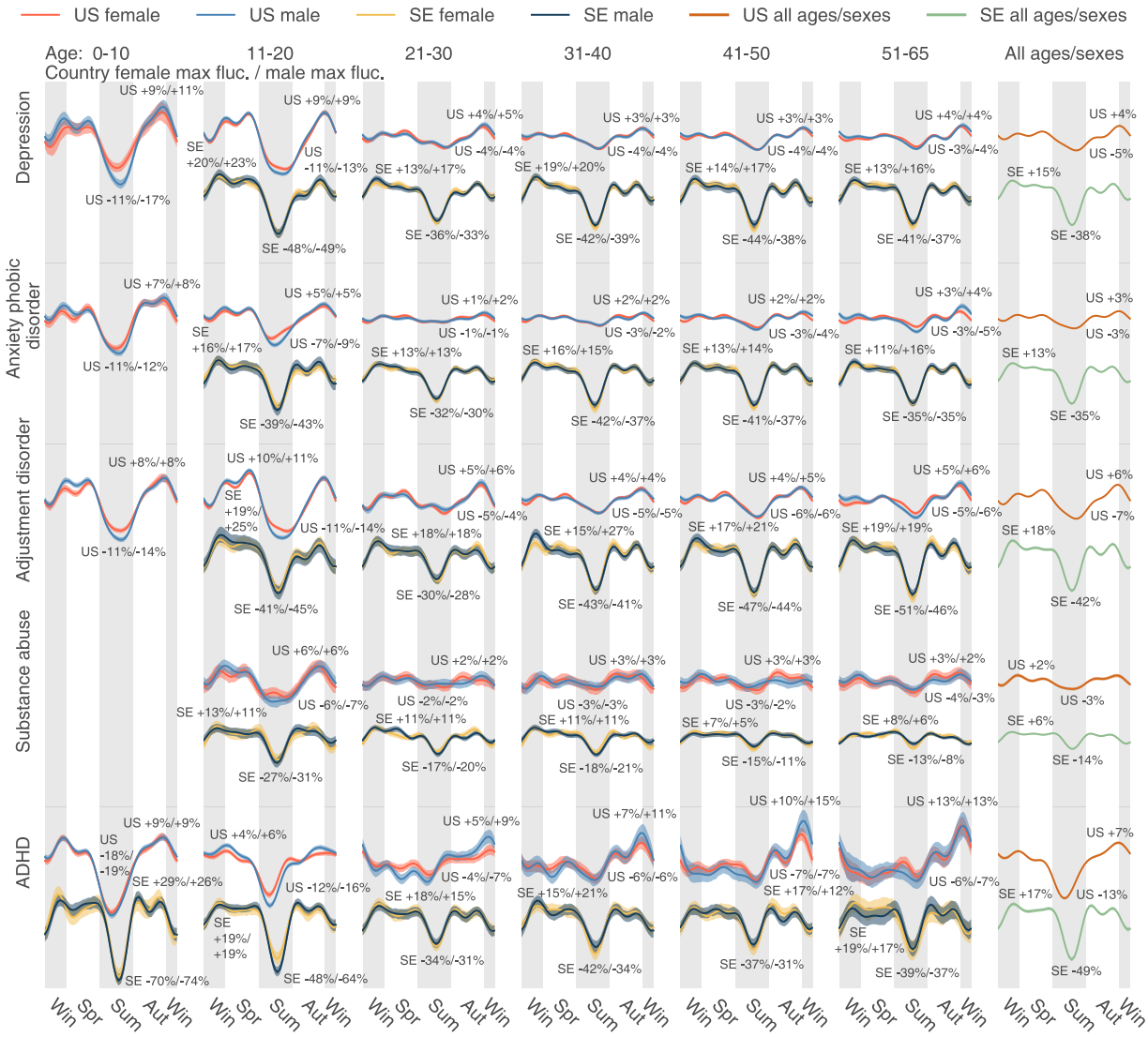

Figure 3.4 The uncorrected seasonality plots of the five most-diagnosed psychiatric diseases in the US: depression, anxiety and phobic disorder, adjustment disorder, substance abuse, and attention-deficit/hyperactivity disorder (ADHD).

Figure 3.4 Continued: The results in Sweden (SE) are juxtaposed but scaled by 0.3 in magnitude for clearer comparison. We plotted all lines based on a weekly diagnosis rate estimated as the total number of diagnoses in a week, divided by the total number of enrollees in our database in the week. Positive and negative maximum fluctuations compared to the mean diagnosis rate are text-labeled following a format: Country Female Maximum Fluctuation in Percentage / Male Max Fluctuation in Percentage. We use the meteorological seasons defined as follows: Winter starts from December 1st and ends on February 28th, spring starts from March 1st and ends on May 31st , summer starts from June 1st and ends on August 31st, and autumn is the rest of the year. We discarded the health records of people over 65 because the majority population of the US data switched to Medicare and the remaining data was not representative. A disease could be extremely rare in some age-sex brackets. The plot only shows those with age-sex-specific seasonality that showed a time-average DR (Expression (3-17)) larger than 1×10^{-5}.

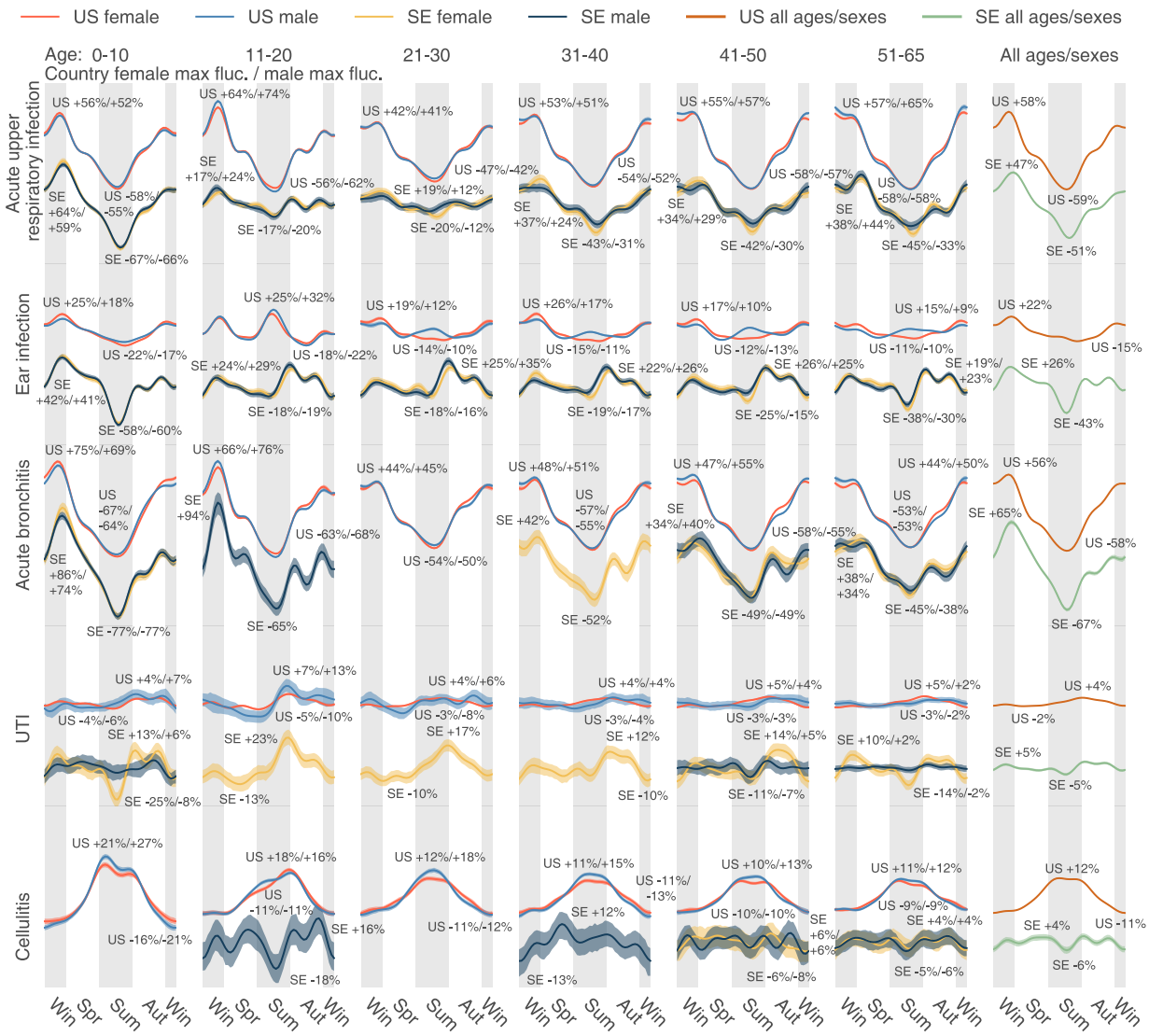

Figure 3.5 The uncorrected seasonality plots of the five most-diagnosed infectious diseases in the US: acute upper respiratory infection, ear infection, acute bronchitis, urinary tract infection, and cellulitis.

The results in Sweden (SE) are juxtaposed without scaling. A disease could be extremely rare in some age-sex brackets. The plot only shows those with age-sex-specific seasonality that showed a time-average DR larger than 1×10^{-5}.

Looking only at psychiatric disorder results, one might conjecture that the observed annual regularities are common for all diseases, and that the cycle dynamics are mainly driven by social factors. This is far from being true, as shown in the annual infectious disease cycles

(Figure 3.5 and Figure 3.6). A low-dimensional embedding of estimated seasonality harmonics using the `Isomap` algorithm [91] (Figure 3.6, https://seasonality-web-app.herokuapp.com) reveals that psychiatric curve shapes are tightly clustered (similar) while curve shapes for infectious diseases are very diverse and therefore scattered in the embedding representation.

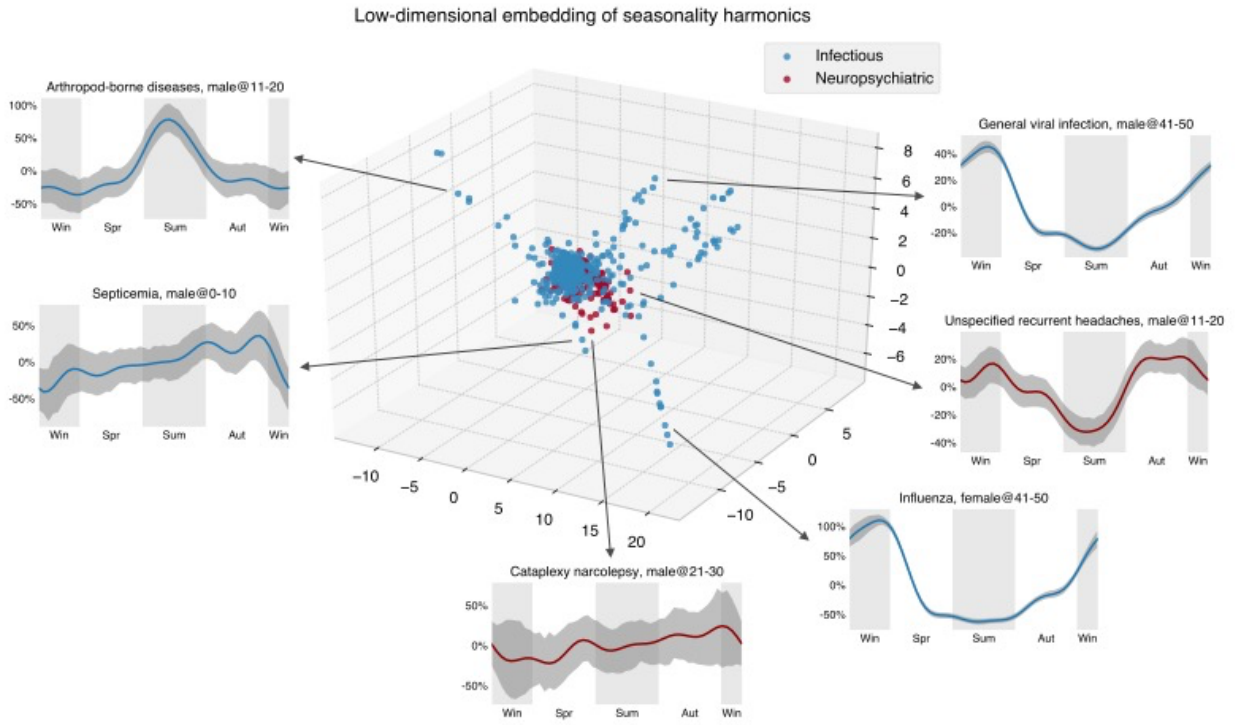

Figure 3.6 Embedding of uncorrected seasonality curves in a low-dimensional space suggests the homogeneity of the psychiatric diseases' seasonal variation.

We used the `Isomap` method to obtain a low-dimensional embedding of seasonality of the first ten Fourier harmonic base estimates $\overline{p}_{j,1}$, $\overline{p}_{j,2}$, …, $\overline{p}_{j,5}$, $\overline{q}_{j,1}$, $\overline{q}_{j,2}$, …, $\overline{q}_{j,5}$ (see Expressions (3-15) and (3-16)). Compared to the infectious diseases, we can see that the embeddings of psychiatric disease harmonics concentrate in a smaller space, implying the relative homogeneity of their seasonality.

If we consider the five most diagnosed infectious diseases in the US (acute upper respiratory infection, ear infection, acute bronchitis, urinary tract infection (UTI), and cellulitis, see Figure 3.5), the patterns are very different. The magnitudes of seasonal variation are comparable between the US and Sweden for infectious diseases, so the curves are scaled in the

same way. As expected, in the US diagnoses, the two respiratory infections (acute upper respiratory infection and bronchitis) rise in colder times, peaking in the early spring, and subside in warmer days, with the lowest rate at the end of summer. On the contrary, cellulitis, a deep skin infection, rises in warmer periods and subsides in the winter in the US – similar to general skin infections (Figure 3.7 and Figure 3.8). In Sweden, cellulitis is extremely rare in children and young females; in males and older adults (over 40-years old) it shows no obvious patterns, possibly because cases of this disease are sparse in this northern and relatively small country.

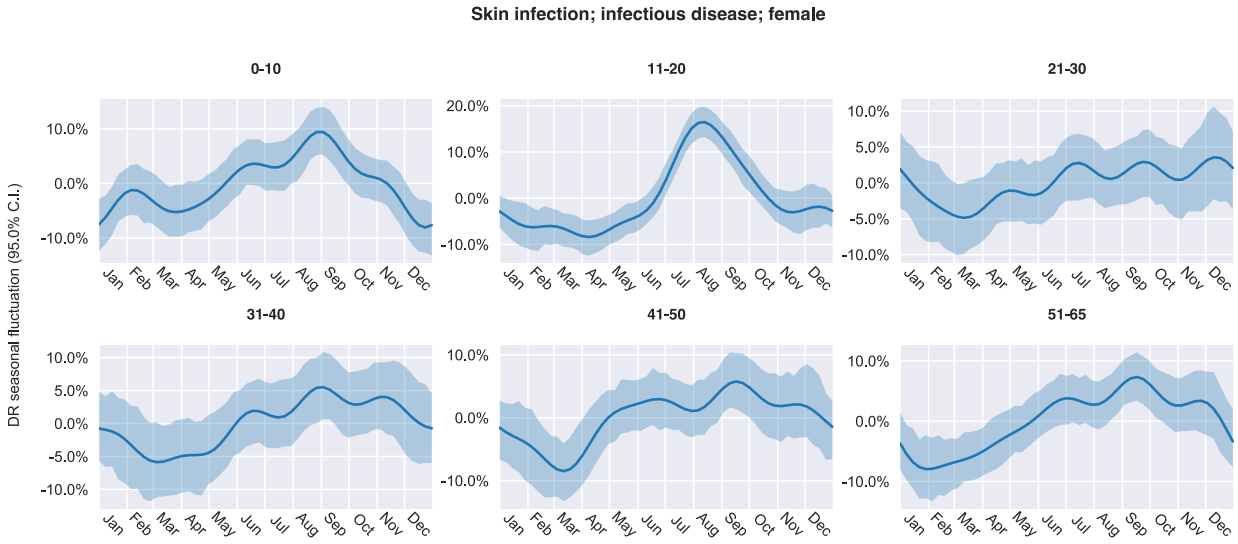

Figure 3.7 The uncorrected seasonality of skin infection in the US females.

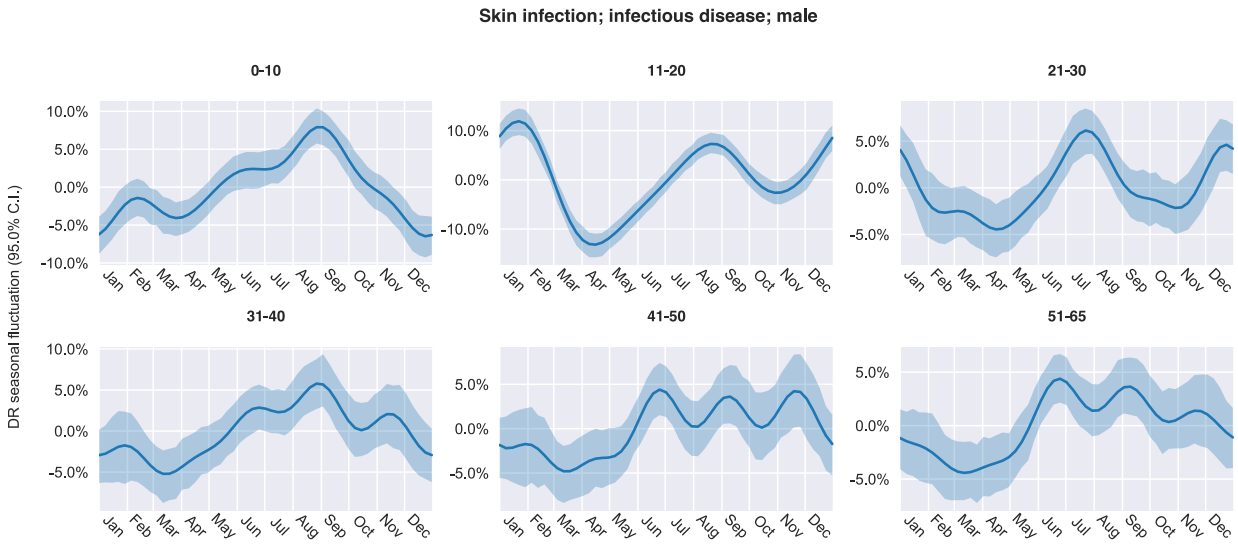

Figure 3.8 The uncorrected seasonality of skin infection in US males.

Ear infections in children (newborns to ten years old) are more common in winter and less common in summer in both countries, as expected. Unlike psychiatric disorder trends, we discern a distinct "peak triplet" pattern of ear infection in US teenagers (eleven to twenty years old) with high diagnosis rates in both the summer and winter and low diagnosis rates in spring and autumn. This "triplet" pattern extends to older US age groups and is visible with considerable variation in the Swedish cohort. Finally, UTI seasonality in the US tends to be level except in teenagers, but it grows from summer to autumn and goes down in winter and spring in Sweden, particularly in females aged between eleven and 40 years.

We conducted an additional analysis over the MarketScan® data to probe the seasonal variation differences among higher- and lower-latitude geographic regions. First, we conducted separate analyses using data exclusively representing the four high-latitude states in the US: Alaska, Washington, Montana, and North Dakota (AK, WA, MT, and ND, respectively). We did not include Maine due to its relatively lower latitude (Portland, ME 43.7° N versus Seattle, WA

47.6° N) as compared to the selected four states. We found that all five most prevalent psychiatric diseases demonstrate larger seasonal oscillation in the four high-latitude states (AK, WA, MT, and ND, or AWMN) than in the whole country (Figure 3.9). For example, in the summer, depression goes down about 23 percent in females aged eleven to twenty in each of the four states, contrasted to an eleven percent decrease for the country on average. In eleven- to twenty-year old males, ADHD decreases by 16 percent in the whole country, but 24 percent in the four high-latitude states. In general, the fluctuation magnitude is around 1.5 to two times larger in AK, WA, MT, and ND, but it is still much smaller than the variation in Sweden, which is at an even more northern latitude. Second, we observed that, for infectious diseases, the magnitude of seasonal variation was similar between the whole country and the four high-latitude states (Figure 3.10). We then examined two large states in the South: Texas and Florida (TX and FL). We did not include other southern states, such as Hawaii or Louisiana, due to the smaller population size represented in our data. Louisiana and other continental southern states are also not as south in latitude as Texas and Florida. Likewise, we did not consider California because a large part of it spans more northern areas. For both psychiatric diseases or infections, the results are similar to those of the whole US (Figure 3.11 and Figure 3.12). It is remarkable that for psychiatric conditions such as ADHD in males aged zero to ten and eleven to 20, the variation in TX and FL is smaller. We saw this seemingly smaller-variation tendency in other psychiatric diseases as well, but it is not as significant as the comparison between the US and Sweden or between the US and the four high-latitude states.

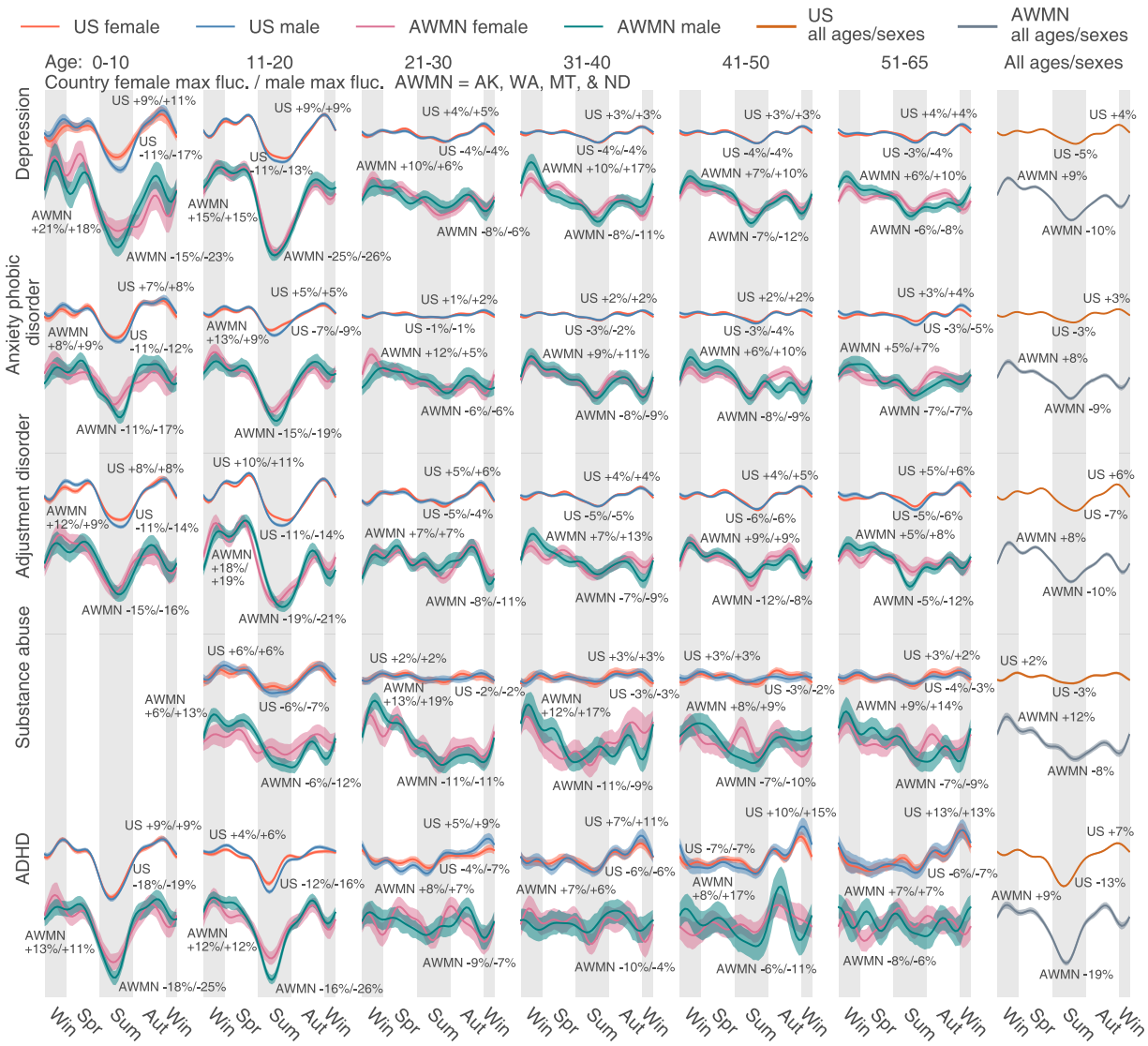

Figure 3.9 The uncorrected seasonality of psychiatric diseases in the four high-latitude states: Alaska, Washington, Montana, and North Dakota (AK, WA, MT, and ND).

To summarize, we observed a consistent, seasonal pattern in psychiatric diseases, with a shared recess in the summer, as well as a shared increase in the fall, in both the US and Sweden (Figure 3.4). Diverging from the conclusions of an earlier, smaller-scale study, which found only limited seasonal changes in general mental disorders [92], we observed that seasonality is shared by a large number of psychiatric disorders – in spite of their diverse symptomatology and

prevalence. In addition to observing the above-mentioned seasonality in depression, anxiety, and adjustment disorders (Figure 3.4), we detected similar patterns in many other psychiatric disorders such as schizophrenia and related psychoses (Figure 3.13–Figure 3.16) and migraine (Figure 3.17–Figure 3.20). By contrast, we found heterogenous seasonality patterns across infectious diseases.

