Autism Prevalence Trends by Birth Year and Diagnostic Year: Indicators of Etiologic and Non-Etiologic Factors – an Age Period Cohort Problem

A Thesis
Submitted to the Division of Epidemiology
Department of Health Research and Policy
In Partial Fulfillment of the Requirements
For the Master of Science in Epidemiology and
Clinical Research Degree
Stanford University School of Medicine

Alexander G. MacInnis
June 2017

Autism Prevalence Trends by Birth Year and Diagnostic Year: Indicators of Etiologic and Non-Etiologic Factors – an Age Period Cohort Problem

Alexander G. MacInnis
June 2017

Approved for Submission to the Division of Epidemiology
Department of Health Research and Policy
Stanford University School of Medicine

Epidemiology reader: Lorene Nelson Date: 5/31/17
Lorene Nelson
Associate Professor, Division of Epidemiology
Department of Health Research and Policy

Co-Reader: Kristin L. Sainani Date: 5/31/17
Kristin Sainani
Associate Professor (Teaching)
Department of Health Research and Policy

Abstract

The primary objective of this study is to characterize the etiological and non-etiological components of the observed increase in incidence of diagnosis of autism in California from 1980 to 2015. We show that the time trends of autism prevalence by birth year and diagnostic year correspond directly to trends in etiologic (causal) and non-etiologic (non-causal) factors respectively and endeavor to estimate the coefficients of both trends. The primary dataset is incidence of autism diagnosis data from the California Department of Developmental Services (CA-DDS). It provides the numbers of clients newly enrolled for services under an autism classification for each diagnostic year from 1980 through 2015 with separate observations for each birth year and gender. The analysis estimates cumulative incidence to age 10 as a more appropriate measure than prevalence. Knowledge of the birth year and diagnostic year trends is important for elucidating the combined effect of variable etiologic factors, that is, environmental effects broadly defined, which may lead to an understanding of potential prevention and treatment strategies. The birth year trend, controlling for diagnostic year trend, could also be used to predict the future case load of adults with autism needing support, which may inform policy decisions and associated funding requirements for care of these individuals, which already consumes an estimated 1.5% of US GDP. It is straightforward to estimate the sum of the birth year and diagnostic year coefficients, which corresponds to a growth of 12.82% per year from 1980 to 2015, but intractable to estimate the allocation of individual coefficients within that sum. The problem of estimating the birth year and diagnostic year trends falls within the class of age period cohort (APC) problems, because the age factor affects the analysis and the three variables are collinear; there is a lack of identifiability, which prevents reliable estimation of the key variables. We investigated novel methods of analyzing this type of problem and demonstrate a

new way to understand the problem. We show that estimating the age factor correctly is both more important and more difficult than indicated in previous APC literature because estimates of the age factor are inherently functions of the coefficients of the period (diagnostic year) and cohort (birth year) effects, which are unknown, and biases in the age factor estimate based on implicit assumptions of these two effects directly affect the resulting estimates.

Introduction

The change of prevalence of autistic disorder and related autism spectrum disorders (ASDs) over time by birth year is controversial and has not been established in the literature. There is indisputably a very large increase in measured prevalence over time, both independent of birth year and by birth year, and it's clear that non-etiologic factors – factors that affect the probability of an autism or ASD diagnosis while not affecting the true prevalence of the disorder – have contributed to this increase.

The degree to which the true prevalence of the disorder has increased is still a matter of ongoing debate. To the best of our knowledge, no peer-reviewed reports have estimated the trend of autism case prevalence by birth year while accounting for changes in diagnostic and other non-etiologic factors. The resulting uncertainty is exploited in the journal literature and to a larger degree in the popular press, including award-winning books, scientific magazines and web sites. As long as this question remains unanswered, people in positions of authority may claim that there is no true increase in prevalence. This belief is reflected in funding priorities: Research funding so far has been primarily directed to studies of genetics and to description of behavioral characteristics of autism rather than toward identifying environmental factors that may contribute to the disorder and which might be leveraged to treat or prevent the disorder.

Here is an illustration by Autism Speaks (2010) showing the time trend in measured autism prevalence by study publication year from 1975 to 2009. While generally representative, it does not show the uncertainty within studies and the variability between them, does not include the most recent results, and does not show prevalence by birth year.

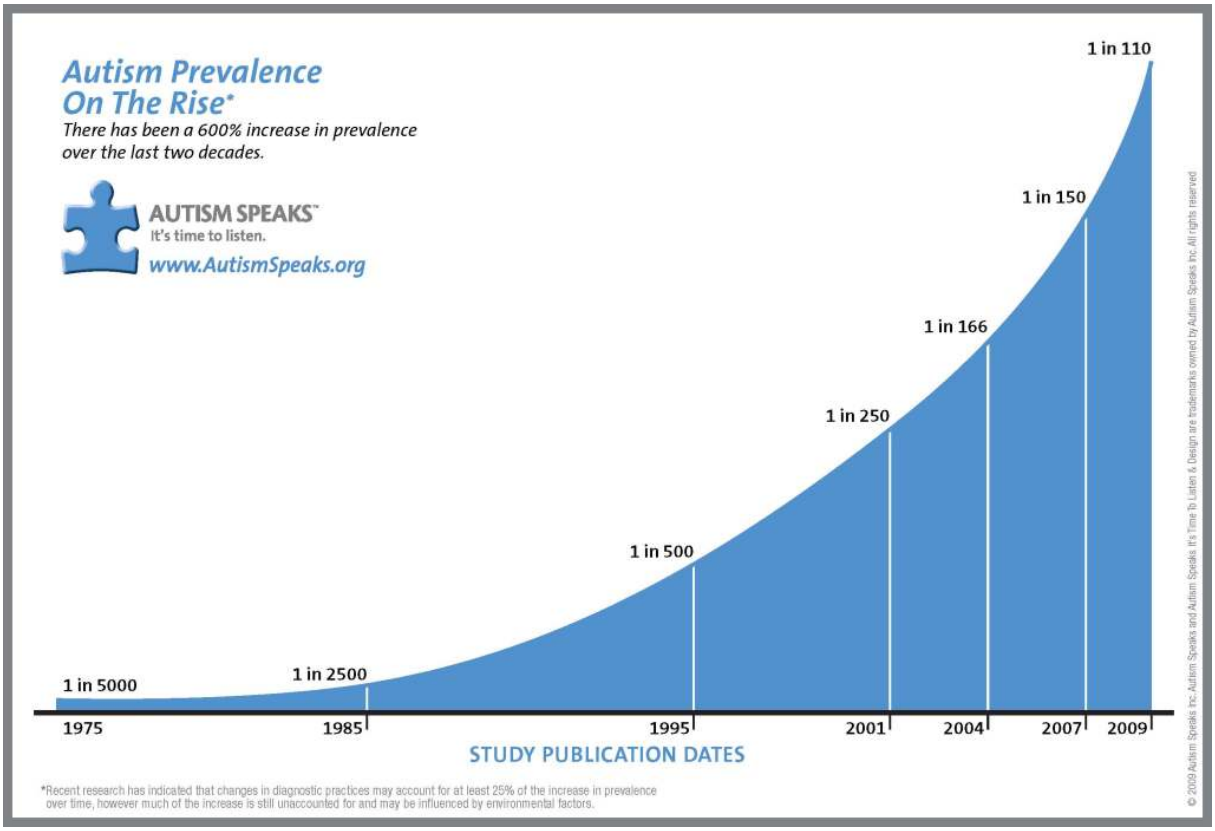

Figure 1 Graph of Overall Autism Prevalence by Year

Some investigators express skepticism that there is any significant true increase in prevalence, arguing that various factors such as awareness of autism, social factors and availability of services, as well as changes in diagnostic criteria and practices, might account for all of the increase. They have studied some of these factors individually (Table 1). However, no studies have accounted for all of these factors at once and no studies have shown that the true increase is zero. It is not feasible to estimate accurately the effects of all possible non-etiologic factors at

once by studying them individually because some hypothesized factors (such as awareness) are unmeasurable. Therefore, in this study, all non-etiologic factors, which is a superset of all diagnostic factors, whether known or unknown, and without making any assumptions about the start or end times or time trends of these factors, are grouped together and treated as a lumped effect on diagnoses of cases which varies with time and which may have different effects on persons of different ages at the time of diagnosis.

This study utilizes data from the California Department of Developmental Services (CA-DDS) providing counts of incident acceptance of clients to the CA-DDS Regional Center system with the classification of autism. In this study, we refer to the incidence of determination of eligibility for services with an autism classification as the incidence of diagnosis of autism within the CA-DDS. Acceptance is based on a determination of eligibility for services by the Regional Center that serves the geographic region where the prospective client resides. Cases are ascertained by a team within the serving Regional Center when a prospective client applies to the Regional Center for services. The determination is similar to a formal diagnosis and follows the diagnostic criteria established by the current version of the Diagnostic and Statistical Manual (DSM), that is, DSM-III, DSM-III-R, DSM-IV, DSM-IV-TR or DSM-V, depending on the year of the evaluation. The CA-DDS (2002) published best practices guidelines, which include considerable detail on the diagnostic process. The Regional Centers must also follow the requirements of the current version of the California Lanterman Act. The CA-DDS is required by the Lanterman Act to provide services to all residents, regardless of age, who are disabled under the categories of autism, epilepsy, cerebral palsy and intellectual disability (mental retardation). While the CA-DDS does not publish estimates of the percentage of eligible California residents that are enrolled with the CA-DDS, Croen et al. (2002) estimated that at least 75-80% of children eligible

for CA-DDS services for autism are enrolled, based on an analysis of a linkage of CA-DDS and California Department of Education Special Education databases. We do not have direct evidence of the proportion of autistic adults in California who are enrolled, nor do we have information regarding changes in the proportion of children enrolled over time.

We obtained population data by both year of estimate and year of birth (age) from the US Census Bureau and use it as the denominator in combination with the incidence data to obtain a probability of diagnosis for each diagnostic year and birth year. This census data includes the effects of in-migration to and out-migration from California, and therefore is more appropriate than California birth data for this purpose. Probability of diagnosis is the main dependent variable in analyses that seek to estimate both the birth year and diagnostic year effects.

While the literature generally refers to autism prevalence, here we focus on cumulative incidence to a specified age, as this provides a consistent set of ages over which incident diagnoses occur and it avoids some issues associated with prevalence estimates. The cumulative incidence value is similar to prevalence at the same age, particularly for childhood, since autism is generally assumed to be either present or pre-determined from birth or shortly thereafter, and it is generally assumed that recovery from autism is either rare or non-existent. However, there have been some documented cases of individuals previously diagnosed with autism no longer qualifying for the diagnosis; when removing such individuals from the numerator the prevalence is slightly less than the cumulative incidence. If individuals with autism were to have a shorter life expectancy than those without, that would cause a reduction in prevalence compared to cumulative incidence. Here we are interested in the probability of ever having been diagnosed by a certain age, and we are not studying rates at which individuals lose an autism diagnosis, which is an

interesting but distinct topic, nor do we study life expectancy. Note that if the rates at which individuals recover from autism (i.e., become reclassified as no longer meeting the criteria for an autism diagnosis) were to decrease over time that could lead to an apparent increase in prevalence. Population changes including death and moving into or out of a geographic region under study also affect prevalence and cumulative incidence.

Estimating the true trend of autism prevalence gives us insight into the disorder's etiology. If the true prevalence of the disorder is actually increasing, this tells us that there must have been a net change in the effects of all environmental factors, broadly defined, since there cannot have been a large change in inherited genetics over this time frame. (If the true prevalence has not changed, environmental factors may still play a role, but that would imply that the net effect of all environmental factors has been effectively constant over the time frame of interest.) If it were to be established that there has been a significant increase in autism prevalence by birth year after controlling for all non-etiologic factors, and hence that this increase was caused at least in part by environmental factors, that information may serve as evidence to influence funding agencies including the US NIH and the US federal Interagency Autism Coordinating Committee (IACC) to prioritize research that would be expected to lead to better understanding of preventable etiology and to practical improvements in treatment and prevention.

The true trend of autism prevalence by birth year translates directly into the future caseload of adults who will require government funded service in future years. Many adults with autism require substantial support services. Frequently the individuals with autism and their families do not have the resources to provide all of the support needed by the affected individuals and they tend to rely on governments to provide or pay for a significant portion of the needed support. If

there is a significant rate of increase in the prevalence of autism by birth year that would translate into a significant exponential increase in the costs of services that will be required in the future. If this is the case, it would be important to ensure that federal and state governments are aware of such a looming problem well in advance, for purposes of planning fiscal policy.

Previous Studies

Many hypotheses, both non-etiologic and etiologic, have been proposed in attempts to explain the large measured increase in autism prevalence over the last few decades. While some hypotheses have some support from quantitative analysis, there is currently no evidence that any particular combination of hypotheses can fully explain the increase. Further, some of the hypotheses claimed in the literature are not accompanied by analytical support, and some that do have such support are dubious based on published letters and logical analysis. It is important to note that one of the factors commonly used to explain the etiology of autism – inherited genetic susceptibility – is not a suitable explanation for the large increase since approximately 1980, because changes in DNA do not occur at a rate consistent with the measured increase. Some investigators, for example Goldani et al. (2014) state that autism is caused by interaction between genetic vulnerability and environmental factors. Lyall et al. (2017) also state that gene and environment interaction contributes to etiology, along with de novo mutations, epigenetics, advanced parental age, air pollution, and short inter-pregnancy intervals.

Non-Etiologic Hypotheses

Table 1

Proposed Non-Etiologic Explanations for the Rise in Measured Autism Prevalence

Hypothesis	**Description**	**Representative Publications**	**Notes**
Diagnostic criteria change Introduction of new criteria e.g., DSM-III-R, DSM-IV, ICD-10		Hansen, Schendel and Parner (2014)	Found significant effect of criteria change in 1994 on autism prevalence in Danish cohort.
		Hertz-Picciotto and Delwiche (2009)	Found that change in diagnostic criteria change in California could explain approximately 20% of the 585% increase in California from 1990 to 2001 birth years.
Diagnostic practice change	Identifiable changes in diagnostic practice separate from criteria	Hansen et al. (2014)	Found significant effect of diagnostic practice change in 1995 on autism prevalence in Danish cohort.
Diagnostic substitution of ID with autism, diagnostic accretion, and diagnostic recategorization	Individuals already diagnosed with intellectual disability changed to autism as primary diagnosis (substitution), autism added to an ID diagnosis (accretion), or individuals diagnosed with autism would have been diagnosed with ID in previous periods (recategorization)	King and Bearman (2009)	Found 631 patients born before 1987 whose diagnosis changed; extrapolated to estimate that 26.4% of California autism patients through 2005 had diagnostic accretion or substitution.
		Shattuck (2006)	Concluded national ecological reduction in mental retardation can explain increase in autism.
		Polyak, Kubina and Girirajan (2015)	Concluded national ecological reduction in mental retardation can explain increase in autism.