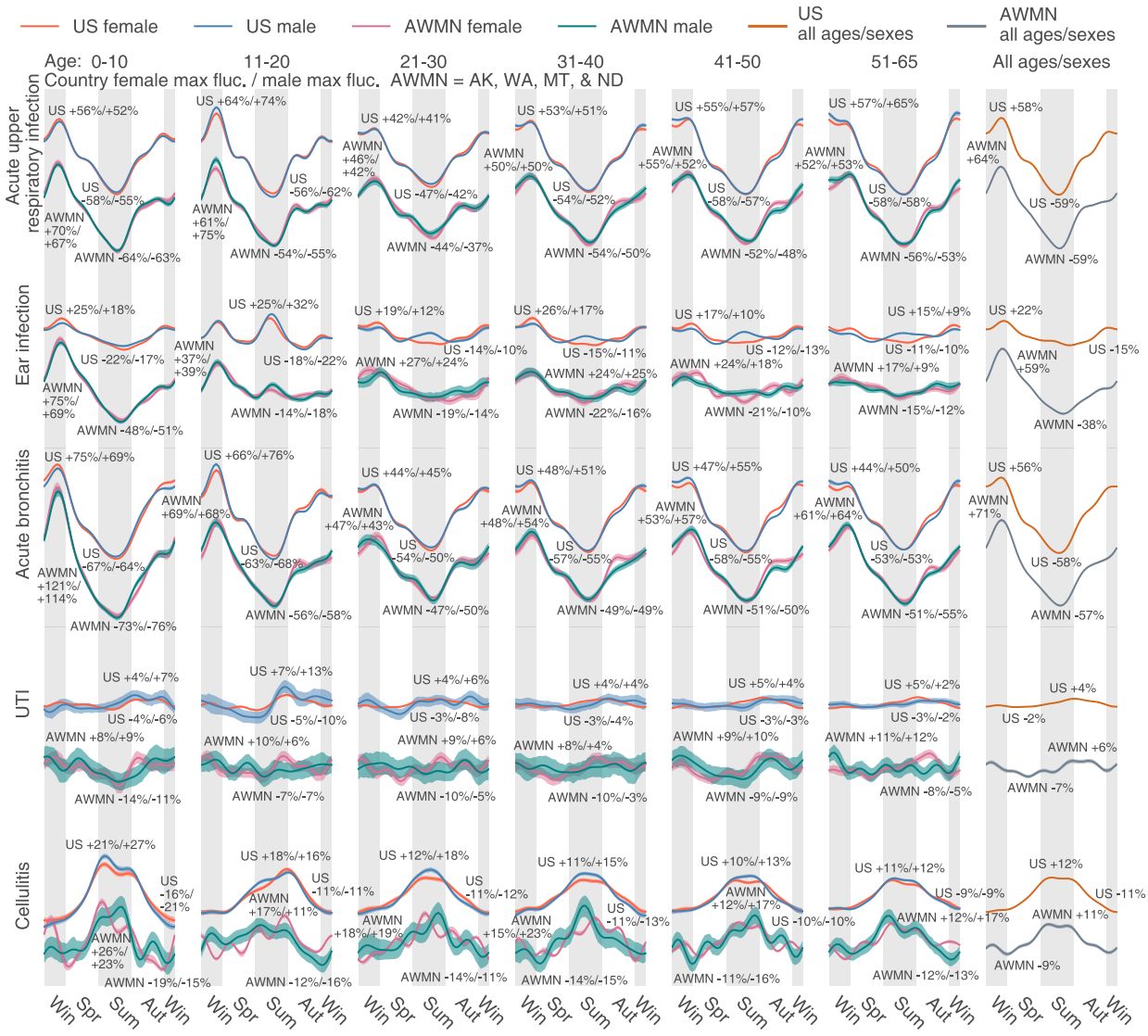

Figure 3.10 The uncorrected seasonality of infectious diseases in the four high-latitude states: Alaska, Washington, Montana, and North Dakota (AK, WA, MT, and ND).

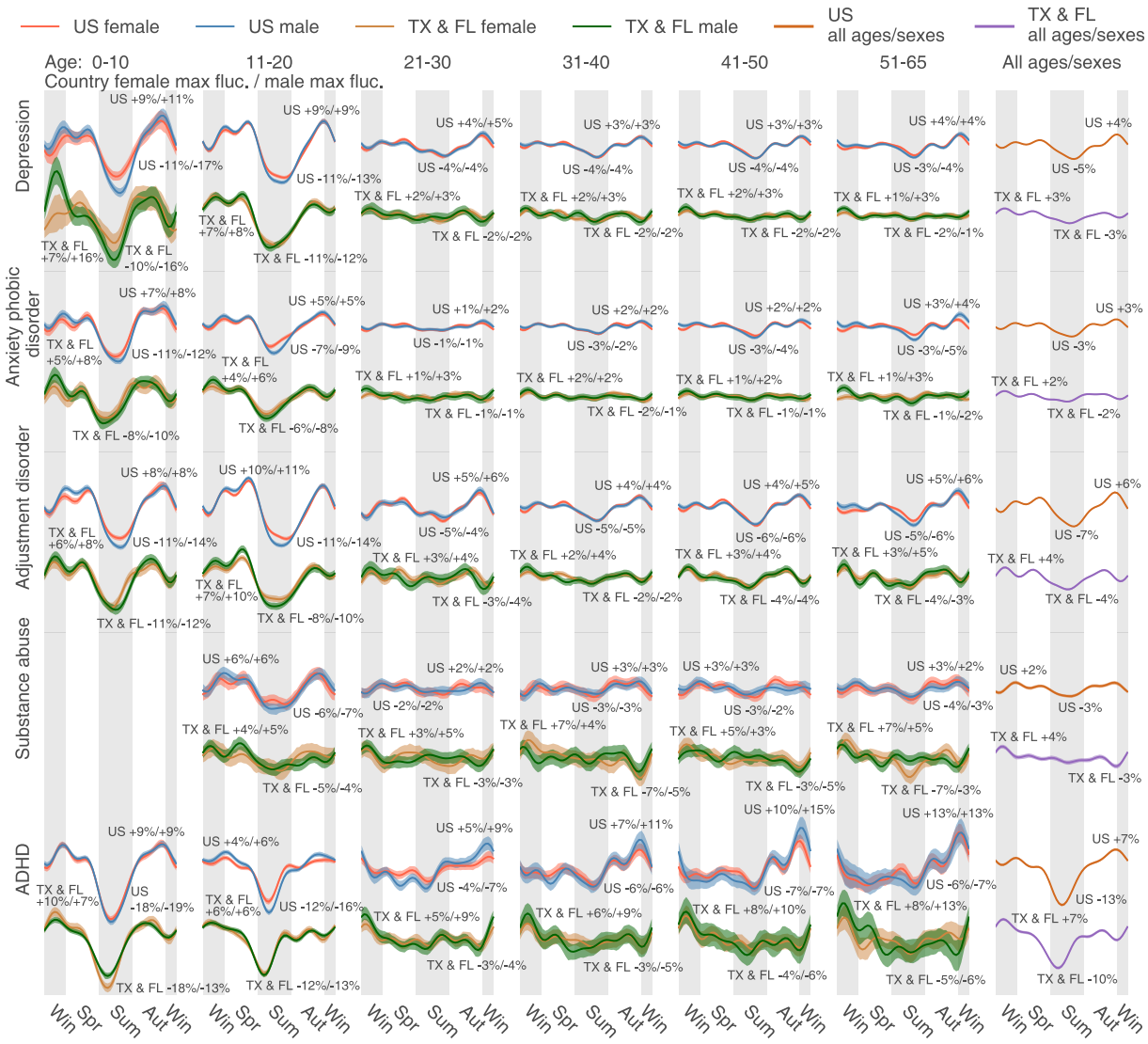

Figure 3.11 The uncorrected seasonality of psychiatric diseases in the two low-latitude states: Texas and Florida (TX and FL).

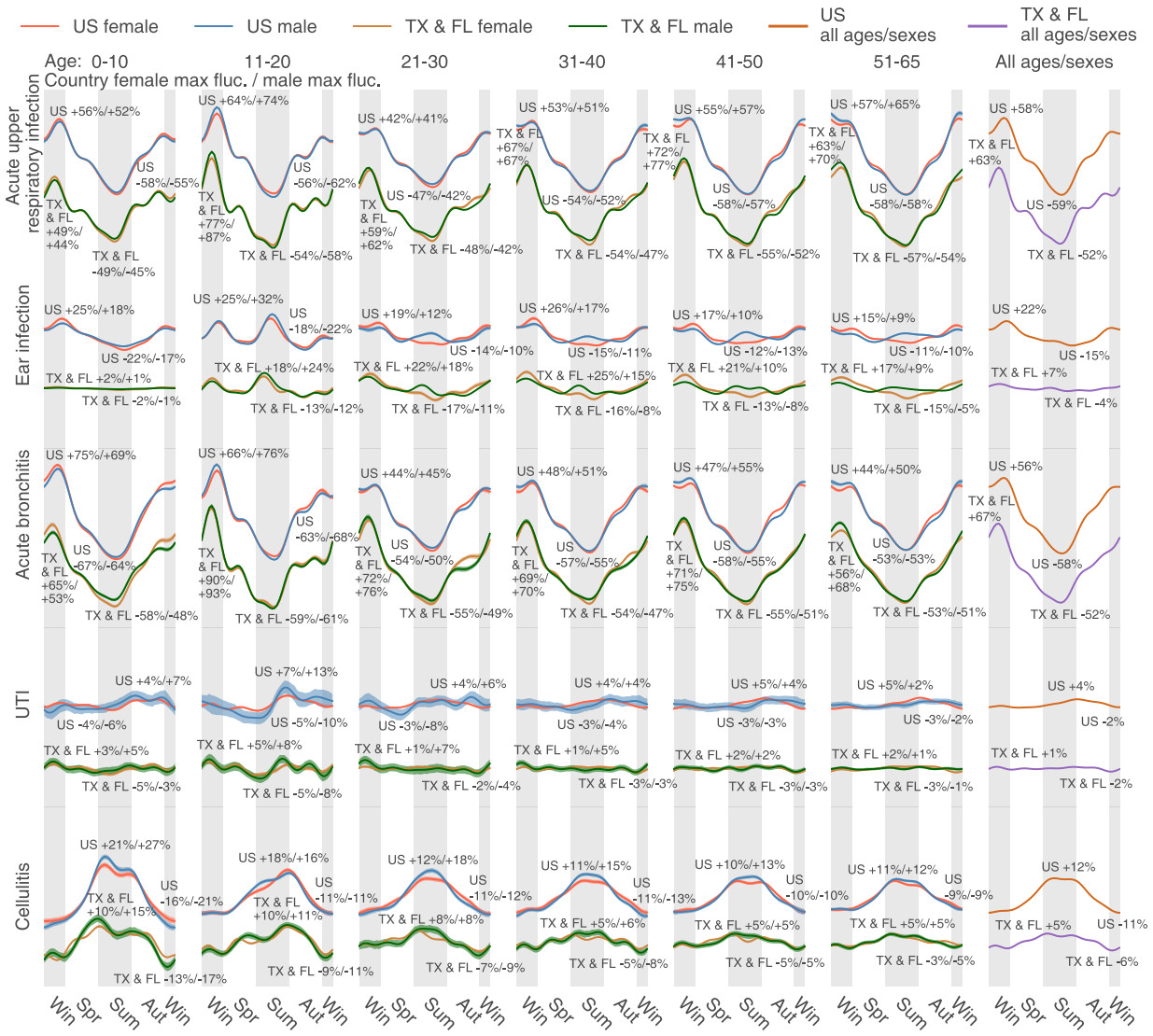

Figure 3.12 The uncorrected seasonality of infectious diseases in the two low-latitude states: Texas and Florida (TX and FL).

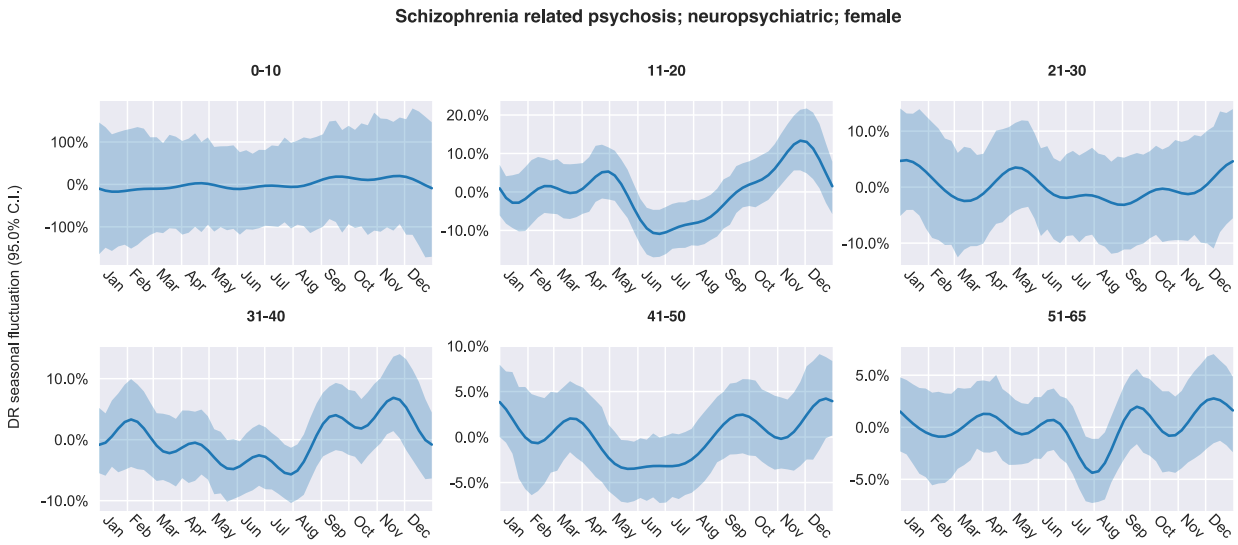

Figure 3.13 The uncorrected seasonality of schizophrenia-related psychosis in US females

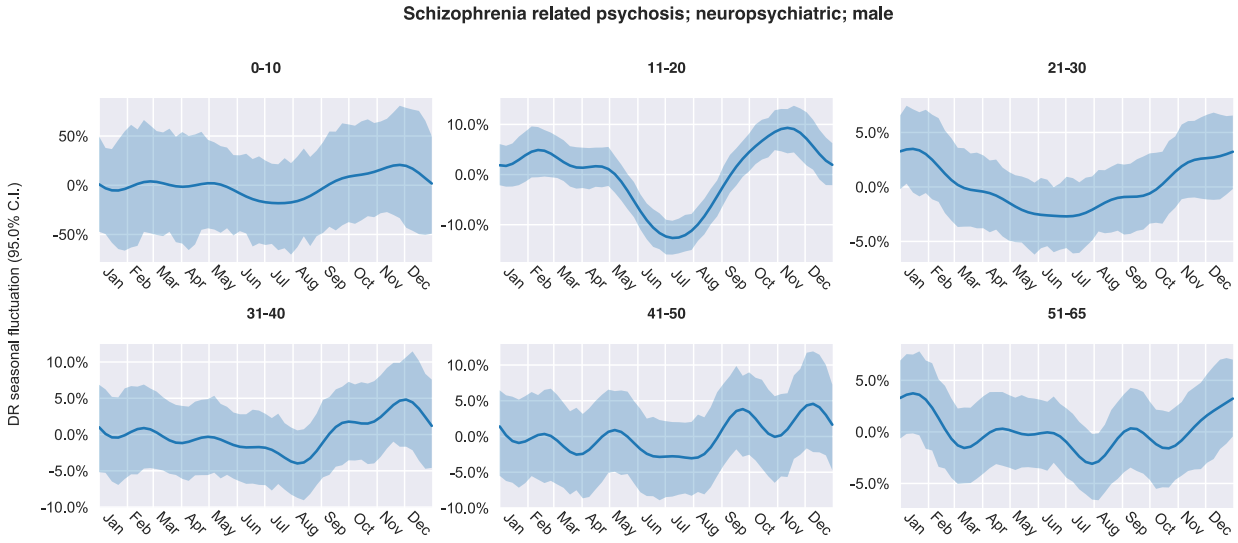

Figure 3.14 The uncorrected seasonality of schizophrenia-related psychosis in US males

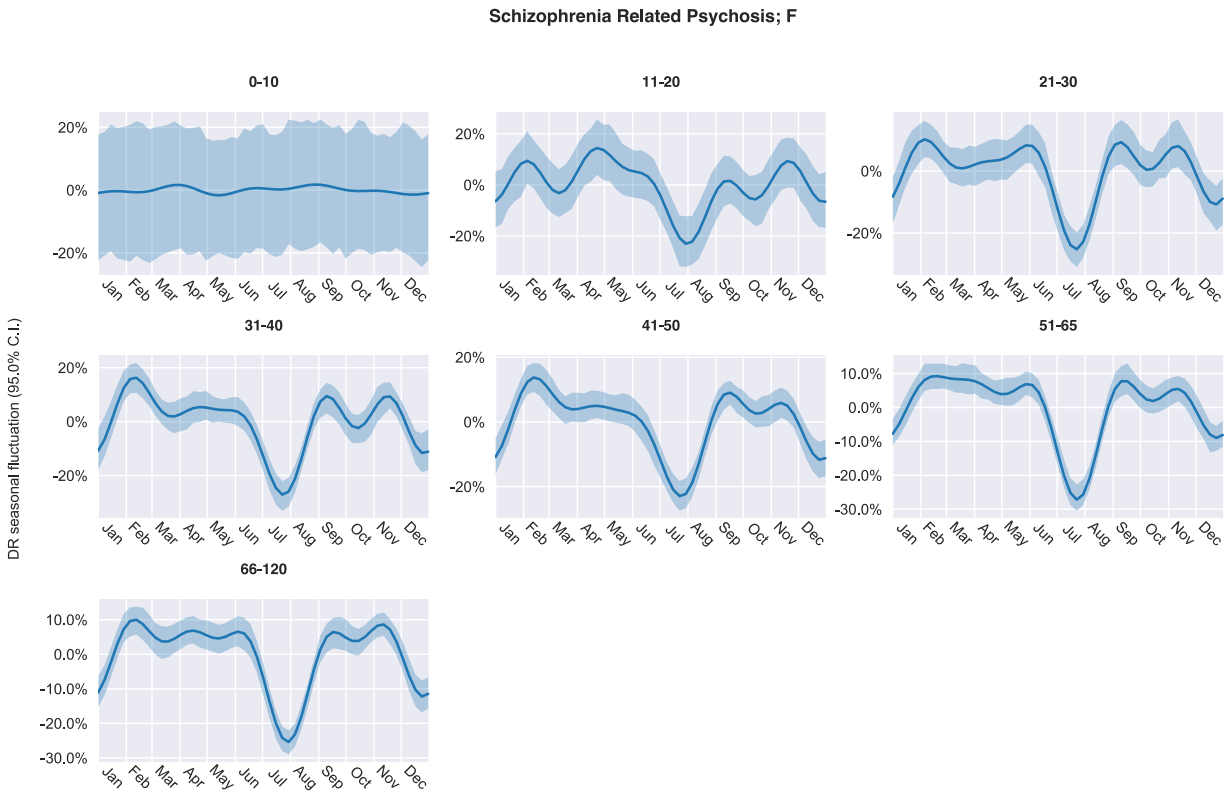

Figure 3.15 The uncorrected seasonality of schizophrenia-related psychosis in Swedish (SE) females

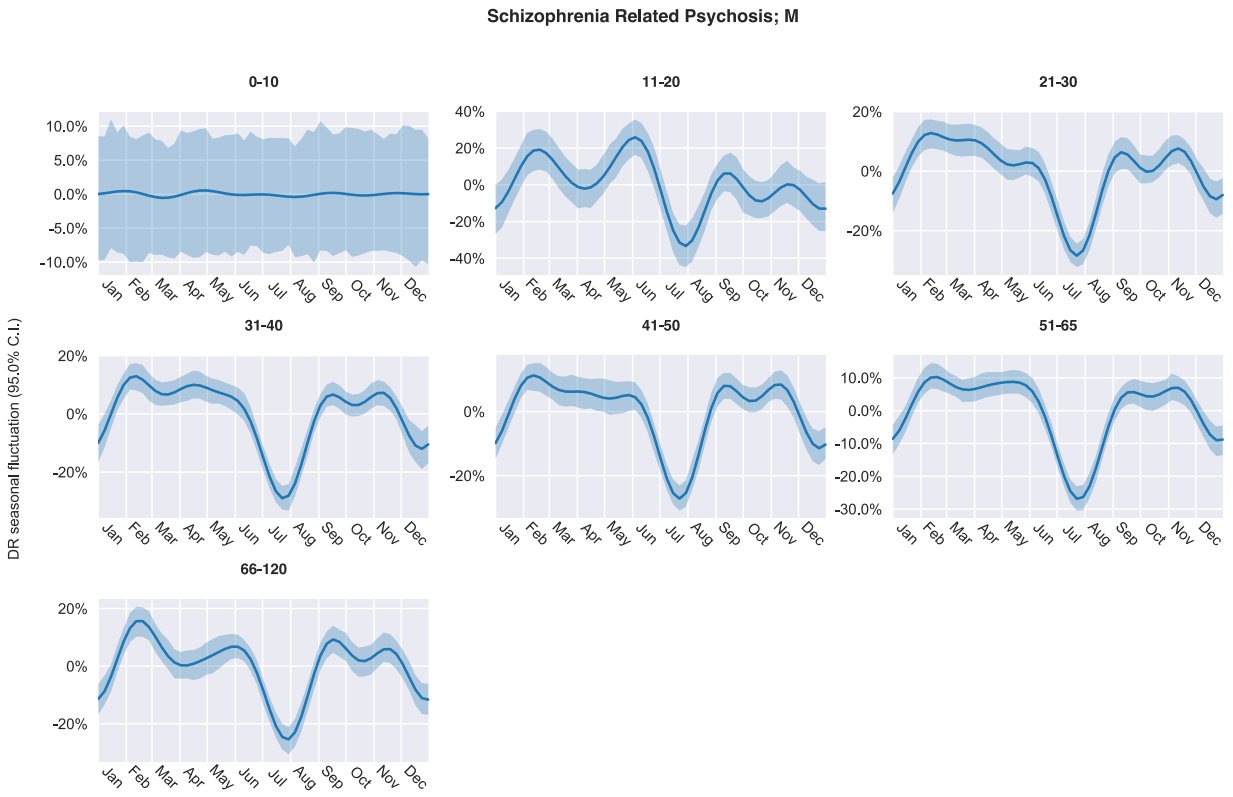

Figure 3.16 The uncorrected seasonality of schizophrenia-related psychosis in Swedish (SE) males