Increased awareness	Increased general awareness of autism may lead to increased probability of diagnosis	Fombonne (2009)	Suggested awareness is a factor, not studied.
		Elsabbagh et al. (2012)	Indicates increased awareness may be a factor.
		Keyes et al. (2012)	Age period cohort analysis found a purely cohort (birth year) effect, suggested awareness could explain the finding; described below.
Trend towards earlier age of diagnosis	Earlier diagnosis in later birth cohorts could lead to higher prevalence measured at young ages	Hertz-Picciotto and Delwiche (2009)	Found that earlier age of diagnosis could explain 12% of the 585% increase in California from 1990 to 2001 birth years.
Availability of diagnostic resources	Geographic locations with higher levels of diagnostic resources may increase likelihood of diagnosis	Mazumdar et al. (2013)	Found effect of neighborhood clusters, indicate a causal effect of neighborhood on autism diagnosis.
Availability of services	Availability of services for autism may make individuals more likely to be diagnosed with autism	CDC ADDM (n.d.), CDC (2016)	Suggested, not studied.

Hansen et al. (2014) estimated the separate and joint effects of a change in diagnostic criteria from ICD-8 to ICD-10 in 1994 and a change in diagnostic practice from in-patient only to the combination of in-patient and out-patient diagnoses in Denmark in 1995 on the rate of being diagnosed with ASD in the Danish population of individuals born between 1980 and 1991. Using a Cox proportional hazards analysis with stratification by pairs of birth years, they found that 60% of the measured change in prevalence (95% CI 33% - 87%) in the population studied could be could explained by the combination of these two non-etiological factors. Individual factor effects were 33% increase due to diagnostic criteria and 42% from inclusion of outpatient diagnoses. Figures in the paper showed an increase in prevalence measured up to age 22 by birth

year from approximately 16 per 10,000 for birth year 1980 to approximately 116 per 10,000 for birth year 1991, however this aspect is not mentioned in the text. It is not clear whether similar effects are operating in the California population and the CA-DDS data, nor to birth years after 1991 when the bulk of the increase occurred in California. Cumulative incidence to age 10 increased by a factor of 4.18 from birth years 1991 to 2005 in the CA-DDS dataset. Denmark and California differ in the characteristics of the populations, medical systems and aspects of the environment. Denmark used ICD diagnostic criteria and California used DSM criteria, where DSM-IV was introduced in 1994, the same year Denmark introduced ICD-10. The change from in-patient only to and both out-patient and in-patient diagnoses did not occur in California.

King and Bearman (2009) examined the CA-DDS data and found 631 individuals (9% of the caseload) born before 1987 who initially had a diagnosis (classification for services) of mental retardation (MR) whose classification was later changed either to include both autism and MR (diagnostic accretion) or solely autism (diagnostic substitution). Of these 631 cases, 87% experienced diagnostic accretion and 14% experienced diagnostic substitution. They created a simulation model under which they estimated that 26% of the increased CA-DDS caseload from 1992 to 2005 was due to individuals being re-classified via either diagnostic accretion or diagnostic substitution. They assumed that the same process of diagnostic change that applied to individuals born before 1987 also applied to those born after 1992. Details of the calculation are not in the paper and are said to be in an appendix, however the appendix is not available from the journal and I contacted the authors who were unable to provide a copy. The process of diagnostic change they described is based on the introduction of DSM-IV criteria in 1994; the authors did not explain how that change would affect individuals born after 1992, and such an effect seems

unlikely. The authors stated that the previously existing evidence indicated that diagnostic substitution was not occurring in California.

Shattuck (2006) examined administrative data from public schools in the US from 1984 to 2003 and found a decrease in prevalence of mental retardation and learning disabilities from 1994 to 2003 in contrast to the trends from 1984 to 1993, coincident with an increase in autism prevalence from 1993 to 2004. They found this in many but not all states, and in particular not in California. They refer to this as an ecologic analysis as it operates at the level of states or the entire nation rather than individuals. In particular, a state level reduction in prevalence of mental retardation is compared to an increase in the same state in autism prevalence, without considering whether the diagnoses of individuals changed. They concluded that there is not an autism epidemic because the administrative prevalence figures for most states were lower than epidemiologic estimates, and suggested that diagnostic substitution may explain at least part of the administrative increase of autism in most states. In the Shattuck paper the expression "diagnostic substitution" means diagnosing an individual with autism who would otherwise have received a different diagnosis in earlier years, in contrast with the definition used by King and Bearman (2009). The author mentioned that Mandell and Palmer (2005) had performed a nationwide cohort study from 1992 to 2001 and found no decreases in prevalence of mental retardation or speech-language impairment. They did find a nationwide increase in the combination of autism and ID over this interval, the increase being substantially less than the increase in autism alone.

Polyak et al. (2015) examined special education administrative data in the US and compared the nationwide increase in prevalence of autism from 2000 to 2010 to the nationwide decrease in the

prevalence of intellectual disability (ID, elsewhere referred to as mental retardation or MR), combined with no significant change in the overall proportion of children in special education under the federal IDEA rules during this same interval, with classifications of autism, ID, specific learning disability or other health impairment. That is, they found a negative correlation of trends of autism prevalence versus trends of other disabilities, notably ID, nationwide. They found this correlation in some but not all states, with a wide variation of correlation values across states. Based on these findings they concluded that diagnostic recategorization from comorbid conditions to autism might explain some portion of the increase in autism prevalence. The use of diagnostic recategorization implies a similar meaning to diagnostic substitution as used by Shattuck (2006) but not King and Bearman (2009).

Croen et al. (2002) concluded that a reduction in the prevalence of MR diagnoses could explain the increase in the prevalence of autism in the US. However, a re-analysis by Blaxill, Baskin and Spitzer (2003) showed that this conclusion was not supported by the data. Blaxill et al. raised several arguments, the primary one being that within each individual 2- or 3-year sub-interval of the overall interval 1987 to 1994 the changes in prevalence of MR do not correspond to opposite changes in prevalence of autism. While we were not able to obtain a copy of the published reply from Croen et al. (2003), Croen stated (2017) that that she disagreed with many of the points made by Blaxill et al. but "we recalculated prevalence just among kids who had 6 years of follow-up and found that the increasing prevalence of autism was not matched by a decreasing prevalence of MR".

Fombonne (2009) reviewed 43 studies that estimated the prevalence of autism and related disorders. He concluded that there is evidence that broadening of the autism phenotype, changes

in diagnostic criteria, availability of services and improved awareness explain at least some of the measured increase in autism prevalence, and that there was not sufficient evidence to attribute the increase directly to an increase in incidence of cases (not of diagnoses) of the disorder. Nevertheless, he concludes that the possibility of a true increase in disorder incidence cannot be ruled out based on existing epidemiological evidence.

Elsabbagh et al. (2012) performed a systematic review of epidemiological surveys worldwide of autism and pervasive development disorders. They summarize the findings of the papers included in the review, including prevalence estimates and clinical presentation. They concluded that the increase in reported autism prevalence is likely to be the result of broadening of the concept of autism, diagnostic switching (substitution or recategorization), availability of services and awareness, however without citing specific evidence supporting that conclusion.