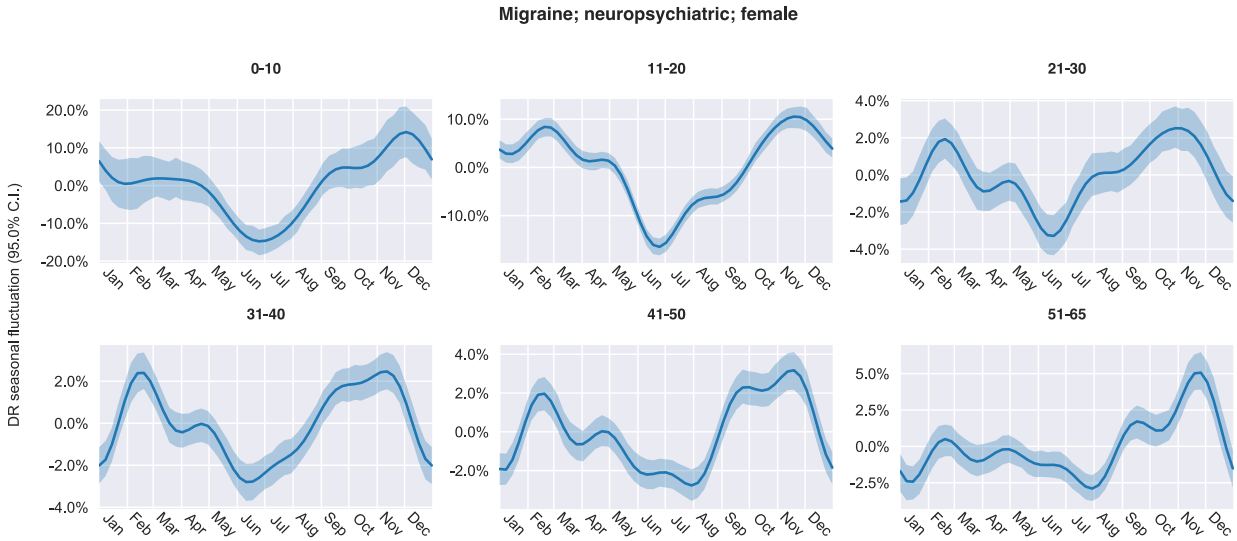

Figure 3.17 The uncorrected seasonality of migraine in the US females

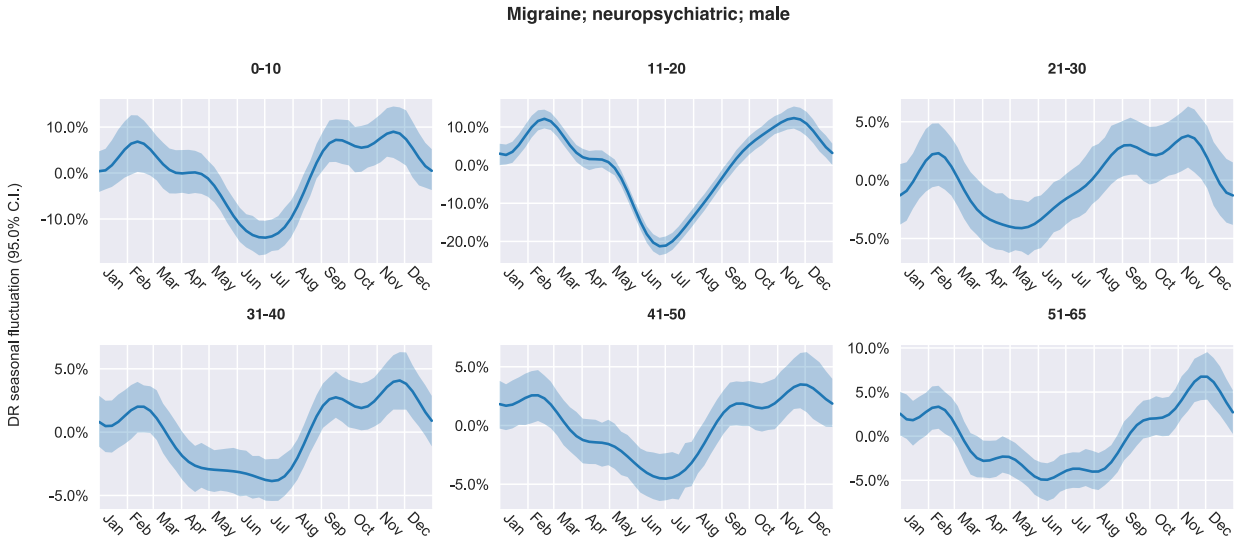

Figure 3.18 The uncorrected seasonality of migraine in US males

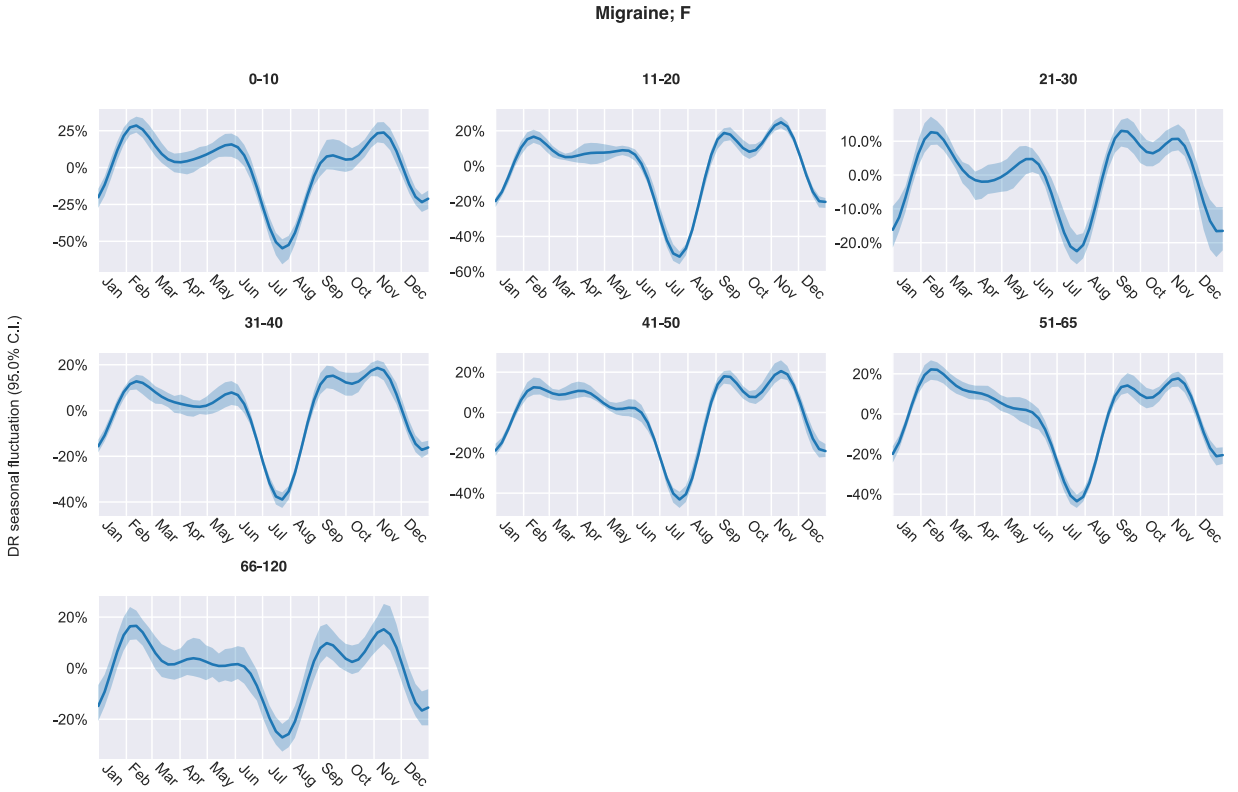

Figure 3.19 The uncorrected seasonality of migraine in Swedish females

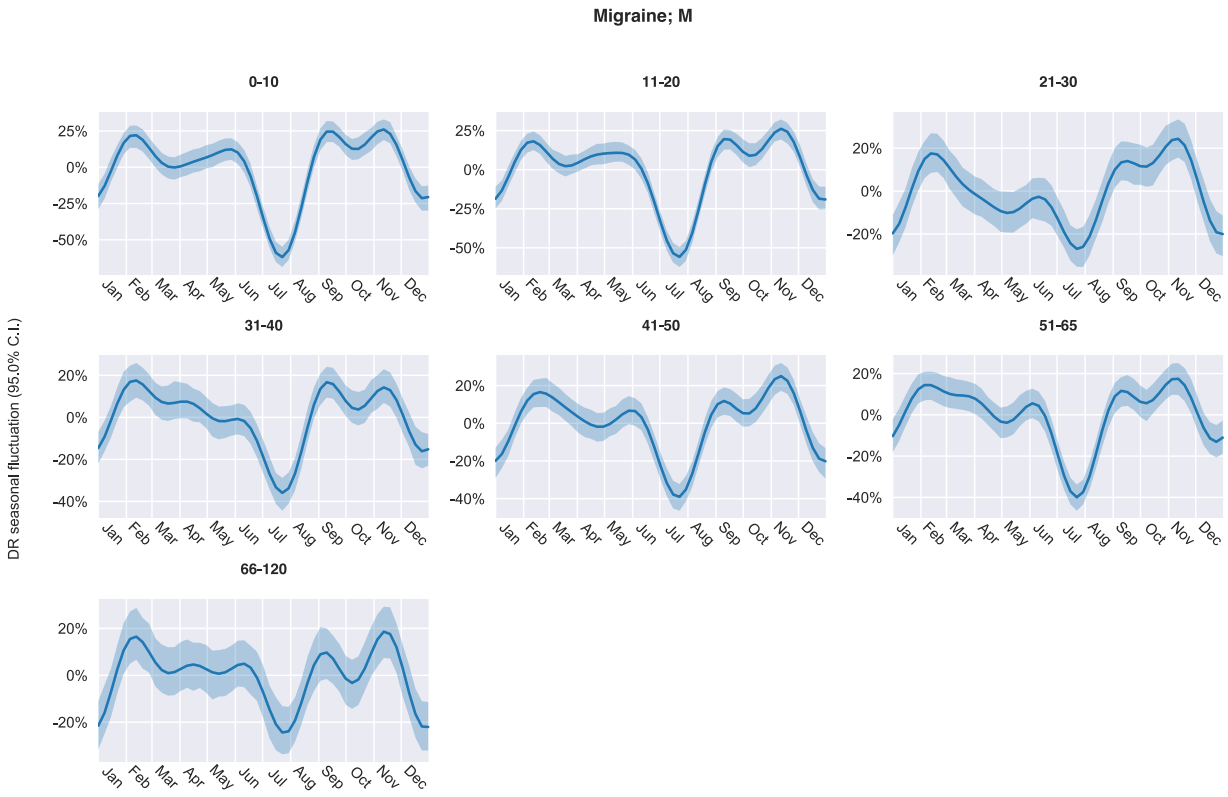

Figure 3.20 The uncorrected seasonality of migraine in Swedish males

3.3.2 Corrected seasonality analysis

While computing corrected seasonality plots, we grouped different age group curves together on the same subplots to compare the variability of seasonality across ages. We set the y-axis limits to be identical (for the same geographic area), so it is easier to compare seasonality across diseases. For each analyzed disease, we also gave its overall seasonality, aggregating all ages and sexes (the third and sixth columns of Figure 3.21–Figure 3.26). The time-average DR, ⟨DR⟩ on the plots, indicates disease prevalence in a particular sex-age bracket, and it could suggest what sub-populations are the most representative groups for a disease.

For the five most diagnosed psychiatric diseases in our US data (Figure 3.21), most sex-age groups' seasonality flattened after correcting for the baseline fluctuation of all medical visits. US patients aged eleven to 20 are exceptional, who still show evident DR decrease in the summer and upward trends in the spring, autumn, and winter after adjusting for the baseline seasonality of all medical visits. Depression, for example, decreases 20 percent more than the all-medical-visit baseline in the summer for both females and males aged eleven to 20 in the US. The age-sex aggregated curves do not suggest much seasonal variation for depression, anxiety phobic disorder, adjustment disorder, and substance abuse in the US, as the age eleven to 20 group is not dominant in terms of disease prevalence. By contrast, for ADHD, the population aged eleven to 20 is the most representative, so we can observe the summer's seasonal decrease in this condition in the age-sex aggregated plot shown in the third column of Figure 3.21.

In Sweden, the correction strongly adjusted the observed seasonality in psychiatric diseases (Figure 3.4 and Figure 3.21) because the baseline variation of all medical visits is large (the fourth row of Figure 3.3). After correction, we found only a minor decreasing trend in the summer for depression in eleven- to 20-year-old patients. It seems that DR for substance abuse goes up in the summer in Sweden, opposite to the trend in the US. Note that before applying the baseline correction, the substance abuse DR decreases in the summer (Figure 3.4). Therefore, the peak in the Swedish summer only suggests that such a seasonal decrease in substance abuse does not exceed the baseline variation. Besides, in Sweden, we also observed a decreasing DR in ADHD in the summer. Finally, for psychiatric diseases, we noticed there is a uniform decrease in DR at the beginning and end of the year (Figure 3.4 and Figure 3.21), possibly due to the winter break or vacation, which is even more obvious in Sweden.

After correction, the five most diagnosed infectious diseases in the US maintain significant seasonality (Figure 3.5 and Figure 3.22), almost consistent across age groups, sexes, and two countries (the US and Sweden). The seasonal trends are comparable to the uncorrected trends (Figure 3.5). In the summer, we found decreased DR for acute upper respiratory infection and increased DR for cellulitis. The distinct peak in summer ear infections still exists for US teenagers (eleven to 20 years old) and some older groups in Sweden.

Additionally, we can see the seasonal variation differences across higher- and lower-latitude regions after correction (Figure 3.23–Figure 3.26). Similar to what we found in the uncorrected analysis, psychiatric diseases in eleven to 20 year-olds demonstrate larger-than-national-average seasonal oscillation in the four high-latitude states (AK, WA, MT, and ND, or AWMN, Figure 3.23), and smaller-than-national-average seasonal oscillation in the two southern states (TX and FL, Figure 3.25). Figure 3.27 merges all psychiatric diseases in eleven- to 20-year-olds and shows the seasonal fluctuation differences across the four northern states (AWMN, largest fluctuation), the whole US (middle), and the two southern states (smallest fluctuation).

Figure 3.21 The corrected seasonality plots of the five most-diagnosed psychiatric diseases in the US and Sweden (SE): depression, anxiety and phobic disorder, adjustment disorder, substance abuse, and attention-deficit/hyperactivity disorder (ADHD)

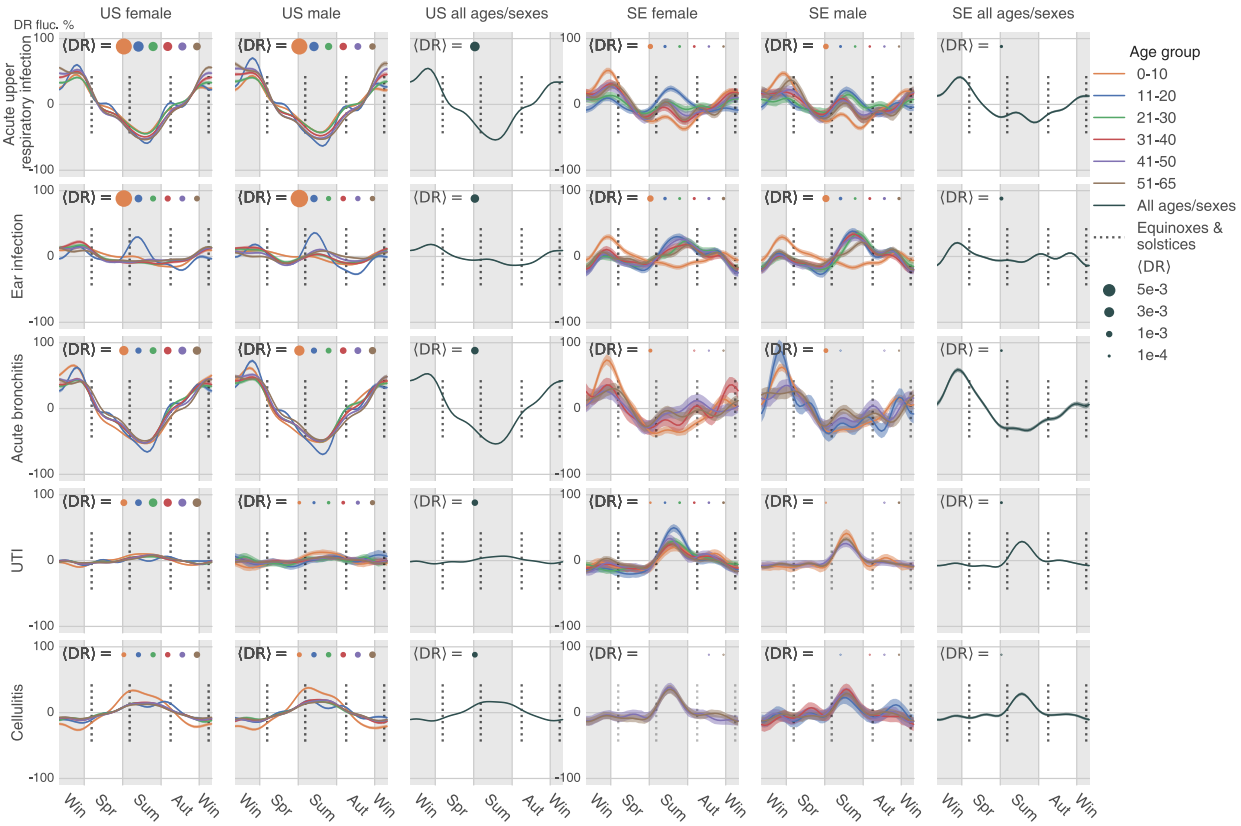

Figure 3.22 The corrected seasonality plots of the five most-diagnosed infectious diseases in the US and Sweden (SE): acute upper respiratory infection, ear infection, acute bronchitis, urinary tract infection, and cellulitis

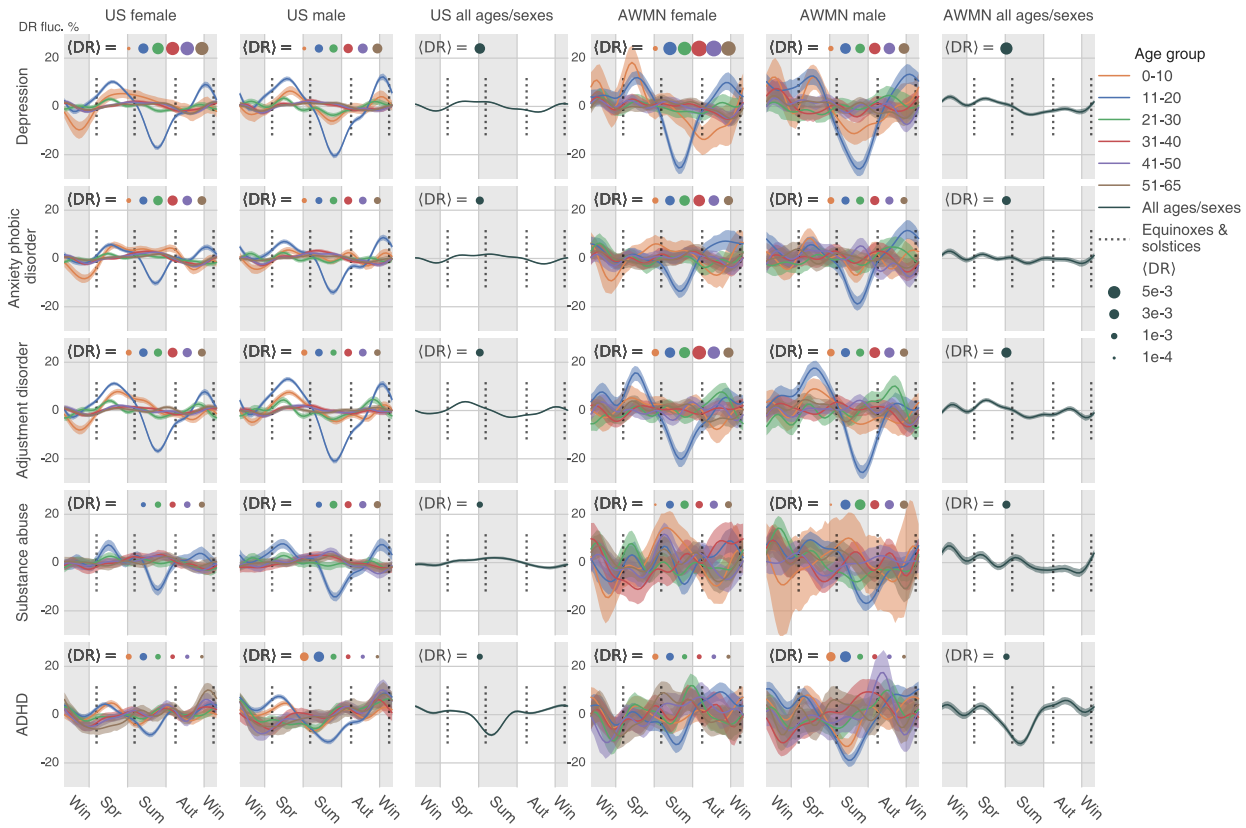

Figure 3.23 The corrected seasonality of psychiatric diseases in the four high-latitude states: Alaska, Washington, Montana, and North Dakota (AK, WA, MT, and ND)

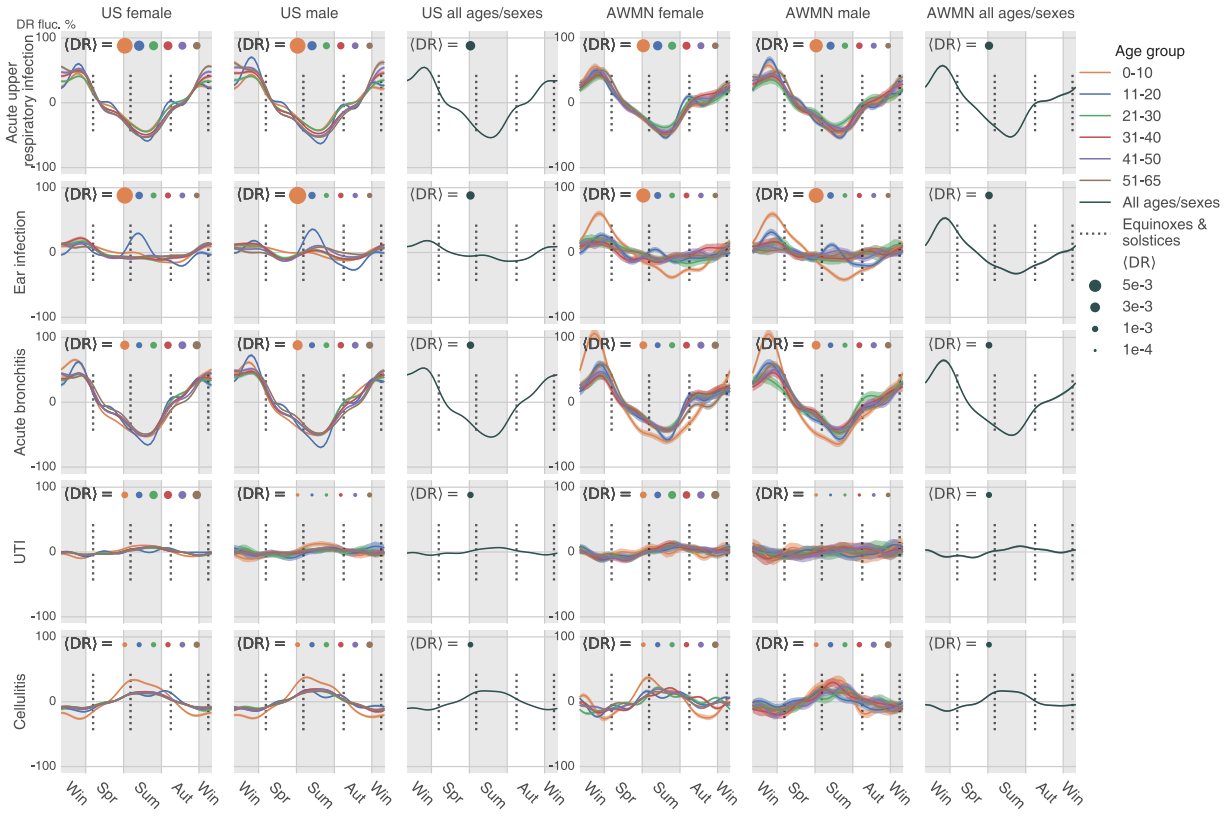

Figure 3.24 The corrected seasonality of infectious diseases in the four high-latitude states: Alaska, Washington, Montana, and North Dakota (AK, WA, MT, and ND)

Figure 3.25 The corrected seasonality of psychiatric diseases in the two low-latitude states: Texas and Florida (TX and FL)

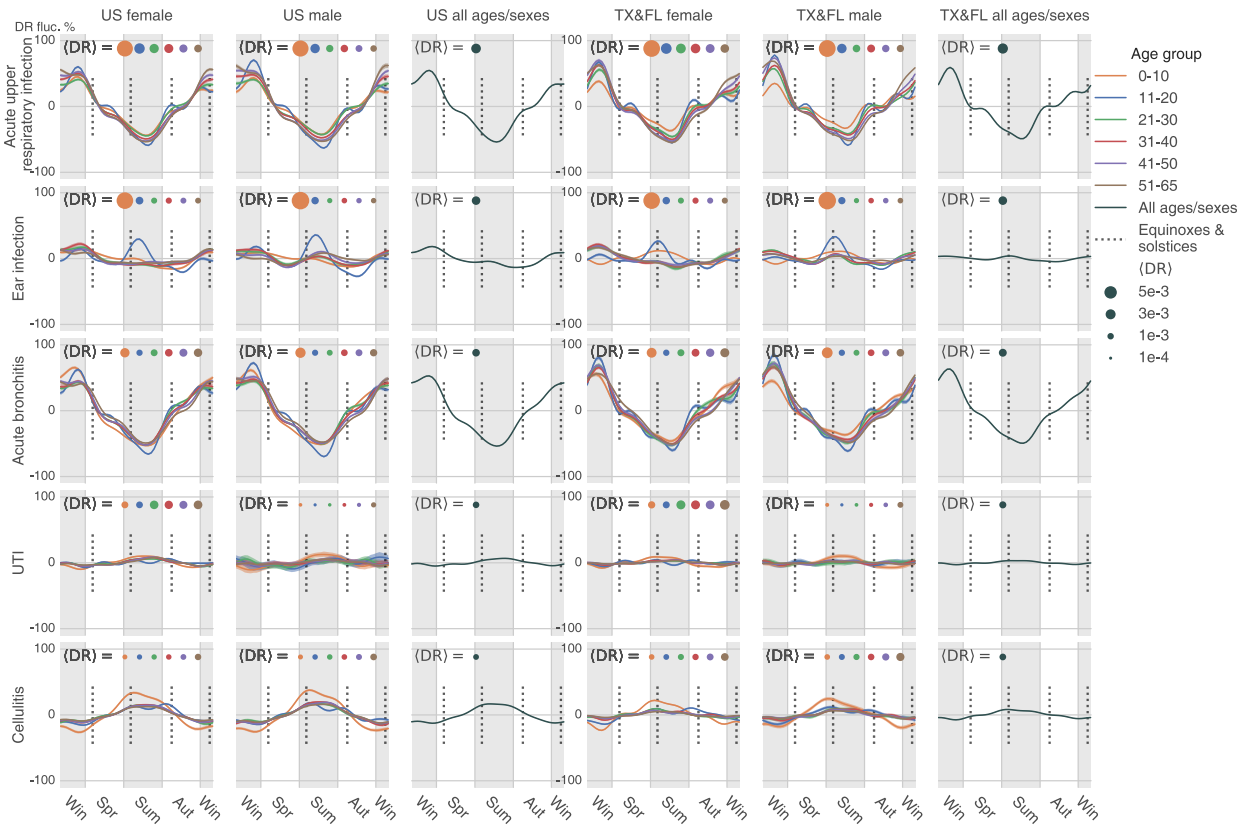

Figure 3.26 The corrected seasonality of infectious diseases in the two low-latitude states: Texas and Florida (TX and FL)

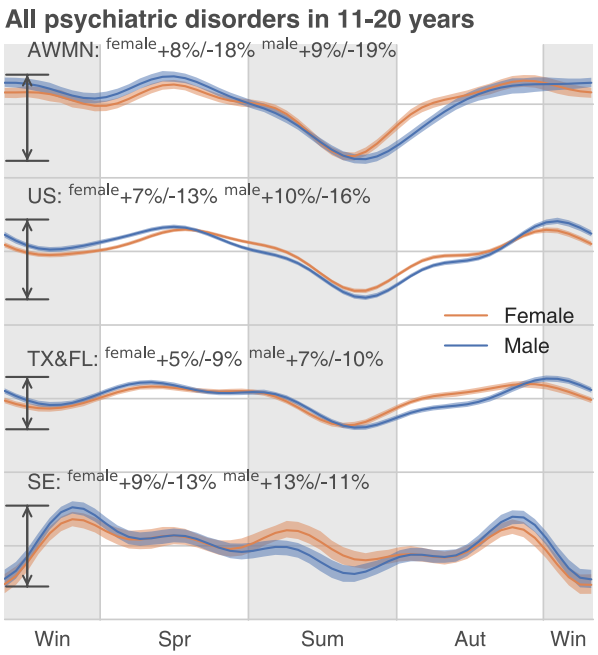

Figure 3.27 The corrected seasonality of all psychiatric disorders in eleven to 20 year-olds across four regions.

In the US, the annual oscillation of psychiatric disease DR is larger in the high-latitude areas (AK, WA, MT, ND, or AWMN) than in the low-latitude areas (TX & FL).

3.4 Discussion

An infectious disease, in its acute manifestation, requires the immediate attention of a physician. Most infectious diseases are transient (aside from a few that are chronic, such as malaria, AIDS, and herpes). Therefore, we expect that annual encounter rates of healthcare-seeking patients suffering with infections reflect real seasonality rather than exclusively annual patterns of recreational activities and vacations.

The situation might be different with patient visits associated with care for chronic diseases, elective procedures, and routine health checkups; harsh weather and vacation time may delay a visit and may require adjustment in analysis. Therefore, we produced two versions of

analysis of annual disease-related visit rates: "uncorrected" and "corrected." The former type of analysis answers the question "How are healthcare-seeking visits of patients with disorder *X* distributed across seasons?" The latter type of analysis answers a different question: "How are the proportion of healthcare-seeking visits of patients with disorder X with respect to all healthcare-seeking visits for this time interval time distributed across seasons?"

First, we argue that the psychiatric conditions we chose to examine in this study behave more like acute infections rather than elective procedures. For example, depression cases recorded in the Swedish registries are the most severe cases, where the patient needs immediate hospitalization (for example, because she stops eating). Milder depression cases, where a patient requires an antidepressant prescription, are handled by primary care providers and do not show up in health registries.

Second, our uncorrected seasonality results better resonate with existing, published, smaller-scale observations than our corrected results: Past studies report anxiety and depression rates higher during the spring than in summer in both Europe [93] and the US [94]; bipolar disorder symptoms receded in the summer in patients in Arctic areas of Northern Fennoscandia [95], and; depression and suicide in the US were reported to be higher in the spring than in the summer [96]. Past studies also reported an increase in substance abuse-related admissions to hospitals in the spring compared to the summer in Vietnam [97]. Studies in Vietnam have also reported that mood disorder-related hospital admissions were higher in the spring and fall compared to the summer, in agreement with our uncorrected seasonality results [97]; in corrected results, a subset of age groups displayed higher rates of depression in the summer than in the fall, as shown in Figure 3.4 and Figure 3.21.

Third, our attempt to "correct" the seasonality of individual disease rates by adjusting for the total rate of healthcare-seeking events (see Figure 3.21 and Figure 3.22) produced nonsensical results. For example, while uncorrected substance abuse rates in Sweden [98, 99] have been reported to be higher in the spring than in the summer (and this is what our uncorrected seasonality data shows), in the corrected Swedish results, the substance abuse rate completely reversed this trend, with the disorder rate higher in the summer than in the spring. In the US, corrected data also changed drastically with respect to the uncorrected – for psychiatric conditions only one age group, eleven- to 20-year-old patients, preserved their behavior across all psychiatric conditions compared to the uncorrected seasonality results. The rest of age groups acquired additional idiosyncratic properties that are not aligned with anything known about these diseases.

Our "uncorrected" version of this study suggested the existence of a uniform seasonality in the psychiatric disease diagnosis rate. This reported regularity was discovered via an analysis of a very large volume of health data, eliminating the possibility of noise-driven, spurious results. However, interpreting these statistically stable trends requires caution. Seasonal patterns for other psychiatric disorders closely follow those for SAD and depression, which suggests a plausible link to the annual daylight cycle, in turn affecting human circadian rhythms, yet we cannot completely rule out the influence of societal and economic factors. Furthermore, the other causality direction cannot be eluded at this stage: Psychiatric symptomatology due to light/dark cycle changes may lead to decreased social activity.

Most importantly, our analyses were based on the interpretation of diagnostic code timestamps, entered by physicians or psychiatrists after professional visits with patients. Here, we implicitly interpret the frequent psychiatric visits of a number of patients at the same time

interval as evidence of the population's deteriorating mental health. (This assumption is reasonable with infectious diseases, requiring a doctor's immediate attention after symptoms manifest themselves.) One may argue that lower psychiatric diagnostic rates in the summer are caused by vacations taken by either the patient or the psychiatrist, and not by the disease itself. This explanation is made less likely by the replication of the same annual disease registration pattern across different latitudes in the US and Sweden – because the vacation cultures of the two countries (and of the North and the South of the US) are drastically dissimilar; US vacations are typically shorter than their Swedish counterparts and tend to occur asynchronously around the year.

Past psychiatric seasonality studies [100, 101] used relatively small cohorts, typically insufficient to distinguish seasonal variation in disease prevalence from the background noise – with a few important exceptions. SAD is the most recognized seasonality-related psychiatric condition, a subtype of a unipolar depression [84-89]. Etiological hypotheses and experimental data have connected SAD to human circadian rhythms, the daily duration of exposure to sunlight, a patient's individual genetic variation, and her neurotransmitters' biochemistry [80-83]. Another plausible mechanism of seasonal changes in psychiatric conditions is seasonal immune dysregulation, due to seasonal allergies and infections [102]. We observed that anxiety and phobic disorders follow annual cycles nearly identical to those of SAD and depression – contrary to earlier studies' conclusions [100, 101] regarding anxiety's lack of seasonality. Previous examination of anxiety in smaller cohorts in the Netherlands found virtually no seasonality effects [86, 100]. Bipolar disorders' previously-reported seasonality properties concerned patterns of individual symptoms, with manic episodes peaking during spring-summer, and depressive episodes rising in the early winter [103]. Similarly, previous schizophrenia-

related studies had a different focus: They examined the risk associated with the season of a patient's birth, rather than the seasonality of disease relapse [101, 104]. Schizophrenia's seasonality, as well as that of migraine, was hardly covered in the literature, yet many studies provide evidence for a connection between circadian rhythms and these two diseases [105-109]. Our results suggest that both conditions follow the characteristic annual cycle with summer recess, similar to other psychiatric conditions.

In contrast to our scarce understanding of psychiatric disorders' seasonality, annual prevalence variation is a well-established phenomenon in infectious disease epidemiology. Seasonal infection waves are driven by typically well-understood factors, such as seasonal transmission of infection, host behaviors, and seasonal variation of host susceptibility to infections [110, 111]. Pronounced infectious disease seasonality aligns well with *a priori* expectation. In the present study, we used an analysis of infection seasonality as a positive control to corroborate the validity of both our method and our data. We still noticed some less obvious discordance. Verified in two nations, children smaller than ten are less affected by ear infection in summer. Conversely, teenagers (eleven to 20) are more likely to be infected in the summer (Figure 3.5). This tendency continues in Sweden's older population in Sweden (21 to 50) and older males in the US (11 to 65). Another interesting observation is that UTI is seasonal only in some population groups: US males between eleven and twenty years old and Swedish females.

We observed a large difference in the magnitude of seasonality's annual variation between the US and Sweden – but only in psychiatric – not infectious – diseases. Diagnostic rate fluctuation is much greater in Sweden than in the US – for example, depression diagnoses rates plunge about 48 percent in Swedish females compared to 14 percent in US females during the

summer. The difference in summer depression rates might be due to daylight exposure [80-83, 112, 113], as summer daylight extension (and daylight shortening in the winter) is much more extreme in proximal to the polar circle, as in Sweden, than in the continental US.

To conclude, it appears that psychiatric disorders follow a strong seasonal prevalence variation, closely resembling that previously described for unipolar depression. The most probable explanation for this observed seasonality, we believe, involves cyclic changes of exposure to solar light, which in turn affects circadian clock rhythms. In addition, this seasonality reflects a society's social rhythms, such as the patterns of summer vacations. The uncloaked, pervasive, homogeneous seasonality encourages us to contemplate the influence of the sleep-wake cycle, light exposure, and circadian rhythms on the development of neuropsychiatric diseases and be aware of the seasonality of mental health and its implication on the healthcare system.

3.5 COMPLETE DETAILS OF MATERIALS AND METHODS

3.5.1 Data and assumptions

The primary goal of this study is to model disease seasonality (and trend) in the US in recent years. To probe this question, we used the IBM Health MarketScan data set, one of the largest and most comprehensive collections of US insurance claims. This data set describes the healthcare encounters of over 150 million unique patients across the US. While the data set is very large, it was generated by merging reports collected by numerous private health insurers. Thus, the data set and the diseases themselves bear some properties that complicate our analysis:

- Though MarketScan follows population health statistics for over a decade (2003 to present), most of the people are "visible" to insurance records for a shorter time interval. Most patients

were only enrolled in the insurance records for a few years, a few months, or sometimes even a few weeks, which is known as "asynchronous enrollment." Due to heterogeneous insurance recruitment practices, groups of people simultaneously joining an insurer by no means resembles a random sample from the general US population.

- Possibly due to changes in coding standards, the data set contains annual shifts (systematic jumps) of diagnosis rates (DR) at the beginning of every year for population strata of consistent composition. These shifts were observed for a subset of diseases.
- US holidays result in a general disruption of both health practice and reporting, and these disruptions are visible in the raw disease prevalence plots.
- Most diseases manifest annual periodic fluctuations of prevalence. For example, skin infections are on the rise in the summer, while upper respiratory system infections are more frequent in the winter.
- Disease prevalence seasonalities and trends vary greatly across sex and age groups.
- The data contain stochastic noise (temporal fluctuations in the recorded disease diagnoses) and, possibly, diagnosis encoding errors.

All these factors influence the raw observations of diagnoses rates over time. A naive approach is to estimate the raw trend, treating the population as a whole and fitting a line of DR which we then calculate as the total number of diagnoses over the total number of enrollments across the whole database. This produces largely nonsensical results. Therefore, we designed a Bayesian model that addresses the issues discussed above (Figure 3.2), which allowed us to infer latent disease trends for any specific age and sex group.

Additionally, we modeled the disease seasonalities (and trends) in Sweden based on their national register that incorporates almost all nine million Swedes. Though there is no exact

enrollment information supplied, we can safely assume static enrollment in years because Swedish patients were disenrolled only if they died or left Sweden. In other words, unlike the US data set, the Swedish database is immune to the "asynchronous enrollment" problem.

3.5.2 Methods and techniques

We modeled male and female trends separately. To make corrections for asynchronous patient enrollments, we first grouped all 150 million enrollees in the database by: 1) their enrollment dates (starts and ends); 2) patients' ages at the middle of enrollment, and; 3) patients' sexes.

We then obtained nearly a million-enrollment range-, age-, and sex-specific population strata. Each stratum was characterized by a unique enrollment interval, for example, January 1, 2003 to December 31, 2004. We also placed different sexes into separate strata, and further subdivided patients by age groups. Specifically, we used the following approximate decade-long age subdivision: 0-10, 11-20, 21-30, 31-40, 41-50, 51-65. Claims occurred at over 65 years-old were discarded because the majority of the enrollees supposedly switched to Medicare, and the remained records over 65 were likely to be erroneous or nonrepresentative.

For each population stratum, assuming a latent linear trend and yearly-repeated seasonality of disease prevalence, we defined a model and estimated its parameters. However, this approach would become practically intractable if we were to fit this model on close to a million population strata simultaneously. Therefore, we reduced the number of population strata by merging them into a smaller number of bigger strata, based on the proximity of enrollment boundaries, using K-means clustering. Population strata with close enrollment boundaries were pooled together. In this way, we obtained around 600 "soft-bounded" population strata (100 for each age group) for each sex. Each composite stratum is a combination of hundreds of raw,

"hard-bounded" populations. These composite strata vary slightly in terms of date of enrollment beginning and end, typically in the range of a few weeks, and are rather homogeneous inside the shared enrollment window. Considering the Bayesian consistency of our model, a more robust and powerful way is to cluster and merge population strata using a Bayesian Gaussian mixture model. However, such a method would soon exhaust a computer's random-access memory because of the need to track a large number of variables. It is well known that the K-means method is equivalent to the hard-EM (Expectation-Maximization) implementation of the Gaussian mixture model in some limiting formulation. Here, we argue that K-means clustering, chosen for its scalability, is good enough to mimic the behavior of the Bayesian Gaussian mixture model and accomplish population number reduction.

For each disease, we decomposed the DR trend, for a given soft-bound population stratum *i* of age group *j* for a sex-specific condition, into four parts: a linear trend, possible shifts at the beginning of every year, a seasonality term modeling periodic patterns, and an error term incorporating all other effects.

$$y_{i,j}(t) = \text{Linear trend} + \text{Yearly shifts} + \text{Seasonality} + \text{Error}, \tag{3-1}$$

where

$$\text{Linear trend} = \alpha_{i,j} + \beta_{i,j}t, \tag{3-2}$$

$$\text{Yearly shifts} = \sum_{k} \mathbf{1}(t > s_k) \cdot \gamma_{i,j,k}, \tag{3-3}$$

$$\text{Seasonality} = \sum_{n=1}^{N} p_{i,j,n} \cos\frac{2n\pi t}{W} + \sum_{n=1}^{N} q_{i,j,n} \sin\frac{2n\pi t}{W}, \quad (3\text{-}4)$$

$$\text{Error } = \epsilon_{i,j}. \quad (3\text{-}5)$$

In the above equations, $y_{i,j}(t)$ is the diagnosis rate of a soft-bound population stratum *i* of age group *j* at time point (week) *t*. Parameters $\alpha_{i,j}$ and $\beta_{i,j}$ are the intercept and the slope, respectively, of the latent linear trend. $\mathbf{1}(\text{condition})$ is an indicator function that evaluates to 1 only if the input condition is true. s_k and $\gamma_{i,j,k}$ is are the k^{th} separation and shift, respectively. The separations are when the shift could happen, and we assumed they are all year starts (s_1 = January 1, 2003, s_2 = January 1, 2004, …).

Note that for the Swedish database, all enrollees are visible from the start, so there is no "asynchronous enrollment" problem and only one all-inclusive population stratum was considered for each age group and sex.

We used a Fourier series with period $W = \frac{365.25}{7}$ weeks to model the potential seasonality of some conditions. $p_{i,j,n}$and $q_{i,j,n}$ are harmonic bases.

The traditional parametrization of the "seasonality term" (a Fourier series) is convenient for the estimation phase of analysis. However, to interpret estimates, it is intuitive to use the following re-parametrization of the Fourier term:

$$\text{Seasonality} = \sum_{n=1}^{N} A_{i,j,n} \sin\left(\frac{2n\pi t}{W} + \phi_{i,j,n}\right), \quad (3\text{-}6)$$

Where

$$A_{i,j,n} = \sqrt{{p_{i,j,n}}^2 + {q_{i,j,n}}^2}, \text{ and} \tag{3-7}$$

$$\phi_{i,j,n} = \text{Arctan2}\left(p_{i,j,n}, q_{i,j,n}\right). \tag{3-8}$$

$A_{i,j,n}$ and $\phi_{i,j,n}$ in Expression (3-6) are amplitudes and phases for the *n*'s harmonic. Arctan2 corresponds to a two-argument arctangent function.

We estimated all parameters under a Bayesian framework. We sampled the prior values of $\alpha_{i,j}$ and $\beta_{i,j}$ from skew-normal distributions with age-group-specific locations, scales, and shapes. A skew-normal distribution density function is defined in the following way:

$$f(x; \text{loc} = \mu, \text{scale} = \sigma, \text{shape} = h) = \frac{2}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right)\Phi\left(h\left[\frac{x-\mu}{\sigma}\right]\right), \tag{3-9}$$

where $\phi(\text{x})$ and $\Phi(\text{x})$ are density and cumulative distribution functions, respectively, for a standard normal distribution. Our choice of prior distribution was motivated by an analysis of the parameter estimate distributions for various groups of patients – they indeed resemble the skew-normal shape.

$$\alpha_{i,j} \sim \text{SkewNormal}\left(\text{loc} = \mu_j^{\alpha}, \text{scale} = \sigma_j^{\alpha}, \text{shape} = h_j^{\alpha}\right), \tag{3-10}$$

$$\beta_{i,j} \sim \text{SkewNormal}\left(\text{loc} = \mu_j^{\beta}, \text{scale} = \sigma_j^{\beta}, \text{shape} = h_j^{\beta}\right). \tag{3-11}$$

To allow information flow through different age groups, we sampled the location parameters from a zero-mean Gaussian process prior:

$$\mu_j^{\alpha} = \mu^{\alpha}(j) \sim \text{GaussianProcess}\left(0, k^{\alpha}(j, j')\right), \tag{3-12}$$

$$\mu_j^{\beta} = \mu^{\beta}(j) \sim \text{GaussianProcess}\left(0, k^{\beta}(j, j')\right), \tag{3-13}$$

where $k^{\alpha}(j, j')$ and $k^{\beta}(j, j')$ are exponentiated quadratic kernels with scale and length drawn from flat hyperpriors. The prior of the Gaussian process "linked" different age groups

within a unified estimation procedure and allowed information about disease trend flow across age groups – by assuming that similar-age groups share similar trend parameters.

We drew the scale parameters of $\alpha_{i,j}$ and $\beta_{i,j}$ from flat, half-Cauchy hyperpriors, and we restricted the shape parameters h_j^{α} and h_j^{β} by zero-mean Laplace distributions so that the scale parameter would not compete with the shape parameter. In our experiments, we found a skew normal with a large shape could behave similarly as a skew normal with a large scale. This pathological behavior would result in inefficient sampling.

We sampled the population- and age-specific shifts $\gamma_{i,j,k}$ from a zero-mean Laplace distribution, incorporating our prior belief that shifts should not mask the linear trend effect. We sampled the bases for seasonality from zero-mean normal distributions.

Finally, to offset the effect of holidays and celebrations, we applied a holiday-smooth function that took the average diagnosis rates around US federal holidays and Easters/Good Fridays. We overcame the presence of outliers caused by other unknown forces by a Student-*t* distribution sampling:

$$\text{HolidaySmooth}\left[y_{i,j}(t)\right] \sim \text{StudentT}\left(\mu_{i,j}^{y}, \sigma_{i,j}^{y}, \nu_{i,j}^{y}\right), \tag{3-14}$$

where the location parameter to $\mu_{i,j}^{y}$ = Linear trend + Yearly shifts + Seasonality. $\sigma_{i,j}^{y}$ and $\nu_{i,j}^{y}$ are scales and degrees of freedoms sampled from flat half-Cauchy hyperpriors. Figure 3.1B illustrates the outcome of the holiday-smooth function.

We approximated the model using a No-U-Turn sampler [12] initialized by variational inference [67]. In general, for one sex-specific condition, it would take hundreds of CPU hours to attain a reasonable estimation due to the high-dimensional searching space – we were sampling thousands of parameters simultaneously. We applied the model to 33 neuropsychiatric and 47

infectious conditions of two sexes and tried to reproduce and make corrections for trends in different age groups.

Once we obtained the estimation of harmonic bases $p_{i,j,n}$and $q_{i,j,n}$ as in Expression (3-4), we computed the posterior harmonic base estimates for the whole population as:

$$\overline{p}_{j,n} = \sum_i w_i \cdot p_{i,j,n}, \tag{3-15}$$

$$\overline{q}_{j,n} = \sum_i w_i \cdot q_{i,j,n}, \tag{3-16}$$

where w_i is the weight according to the size of population stratum *i*, so $\overline{p}_{j,n}$ and $\overline{q}_{j,n}$ could be interpreted as the estimate of n^{th} harmonic bases for age group j for the whole population.

After Bayesian inference, we obtained the posterior distributions of all parameters, and we could then estimate the annual seasonality free from the influence of the trends, sudden shifts, and noises such as holiday effects. For each age/sex-specific condition, we divided the raw DR seasonality to its time-average DR over 570 weeks (from Jan 1. 2003, Expressions (3-17) and (3-18)) and obtained the relative fluctuation in percentage, as shown in the main figures.

$$\text{time-average DR} = \langle \text{DR} \rangle = \frac{1}{570} \sum_{w=1}^{570} \text{DR}_\text{w} \tag{3-17}$$

$$s(t) = \frac{\text{Seasonality estimate}}{\langle \text{DR} \rangle} \tag{3-18}$$

We can possibly correct the baseline fluctuation of all-medical visits by deducting $s(t)$ from the $s_{\text{all}}(t)$ that represents the yearly variation of all conditions and diseases (Figure 3.3 and Figure 3.2 Step 2 lower center plot):

$$\text{Corrected } s'(t) = s(t) - s_{\text{all}}(t) \tag{3-19}$$

The procedure also revealed the disease trends. However, as we carefully examined these estimates, it was clear they might not reflect real disease trends over time simply because we estimated the trend with cohorts of quasi-static enrollments – the same people joined and left and their age went up accordingly. The change of age drastically impacted disease trend estimations. For example, we observed that incidents of some infectious diseases went down because the prevalence of some pediatric infections decrease as children grow older. By contrast, the trends of many cardiovascular conditions are positive because older people are more prone to them.

Figure 3.2 summarizes our model where data corresponding to a "raw" trend is an input to our model. The "raw" trend is deconvoluted into trends within hundreds of population strata based on enrollment dates (the left panel and the center panel). The model fits each population stratum separately, but still allows certain information shared across population strata, in a hierarchical framework (the center panel). Finally, we make corrections and estimate the seasonality and trend for specific age groups (the right panel).

Last, it is worth mentioning that we dropped all higher-order harmonics in the Fourier series after the first 5 ($N = 5$) for approximation based on model selection results (Expressions (3-4) and (3-6)). We tested $N = 5$, 15, and 25 to find the best approximation model. To evaluate the model, we computed the sum of Watanabe–Akaike information criteria (WAIC) [114] over 33 neuropsychiatric and 47 infectious diseases in the two sexes and found that $N = 5$ is simple and good enough to model the seasonality (Figure 3.28).

The Bayesian procedure we designed helped to mitigate multiple confounding factors with a multilevel model, but it could also be problematic given its complexity. First, we could not certify the convergence of the Markov Chain Monte Carlo since we were not estimating one single parameter, whereby a diagnostic statistic like Gelman-Rubin [69] or Geweke [115] would

be applicable to determine the mixing of that parameter. The approximation of each disease's seasonality involved thousands of parameters, making it difficult to determine how many iterations were needed to reach a stationary point. To alleviate this concern, we employed the No-U-Turn sampler, which is able to mix rapidly [12]. More importantly, we inspected the disease trend's posterior expectation curves and seasonality, restored from the posterior estimation of parameters (like the green lines on Figure 3.2 Step 2 upper panel), and made sure they were aligned with the input raw trend and seasonality. Collectively, using all the available tools, the intrinsic seasonality is reflected in the results insofar as we able.

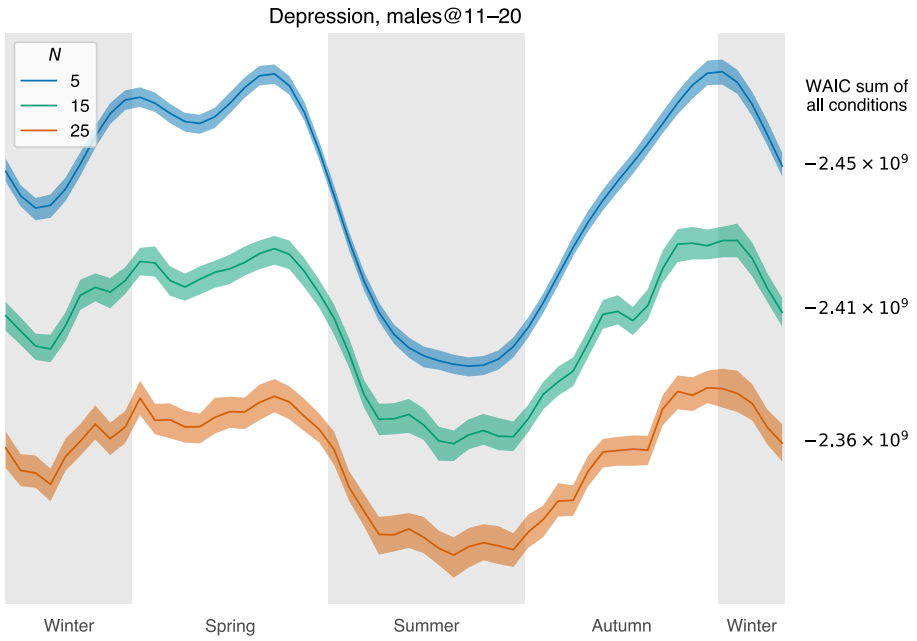

Figure 3.28 Model selection for choosing the number of harmonics.

The model with N = 5 has the lowest sum of WAIC over 33 psychiatric and 47 infectious diseases. It suggests the simpler model is good enough to model disease seasonality. In the example of depression in young males, adding up harmonics would not help the estimation, given the intrinsic simplicity of seasonality.