Keyes at al. (2012) used an Age Period Cohort (APC) approach to estimate the magnitude of the cohort (birth year) effect for birth years 1992 to 2003 in California, and found that the cohort effect explains essentially all of the observed increase. The analysis used a constraint that the age factor is constant after age 8, without stating a justification for that assumption; this is one type of APC analysis. The Statistical Analysis section below explains the use of age factor constraints in APC analysis. Spiers (2013) pointed out that the method used for the analysis is extremely sensitive to the details of the chosen constraint and with a slightly different constraint the conclusion could have been that the period (i.e., diagnostic year) effect explained the measured increase, as explained by Rodgers (1982). Keyes (2013) replied that the simple constraint that diagnosis (rate) is constant after age 8 is reasonable because diagnoses are more common among 3 and 4 year olds, without further explanation. We note that other constraints are possible

following the same logic, for example one could assume a reduction in diagnosis rate starting at age 8. The reasoning for specifying the constraint is not sufficient, as pointed out in Spiers (2013) and Rodgers (1982), and later sections of this paper show that the way Keyes at al. estimated the age factor using a cohort analysis inherently tends to cause the result to indicate primarily a cohort effect; it is just as possible that data are explained by any combination of cohort and period effects. Keyes and Bearman's (2013) reply agreed with this, saying that their evidence is consistent with both period and cohort based explanations. Keyes et al. (2012) argue that linearly increasing awareness of autism could be the reason for the observed cohort effect, however increasing awareness is a factor that affects all ages albeit not necessarily equally and not just one birth cohort, hence it is consistent with a period (diagnostic year) effect, not with a cohort (birth year) effect. A later section of this paper shows this aspect of awareness more fully. In other words, if Keyes et al. (2012) had successfully shown that cohort (birth year) effects fully explained the observed increase in autism prevalence it would have shown that rising awareness could not have been a significant factor and rather that there must be a strong increase in true autism case prevalence caused by significantly increasing environmental factors, the opposite of their conclusion.

Mazumdar et al. (2013) studied the CA-DDS data and looked for spatial clusters at time of birth and time of autism diagnosis, restricting the analysis to only those children diagnosed at three or four years of age and only those children with a sole diagnosis of autism, that is, without mental retardation (MR or ID). They found that those children who move into a neighborhood with more diagnostic resources from a neighborhood with less such resources are more likely to receive a diagnosis than those who remain in a less resource-rich neighborhood, and the association was strongest among children with higher levels of functioning. They identified three

regions in the greater Los Angeles area that were associated with significantly increased risk of autism diagnosis. They state "Our findings implicate a causal relationship between neighborhood-level diagnostic resources and spatial patterns of autism incidence but do not rule out the possibility that environmental toxicants have also contributed to autism risk." They do not provide any reasons why their findings might even plausibly rule out environment toxicants as a significant cause, only that environmental toxicants are unlikely to fully explain their findings. The restriction of subjects to only three or four years of age omits a large portion of all diagnoses, implying they may have simply observed an earlier age of diagnosis for some subjects, particularly those with less severe symptoms, in resource-rich neighborhoods. The effects of earlier ages of diagnosis and inclusion of milder cases is roughly consistent with the findings of Hertz-Picciotto and Delwiche (2009). The earlier age of diagnosis combined with the restriction of no ID and the finding that the association is strongest for higher functioning children implies that the association may not hold for children with more severe core autism symptoms or lower levels of mental function.

Hertz-Picciotto and Delwiche (2009) appears to be the most rigorous and objective study of those we identified that seek to explain the measured increase in autism in California as a function of non-etiologic factors. They examined the CA-DDS data from 1990 through 2006. They considered incident cases (here referring to incidence of diagnosis), cumulative incidence, the distribution of ages at diagnosis and the inclusion of milder cases. They found that cumulative incidence to age 5 increased from 6.2 per 10,000 1990 births to 42.5 per 10,000 for 2001 births, an increase of 585%. They found via quantitative analysis that changes in diagnostic criteria could explain a 120% increase (approximately 20% of the total increase) over this interval, earlier ages of diagnosis could explain a 12% increase and the inclusion of milder

cases could explain at 56% increase, and together these factors cannot explain the observed increase. They found that cumulative incidence to age 10 minimizes the effect of decreasing age at diagnosis and diagnoses above that age are infrequent.

One specific non-etiological factor that is an obvious candidate for explaining at least part of the increased in autism prevalence or cumulative incidence results from changes in the criteria and practice used by the CA-DDS to enroll individuals under the classification of autism. As noted above Hertz-Picciotto and Delwiche (2009) found that changes in diagnostic criteria, such as from DSM-III-R to DSM-IV, could explain approximately 20% of the increase in CA-DDS served autism cases for birth years from 1990 to 2001. The California legislature amended the Lanterman act effective August 2003 to require significant functional limitations in three or more of a specified set of areas of major life activity, which was a stricter requirement than that which applied previously.

Etiologic Hypotheses

Ng et al. (2017) performed a scoping review of published literature with information regarding potential environmental etiological factors for autism. After screening over 50,000 papers they identified 315 articles to retain. They excluded studies whose epidemiological associations were not directly related to etiology, those that used animal models, studied cells or were purely genetic studies. They also excluded commentaries, editorials, letters, news articles and studies whose main focus was not ASD etiology. They found consistent support for several factors: chemical factors including air pollution related to traffic, advanced parental age, preterm birth, low birth weight, pregnancy complications, and maternal immigrant status. They found a substantial amount of literature with some support, however inconsistent, for effects from

mercury exposure. They did not find support for an association with vaccines. They found that inconsistencies and lack of specificity in the literature made it difficult to draw conclusions regarding causality. Ng et al. did not quantify the magnitude of the effects of the factors considered.

Kamowski-Shakibai, Kollia and Magaldi (2017) examined the influences of advanced parental age and assisted reproductive technology on the relative risk of bearing children who would become diagnosed with ASD. They found a significantly elevated risk from the combination of maternal age over 35 and infertility but not from advanced maternal age alone. They also considered communication disorders and found associations with advanced maternal and/or paternal age, but not an association with paternal age and ASD alone.

Becerra et al. (2013) studied the influence of traffic related air pollution in the Los Angeles area during pregnancy on the risk of autism. This is an ecological study that studied exposures by populations not individuals, and as such the findings are not as strong as they would be if they had measured individual level exposure. They found significant risks from exposure to ozone, particulate matter <= 2.5μm, nitric oxide and nitrogen dioxide. Volk et al. (2013) performed a case control study in California using the CHARGE (Childhood Autism Risks from Genetics and the Environment) study sample. They found significantly elevated risks associated with exposure to traffic related air pollution during gestation and especially during the first year of life. They also found elevated risks from exposure to nitrogen dioxide, particulate matter ≤ 2.5μm and ≤ 10μm.

Sanders et al. (2012) performed whole exome sequencing of 928 individuals including 200 discordant sibling pairs from 238 families and found large effects on autism status from de novo

mutations in brain-expressed genes. For 279 de novo mutations they found an instance in autistic probands and none in unaffected siblings. What is particularly interesting about de novo mutations is that they are not inherited, which raises questions about what is causing the de novo mutations. Few if any studies have examined whether the rate of de novo mutations associated with autism has increased over decades of birth years. If de novo mutations cause a significant portion of the observed increase in autism prevalence, that would tend to imply that some environmental factor is increasingly causing the incidence of the de novo mutations that are factors in autism.