CHAPTER 4. GENE-ENVIRONMENT INTERACTIONS IN PSYCHIATRIC DISORDERS

This chapter is adapted from the manuscript "**Gene-environment interactions explain a substantial portion of the variability of common psychiatric disorders**" authored by Hanxin Zhang, Atif Khan, and Andrey Rzhetsky.

4.1 INTRODUCTION

The study of a phenotypic trait's heritability is one of the most important topics in biology. By definition, heritability is the proportion of the overall trait variance that can be explained by the genetic variation under an explicitly defined genetic penetrance model. In practice, heritability has been used to indicate the strength of a trait's response to artificial selection in domesticated organisms. As expected, it would not be very productive when attempting to artificially select for traits with low heritability. From the earliest days in the field, geneticists imagined the possibility of the existence of complex interactions between genetic variation and environmental stimuli, often denoted with G × E (the G-by-E effect). However, measuring such interactions in real-world data has been proved rather difficult [116-121]. The limiting factor in such studies is the availability of sufficiently large and rich data, combining genetic, environmental, and phenotypic information for the same individual. A hypothetically ideal data set for analyzing how environmental and genetic factors affect human disease would have the following properties:

First, the data would need to contain the full germ-line genomic sequence for a large number of individuals. While such human data sets are already available today, our ideal data would link all study individuals into a unified genetic pedigree, which would allow inference of genetic models associated with the vertical trait transmission.

Second, for each study participant, the data would contain the complete phenotypic profiles for the trait of interest. This is very challenging, as myriad phenotypic characteristics, which change along individual lives, have to be recorded. Currently – at best – human observational data contains snapshots of medical histories and collections of biometric, metabolic, and clinical measurements for each individual for a limited window of their ontogenetic timeline, usually continuing throughout a few years.

Last, for each individual, the data would document a complete record of exterior environment and intervention, including exposures to changes in climate, tidal waves of uncountable bacterial, viral, fungal, and protozoic encounters, billions of environmental molecular species entering the individual's body over time, exposure to electromagnetic fields and to the bombardment of elementary particles, the individual's diet, exercise routines, and their social support structure. These factors are typically the worst-documented part of human life, because the set of relevant stimuli is astronomically vast and must be recorded continuously from conception to death.

Having generated a gigantic genotype × environment × phenotype × time data matrix, we would use it to fit a battery of increasing-complexity mathematical models, encapsulating the probability of observing a specific phenotype, given genetic and environmental inputs. The quality of fit to data of each candidate model, entailing a theory of disease etiology, would then explicitly quantify the theory's value. Generating such a gargantuan data set for a practical-size cohort of at least a few thousands of individuals is prohibitively expensive at this time (genetics alone would require a budget of tens of millions of US dollars). Because this imaginary, ideal data set is not available and is not likely to be generated in the near future, the real-life heritability estimates of human disease rely on: (a) access to simplified-structure, cheaper data

sets, and; (b) mathematical models with strong, simplifying assumptions permitting inference from incomplete data. There are three major groups of mathematical approaches, that are twin-, pedigree-, and genetic association studies, constituting the modern toolbox for dissecting disease etiology.

Twin Studies: In these studies, the data includes limited phenotypic descriptors (disease or no disease), and limited genetic information (monozygotic or dizygotic, abbreviated as DZ and MZ) about twin individuals. The genetic data restrictes the knowledge to100 percent and 50 percent genetic similarity of MZ and DZ twins, respectively. Then, with the strong assumption that the environment is identical for both individuals in each twin pair, we can use a simple model to attribute disease status in concordance to twins' genetic similarity [122-126].

Pedigree Studies: The limited phenotypic data in these studies includes disease or no disease status for each family member. Genetic data is reduced to a pedigree structure, where siblings share half of the genetic variants with each parent and with each other. No explicit environmental data would be provided. Model assumptions can include the same environment for all individuals described in the data set (earlier models), or distinct shared environments for siblings, for spouses, families, and individuals (in the more recent and sophisticated models) [127].

Whole-Genome Sequences and Genetic Association Studies: In this case, genetic data includes individual-level genotypes for a large number of participants, possibly with knowledge of family structure for a subset of individuals. Phenotypic data involves disease status. Typically, there is no environmental data available, and individual-specific environments are implicitly assumed identical for all participants [128].

One should expect that each simplifying assumption in a mathematical model of heritability results in biased estimates of parameters. In twin analyses, all concordant disease states between twins are explained genetically. Therefore, we would expect that heritability estimates from twin data would be the highest. On one hand, association studies focus on the explanatory power of common genetic variation – so, by design, heritability estimated with association data is smaller than estimates than would be obtained from total (meaning, common plus rare plus ultra-rare/individual) genetic variation. On the other hand, association analysis typically ignores environmental data, which means that heritability estimates should be inflated with respect to hypothetical association studies with explicit environmental data included. Family pedigree analyses incorporate both limited genetic and limited environmental data, and thus promise better-balanced heritability estimates. Keeping the strengths and weakness of current tools in mind, we expect that richer data and more complex models are needed for dissecting the influence of environment and genetics on human traits.

What We Propose Here: To represent genetics, we will use family trees, as in pedigree studies mentioned above. Environmental influences will be implicitly represented in random effects, associated by the predetermined environmental relationship (familial, environment shared by couples or siblings, etc). More interestingly, we will construct models that consider the gene-environment interactions and compare them to their counterparts in the absence of the interaction term. On the phenotypic side, we will use disease presence or absence status for a collection of diseases, rather than a single disease. Furthermore, we will incorporate explicit environmental data – each family's geographic position will be associated with multidimensional environmental measurements, systematically collected over US territories. For each family, we will also include sociodemographic, economic, and topographic parameters linked to their

geographic position. In addition, by incorporating families' geographic proximity in the model, we will be inferring implicit environmental variation that is not explained by the explicit environmental measurements.

In this study, we focused on dissecting the etiology of ten major psychiatric disorders, documented for 138 thousand US families, containing nearly half a million unique individuals. We built five Bayesian regression models from simple linear models that contains only genetic and random environmental factors to complex interaction models. For ten representative psychiatric diseases, we then estimated how much of the phenotypic variation comes from the genetic factor, different environmental factors, or interactions between them, using the insurance claim data of the 138 thousand families. Our results suggest that gene-environment interactions account for substantial risk variability of common psychiatric disorders.

4.2 METHODS

4.2.1 Data

Our main data resource is the IBM Health MarketScan data set [19], which records the health histories of over 150 million unique patients across the US from 2003 to the present. The data also documents both the kinship and county-level place of residence for patients. We chose to analyze individuals who consistently lived in the same geographical location and were enrolled in our data for at least six years, so we could assume they had been exposed to the same environment for some time. From this filtered population, we selected 138 thousand families (404 thousand individuals), including parents and their above-16-year-old children.

We used a comprehensive environmental quality index (EQI) developed by the US Environmental Protection Agency (EPA) to quantify the fixed environmental factor in our

mixed-effects models [129, 130]. For every county, the EQI gives numerical quality estimation in five domains: air, water, land, sociodemographic, and built environment. It is worth noting that each domain index already summarizes a large number of relevant variables, and lower scores suggest better environmental quality in general. For the air, water and land domain, the EPA summary indices represent the overall quality considering numerous pollutants and contaminants. The sociodemographic domain represents environmental quality related to income, education, employment, crime, and other socioeconomic elements, while the build-environment domain summarizes the housing quality, traffics, roads, etc.

4.2.2 Modeling

We employed a mixed-effects linear regression model, using logit-link for binary disease outcome. The fixed effects included basic demographic factors (such as sex and age groups) and environmental quality indices split by categories: air, water, land, sociodemographic, and built-environmental (e.g., housing and highway safety). For the random effects, we designed specific relationship matrices which determined correlation structures across individuals. For example, we specified genetic random effects with a multivariate normal distribution using the genetic relationship matrix (GRM), which indicated the relatedness between individuals. Similarly, environmental random effects shared by family members, couples, or siblings also follow distributions controlled by corresponding relationship matrices (see the Models part of the Complete details of materials and methods section). In addition, we considered the geographic random effects modelled by a Gaussian process (GP). Specifically, the GP's kernel defines the strength of the correlation between two patients' geographic random effects according to the distance between them. We assumed people living closely would share similar environments

and, thus, highly-correlated geographic random effects (see the Models part of the Complete details of materials and methods section).

We also considered the interaction between genetic and environmental random effects. By forwardly adding variables, we constructed a group of models from the simple to the more complex ones. Table 4.1 summarizes what additive effects are considered in each (named) model. We categorized the models into two forward selection traces: Linear model 0 (LM0) ⟶ Linear model 1 (LM1) ⟶ Interaction model 1 (IM1) and Linear model 2 (LM2) ⟶ Interaction model 2 (IM2). Within each trace, the later models encompass all variables of preceding models. The forward selection traces allowed us to determine whether models with additional variables could describe the data better by comparing the information criterion estimates.

Finally, it was possible for us to define statistics that quantified how much outcome variation (logit-probability of disease) could be explained by these additive random effects. For example, heritability (h^2) corresponds to the genetic random effect. In Table 4.1, we show the statistics p^2, h^2, f^2, c^2, s^2, e^2, hf^2, hc^2, hs^2, and he^2, which represent how much outcome variation can be attributed to geographic or genetic factors, familial environments, couple-shared environments, sibling-shared environments, individually-independent environments, interactions between genetics and familial environments, interactions between genetics and couple-shared environments, interactions between genetics and sibling-shared environments, and interactions between genetics and individual environments. Please refer to the Models part of the Complete details of materials and methods section for the definitions.

Table 4.1 Model set-ups and statistics

Model	Fixed effects	Random effects	Statistics
Linear model 0	Demo + Env	G + E	h^2, e^2
Linear model 1	Demo + Env	Geo + G + E	p^2, h^2, e^2
Interaction model 1	Demo + Env	Geo + G + E + GE	p^2, h^2, e^2, he^2
Linear model 2	Demo + Env	Geo + G + F + C + S + E	p^2, h^2, f^2, c^2, s^2, e^2
Interaction model 2	Demo + Env	Geo + G + F + C + S + E + GF + GC + GS + GE	p^2, h^2, f^2, c^2, s^2, e^2, hf^2, hc^2, hs^2, he^2

The fixed-effect terms are
Sex + Age (Demo)
Environmental quality indices (Env)

The random effect terms are defined by the partition of the phenotype explained by:
Geo: geographic position, described by coordinates (latitude and longitude)
G: genetics
E: the individually independent environment
F: the environment shared by family members
C: the environment shared by couples
S: the environment shared by siblings
GE: the interaction between genetics and the individually independent environment
GF: the interaction between genetics and the family-shared environment
GC: the interaction between genetics and the couples-shared environment
GS: the interaction between genetics and the siblings-shared environment

4.3 RESULTS

We fit the models in a Markov chain Monte Carlo (MCMC) procedure with a Bayesian framework (see the Supplementary Materials) for the ten most common psychiatric disorders in our data: anxiety phobic disorder, depression, migraine, adjustment disorder, and substance abuse. Stratified by models, Table 4.2 shows the mean estimates of the heritability and environmental statistics indicating how much of the outcome variation can be explained by

individual environment (e^2), familial environment (f^2), environment shared by couples (c^2), siblings (s^2), and shared geographic location of residence (p^2). Table 4.2 also shows how much variation can be accounted for by the interactions between the above-mentioned environmental and genetic effects. The widely-applicable information criterion (WAIC [131]) in Table 4.2 is a Bayesian information criterion that estimates the out-of-sample prediction error. Like the more commonly known Akaike information criterion, AIC [132], the Bayesian WAIC rewards goodness of fit but penalizes more complex, parameter-rich, models. It is commonly used to compare a collection of competing models fit to the same data. Models with the smallest WAIC provide the optimum balance between complexity and explanatory powers.

Figure 4.1 shows heritability estimates juxtaposed with model-specific WAIC values, also provided in Table 4.2. For each psychiatric disease we discussed, we plotted a model selection graph on the top panel, indicating the WAIC change along two forward selection traces (the green and golden points). The bar plot on the bottom panel shows how much variance might be explained by each effect variable, grouped in four categories: heritability (gray bars), geographic location (violet bars), environmental factors (yellow-orange bars), and gene-environment interactions (blue bars).

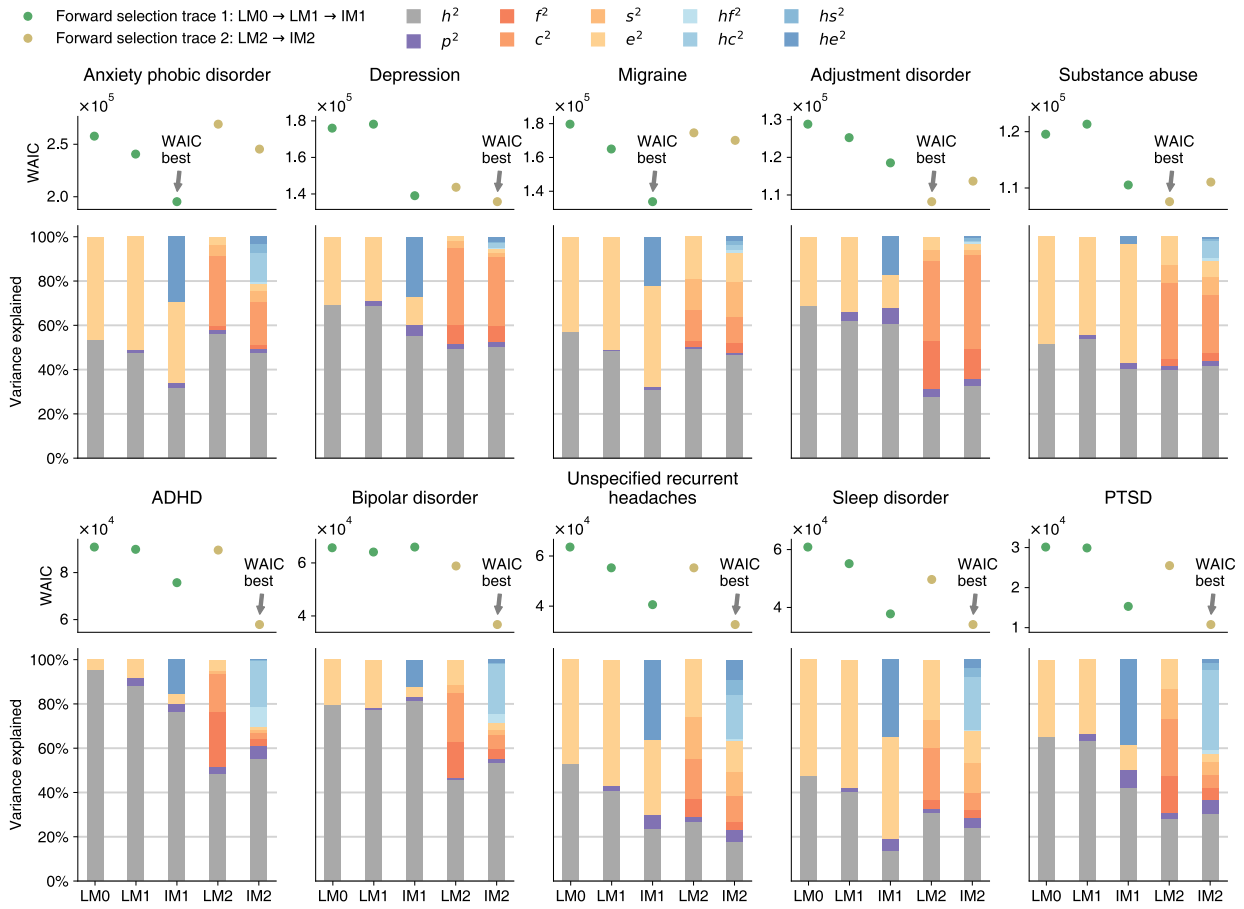

Figure 4.1 Model WAIC estimates and the mean estimates of heritability and environmental statistics

The bar plots show the posterior mean estimates of heritability (h^2, gray), variance explained by the geographic location (p^2, violet), variances explained by shared environments (f^2, familial, c^2, couple-shared, s^2, sibling-shared, painted in yellow-orange bars), and variances explained by gene-environment interactions (hf^2, gene-familial, hc^2, gene-couple-shared-environmental, hs^2, gene-sibling-shared environmental, painted in blue bars) given in the five models for the ten most diagnosed psychiatric diseases in our data. LM0, LM1, and LM2 are the three linear models that only consider the additive effects of genetics and shared environments as defined in Table 4.1. IM1 and IM2 are the two interaction models. Corresponding to their linear counterpart (LM1 and LM2), these two models take into account the gene-environment interactions as defined in Table 4.1. The five models form two forward selection traces: Linear model 0 (LM0) ⟶ Linear model 1 (LM1) ⟶ Interaction model 1 (IM1) and Linear model 2 (LM2) ⟶ Interaction model 2 (IM2). Within each trace, the later models encompass all variables of preceding models. The widely applicable information criterion (WAIC) rewards goodness of fit but penalizes more complex models. The lower the WAIC, the better the model. The scatter plot above each bar plot illustrates which model could be considered as the best one for each disease.

Our analyses allow us to draw the following conclusions.

First, interactions models (IM1 and IM2) fit data better than their linear counterparts (LM1 and LM2) in 17 out of 20 comparisons. The three exceptions are LM2 *vs*. IM2 in adjustment disorder and substance abuse, and LM1 *vs.* IM1 in bipolar disorder. The lower WAIC estimates suggest the interaction models fit the data better than the linear model with smaller information loss [131].

Correspondingly, gene-environment interactions seem to explain a significant portion of the total phenotypic variance. For example, gene-environment interactions explain over 20 percent of the total variance of anxiety / phobic disorder (blue bars in the upper left bar plot of Figure 4.1). Similarly, the interaction terms explain close to 30 percent of the total variance for unspecified recurrent headaches, sleep disorders, and posttraumatic stress disorder (PTSD).

By contrast, for some diseases, such as substance abuse and adjustment disorder, the gene-environment interactions do not appear to explain a significant portion of trait variance, so that interaction models are outcompeted by linear models according to WAIC.

Additionally, heritability estimates differ significantly between linear and interaction models. For attention deficit hyperactivity disorder (ADHD), the simplest G+E (LM0) model corresponds to an extremely high heritability estimate (95 percent). However, the best in WAIC model, IM2, incorporating all environmental and interaction effects, corresponds to much lower heritability of 55. Similarly, for bipolar disorder, the heritability is as high as 80 percent for LM0, while the WAIC-best model, IM2, it drops to 53 percent. This pattern is common across all the psychiatric disorders that we studied: The simpler models (LM0, LM1, and IM1) tend to provide higher heritability estimates compared to the more complex models (LM2 and IM2).

Besides the findings regarding the G-by-E effects and heritabilities, estimates of p^2, the proportion of variance contributed by the geographic position of each family, are remarkably variable across disorders. For anxiety / phobic disorder, migraine, substance abuse, and bipolar disorder, the geographic position does not explain much of the variance (under two percent in all models including the WAIC-best model). By contrast, for ADHD and PTSD, the geographic variation contributes more profoundly to the overall variance. The p^2 estimates for ADHD and PTSD are close to six percent, according to their WAIC-best model LM2. Figure 4.2 shows the geographic random effects' posterior mean estimates $f_p(\mathbf{x})$ (see Expression (4-8) in the Complete details of materials and methods, in the Models part) given the geographic coordinates vector $\mathbf{x}$; $f_p(\mathbf{x})$ was described by a Gaussian process assuming that adjacent geographical locations have closer-value random effects (assumption of smoothness).

For both ADHD and PTSD, all models across the complexity spectrum (LM1, IM1, LM2, and IM2) give nearly identical patterns of $f_p(\mathbf{x})$ estimates across the continental US (see Figure 4.2). This suggests that individuals living in the southern US and near the Great Lakes region may have been exposed to a higher risk of ADHD due to (yet unidentified) geographic/environmental effects. Similarly, residents of the US west coast and of the New England region bear higher PTSD rates. Note that we explicitly tried to account for environmental factors by incorporating geographically-associated environmental quality indices Therefore, we have already adjusted for known air, water, and land pollutants have already been adjusted for, see Table 4.1 and the Models part of the Complete details of materials and methods.

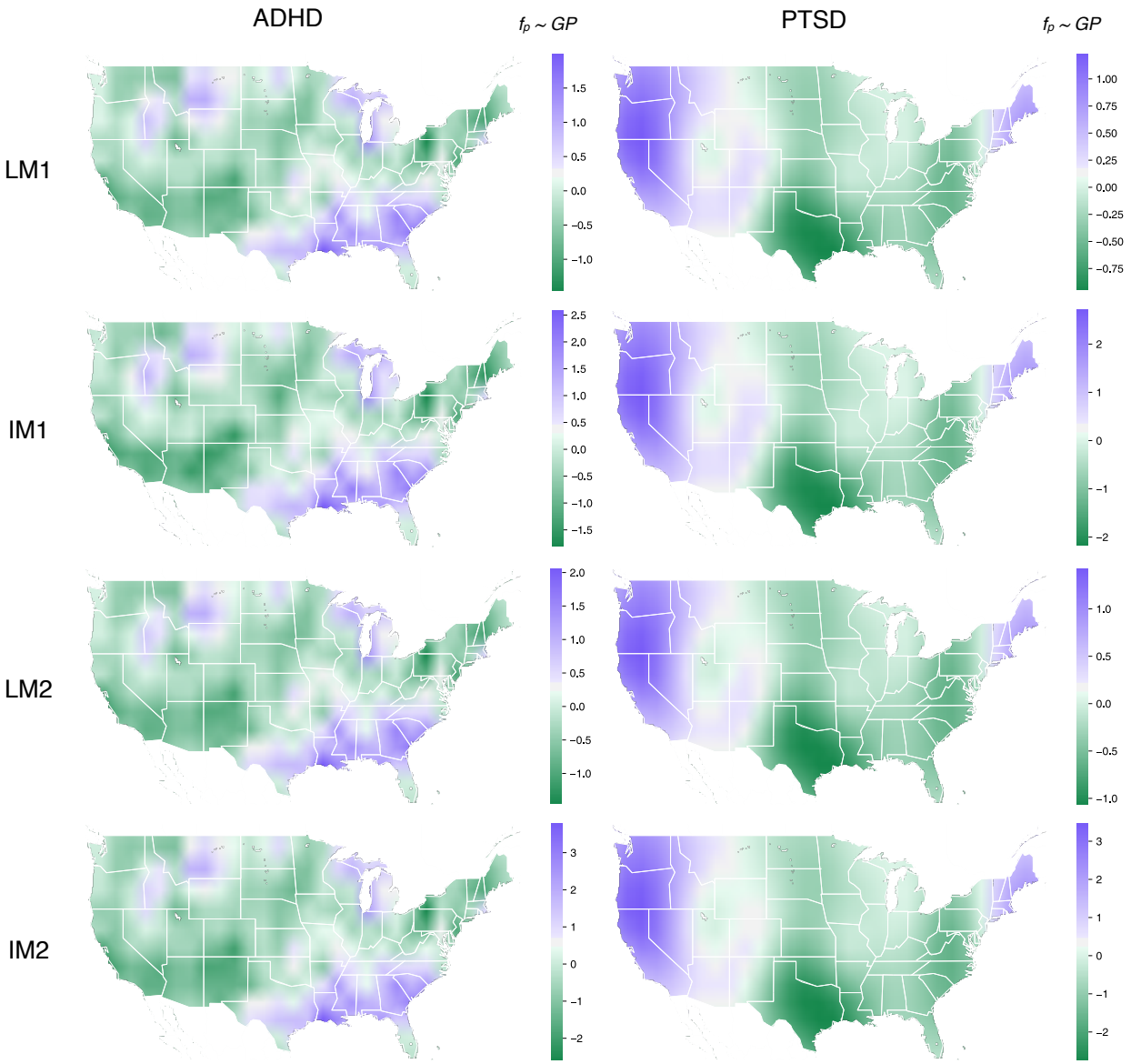

Figure 4.2 The mean estimates of the geographic random effects for ADHD and PTSD

Figure 4.2 Continued: The figure plots the posterior mean estimates of the geographic random effects for ADHD and PTSD ($f_p(\mathbf{x})$ (see Expression (4-8) in the Models part of the Complete details of materials and methods section). We modeled the geographic random effects $f_p(\mathbf{x})$ by a Gaussian process assuming adjacent geographical locations have closer-value random effects (assumption of smoothness). For both ADHD and PTSD, all models across the complexity spectrum (LM1, IM1, LM2, and IM2; note that LM0 does not incorporate the geographic random effect term) give nearly identical patterns of $f_p(\mathbf{x})$ estimates across the continental US. Our estimation process does include residents of Alaska and Hawaii, but the results were not shown here. That is because the discontinuity between the geographic locations disobeys the Gaussian process's presumptions. The poor extrapolation power of Gaussian process also makes it difficult to estimates the random effects related to outlying territories like Alaska and Hawaii. Our data does not record any residents of other non-continental US islands.

We also inspected the fixed effect estimates of demographic (sex and age) and environmental factors (fixed EQI scores) for these ten most common psychiatric diseases in our data. Figure 4.3 shows the posterior distribution of the log-odds (logit) change contributed by one's sex according to each disease's WAIC-best model. We code females in zero and males in one, so a greater log-odds change suggests a higher risk in males. After controlling for other effects, we observed high risks of ADHD and substance abuse in males and high risks of migraine, unspecified recurrent headaches, PTSD, etc., in females (Figure 4.3). Similarly, Figure 4.4 summarizes the log-odds change contributed by the numeric age according to each disease's WAIC-best model. Higher values in Figure 4.4 indicate higher risks predicted by the older age. As expected, ADHD and bipolar disorder are more common in the younger population, and sleep disorder is more common in seniors (Figure 4.4).