Hallmayer et al. (2011) considered levels of concordant and discordant autism among monozygotic and dizygotic twin pairs and estimated the degrees of heritability and shared environment as risk factors. They found that approximately 55% of the autism liability can be explained by shared environmental factors and 37% can be explained by heritability. As in other twin studies, the contributions from these factors were estimated using a model of additive genetics, common environment and unique environment, which is generally referred to as an ACE model. ACE models inherently assume there is no interaction between genetic and environmental factors, and as the paper points out, if there were such an interaction the actual environmental component would be greater than stated in the paper and the genetic component would be over-estimated. While this and other twin studies are often cited as evidence that autism is inherited in the classic sense, heritability may result from de novo mutations which are not inherited. Goldani et al. (2014) point out that differences in symptom severity in concordant twin pairs is strong evidence for non-genetic epigenetic factors.

Shelton et al. (2014) studied the relationship between autism and exposure during pregnancy to agricultural pesticides in the CHARGE study. The found significantly increased risk of autism from exposure to organophosphates, chlorpyrifos and pyrethroid insecticides.

Croen et al. (2011) studied the effects of prenatal maternal exposure to selective serotonin reuptake inhibitor (SSRI) antidepressants on ASD outcomes in a case control study. Six-point seven percent of the cases had prenatal exposure to antidepressants. They found an adjusted odds ratio of 2.2 for SSRI exposure during pregnancy, and an adjusted OR of 3.8 for SSRI exposure during the first trimester. They found no increased risk for mothers with a history of mental health treatment without prenatal SSRI exposure.

Some investigators have proposed epigenetic factors as a possible explanation for the increase in autism prevalence. For example, Goldani et al. (2014) found evidence for epigenetic factors in the etiology of autism, and explain the various epigenetic influences. One source of evidence cited is the substantial difference in symptom severity within monozygotic twin pairs concordant for ASD. One type of epigenetic change occurs via differences in methylation, and Goldani et al. cite studies showing DNA methylation differences at numerous loci, and those showing an association between ASD symptom severity and DNA methylation at numerous sites. Epigenetic changes have been shown to be associated with environmental exposures. Lyall et al. (2017) also cite evidence for epigenetics as an etiological factor, including finding epigenetic changes in the brains of individuals with ASD and in more readily accessible tissues. They also state that rare genetic variants associated with ASD implicate chromatin remodeling, a type of epigenetic modification.

Specific Aims

The primary specific aim of this study is to estimate, as accurately as possible, the true change in the cumulative incidence of autism to a specified age (e.g., 10) by birth year, accounting for all non-etiologic factors, among California residents using the primary dataset, which covers the years 1980 through 2015. In general, autism can mean either autistic disorder or the broader category of ASD. The specific definition that applies to the analysis depends on the available datasets. The dataset used for this study is specific to autism (i.e., autistic disorder) and hence that is the meaning of autism in this study. Our objective was to have the estimating process account for all non-etiologic factors including but not limited to diagnostic criteria, diagnostic practices, age of diagnosis, awareness and availability of services, and interactions between such factors and age at diagnosis. The choice of age 10 as the last year for analysis of cumulative incidence is based on Hertz-Picciotto and Delwiche (2009), which analyzed the effect of earlier ages of diagnosis on the measured prevalence and found that diagnoses up to age 10 cover the vast majority of diagnoses in recent decades.

Diagnostic criteria have changed over time. In the United States, autism criteria have been specified in different time intervals by DSM-III, DSM-III-R, DSM-IV, DSM-IV-TR and DSM-V. While there have been some identified changes in diagnostic practice, in general such changes are not readily identified. Awareness is a general concept that is not measured with accuracy. Availability of services is likewise amorphous, although there have been some policies such as the US federal IDEA (Individuals with Disabilities Education Act) which mandates services to qualified public school students and state level laws which mandate insurance coverage for some autism treatments. Rather than attempting to quantify the effects of each of these and other non-

etiologic factors, we attempt to analyze the aggregate effect of all non-etiologic factors statistically, as explained in the Methods section.

The primary hypothesis is that there has been a significant increase in the cumulative incidence of autism by birth year since approximately 1980, and changes in non-etiologic factors explain only a portion of the measured increase in prevalence. The shape of the trajectory of cumulative incidence by birth year, including possible inflection points, may indicate candidate causes suitable for further investigation, however such inflections are beyond the scope of this study.

Methods

Overview

This study attempts to estimate the cumulative incidence of autism diagnoses by birth year and gender while accounting for all non-etiologic factors. We examine the use of multiple methods to perform this analysis. The methods considered include regression, regression with constraints on the effect of age based on Age Period Cohort (APC) theory, and novel direct analytical methods. The Statistical Analysis section below discusses the details, advantages, disadvantages and implications of each method.

The approach uses statistical methods to attempt to isolate the effects of the aggregate of all non-etiologic factors from incidence by birth year. By treating the aggregate of non-etiologic factors as a single variable, it is not necessary to know specifically what the individual factors are nor is it necessary to separate the effects of individual diagnostic and other non-etiologic factors. Many non-etiologic factors are unobservable and hence not amenable to direct analysis. The effects of individual non-etiologic factors would be nuisance variables in terms of the primary aim of this

present study. However, there may be interest in estimates of the effects of certain identifiable changes in diagnostic and other non-etiologic factors. For example, at years where there were known changes in diagnostic criteria, it may be possible to evaluate terms that represent those changes.

The null hypothesis H_0 is that the cumulative incidence by birth year of autism has not changed over the range of birth years of interest and therefore the entire measured increase in prevalence is due to the complete set of non-etiologic factors. The alternative hypothesis H_A is that the cumulative incidence of autism by birth year has changed significantly over the study period. If H_A is correct then logically the increase in cumulative incidence is due at least in part to changes in etiologic factors.

This study uses a retrospective cohort design using an available dataset.

Study Population

The study population consists of California residents born in years from 1980 to 2015 for the primary dataset. The earliest and latest diagnosis years studied are bounded by availability of data.

Data Sources

The primary dataset is a tabulation of the incidence of diagnosis of autism from the California Department of Developmental Services (CA-DDS). The Introduction section above describes the CA-DDS dataset and how cases are ascertained. The CA-DDS provided a table giving diagnosis year, birth year, sex and number of new diagnoses for each combination of values of those variables, in response to a public document request.

The denominator group is the population in California corresponding to each diagnosis count by birth year, sex and diagnosis year in the CA-DDS dataset. These population counts include the effects of immigration and emigration as well as deaths, as opposed to counting only live births from each birth year minus later deaths. Immigrants to and emigrants from California could include a greater or lesser proportion of individuals with autism, some of whom may have previously been diagnosed in other geographic regions, than those born in California who stayed in California. The secondary dataset is a set of tables provided by the US Census Bureau specifying the estimated population of each state for each applicable year and each applicable age and gender. These tables are provided in different formats for different decades. We filtered and re-arranged these tables, merged them into one series for all of the applicable years, and merged the result with the CA-DDS data to create the primary combined dataset that includes the probability of diagnosis for each year and age as the ratio of the number of new cases divided by the population for that year, age and gender, or combined genders as applicable.