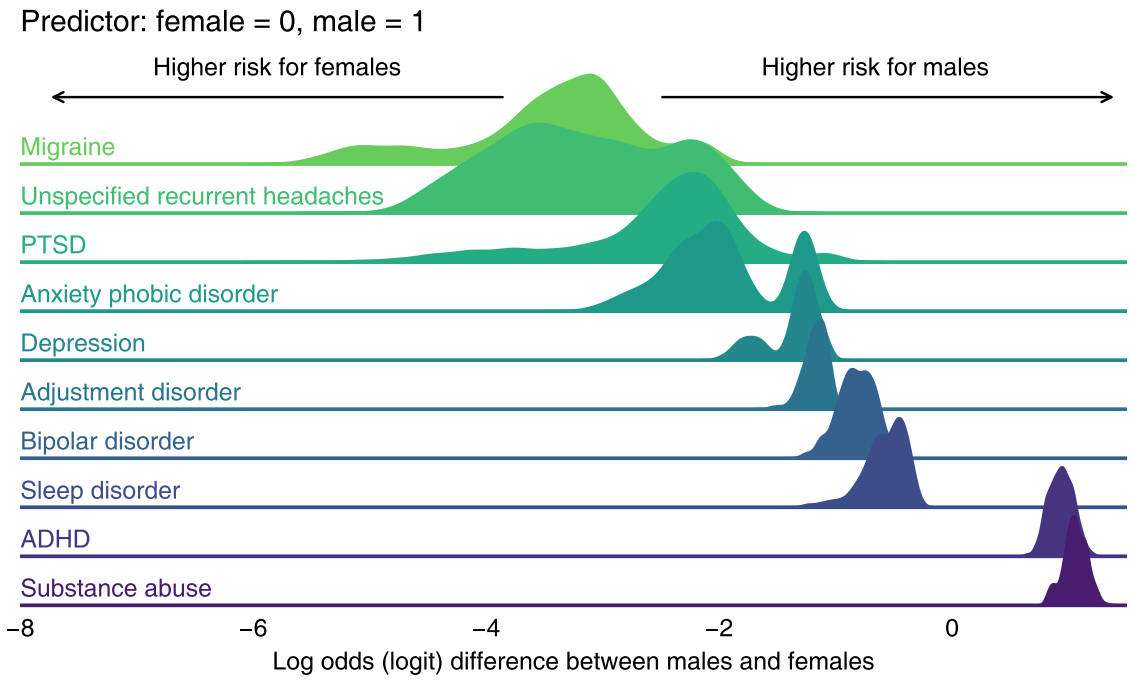

Figure 4.3 The posterior distribution of the log-odds (logit) change contributed by one's sex according to each disease's WAIC-best model

For each disease's WAIC-best model, this figure delineates the posterior distribution of the regression coefficient estimate associated with the dummy-coded sex (female = 0, male = 1). Because we used a logit link, the coefficient estimate represents the log odds difference of the risk (in terms of the probability of diagnosis) between males and females. High values in this figure indicate high risks for males, and low values indicate high risks for females – after controlling for other effects in the regression.

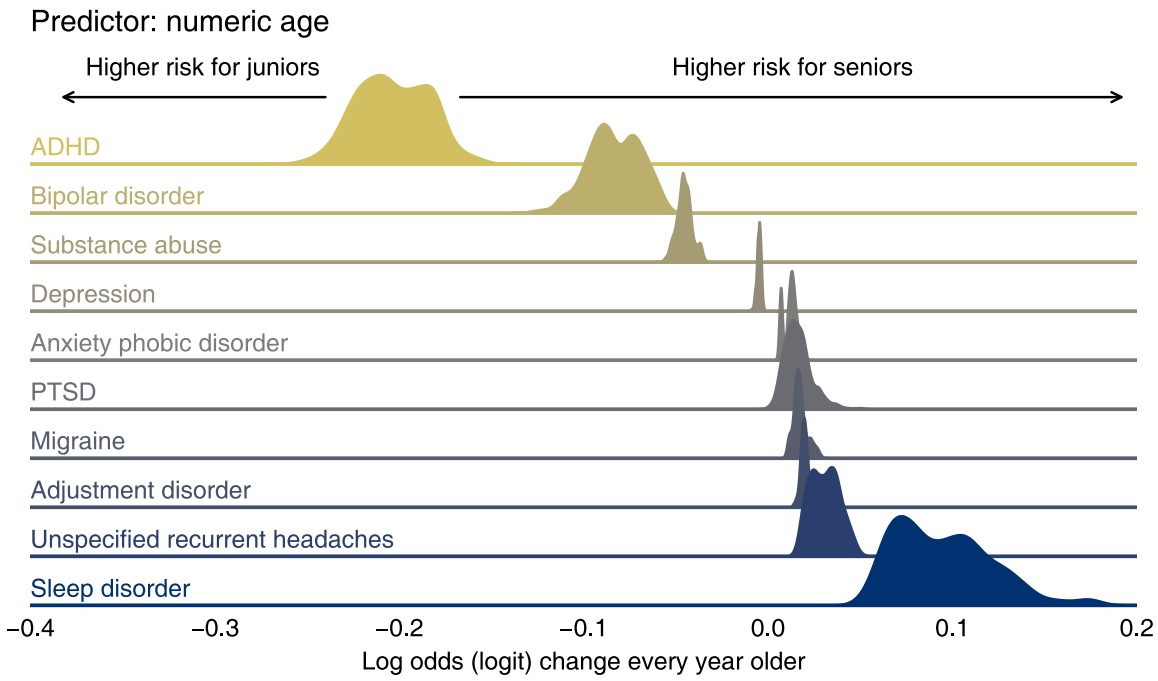

Figure 4.4 The posterior distribution of the log-odds (logit) change contributed by one's numeric age according to each disease's WAIC-best model

For each disease's WAIC-best model, this figure delineates the posterior distribution of the regression coefficient estimate associated with the numeric age. Because we used a logit link, the coefficient estimate represents the log odds change of the risk (in terms of the probability of diagnosis) every year older. High values in this figure indicate high risks for seniors, and low values indicate high risks for juniors – after controlling for other effects in the regression.

We then analyzed how the fixed-effect environmental qualities affect the log-odds of psychiatric diseases. Figure 4.5 shows the nonlinear effects estimated for the five EQI domains (air, water, land, sociodemographic, and built environment) based on each disease's WAIC-best model. Each EQI domain was summarized and standardized by a principal component analysis (PCA) procedure into one-dimensional PCA scores. Higher scores generally reflect worse environmental quality. To visualize the result, we converted the PCA scores to percentiles based on the score distribution in the whole population and used the 2.5th to 97.5th percentiles as the x-axis limits of the plots. The blue segments of the lines indicate EQI regions that do not change

the log odds significantly. The olive segments are EQI regions that influence the log odds of diseases significantly. We also fit a simple linear regression for each curve in Figure 4.5 and plotted the slopes to better visualize the overall trends associated with each EQI domain (Figure 4.6). The figures demonstrate the miscellaneous effects of various environmental qualities. For air quality, worse environments (high scores) may be associated with increased risks of depression, adjustment disorder, ADHD, and bipolar disorder (Figure 4.5 and Figure 4.6, the first columns). On the contrary, residents in areas of worse air seem to have lower risks of anxiety phobic disorder, unspecified recurrent headaches, sleep disorder, and PTSD (Figure 4.5 and Figure 4.6, the first columns). Compared to air quality, water quality has only minor effects on the log odds of the studied psychiatric diseases (Figure 4.5 and Figure 4.6, the second columns). Low-quality land environments may be associated with higher risks of adjustment disorder and unspecified recurrent headaches (Figure 4.5 and Figure 4.6, the third column). In addition, lower sociodemographic quality scores may be associated with elevated risks in depression, adjustment disorder, ADHD, and decreased incidence in unspecific recurrent headaches (Figure 4.5 and Figure 4.6, the fourth columns). Finally, the last columns of Figure 4.5 and Figure 4.6 suggest the broad impact of built-environmental quality, which appears linked to significantly higher risks in depression, adjustment disorder, substance abuse, ADHD, bipolar disorder, and sleep disorder.

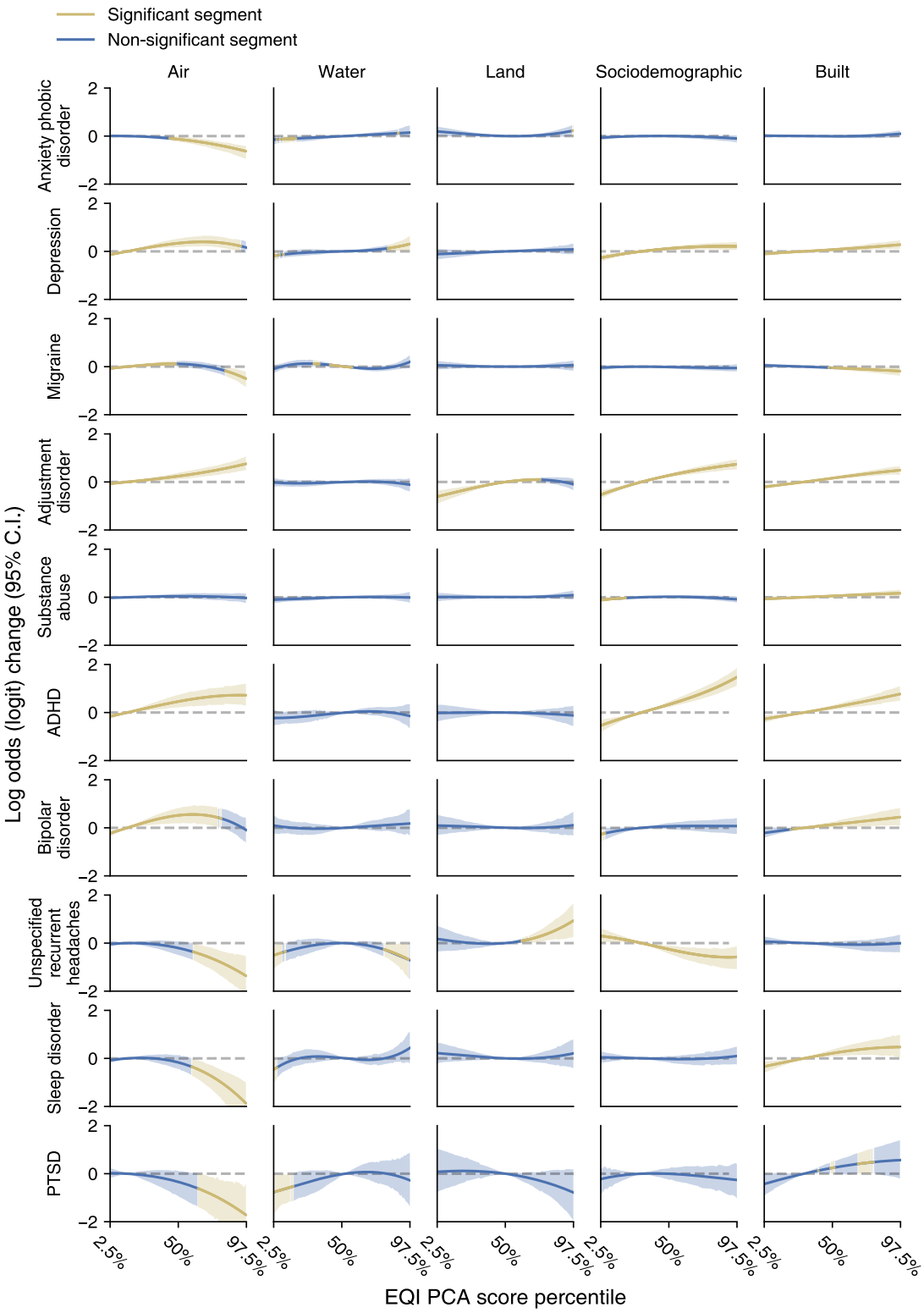

Figure 4.5 The nonlinear effects estimated for the five EQI domains (air, water, land, sociodemographic, and built environment) given by each disease's WAIC-best model

Figure 4.5 Continued: Each EQI domain contains multiple relevant variables. A PCA procedure was employed to summarize and standardize domain-specific variables into one-dimensional PCA scores. According to the publisher (EPA), higher scores generally reflect worse environmental quality. We converted the PCA scores to percentiles based on the score distribution in the whole population. We used the 2.5th to 97.5th percentiles as the plot's x-axis limits of the plots. In the line plots, the blue segments indicate EQI regions that do not change the log odds significantly (statistically equal to zero, 95 percent credible intervals). The olive segments are EQI regions that significantly alter the log odds of diseases (statistically non-zero regions, 95 percent credible intervals).

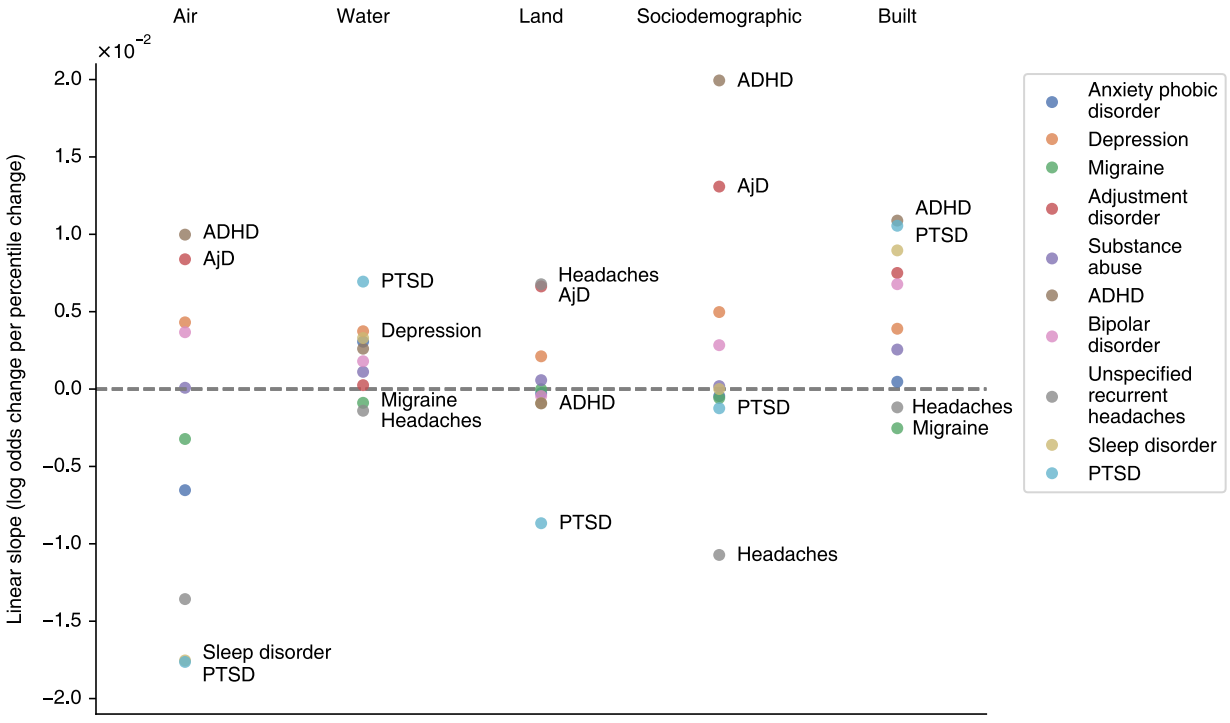

Figure 4.6 Linear slopes of the EQI effect curves

We fit a simple linear regression for each curve in Figure 4.5 and plotted the slopes to visualize the overall trends associated with each EQI domain (air, water, land, sociodemographic, and built environment). For each domain, the top two and bottom two diseases with the largest absolute linear slopes are marked. "AjD" stands for adjustment disorder, and "Headaches" stands for unspecified recurrent headaches.

Table 4.2 Mean estimates of the heritability and environmental statistics

ADHD: Attention Deficit Hyperactivity Disorder
PTSD: Posttraumatic Stress Disorder

The statistics are defined by the partition of the phenotype explained by
p^2: geographic position, described by coordinates (latitude and longitude)
h^2: genetics
e^2: the individually independent environment
f^2: the environment shared by family members
c^2: the environment shared by couples
s^2: the environment shared by siblings
he^2: the interaction between genetics and the individually independent environment
hf^2: the interaction between genetics and the family-shared environment
hc^2: the interaction between genetics and the couples-shared environment
hs^2: the interaction between genetics and the siblings-shared environment

WAIC
The widely-applicable information criterion rewards goodness of fit but penalizes more complex models. The lower the WAIC, the better the model.

Anxiety phobic disorder											
	p^2	h^2	f^2	c^2	s^2	e^2	hf^2	hc^2	hs^2	he^2	WAIC
LM0		54%				46%					257729.31
LM1	1%	47%				51%					240648.05
IM1	2%	32%				36%				29%	195306.87
LM2	2%	56%	2%	31%	5%	3%					269150.17
IM2	2%	47%	2%	19%	5%	3%	1%	13%	4%	3%	245390.66
Depression											
	p^2	h^2	f^2	c^2	s^2	e^2	hf^2	hc^2	hs^2	he^2	WAIC
LM0		69%				31%					176007.77
LM1	3%	69%				29%					178182.13
IM1	5%	55%				13%				27%	139055.19
LM2	2%	50%	9%	35%	3%	2%					143671.22
IM2	2%	50%	7%	31%	2%	2%	0%	2%	1%	2%	135787.57

Table 4.2 Continued

Migraine

	p^2	h^2	f^2	c^2	s^2	e^2	hf^2	hc^2	hs^2	he^2	WAIC
LM0		57%				43%					179692.88
LM1	1%	48%				51%					165004.39
IM1	1%	31%				46%				22%	133837.58
LM2	1%	49%	3%	14%	14%	19%					174572.51
IM2	1%	47%	5%	12%	16%	13%	1%	2%	2%	2%	170124.00

Adjustment Disorder

	p^2	h^2	f^2	c^2	s^2	e^2	hf^2	hc^2	hs^2	he^2	WAIC
LM0		69%				31%					128812.67
LM1	4%	62%				34%					125231.76
IM1	8%	61%				15%				17%	118529.41
LM2	4%	28%	22%	36%	5%	6%					108221.89
IM2	3%	33%	14%	43%	2%	3%	1%	1%	1%	1%	113680.71

Substance Abuse

	p^2	h^2	f^2	c^2	s^2	e^2	hf^2	hc^2	hs^2	he^2	WAIC
LM0		52%				48%					119551.43
LM1	2%	54%				44%					121334.21
IM1	2%	41%				54%				3%	110533.35
LM2	2%	40%	3%	34%	8%	13%					107571.67
IM2	2%	42%	4%	26%	8%	7%	2%	8%	1%	1%	111049.17

Table 4.2 Continued

ADHD

	p^2	h^2	f^2	c^2	s^2	e^2	hf^2	hc^2	hs^2	he^2	WAIC
LM0		95%				5%					90855.79
LM1	3%	88%				8%					89893.81
IM1	4%	77%				4%				16%	75706.35
LM2	3%	48%	25%	17%	2%	5%					89568.83
IM2	6%	55%	3%	2%	2%	1%	9%	21%	0%	0%	57894.43

Bipolar Disorder

	p^2	h^2	f^2	c^2	s^2	e^2	hf^2	hc^2	hs^2	he^2	WAIC
LM0		80%				20%					65719.01
LM1	1%	77%				22%					64106.44
IM1	1%	82%				5%				12%	65983.85
LM2	1%	46%	16%	22%	4%	11%					58846.71
IM2	2%	53%	4%	7%	2%	3%	4%	22%	0%	2%	36754.75

Unspecified Recurrent Headaches

	p^2	h^2	f^2	c^2	s^2	e^2	hf^2	hc^2	hs^2	he^2	WAIC
LM0		53%				47%					63569.71
LM1	2%	41%				57%					55287.67
IM1	6%	24%				34%				36%	40592.46
LM2	2%	27%	8%	18%	19%	26%					55280.79
IM2	5%	18%	4%	12%	11%	14%	1%	20%	7%	9%	32654.79

Table 4.2 Continued

Sleep Disorder											
	p^2	h^2	f^2	c^2	s^2	e^2	hf^2	hc^2	hs^2	he^2	WAIC
LM0		47%				53%					60866.74
LM1	2%	40%				58%					55129.58
IM1	6%	14%				46%				35%	37796.65
LM2	2%	31%	4%	23%	13%	27%					49665.71
IM2	4%	24%	4%	8%	14%	14%	0%	24%	4%	4%	34098.02
PTSD											
	p^2	h^2	f^2	c^2	s^2	e^2	hf^2	hc^2	hs^2	he^2	WAIC
LM0		65%				35%					30127.48
LM1	3%	63%				34%					29902.80
IM1	8%	42%				11%				38%	15312.56
LM2	3%	28%	17%	26%	13%	13%					25494.21
IM2	6%	30%	5%	6%	6%	4%	1%	36%	3%	2%	10777.49

4.4 DISCUSSION

For selectionists working with domesticated species, the notion of heritability is an instrument, while the remaining environmental variability is a nuisance. In case of human disease, the situation could be quite the opposite. Associations between genetic variation and disease may lead to genetic tests and indicate molecular pathways that should be targeted by treatment. However, for most complex diseases genetic variation is not very instrumental in disease treatment; predicted genetic predisposition to complex disease is often seen as a verdict rather than an actionable advice.

This is not the case with environmental predisposition to disease: If we knew environmental stimuli that might trigger particular disease, we could take preventive measures or design life strategies aimed at avoiding dangerous triggers. Similarly, with interactions between genetic and environmental signals, this information can become practical and clinically actionable. If we are able to ascertain a catalog of genetic variants interacting with specific environmental stimuli, we could design personalized environmental plans for the patient at risk.

For the ten most diagnosed psychiatric conditions in our data, we designed a group of Bayesian regression models that incorporate various genetic, environmental, and demographic effects. By comparing models with G-by-E interactions to their linear counterparts, we established the existence of interactions with considerable strength between the genetic variation and environmental stimuli affecting the manifestation of psychiatric disorders. We also jointly estimated the effects of the geographic location of residency, sex, age, and pre-determined environmental qualities through the regression procedure.

Unexplained geographic-environmental variations (Figure 4.2) can serve as a basis for future association studies. For example, we can scan myriad environmental factors to explain the difference in ADHD rates in California vs. Georgia. Furthermore, it is reassuring to "re-discover" known factors affecting the phenotypic variation of psychiatric disorders. For example, ADHD is associated with younger age, while sleep disorder is a feature of much older people (Figure 4.4) [19, 20]. Similar, sex bias in diseases is well-reported: Migraine affects females more often, while substance abuse is biased towards males (Figure 4.3) [21-24].

Fixed effects estimated for environmental quality indices (EQIs, see Figure 4.5) indicate curious associations between the quality of immediate environment and the rate of psychiatric disorders. In the observational setting of our analyses, we can talk about the predictive effects of

environmental quality on disease rates but avoid any statements implying causality. For example, we can observe that deteriorating air quality is associated with increased rates of ADHD and decreased rates of sleep disorders. Specially designed observational studies (for example, involving an instrumental variable) can closely approximate randomized clinical trial causality analysis, but we will defer this to follow-up studies.

Our study finds substantial gene-environment interactions in common psychiatric diseases through elegant methods. Yet, there remain many unresolved puzzles. First, it is not definite what, precisely, the G-by-E effects represent. One well-liked interpretation reads the gene-environment interaction as heterogeneous genetic effects conditioned on different environments or exterior exposures [25-27]. This interpretation provides a solution for the "missing heritability" puzzle, given that the G-by-E effect may comprise a considerable part of the heritable factor, undetectable in former linear models [26, 27]. The one-directional interpretation (E influences G, heterogeneous genetics in response to the environment) is far from the only plausible interpretation. Conversely, it is also possible that genetics may affect the environment or the exterior exposure. Certain genotypes might influence one's preference for diets and living conditions. People are also continuously transforming their surrounding environments – in which process, behavior-related genes might be involved.

We also only estimated the across-the-genome G-by-E effects with a very simplified mathematical form and mixed-effects model. Our analyses did not determine what genetic variants and environmental factors participate in the G-by-E mechanism. It is also unclear if the gene-environment interactions act as a simple $G \times E$ multiplication. Specific genes and environmental factors could engage themselves in the G-by-E effects in more complicated forms. Epistasis and environment-environmental effects could join in to produce G-by-G-by-E or G-by-

E-by-E effects. The next important target would be to map genic variant–environmental stimulus pairs for each disorder and ascertain the mechanism of the interactions. The goal is likely to require the generation of new data approaching the "ideal" data set described in our introduction or bench experiments focusing on particular genes.

Additionally, our methods may suffer from many known problems and give erroneous estimates. We applied a Bayesian MCMC procedure with multiple checks to ensure accurate estimation. First, we warmed up the Bayesian sampling using variational inference (see the Bayesian Inference part of the Complete details of materials and methods) [28], which essentially fit the mixed-effects model similar to the Frequentist's optimization method (e.g., maximum likelihood). In addition, the state-of-the-art sampler we employed (the No-U-Turn Sampler, NUTS [29]) has been proven to be efficient in exploring high-dimensional parameter space with large curvature and complex geometry. We also sampled four MCMC chains independently and monitored the convergence and space exploration actively by comparing the chains. Despite applying all the tactics, we still could not assert that our model calculated perfect posterior distributions for all parameters, especially if we acknowledge that it is difficult to separate the random effects (G, F, C, S, E in Table 4.1) completely in a large regression analysis of almost half million individuals.

Nonetheless, challenges imply promising opportunities. The present work has demonstrated that much is waiting for investigation. We can compute the interactions between genetics and concrete environmental indices, such as geographic locations, pollutants, and socioeconomic factors. This analysis could produce more interpretable results than ours, which only considered interactions between random, hard-to-interpret genetic and environmental effects. Moreover, we can exploit experiments and sequencing data to dig deep into interactions

between particular genes and exterior conditions. By designing specific assays and methods, it is also possible to determine the interactions' directions and mechanisms. Does the G act as a modifier for the environmental effects E? Does the E act as a modifier for the G? How are epistasis and E-by-E effects involved? Future efforts could answer these questions and allow us to map out strategic plans for preventive and precision medicine.