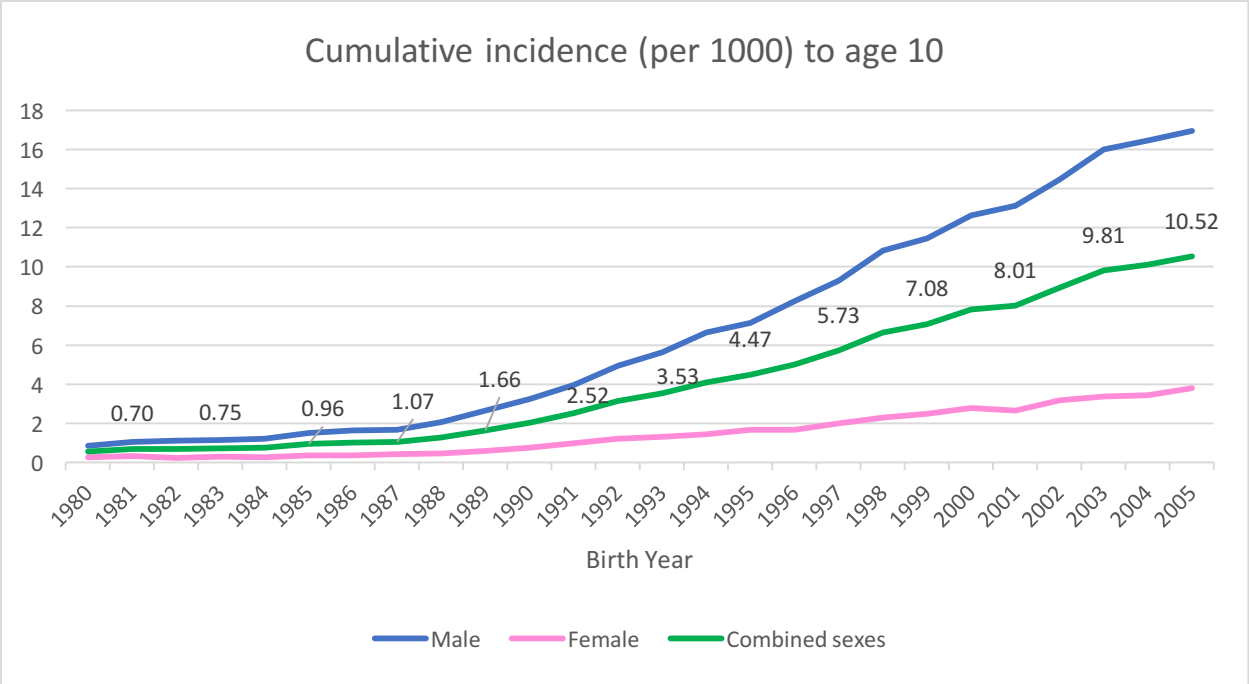

Figure 2 Cumulative incidence (x1000) to age 10 for males, females and combined

Figure 2 illustrates summary cumulative incidence of autism diagnosis in the CA-DDS dataset. Cumulative incidence to age 10 approximates prevalence and is measured consistently for all birth years. We can compare this with *Figure 1*, which shows overall prevalence by study year, not by birth year, from a different analysis. In *Figure 2* the cumulative incidence of the combined sexes for birth year 2005 is 10.52 per 1000, and in *Figure 1* the prevalence for study year 2009 is shown as 1/110 which is 9.09 per 1000. These are roughly comparable but cannot be compared directly due to methodological differences.

Statistical Analysis

Relationship of Birth Year and Diagnostic Year Effects to Etiologic and Non-Etiologic Factors

Our goal is to directly estimate the probability P(diag) of being diagnosed with autism as a function of birth year while controlling for diagnostic year using various analytic methods. We can readily convert this result into cumulative incidence by birth year and hence into changes of cumulative incidence over the range of years studied and test the null and alternative hypotheses H_0 and H_A respectively. We are interested in changes in true case prevalence by birth year. Prevalence is not a suitable measure, however, since we can only measure diagnoses rather than all true cases, and diagnoses occur at different ages for different individuals. Cumulative incidence for all ages would represent prevalence if there were no effects from death, immigration, emigration, recovery from autism or late onset cases. However cumulative incidence up to a specified year would mean different terminal ages for different birth years, which would introduce a bias. Therefore, we use cumulative incidence up to a consistently specified terminal age, in this case age 10.

Figure 3 shows a directed acyclic graph (DAG) which illustrates the relationships between birth year and etiologic (causal) factors, and between diagnostic year and non-etiologic factors. Any factor that affects changes over time in the probability of being diagnosed must act via the pathways illustrated here, specifically via either non-etiologic factors or time-dependent causal factors.

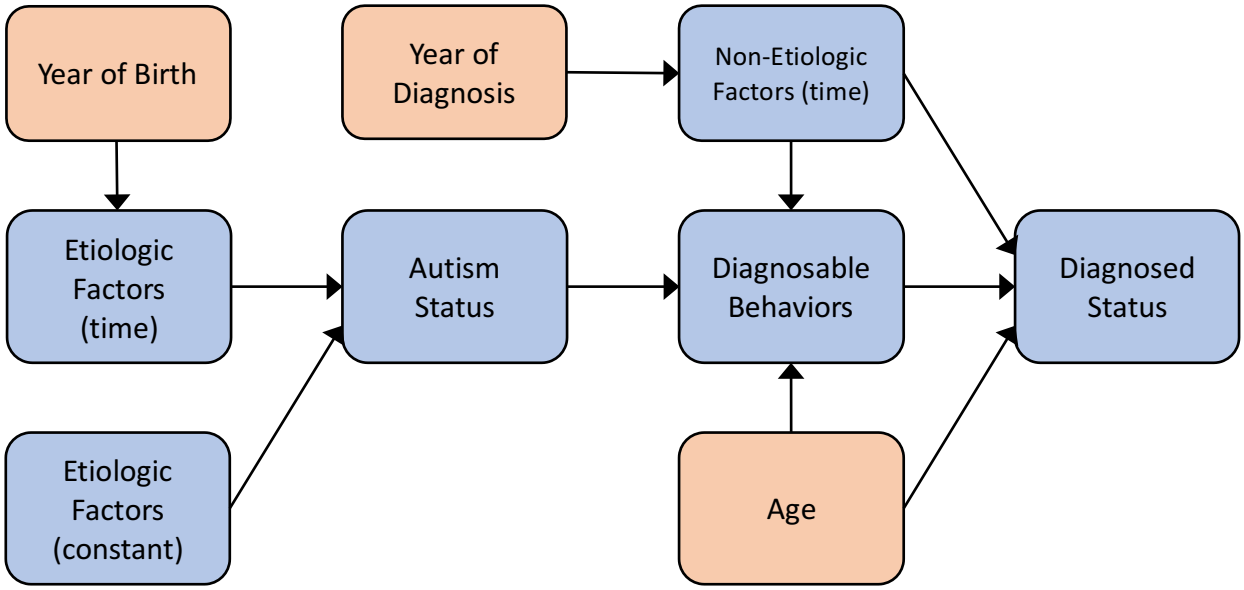

Figure 3 DAG Representing Year of Birth and Diagnostic Factors

Whether or not an individual is ever diagnosed with autism, and at what age, depends on the individual's symptoms, the time trajectory of those symptoms, and non-etiologic factors in effect over time relative to the individuals' birth, as well as the age factor. The non-etiologic factors include diagnostic criteria, diagnostic practices, and all other factors related to diagnosis such as whether and when the individual is evaluated for autism, including general awareness of the disorder.

Here we are interested in the extents to which each of the complete set of non-etiologic factors and the year of birth explain the incident diagnosed status. We are not addressing the causes, or reasons, why an individual has or does not have the status and hence symptoms of autism, nor are we addressing individual non-etiologic or etiologic factors. As *Figure 3* illustrates, each individual's diagnosed autism status depends only on diagnosable behaviors, the complete set of non-etiologic factors, and age. The set of non-etiologic factors is only partially observable;

however, it is a function of time. Age may interact with the non-etiologic factors to produce the outcome, and diagnosable behaviors may change with age. Changes in non-etiologic factors may have different effects on individuals of different ages. For example, if diagnostic criteria were broadened at a specific calendar year, subjects who were 2 years of age at that year would be expected to be more affected by the change of criteria than subjects who were 10 years old that same year, as some of the latter subjects would have already received a diagnosis by age 10, while others of that age not already diagnosed might not be brought in for diagnosis after the change in criteria. Similar comments apply to changes in awareness. The age range of approximately 2 to 6 is the prime age for initial diagnosis and hence changed criteria are more likely to affect individuals at these ages. Beyond specific identifiable factors such as diagnostic criteria, other hypothesized non-etiologic factors such as general awareness, availability of services and social factors are plausibly more likely to affect younger individuals than other ones, while aggressive outreach for adults combined with looser criteria may have a significant effect on adults. Therefore, there may be an interaction effect between age and changes in the set of non-etiologic factors over time.

Diagnosable behaviors are only observable via diagnosis and are caused only by autism status. Autism status manifests only in diagnosable behaviors. Autism status is caused by an unknown set of causes, which are partitioned into those that are a function of time and those that are constant over time. If year of birth has an effect on the outcome it is via the time-dependent causal factors. Constant causal factors constitute an unobservable fixed effect.

The statistical analysis is based on the DAG of *Figure 3*. The independent explanatory variables are year of birth (birthyear) and year of diagnosis (diagyear) and the dependent (i.e., outcome)

variable is the probability of being diagnosed with autism, P(diag). Age is a necessary independent variable, due to its collinearity with birthyear and diagyear. The time-invariant causal factors affect only the baseline value of P(diag). They are represented by the intercept in regression analysis, and they do not affect differences between values of P(diag) for different birth years nor diagnostic years.

Age Period Cohort (APC) Problem Class

The analysis of the distinct effects of birth year and diagnosis year on the probability of diagnosis is in the class of problems known as Age Period Cohort (APC) problems. The birth year represents the cohort; period in this case refers to diagnosis year, and age is the effect of age on the outcome. Age is an important part of the analysis even if we are not interested in finding the value of the age effect. Birthyear + age = diagyear so these three terms are intimately related, and only two of these three can be independent in any given model. As explained in later sections, it is essential to establish a valid value for the age effect in order to estimate the birth year and diagnostic year effects. It turns out that this is both a critical problem and an intractable one, as explained in later sections.

Data Generating Process Model

We assume the data generating process for autism cases and diagnoses operates according to the following model, where the birth year and diagnostic year effects are exponential and the age effect is discrete, (i.e., non-parametric).

Some portion of the children born each year either already have or will have autism, whether or not they are ever diagnosed for the disorder. This is the probability of being a case given the

value of birthyear, p(case|birthyear). This probability is affected by etiologic factors, which are factors that are causal in the disorder, and it is not affected by non-etiologic factors, which can affect the probability of being diagnosed. We recognize that it is possible that some individuals who eventually have autism do not have it at birth and are not predestined to have it later, that is, that there may be causal factors to which they are exposed after birth that influence the probability of becoming a case. The assumption used here is a simplifying assumption and one that we expect most experts would agree with. If it were the case that a post-natal exposure caused an increased probability of becoming a case, and that exposure changed with year, the result could be an interaction between birth year and age, since such cases would not be eligible for diagnosis until symptoms appeared, which might be at a later age than for other cases without such a post-natal causal effect. If such a hypothesized exposure occurred at sufficiently early ages that the onset of diagnosable symptoms were not significantly delayed compared to cases without such an effect, the result on the analysis would be negligible.

Some portion of individuals who in fact have autism (i.e., cases) are diagnosed in each diagnostic year. This is the probability of being diagnosed given that one is a case as a function of diagnostic year, p(diag|case, diagyear). This term represents the set of non-etiologic factors other than the age at diagnosis; it does not represent any etiologic factors.

For those cases who are eventually diagnosed, there is a factor describing the probability of being diagnosed given that one is a case as a function of age, p(diag|case, age). For purposes of the data generating process model we assume that this age factor is independent of both the diagnostic year term and the birth year term. For the initial analysis, we assume this age factor of diagnosis probability does not vary significantly with diagnostic year or birth year; if this is true

then the independence assumption is justified. We have chosen to model and analyze diagnoses up to age 10 because we are measuring cumulative incidence to age 10. We assume the age factor is best described as a discrete function, because prior analysis and visualization of the primary dataset indicated that this function is not readily represented as a continuous function, and because there is no apparent advantage in the present analysis to using a parametric model for the age factor. The age factor is not observable, as explained in a later section. Estimates of the age factor from real data are inherently affected by the birth year and/or diagnostic year effects. If we could determine the age factor correctly, we could specify its values as a restriction in a regression on birth year and diagnostic year, but unfortunately this is not possible.

Combining these three factors, which we assume are independent, we can write the data generating process (DGP) for autism cases and diagnoses as:

$$P_{diag} = P_{case|birthyear} * P_{diag|case,diagyear} * P_{diag|case,age} \quad [1]$$

The first two terms are exponential:

$$\mathrm{P}_{case|birthyear} = \mathrm{e}^{(\beta_{0BY} + \beta_{1BY} * birthyear)}$$

$$P_{diag|case,diagyear} = e^{(\beta_{0DY} + \beta_{1DY} * diagyear)}$$

$$P_{diag|age} = \{AgeFactor_{age}\}, age \in [0, 10]$$

This model implicitly assumes that the age factor itself is not a function of calendar year. It is possible that the age factor actually is a function of calendar year. If we had evidence of a dependence on calendar year we could model the age factor accordingly. Estimating the age factor from sample data is problematic, as discussed below, and estimating changes in the age factor by either birth year or diagnostic year is similarly problematic.

The assumption of exponential terms, which is consistent with log-linear analysis as opposed to logit or linear analysis, is the conventional assumption in APC analysis, it appears to be logically consistent with what we would expect in the real world and it appears to be consistent with the primary dataset as illustrated in the following figures. There may be an exception to this assumption if the value of the diagyear term P(diag|case, diagyear) approaches 1 in the interval of interest, that is, the probability cannot continue to increase exponentially as it approaches 1. However, we do not have evidence of this effect in the present analysis.