4.5 COMPLETE DETAILS OF MATERIALS AND METHODS

4.5.3 Models

4.5.3.1 Linear model 0 (LM0)

For a phenotypic disease, we assumed that the probability of presenting this disease is p. The logit-probability, defined as $l = \log\left(\frac{p}{1-p}\right)$, could be expressed in an additive mixed-effects model:

$$l = X\beta + f_q(\mathbf{q}) + G + E, \tag{4-1}$$

where X is the design matrix of the demographic fixed effects, including the effects of age and sex. For the fixed effect controlled by environmental quality indices $\mathbf{q} = (q_1, q_2, \dots, q_m)$, we assumed a polynomial model:

$$f_q(\mathbf{q}) = f_q^{(1)}(q_1) + f_q^{(2)}(q_2) + \cdots + f_q^{(m)}(q_m). \tag{4-2}$$

Our model includes five ($m = 5$) different types of environmental qualities quantified by summary indices: air, water, land, sociodemographic, and build-environment domain. We set the degree of the polynomial function f_q and its components $f_q^{(1)} \dots f_q^{(m)}$ to be three (cubic).

G is the genetic effect contributing to the phenotype, and E is the individually-independent environmental effect. For an individual group, their genetic effects were associated

by the genetic relationship matrix (GRM, Σ_G). As an example, the GRM for a family of two parents and one child should be close to

$$\Sigma_G = \begin{bmatrix} 1.0 & 0.0 & 0.5 \\ 0.0 & 1.0 & 0.5 \\ 0.5 & 0.5 & 1.0 \end{bmatrix}, \tag{4-3}$$

where the first two rows and columns represent the parents, and the last row and column represents the child. Then, we assumed the genetic effects followed a multivariate normal distribution:

$$G \sim \text{MvNormal}(\text{mean} = 0, \text{cov} = \sigma_G{}^2 \Sigma_G). \tag{4-4}$$

The individually-independent environmental effects also followed a multivariate normal distribution with its covariance equal to a multiple of the identity matrix I:

$$E \sim \text{MvNormal}(\text{mean} = 0, \text{cov} = \sigma_E{}^2 I). \tag{4-5}$$

In these expressions, $\sigma_G{}^2$ and $\sigma_E{}^2$ are the constants we wanted to find. For a population, the variance-covariance of l is

$$\text{Var}(l) = \sigma_G{}^2 \Sigma_G + \sigma_E{}^2 I. \tag{4-6}$$

Finally, we defined the heritability h^2 and the independent environmental factor e^2 as

$$h^2 = \frac{\sigma_G{}^2}{\sigma_G{}^2 + \sigma_E{}^2}, \tag{4-7a}$$

$$e^2 = \frac{\sigma_E{}^2}{\sigma_G{}^2 + \sigma_E{}^2}. \tag{4-7b}$$

4.5.3.2 Linear model 1 (LM1)

Besides the effects contributed by the basic demographic information (sex and age) and environmental quality, we acknowledged that geographic position might also play a part in disease etiology.

Given a patient's dwelling's latitude and longitude (coordinates) $\mathbf{x} = (x_1, x_2)$, the random effect, as a part of the logit-probability l, is $f_p(\mathbf{x})$ that follows a Gaussian process (GP):

$$f_p(\mathbf{x}) \sim \mathcal{GP}\big(\text{mean} = 0, \text{cov} = k(\mathbf{x}, \mathbf{x}')\big). \tag{4-8}$$

The Gaussian process model constrains the distribution of $f_p(\mathbf{x})$ so that the joint distribution of two data points $f_p(\mathbf{x})$ and $f_p(\mathbf{x}')$ is multivariate normal:

$$\begin{bmatrix} f_p(\mathbf{x}) \\ f_p(\mathbf{x}') \end{bmatrix} \sim \text{MvNormal}\left(\text{mean} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \text{cov} = \begin{bmatrix} k(\mathbf{x}, \mathbf{x}) & k(\mathbf{x}, \mathbf{x}') \\ k(\mathbf{x}, \mathbf{x}') & k(\mathbf{x}', \mathbf{x}') \end{bmatrix}\right). \tag{4-9}$$

Therefore, if we choose the kernel function $k(\cdot,\cdot)$ appropriately, we can fit a function $f_p(\mathbf{x})$ that makes two random effects $f_p(\mathbf{x})$ and $f_p(\mathbf{x}')$ correlated according to the proximity of coordinates $\mathbf{x}$ and $\mathbf{x}'$. Here, we used the Exponentiated Quadratic kernel function, which is commonly accepted in geo-statistics and applications in a smooth metric space:

$$k(\mathbf{x}, \mathbf{x}') = {\sigma_P}^2 \cdot \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\ell^2}\right), \tag{4-10}$$

where ${\sigma_P}^2$ and ℓ were the scale parameters we fit through the Markov chain Monte Carlo (MCMC) process.

In all, as a forward addition to the simplest linear model 0 (LM0), we expressed the logit-probability of presenting a disease for an individual as:

$$\begin{aligned} l &= \beta_0 + \beta_{\text{sex}} \cdot \text{Sex} + \beta_{\text{age}} \cdot \text{Age} + f_q(\mathbf{q}) + f_p(\mathbf{x}) + G + E \\ &= X\beta + f_q(\mathbf{q}) + f_p(\mathbf{x}) + G + E \end{aligned} \tag{4-11}$$

The variance-covariance of l for a group of individuals was

$$\text{Var}(l) = {\sigma_P}^2 K_P + {\sigma_G}^2 \Sigma_G + {\sigma_E}^2 I, \tag{4-12}$$

Where ${\sigma_P}^2 K_P$ was the covariance matrix associating individuals based on their geographic locations. It can be derived from the kernel function, Expression (4-10), as we did in Expression (4-9).

Consequently, the geographic position factor p^2 that quantified how much variation could be explained by one's coordinates, the heritability h^2, and the independent environmental factor e^2 were

$$p^2 = \frac{{\sigma_P}^2}{{\sigma_P}^2 + {\sigma_G}^2 + {\sigma_E}^2}, \tag{4-13a}$$

$$h^2 = \frac{{\sigma_G}^2}{{\sigma_P}^2 + {\sigma_G}^2 + {\sigma_E}^2}, \tag{4-13b}$$

$$e^2 = \frac{{\sigma_E}^2}{{\sigma_P}^2 + {\sigma_G}^2 + {\sigma_E}^2}. \tag{4-13c}$$

4.5.3.3 Interaction Model 1 (IM1)

We then considered the interaction between the genetic effect and the environmental effect. The logit-probability of having a disease can be expressed as

$$l = X\beta + f_q(\mathbf{q}) + f_p(\mathbf{x}) + G + E + k_{GE} \cdot G \cdot E, \tag{4-14}$$

The scale factor k_{GE} determines how much effect the interaction contributes to the phenotype. G and E were the same genetic and environmental effects as they were given in the linear models 0 and 1. The variance-covariance of l was

$$\mathrm{Var}(l) = {\sigma_P}^2 K_P + {\sigma_G}^2 \Sigma_G + {\sigma_E}^2 I + {k_{GE}}^2 {\sigma_G}^2 {\sigma_E}^2 \Sigma_G \odot I \tag{4-15}$$

, if we assume the genetic effect and the environmental effect are statistically independent. The operator $\odot$ represents the elementwise (Hadamard) product. The geographic position factor p^2, the heritability h^2, and the independent environmental factor e^2 are

$$p^2 = \frac{{\sigma_P}^2}{{\sigma_P}^2 + {\sigma_G}^2 + {\sigma_E}^2 + {k_{GE}}^2 {\sigma_G}^2 {\sigma_E}^2}, \tag{4-16a}$$

$$h^2 = \frac{{\sigma_G}^2}{{\sigma_P}^2 + {\sigma_G}^2 + {\sigma_E}^2 + {k_{GE}}^2 {\sigma_G}^2 {\sigma_E}^2}, \tag{4-16b}$$

$$e^2 = \frac{{\sigma_E}^2}{{\sigma_P}^2 + {\sigma_G}^2 + {\sigma_E}^2 + {k_{GE}}^2 {\sigma_G}^2 {\sigma_E}^2}. \tag{4-16c}$$

Furthermore, the interaction between genetics and the environment can also explain the variance of the logit-probability. We have defined a new interaction factor

$$he^2 = \frac{{k_{GE}}^2 {\sigma_G}^2 {\sigma_E}^2}{{\sigma_P}^2 + {\sigma_G}^2 + {\sigma_E}^2 + {k_{GE}}^2 {\sigma_G}^2 {\sigma_E}^2}, \tag{4-17}$$

4.5.3.4 Linear Model 2 (LM2)

The linear models 0 and 1 incorporate the individually-independent environmental effect E only. However, because family members, couples, and siblings may have shared similar behaviors and milieus, we also included other types of environmental effects in our model. Here, the linear model 2 considers the family effect F, the couple effect C, and the sibling effect S in addition to the individually-independent environmental effect E:

$$l = X\beta + f_q(\mathbf{q}) + f_p(\mathbf{x}) + G + F + C + S + E. \tag{4-18}$$

These additional environmental effects were given by multivariate normal distributions

$$F \sim \text{MvNormal}(\text{mean} = 0, \text{cov} = {\sigma_F}^2 \Sigma_F), \tag{4-19a}$$

$$C \sim \text{MvNormal}(\text{mean} = 0, \text{cov} = {\sigma_C}^2 \Sigma_C), \tag{4-19b}$$

$$S \sim \text{MvNormal}(\text{mean} = 0, \text{cov} = {\sigma_S}^2 \Sigma_S). \tag{4-19c}$$

The relationship matrices Σ_F, Σ_C, and Σ_S were determined by how much of the kinship-specific environment they shared on average. The relationship matrices for a family of two parents (first two rows and columns) and two children (last two rows and columns) are

$$\Sigma_F = \begin{bmatrix} 1.0 & 1.0 & 1.0 & 1.0 \\ 1.0 & 1.0 & 1.0 & 1.0 \\ 1.0 & 1.0 & 1.0 & 1.0 \\ 1.0 & 1.0 & 1.0 & 1.0 \end{bmatrix}, \tag{4-20a}$$

$$\Sigma_C = \begin{bmatrix} 1.0 & 1.0 & 0.0 & 0.0 \\ 1.0 & 1.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 1.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 1.0 \end{bmatrix}, \tag{4-20b}$$

$$\Sigma_S = \begin{bmatrix} 1.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 1.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 1.0 & 1.0 \\ 0.0 & 0.0 & 1.0 & 1.0 \end{bmatrix}. \tag{4-20c}$$

The variance-covariance of l can then be estimated:

$$\mathrm{Var}(l) = \sigma_P{}^2 K_P + \sigma_G{}^2 \Sigma_G + \sigma_F{}^2 \Sigma_F + \sigma_C{}^2 \Sigma_C + \sigma_S{}^2 \Sigma_S + \sigma_E{}^2 I. \tag{4-21}$$

We defined the heritability h^2 and other statistics specifying environmental effects as:

$$p^2 = \frac{\sigma_P{}^2}{\sigma_P{}^2 + \sigma_G{}^2 + \sigma_F{}^2 + \sigma_C{}^2 + \sigma_S{}^2 + \sigma_E{}^2}, \tag{4-22a}$$

$$h^2 = \frac{\sigma_G{}^2}{\sigma_P{}^2 + \sigma_G{}^2 + \sigma_F{}^2 + \sigma_C{}^2 + \sigma_S{}^2 + \sigma_E{}^2}, \tag{4-22b}$$

$$f^2 = \frac{\sigma_F{}^2}{\sigma_P{}^2 + \sigma_G{}^2 + \sigma_F{}^2 + \sigma_C{}^2 + \sigma_S{}^2 + \sigma_E{}^2}, \tag{4-22c}$$

$$c^2 = \frac{\sigma_C{}^2}{\sigma_P{}^2 + \sigma_G{}^2 + \sigma_F{}^2 + \sigma_C{}^2 + \sigma_S{}^2 + \sigma_E{}^2}, \tag{4-22d}$$

$$s^2 = \frac{\sigma_S{}^2}{\sigma_P{}^2 + \sigma_G{}^2 + \sigma_F{}^2 + \sigma_C{}^2 + \sigma_S{}^2 + \sigma_E{}^2}, \tag{4-22e}$$

$$e^2 = \frac{{\sigma_E}^2}{{\sigma_P}^2 + {\sigma_G}^2 + {\sigma_F}^2 + {\sigma_C}^2 + {\sigma_S}^2 + {\sigma_E}^2}. \quad \text{(4-22f)}$$

4.5.3.5 Interaction Model 2 (IM2)

Similar to what we did in the experimental models 1 and 2, we added interaction terms into the above-defined linear model 2, thus the logit-probability of having a disease is

$$\begin{aligned} l = X\beta + f_q(\mathbf{q}) + f_p(\mathbf{x}) + G + F + C + S + E + \\ k_{GF} \cdot G \cdot F + k_{GC} \cdot G \cdot C + \\ k_{GS} \cdot G \cdot S + k_{GE} \cdot G \cdot E. \end{aligned} \quad \text{(4-23)}$$

The variance-covariance of l is

$$\begin{aligned} \mathrm{Var}(l) = {\sigma_P}^2 K_P + {\sigma_G}^2 \Sigma_G + {\sigma_F}^2 \Sigma_F + {\sigma_C}^2 \Sigma_C + {\sigma_S}^2 \Sigma_S + {\sigma_E}^2 I + \\ {k_{GF}}^2 {\sigma_G}^2 {\sigma_F}^2 \Sigma_G \odot \Sigma_F + {k_{GC}}^2 {\sigma_G}^2 {\sigma_C}^2 \Sigma_G \odot \Sigma_C + \\ {k_{GS}}^2 {\sigma_G}^2 {\sigma_S}^2 \Sigma_G \odot \Sigma_S + {k_{GE}}^2 {\sigma_G}^2 {\sigma_E}^2 \Sigma_G \odot I \,. \end{aligned} \quad \text{(4-24)}$$

We defined the heritability h^2 and statistics specifying the environmental effects (geographic position: p^2, family: f^2, couple: c^2, sibling: s^2, independent: e^2) as we did in Expressions (4-16) and (4-22). Likewise, statistics quantifying the interactions (genetics-family: hf^2, genetics-couple: hc^2, genetics-sibling: hs^2, genetics-individual-environment: he^2) are given according to their partitions in $\mathrm{Var}(l)$, similar to the Expression (4-17).

4.5.4 Bayesian Inference

4.5.4.1 Priors

To find the posterior distribution of a parameter θ under a Bayesian framework, we first needed to specify the prior according to the Bayes theorem $P(\theta|\mathrm{Data}) \propto P(\theta) \times P(\mathrm{Data}|\theta) = \mathrm{Prior} \times \mathrm{Likelihood}$. We used the horseshoe shrinkage prior[133] for the fixed-effect parameters

(as in the demographic predictor parameter β in the Expression (4-1) and the polynomial function parameter in the Expression (4-2)), so sparsity and regularization were imposed to avoid overly complicated models. For parameters restricted to be positive, such as the variance scale factor ${\sigma_G}^2$ and ${\sigma_E}^2$ in the Expression (4-6), we used the zero-avoiding Gamma prior, as recommended by Chung et al. [134] and the STAN prior choice recommendations [135].

4.5.4.2 Sampling

After specifying the hierarchical models with priors and likelihood functions, Bayesians rely on a sampling algorithm (sampler) to draw from the posterior distribution and approximate the posterior efficiently. In the present study, we used Hoffman and Gelman's No-U-Turn sampler (NUTS) [29] designed for high-dimensional, ill-shaped target distributions. The No-U-Turn sampler tunes the step size automatically and directs sampling by looking to the gradient information. Compared to vanilla Markov chain Monte Carlo methods, the No-U-Turn sampler [29] depicts complex posterior distributions more quickly. We initialized the sampling process using automatic differentiation variational inference (ADVI) [28], providing the MCMC an optimized start similar to the Frequentist's maximum likelihood methods [33].

CHAPTER 5. CONCLUSION

Studies of etiology uncover the hidden cause that lead to the illness and potential ways to treat or even prevent the conditions. Understanding disease etiology has, therefore, paramount importance for clinical medicine. Healthcare workers can then take preventive measures and map out precise individualized strategies fitting the patient's genetic variation and environmental exposures. There are undoubtedly multiple levels of disease description ranging from cellular and molecular mechanisms to socioeconomic forces in the whole human population. The classical equation, $P(t, \mathbf{x}) = G + E(t, \mathbf{x})$, describes how genetic and ever-changing environmental factors contribute to phenotypic traits, where t and $\mathbf{x}$ represent temporal and spatial variables. Due to the availability of modern computational tools and extensive observational data, more complicated and subtle effects, such as the gene-environment interactions, can be added and inferred by our models.

We began our analysis by estimating the potential health implications of changing time biannually to increase the amount of natural light during work hours (daylight-saving time, DST), as an example of demonstrating how we could leverage large observational data to probe the effect of exterior interventions (also an instance of $E(t, \mathbf{x})$). The transition to DST is beneficial for energy conservation, but at the same time, it has been reported to increase the risk of cerebrovascular and cardiovascular problems. We evaluated the effect of the DST shift on the whole spectrum of diseases — an analysis that, we hope, will help to re-evaluate the risks and benefits of DST shifts. Our study relied on a population-based, cross-sectional analysis of the IBM Watson Health MarketScan insurance claim data set, which contains over 150 million unique patients in the US, and the Swedish national inpatient register, which incorporates health information describing more than nine million unique Swedes. For hundreds of sex- and age-

specific diseases, we assessed the effects of the DST shifts forward and backward by one hour in spring and autumn by comparing the observed and expected diagnosis rates after DST shift exposure. We found four prominent, elevated risk clusters, including cardiovascular diseases (such as heart attacks), injuries, mental and behavioral disorders, and immune-related diseases, such as noninfective enteritis and colitis, to be significantly associated with DST shifts in the United States and Sweden. While the majority of disease risk elevations are modest (a few percent), a considerable number of diseases exhibit an approximately ten percent relative risk increase. We estimate that each spring DST shift is associated with negative health effects – with 150,000 incidences in the US and 880,000 globally. We also identify for the first time a collection of diseases with relative risks that appear to decrease immediately after the spring DST shift, enriched with infections and immune system-related maladies. These diseases' decreasing relative risks might be driven by the documented boosting effect of short-term stress (such as that experienced around the spring DST shift) on the immune system.

To further explore the environmental effects on the disease incidence, we expanded our model and studied the disease seasonality. With similar data (MarketScan and the Swedish national register), we built a Bayesian hierarchical model to jointly infer the disease seasonality and trend. We focused this chapter of our study on disease seasonality, ignoring other properties of diseases for the investigation. Here, we conducted two types of analysis, called "uncorrected" and "corrected" ones, were conducted. The former analysis focused on counts of daily patient visits associated with each disease. The latter analysis instead looked at the proportion of disease-specific visits within the total volume of visits for a time interval. In the spirit of full disclosure, we present both unredacted sets of results. In the uncorrected analysis, we found that psychiatric diseases' annual patterns were remarkably similar across the studied diseases in both

countries, with the magnitude of annual variation significantly higher in Sweden than in the US for psychiatric, but not infectious diseases. In the corrected analysis, only one group of patients – eleven to 20 years old – reproduced all the regularities we observed for psychiatric disorders in the uncorrected analysis; the annual healthcare-seeking visit patterns associated with other age groups changed drastically. Analogous analyses over infectious diseases were less divergent over these two types of computation. Comparing two sets of results in the context of published psychiatric disease seasonality studies, we tend to believe that our uncorrected results are likely to capture the real trends. In contrast, the corrected results reflect mostly artifacts generated by idiosyncratically fluctuating volumes of patient health-seeking visits across the year.

Of course, the etiology of most diseases comprises more than environmental influences. Therefore, in our latest project, we considered a modified version of the equation, $P(t, \mathbf{x}) = G + E(t, \mathbf{x}) + G \times E(t, \mathbf{x})$, where the $G \times E(t, \mathbf{x})$ term represents the interactions between genetic and environmental factors. For the ten most diagnosed psychiatric diseases in the MarketScan data, we built five Bayesian mixed-effects models ranging from simple linear models ($P = G + E$) to interactions models ($P = G + E + G \times E$). We compared the estimated heritability and other statistics indicating how much of the total phenotypic variance could be attributed to the shared environments or the gene-environment interactions. The results suggested that gene-environment interactions might explain a substantial portion of the variability of common psychiatric disorders. Among the ten diseases tested, eight were better modeled by the interactions models that considered the G-by-E effects, including anxiety phobic disorder, depression, migraine, ADHD, bipolar disorder, unspecified recurrent headaches, sleep disorder, and PTSD. As by-products, our methods also estimated the fixed effects associated with sex, age, environmental quality indices, and random effects of the geographic location of residency.

As the central topic of the dissertation, the equation $P(t, \mathbf{x}) = G + E(t, \mathbf{x}) + G \times E(t, \mathbf{x})$ already captures a large amount of information about the disease. Those who want to shape any phenotypic trait must consider what they are able to alter on the right side of the equation. Geneticists may be willing to breed species with desired characteristics by selection and change the G. In contrast to the breeding of domesticated species, in the real world of human medicine, it is difficult, if not impossible, to alter the patient's genetics. Available instruments are limited to the environmental effects and the interactions (E and $G \times E$). The dissection of complex disease etiology in the present dissertation suggests that perturbing the environmental and gene-environment factors can affect the downstream phenotype. Even a one-hour time change in spring (the daylight saving time shift) could significantly influence many conditions. As impactful factors, gene-environment interactions are almost certainly involved in the etiology of common psychiatric diseases. In selective breeding, geneticists are interested in traits with fairly large heritability because it implies a considerable response to selection. Similarly, as we demonstrated, both environmental factors and gene-environment interactions could explain significant parts of the trait variance. Therefore, we must be able to intervene in the development of complex diseases by imposing specific exterior conditions. The next important step is to find out certain environmental factors or gene-environment factor pairs and see what they hint about preventive and precision medicine.

REFERENCES

1. Neale M, Cardon LR. Methodology for genetic studies of twins and families: Springer Science & Business Media; 1992.

2. Benchek PH, Morris NJ. How meaningful are heritability estimates of liability? Human genetics. 2013;132(12):1351-60.

3. Sulc J, Mounier N, Günther F, Winkler T, Wood AR, Frayling TM, et al. Quantification of the overall contribution of gene-environment interaction for obesity-related traits. Nature communications. 2020;11(1):1-13.

4. Paszke A, Gross S, Chintala S, Chanan G, Yang E, DeVito Z, et al. Automatic differentiation in pytorch. 2017.

5. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. Pytorch: An imperative style, high-performance deep learning library. arXiv preprint arXiv:191201703. 2019.

6. Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, et al., editors. Tensorflow: A system for large-scale machine learning. 12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16); 2016.

7. Bauer FL. Computational graphs and rounding error. SIAM Journal on Numerical Analysis. 1974;11(1):87-96.

8. Baydin AG, Pearlmutter BA, Radul AA, Siskind JM. Automatic differentiation in machine learning: a survey. Journal of machine learning research. 2018;18.

9. Bottou L. Large-scale machine learning with stochastic gradient descent. Proceedings of COMPSTAT'2010: Springer; 2010. p. 177-86.

10. Brooks S, Gelman A, Jones G, Meng X-L. Handbook of markov chain monte carlo: CRC press; 2011.

11. Betancourt M, Girolami M. Hamiltonian Monte Carlo for hierarchical models. Current trends in Bayesian methodology with applications. 2015;79(30):2-4.

12. Hoffman MD, Gelman A. The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. Journal of Machine Learning Research. 2014;15(1):1593-623.

13. Smith A. Sequential Monte Carlo methods in practice: Springer Science & Business Media; 2013.

14. Blei DM, Kucukelbir A, McAuliffe JD. Variational inference: A review for statisticians. Journal of the American statistical Association. 2017;112(518):859-77.

15. Zhang C, Bütepage J, Kjellström H, Mandt S. Advances in variational inference. IEEE transactions on pattern analysis and machine intelligence. 2018;41(8):2008-26.

16. Salvatier J, Wiecki TV, Fonnesbeck C. Probabilistic programming in Python using PyMC3. PeerJ Computer Science. 2016;2:e55.

17. Carpenter B, Gelman A, Hoffman MD, Lee D, Goodrich B, Betancourt M, et al. Stan: a probabilistic programming language. Grantee Submission. 2017;76(1):1-32.

18. Gordon AD, Henzinger TA, Nori AV, Rajamani SK. Probabilistic programming. Future of Software Engineering Proceedings2014. p. 167-81.

19. IBM Watson Health. IBM MarketScan Research Databases. 2019.

20. Ludvigsson JF, Andersson E, Ekbom A, Feychting M, Kim J-L, Reuterwall C, et al. External review and validation of the Swedish national inpatient register. BMC public health. 2011;11(1):1-16.

21. Administration UEI. How much electricity is used for lighting in the United States? - FAQ - U.S. Energy Information Administration (EIA). 2021.

22. Kantermann T, Juda M, Merrow M, Roenneberg T. The human circadian clock's seasonal adjustment is disrupted by daylight saving time. Current Biology. 2007;17(22):1996-2000.

23. Rivers N. Does daylight savings time save energy? Evidence from Ontario. Environmental and Resource Economics. 2018;70(2):517-43.

24. Kotchen MJ, Grant LE. Does daylight saving time save energy? Evidence from a natural experiment in Indiana. Review of Economics and Statistics. 2011;93(4):1172-85.

25. Monk TH, Folkard S. Adjusting to the changes to and from Daylight Saving Time. Nature. 1976;261(5562):688-9.

26. Prats-Uribe A, Tobías A, Prieto-Alhambra D. Excess risk of fatal road traffic accidents on the day of daylight saving time change. Epidemiology. 2018;29(5):e44-e5.

27. Barnes CM, Wagner DT. Changing to daylight saving time cuts into sleep and increases workplace injuries. Journal of applied psychology. 2009;94(5):1305.

28. Janszky I, Ahnve S, Ljung R, Mukamal KJ, Gautam S, Wallentin L, et al. Daylight saving time shifts and incidence of acute myocardial infarction–Swedish Register of Information and Knowledge About Swedish Heart Intensive Care Admissions (RIKS-HIA). Sleep medicine. 2012;13(3):237-42.

29. Lahti TA, Haukka J, Lönnqvist J, Partonen T. Daylight saving time transitions and hospital treatments due to accidents or manic episodes. BMC Public Health. 2008;8(1):1-4.

30. Heboyan V, Stevens S, McCall WV. Effects of seasonality and daylight savings time on emergency department visits for mental health disorders. The American journal of emergency medicine. 2019;37(8):1476-81.

31. Fuller PM, Gooley JJ, Saper CB. Neurobiology of the sleep-wake cycle: sleep architecture, circadian regulation, and regulatory feedback. Journal of biological rhythms. 2006;21(6):482-93.

32. Roenneberg T, Winnebeck EC, Klerman EB. Daylight saving time and artificial time zones–a battle between biological and social times. Frontiers in physiology. 2019;10:944.

33. Malow BA, Veatch OJ, Bagai K. Are Daylight Saving Time Changes Bad for the Brain? JAMA neurology. 2020;77(1):9-10.

34. Ludvigsson JF, Andersson E, Ekbom A, Feychting M, Kim J-L, Reuterwall C, et al. External review and validation of the Swedish national inpatient register. BMC public health. 2011;11(1):450.

35. World Health Organization. International statistical classification of diseases and related health problems, 10th revision (ICD-10). World Health Organization; 2016.

36. Gelman A, Hill J, Yajima M. Why we (usually) don't have to worry about multiple comparisons. Journal of Research on Educational Effectiveness. 2012;5(2):189-211.

37. Benjamini Y, Yekutieli D. False discovery rate–adjusted multiple confidence intervals for selected parameters. Journal of the American Statistical Association. 2005;100(469):71-81.

38. Janszky I, Ljung R. Shifts to and From Daylight Saving Time and Incidence of Myocardial Infarction. The New England journal of medicine. 2008;359(18):1966-8.

39. Manfredini R, Fabbian F, De Giorgi A, Zucchi B, Cappadona R, Signani F, et al. Daylight saving time and myocardial infarction: should we be worried? A review of the evidence. Eur Rev Med Pharmacol Sci. 2018;22(3):750-5.

40. Meira e Cruz M, Miyazawa M, Manfredini R, Cardinali DP, Madrid J, Reiter R, et al. Impact of Daylight Saving Time on circadian timing system: An expert statement. 2019.

41. Robb D, Barnes T. Accident rates and the impact of daylight saving time transitions. Accident Analysis & Prevention. 2018;111:193-201.

42. Menet JS, Rosbash M. When brain clocks lose track of time: cause or consequence of neuropsychiatric disorders. Current opinion in neurobiology. 2011;21(6):849-57.

43. Hasler BP, Soehner AM, Clark DB. Circadian rhythms and risk for substance use disorders in adolescence. Current opinion in psychiatry. 2014;27(6):460.

44. Webb IC. Circadian rhythms and substance abuse: chronobiological considerations for the treatment of addiction. Current psychiatry reports. 2017;19(2):12.

45. Hasler BP, Smith LJ, Cousins JC, Bootzin RR. Circadian rhythms, sleep, and substance abuse. Sleep medicine reviews. 2012;16(1):67-81.

46. Logan RW, Williams III WP, McClung CA. Circadian rhythms and addiction: mechanistic insights and future directions. Behavioral neuroscience. 2014;128(3):387.

47. Falcón E, McClung CA. A role for the circadian genes in drug addiction. Neuropharmacology. 2009;56:91-6.

48. Martino TA, Young ME. Influence of the cardiomyocyte circadian clock on cardiac physiology and pathophysiology. Journal of biological rhythms. 2015;30(3):183-205.

49. Fletcher EK, Kanki M, Morgan J, Ray DW, Delbridge LM, Fuller PJ, et al. Cardiomyocyte transcription is controlled by combined mineralocorticoid receptor and circadian clock signalling. Journal of Endocrinology. 2019;241(1):17-29.

50. Sulli G, Lam MTY, Panda S. Interplay between circadian clock and cancer: new frontiers for cancer treatment. Trends in cancer. 2019;5(8):475-94.

51. Zhanfeng N, Hechun X, Zhijun Z, Hongyu X, Zhou F. Regulation of Circadian Clock Genes on Sleep Disorders in Traumatic Brain Injury Patients. World neurosurgery. 2019;130:e475-e86.

52. Olaoye OA, Masten SH, Mohandas R, Gumz ML. Circadian clock genes in diabetic kidney disease (DKD). Current diabetes reports. 2019;19(7):1-7.

53. Shi D, Chen J, Wang J, Yao J, Huang Y, Zhang G, et al. Circadian clock genes in the metabolism of non-alcoholic fatty liver disease. Frontiers in physiology. 2019;10:423.

54. Faragó A, Zsindely N, Bodai L. Mutant huntingtin disturbs circadian clock gene expression and sleep patterns in Drosophila. Scientific reports. 2019;9(1):1-6.

55. Chang WH, Lai AG. Timing gone awry: distinct tumour suppressive and oncogenic roles of the circadian clock and crosstalk with hypoxia signalling in diverse malignancies. Journal of translational medicine. 2019;17(1):1-16.

56. Kaneshiro K, Yoshida K, Morii K, Oketani Y, Uchida K, Yaekura A, et al. Expressions of circadian clock genes represent disease activities of RA patients treated with biological DMARDs. Modern rheumatology. 2020;30(2):293-300.

57. Maury E. Off the clock: from circadian disruption to metabolic disease. International journal of molecular sciences. 2019;20(7):1597.

58. Mukherji A, Bailey SM, Staels B, Baumert TF. The circadian clock and liver function in health and disease. Journal of hepatology. 2019;71(1):200-11.

59. Crnko S, Du Pré BC, Sluijter JP, Van Laake LW. Circadian rhythms and the molecular clock in cardiovascular biology and disease. Nature Reviews Cardiology. 2019;16(7):437-47.

60. Davis K, Roden LC, Leaner VD, van der Watt PJ. The tumour suppressing role of the circadian clock. IUBMB life. 2019;71(7):771-80.

61. Li H, Song S, Wang Y, Huang C, Zhang F, Liu J, et al. Low-grade inflammation aggravates rotenone neurotoxicity and disrupts circadian clock gene expression in rats. Neurotoxicity research. 2019;35(2):421-31.

62. Kim P, Oster H, Lehnert H, Schmid SM, Salamat N, Barclay JL, et al. Coupling the circadian clock to homeostasis: the role of period in timing physiology. Endocrine reviews. 2019;40(1):66-95.

63. Dhabhar FS. Effects of stress on immune function: the good, the bad, and the beautiful. Immunologic research. 2014;58(2):193-210.

64. Dhabhar FS, Saul AN, Daugherty C, Holmes TH, Bouley DM, Oberyszyn TM. Short-term stress enhances cellular immunity and increases early resistance to squamous cell carcinoma. Brain, behavior, and immunity. 2010;24(1):127-37.

65. Centers for Disease Control and Prevention (CDC). Diagnosis code set general equivalence mappings – ICD-10-CM to ICD-9-CM and ICD-9-CM to ICD-10-CM.

66. Gelman A. Prior choice recommendations. Wiki for the Stan project on Github. 2019.

67. Kucukelbir A, Tran D, Ranganath R, Gelman A, Blei DM. Automatic differentiation variational inference. The Journal of Machine Learning Research. 2017;18(1):430-74.

68. Brooks SP, Gelman A. General methods for monitoring convergence of iterative simulations. Journal of computational and graphical statistics. 1998;7(4):434-55.

69. Gelman A, Rubin DB. Inference from Iterative Simulation Using Multiple Sequences. Stat Sci. 1992;7(4):457-72.

70. Hinkley DV. On the ratio of two correlated normal random variables. Biometrika. 1969;56(3):635-9.

71. Ederer F, Mantel N. Confidence Limits on the Ratio of Two Poisson Variables. American journal of epidemiology. 1974;100(3):165-7.

72. Wilson EB. Probable inference, the law of succession, and statistical inference. Journal of the American Statistical Association. 1927;22(158):209-12.

73. Brainstorm Consortium, Anttila V, Bulik-Sullivan B, Finucane HK, Walters RK, Bras J, et al. Analysis of shared heritability in common disorders of the brain. Science. 2018;360(6395). Epub 2018/06/23. doi: 10.1126/science.aap8757. PubMed PMID: 29930110.

74. Solberg BS, Zayats T, Posserud MB, Halmoy A, Engeland A, Haavik J, et al. Patterns of Psychiatric Comorbidity and Genetic Correlations Provide New Insights Into Differences Between Attention-Deficit/Hyperactivity Disorder and Autism Spectrum Disorder. Biol

Psychiatry. 2019;86(8):587-98. Epub 2019/06/12. doi: 10.1016/j.biopsych.2019.04.021. PubMed PMID: 31182215; PubMed Central PMCID: PMCPMC6764861.

75. Tylee DS, Sun J, Hess JL, Tahir MA, Sharma E, Malik R, et al. Genetic correlations among psychiatric and immune-related phenotypes based on genome-wide association data. Am J Med Genet B Neuropsychiatr Genet. 2018;177(7):641-57. Epub 2018/10/17. doi: 10.1002/ajmg.b.32652. PubMed PMID: 30325587; PubMed Central PMCID: PMCPMC6230304.

76. Wen Y, Zhang F, Ma X, Fan Q, Wang W, Xu J, et al. eQTLs Weighted Genetic Correlation Analysis Detected Brain Region Differences in Genetic Correlations for Complex Psychiatric Disorders. Schizophr Bull. 2019;45(3):709-15. Epub 2018/06/19. doi: 10.1093/schbul/sby080. PubMed PMID: 29912442; PubMed Central PMCID: PMCPMC6483588.

77. Jia G, Li Y, Zhang H, Chattopadhyay I, Boeck Jensen A, Blair DR, et al. Estimating heritability and genetic correlations from large health datasets in the absence of genetic data. Nat Commun. 2019;10(1):5508. Epub 2019/12/05. doi: 10.1038/s41467-019-13455-0. PubMed PMID: 31796735; PubMed Central PMCID: PMCPMC6890770.

78. Wang K, Gaitsch H, Poon H, Cox NJ, Rzhetsky A. Classification of common human diseases derived from shared genetic and environmental determinants. Nat Genet. 2017;49(9):1319-25. Epub 2017/08/08. doi: 10.1038/ng.3931. PubMed PMID: 28783162; PubMed Central PMCID: PMCPMC5577363.

79. Khan A, Plana-Ripoll O, Antonsen S, Brandt J, Geels C, Landecker H, et al. Environmental pollution is associated with increased risk of psychiatric disorders in the US and Denmark. PLoS Biol. 2019;17(8):e3000353. Epub 2019/08/21. doi: 10.1371/journal.pbio.3000353. PubMed PMID: 31430271; PubMed Central PMCID: PMCPMC6701746.

80. Lam RW, Levitan RD. Pathophysiology of seasonal affective disorder: a review. Journal of Psychiatry and Neuroscience. 2000;25(5):469.

81. Wehr TA, Duncan WC, Sher L, Aeschbach D, Schwartz PJ, Turner EH, et al. A circadian signal of change of season in patients with seasonal affective disorder. Arch Gen Psychiat. 2001;58(12):1108-14.

82. Johansson C, Willeit M, Smedh C, Ekholm J, Paunio T, Kieseppä T, et al. Circadian clock-related polymorphisms in seasonal affective disorder and their relevance to diurnal preference. Neuropsychopharmacology. 2003;28(4):734.

83. Lee H-J, Rex KM, Nievergelt CM, Kelsoe JR, Kripke DF. Delayed sleep phase syndrome is related to seasonal affective disorder. J Affect Disorders. 2011;133(3):573-9.

84. O'Hare C, O'Sullivan V, Flood S, Kenny RA. Seasonal and meteorological associations with depressive symptoms in older adults: A geo-epidemiological study. J Affect Disorders. 2016;191:172-9.

85. Oyane NM, Bjelland I, Pallesen S, Holsten F, Bjorvatn B. Seasonality is associated with anxiety and depression: the Hordaland health study. J Affect Disorders. 2008;105(1-3):147-55.

86. Winthorst WH, Post WJ, Meesters Y, Penninx BW, Nolen WA. Seasonality in depressive and anxiety symptoms among primary care patients and in patients with depressive and anxiety disorders; results from the Netherlands Study of Depression and Anxiety. BMC psychiatry. 2011;11(1):198.

87. Øverland S, Woicik W, Sikora L, Whittaker K, Heli H, Skjelkvåle FS, et al. Seasonality and symptoms of depression: A systematic review of the literature. Epidemiology and psychiatric sciences. 2019:1-15.

88. Lukmanji A, Williams JV, Bulloch AG, Bhattarai A, Patten SB. Seasonal variation in symptoms of depression: A Canadian population based study. J Affect Disorders. 2019;255:142-9.

89. Harmatz MG, Well AD, Overtree CE, Kawamura KY, Rosal M, Ockene IS. Seasonal variation of depression and other moods: a longitudinal approach. Journal of biological rhythms. 2000;15(4):344-50.

90. Hoffman MD, Gelman A. The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. J Mach Learn Res. 2014;15:1593-623. PubMed PMID: WOS:000338420000013.

91. Balasubramanian M, Schwartz EL. The isomap algorithm and topological stability. Science. 2002;295(5552):7-.

92. Medici CR, Vestergaard CH, Hadzi-Pavlovic D, Munk-Jørgensen P, Parker G. Seasonal variations in hospital admissions for mania: Examining for associations with weather variables over time. J Affect Disorders. 2016;205:81-6.

93. Sebestyen B, Rihmer Z, Balint L, Szokontor N, Gonda X, Gyarmati B, et al. Gender differences in antidepressant use-related seasonality change in suicide mortality in Hungary, 1998-2006. The world journal of biological psychiatry : the official journal of the World Federation of Societies of Biological Psychiatry. 2010;11(3):579-85. Epub 2010/03/12. doi: 10.3109/15622970903397722. PubMed PMID: 20218927.

94. Postolache TT, Langenberg P, Zimmerman SA, Lapidus M, Komarow H, McDonald JS, et al. Changes in Severity of Allergy and Anxiety Symptoms Are Positively Correlated in Patients with Recurrent Mood Disorders Who Are Exposed to Seasonal Peaks of Aeroallergens. Int J Child Health Hum Dev. 2008;1(3):313-22. Epub 2008/01/01. PubMed PMID: 19430577; PubMed Central PMCID: PMCPMC2678838.

95. Pirkola S, Eriksen HA, Partonen T, Kieseppa T, Veijola J, Jaaskelainen E, et al. Seasonal variation in affective and other clinical symptoms among high-risk families for bipolar disorders in an Arctic population. Int J Circumpol Heal. 2015;74. doi: ARTN 29671

10.3402/ijch.v74.29671. PubMed PMID: WOS:000369579600001.

96. Schwartz PJ. Chris Cornell, the Black Hole Sun, and the Seasonality of Suicide. Neuropsychobiology. 2019;78(1):38-47. Epub 2019/03/29. doi: 10.1159/000498868. PubMed PMID: 30921807; PubMed Central PMCID: PMCPMC6549453.

97. Trang PM, Rocklov J, Giang KB, Nilsson M. Seasonality of hospital admissions for mental disorders in Hanoi, Vietnam. Glob Health Action. 2016;9:32116. Epub 2016/08/28. doi: 10.3402/gha.v9.32116. PubMed PMID: 27566716; PubMed Central PMCID: PMCPMC5002036.

98. Bradvik L, Berglund M. Seasonal distribution of suicide in alcoholism. Acta Psychiatr Scand. 2002;106(4):299-302. Epub 2002/09/13. doi: 10.1034/j.1600-0447.2002.02234.x. PubMed PMID: 12225497.

99. Levine ME, Duffy LK, Bowyer RT. Fatigue, Sleep and Seasonal Hormone Levels: Implications for Drinking Behavior in Northern Climates. Drugs & Society. 1994;8(2):61-70. doi: 10.1300/J023v08n02_04.

100. De Graaf R, Van Dorsselaer S, Ten Have M, Schoemaker C, Vollebergh WA. Seasonal variations in mental disorders in the general population of a country with a maritime climate: findings from the Netherlands mental health survey and incidence study. American Journal of Epidemiology. 2005;162(7):654-61.

101. 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):587-93.

102. Sperner-Unterweger B. Immunological aetiology of major psychiatric disorders: evidence and therapeutic implications. Drugs. 2005;65(11):1493-520. Epub 2005/07/22. doi: 10.2165/00003495-200565110-00004. PubMed PMID: 16033289.

103. Geoffroy PA, Bellivier F, Scott J, Etain B. Seasonality and bipolar disorder: a systematic review, from admission rates to seasonality of symptoms. J Affect Disorders. 2014;168:210-23.

104. Escott-Price V, Smith DJ, Kendall K, Ward J, Kirov G, Owen MJ, et al. Polygenic risk for schizophrenia and season of birth within the UK Biobank cohort. Psychological medicine. 2019;49(15):2499-504.

105. Wulff K, Dijk D-J, Middleton B, Foster RG, Joyce EM. Sleep and circadian rhythm disruption in schizophrenia. The British Journal of Psychiatry. 2012;200(4):308-16.

106. Rao ML, Gross G, Strebel B, Halaris A, Huber G, Bräunig P, et al. Circadian rhythm of tryptophan, serotonin, melatonin, and pituitary hormones in schizophrenia. Biol Psychiat. 1994;35(3):151-63.

107. Monti JM, BaHammam AS, Pandi-Perumal SR, Bromundt V, Spence DW, Cardinali DP, et al. Sleep and circadian rhythm dysregulation in schizophrenia. Progress in Neuro-Psychopharmacology and Biological Psychiatry. 2013;43:209-16.

108. Solomon GD. Circadian rhythms and migraine. Cleve Clin J Med. 1992;59(3):326-9.

109. Ong JC, Taylor HL, Park M, Burgess HJ, Fox RS, Snyder S, et al. Can circadian dysregulation exacerbate migraines? Headache: The Journal of Head and Face Pain. 2018;58(7):1040-51.

110. Grassly NC, Fraser C. Seasonal infectious disease epidemiology. Proceedings of the Royal Society B: Biological Sciences. 2006;273(1600):2541-50.

111. Martinez ME. The calendar of epidemics: Seasonal cycles of infectious diseases. PLoS pathogens. 2018;14(11):e1007327.

112. Wynchank DS, Bijlenga D, Lamers F, Bron TI, Winthorst WH, Vogel SW, et al. ADHD, circadian rhythms and seasonality. J Psychiatr Res. 2016;81:87-94.

113. Hakkarainen R, Johansson C, Kieseppä T, Partonen T, Koskenvuo M, Kaprio J, et al. Seasonal changes, sleep length and circadian preference among twins with bipolar disorder. BMC psychiatry. 2003;3(1):6.

114. Watanabe S. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. Journal of Machine Learning Research. 2010;11(Dec):3571-94.

115. Geweke J. Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments: Federal Reserve Bank of Minneapolis, Research Department Minneapolis, MN; 1991.

116. Tropf FC, Lee SH, Verweij RM, Stulp G, van der Most PJ, de Vlaming R, et al. Hidden heritability due to heterogeneity across seven populations. Nat Hum Behav. 2017;1(10):757-65. Epub 2017/10/21. doi: 10.1038/s41562-017-0195-1. PubMed PMID: 29051922; PubMed Central PMCID: PMCPMC5642946.

117. Sulc J, Mounier N, Gunther F, Winkler T, Wood AR, Frayling TM, et al. Quantification of the overall contribution of gene-environment interaction for obesity-related traits. Nat Commun. 2020;11(1):1385. Epub 2020/03/15. doi: 10.1038/s41467-020-15107-0. PubMed PMID: 32170055; PubMed Central PMCID: PMCPMC7070002.

118. McAllister K, Mechanic LE, Amos C, Aschard H, Blair IA, Chatterjee N, et al. Current Challenges and New Opportunities for Gene-Environment Interaction Studies of Complex

Diseases. Am J Epidemiol. 2017;186(7):753-61. Epub 2017/10/06. doi: 10.1093/aje/kwx227. PubMed PMID: 28978193; PubMed Central PMCID: PMCPMC5860428.

119. Assary E, Vincent JP, Keers R, Pluess M. Gene-environment interaction and psychiatric disorders: Review and future directions. Semin Cell Dev Biol. 2018;77:133-43. Epub 2017/10/21. doi: 10.1016/j.semcdb.2017.10.016. PubMed PMID: 29051054.

120. Fu J, Nogueira SV, Drongelen VV, Coit P, Ling S, Rosloniec EF, et al. Shared epitope-aryl hydrocarbon receptor crosstalk underlies the mechanism of gene-environment interaction in autoimmune arthritis. Proc Natl Acad Sci U S A. 2018;115(18):4755-60. Epub 2018/04/19. doi: 10.1073/pnas.1722124115. PubMed PMID: 29666259; PubMed Central PMCID: PMCPMC5939100.

121. Rivera NV, Patasova K, Kullberg S, Diaz-Gallo LM, Iseda T, Bengtsson C, et al. A Gene-Environment Interaction Between Smoking and Gene polymorphisms Provides a High Risk of Two Subgroups of Sarcoidosis. Sci Rep. 2019;9(1):18633. Epub 2019/12/11. doi: 10.1038/s41598-019-54612-1. PubMed PMID: 31819081; PubMed Central PMCID: PMCPMC6901455.

122. Arbet J, McGue M, Basu S. A robust and unified framework for estimating heritability in twin studies using generalized estimating equations. Stat Med. 2020;39(27):3897-913. Epub 2020/05/26. doi: 10.1002/sim.8564. PubMed PMID: 32449216.

123. Grasby KL, Verweij KJH, Mosing MA, Zietsch BP, Medland SE. Estimating Heritability from Twin Studies. Methods Mol Biol. 2017;1666:171-94. Epub 2017/10/06. doi: 10.1007/978-1-4939-7274-6_9. PubMed PMID: 28980246.

124. Scheike TH, Holst KK, Hjelmborg JB. Estimating heritability for cause specific mortality based on twin studies. Lifetime data analysis. 2014;20(2):210-33. Epub 2013/02/05. doi: 10.1007/s10985-013-9244-x. PubMed PMID: 23378036.

125. Verweij KJ, Mosing MA, Zietsch BP, Medland SE. Estimating heritability from twin studies. Methods Mol Biol. 2012;850:151-70. Epub 2012/02/07. doi: 10.1007/978-1-61779-555-8_9. PubMed PMID: 22307698.

126. Lopes MC, Andrew T, Carbonaro F, Spector TD, Hammond CJ. Estimating heritability and shared environmental effects for refractive error in twin and family studies. Investigative ophthalmology & visual science. 2009;50(1):126-31. Epub 2008/09/02. doi: 10.1167/iovs.08-2385. PubMed PMID: 18757506.

127. Bochud M. Estimating Heritability from Nuclear Family and Pedigree Data. Methods Mol Biol. 2017;1666:195-210. Epub 2017/10/06. doi: 10.1007/978-1-4939-7274-6_10. PubMed PMID: 28980247.

128. Evans LM, Tahmasbi R, Vrieze SI, Abecasis GR, Das S, Gazal S, et al. Comparison of methods that use whole genome data to estimate the heritability and genetic architecture of

complex traits. Nat Genet. 2018;50(5):737-45. Epub 2018/04/28. doi: 10.1038/s41588-018-0108-x. PubMed PMID: 29700474; PubMed Central PMCID: PMCPMC5934350.

129. Messer LC, Jagai JS, Rappazzo KM, Lobdell DT. Construction of an environmental quality index for public health research. Environ Health. 2014;13(1):39. Epub 2014/06/03. doi: 10.1186/1476-069X-13-39. PubMed PMID: 24886426; PubMed Central PMCID: PMCPMC4046025.

130. Lobdell DT, Jagai JS, Rappazzo K, Messer LC. Data sources for an environmental quality index: availability, quality, and utility. Am J Public Health. 2011;101 Suppl 1:S277-85. Epub 2011/08/13. doi: 10.2105/AJPH.2011.300184. PubMed PMID: 21836111; PubMed Central PMCID: PMCPMC3222503.

131. Watanabe S. Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory. ArXiv. 2010;abs/1004.2316.

132. Akaike H. A new look at the statistical model identification. IEEE Transactions on Automatic Control. 1974;19(6):716-23. doi: 10.1109/TAC.1974.1100705.

133. Carvalho CM, Polson NG, Scott JG, editors. Handling sparsity via the horseshoe. Artificial Intelligence and Statistics; 2009.

134. Chung Y, Rabe-Hesketh S, Dorie V, Gelman A, Liu J. A nondegenerate penalized likelihood estimator for variance parameters in multilevel models. Psychometrika. 2013;78(4):685-709. Epub 2013/10/05. doi: 10.1007/s11336-013-9328-2. PubMed PMID: 24092484.

135. Gelman A. Prior choice recommendations. Retrieved July. 2019;24:2019.