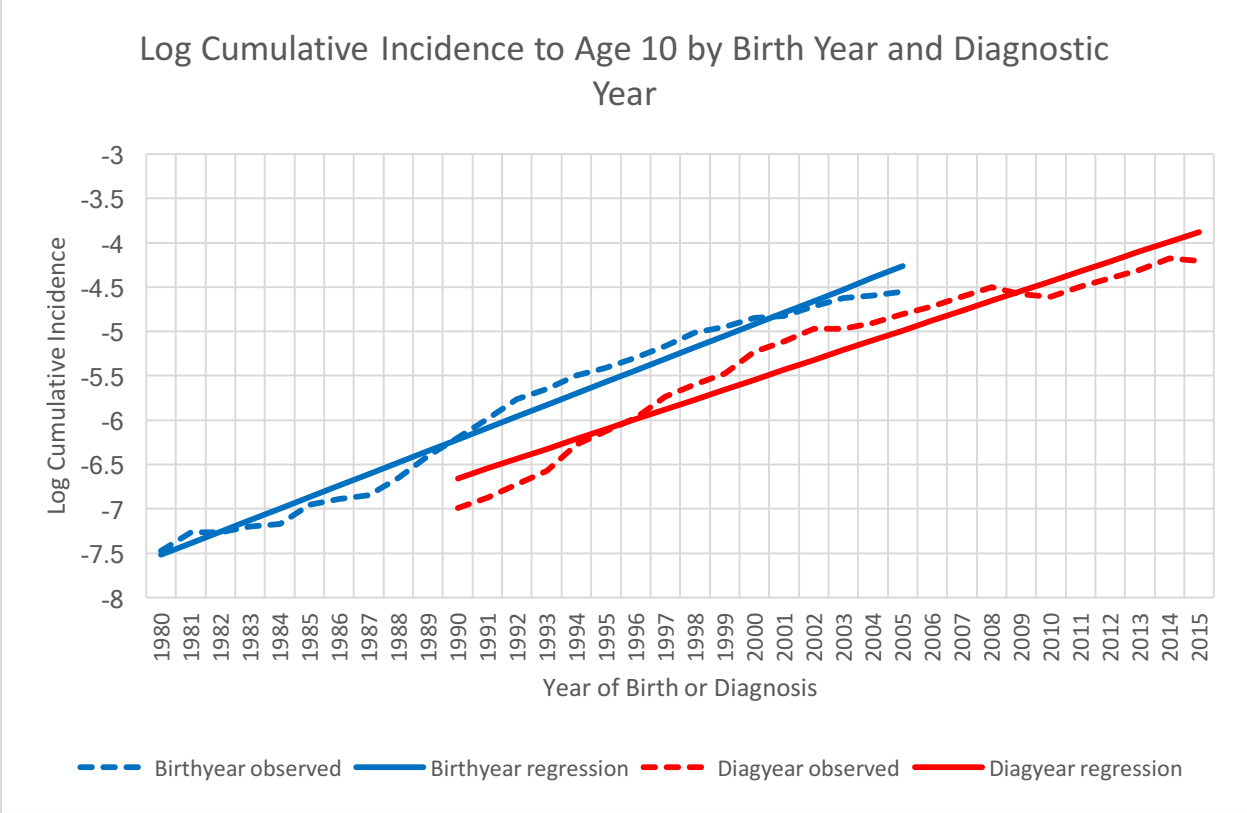

Figure 4 Log probability of CA-DDS cumulative incidence to age 10 by birth and diagnostic years

Figure 4 shows the log probability of cumulative incidence to age 10 by birth year and by diagnostic year from the CA-DDS data. In both cases the data points follow the regression line reasonably well with modest residuals. The R^2 value for birth year is 0.98 and for diagnostic year

the R^2 value is 0.935. These data indicate the data fit fairly well with the assumed exponential data generating process model, where the birth year trend fits this model better than the diagnostic year trend. Residuals from the regression lines could potentially be associated with either etiologic or non-etiologic factors that disturb the simple exponential model. It is beyond the scope of this paper to investigate associations between events and deviations from the model. It may be worth noting that both plots show a residual excursion above the trend line starting at approximately the 1990 birth year and the 1996 diagnostic year.

The slope of the regression line for birth year is 0.130 and the slope for diagnostic year is 0.111. The average of the two slopes is 0.1206, which serves as an estimate of the sum of the coefficients for birth year and diagnostic year. This exponential coefficient corresponds to an increase of 12.82% per year in the probability of being classified with autism in the CA-DDS system among the California population from ages 0 through 10. If the data exactly fit the exponential model, these two slopes would be identical apart from sampling variations, as explained in the section Estimating birthyear and diagyear coefficients given AgeFactor below.

Synthetic and Real Data

We generated various synthetic datasets according to the model specified above using known parameter values in order to test the validity of various analysis approaches. If an analysis performs properly it should generate parameter estimates that correspond to the known true parameter values, and do so across a range of parameter values. We estimated some parameters from the primary dataset and used those in models that generate synthetic data.

Regression

Ideally, we would like to be able to estimate the coefficients for birthyear and diagyear directly using regression. Unfortunately, that approach does not yield reliable results. Nevertheless, a description of the use of regression for this problem serves to introduce the impediments to finding a solution.

Under H_0, all of the measured changes in prevalence are due entirely to changes in the set of non-etiologic factors, with no effect of birth year. In terms of *Figure 3* under H_0 there are no causal factors that are functions of time, since if there were they would differentially affect the autism status of individuals by birth year, which would affect their diagnosable behaviors and hence their probability of being diagnosed with autism even if all non-etiological factors were to remain constant. We designate the complete set of non-etiologic factors as diag(t), which is unobservable. Therefore, under H_0 the function is schematically illustrated as:

$$P_{diag} = f(diag_t, age)$$

where age is the age at diagnosis. Since diag(t) is not observable we cannot analyze it directly. However, it is a function of time, hence we can create a regression model using corresponding observable elements as:

$$P_{diag} = \beta_0 + \beta_1 * diagyear + \beta_2 * age$$

where regression uses the logarithmic link function and diagyear is year of diagnosis. By substitution diagyear corresponds to diag(t).

Under H_A, p(diag) is a function of birth year as well as the terms that contribute to the H_0 form. The regression model is represented as:

$$P_{diag} = \beta_0 + \beta_1 * birthyear + \beta_2 * diagyear + \beta_3 * age$$

Here β_1 corresponds to the effect of birth year on the probability of being diagnosed p(diag) after accounting for all non-etiologic factors and the age factor.

Age at diagnosis is equal to diagyear - birthyear hence it is collinear with them so we cannot estimate a parameter for age. If we omit the age term, that is equivalent to assuming that the effect of age is 0, which is generally not correct. Setting the age coefficient to 0 in a log-linear regression is equivalent to assuming that the age factor has a constant value for all values of age. That is, each case is equally likely to be diagnosed at each of the ages in the range of ages considered. Experiments using synthetic data confirm that estimates produced using a regression model without an age term are incorrect, with very large errors in the estimates. Any assumption of the age factor being constant would not be consistent with most real-world data, particularly the primary dataset of this study.

Another problem with performing regression on APC problem stems from the fact that diagyear = birthyear + age. Substituting this expression into the regression above and omitting the age term yields the follow regression.

$$\begin{aligned} P_{diag} &= \beta_0 + \beta_1 * birthyear + \beta_2 * diagyear \\ &= \beta_0 + \beta_1 * birthyear + \beta_2 * (birthyear + age) \\ &= \beta_0 + (\beta_1 + \beta_2) * \text{birthyear} + \beta_2 * \text{age} \end{aligned}$$

Substituting birthyear = diagyear – age yields a similar type of rearrangement of the regression. The equivalences indicate confounding between age, birthyear and diagyear, implying an assumption that either birthyear or diagyear have a coefficient of zero. Also, the final versions imply an assumption that the age factor is well represented by an exponential function, which may not be the appropriate. In the present study, it is apparent that the age factor does not fit an exponential model.

We would like to evaluate potential interaction between age and diagyear using standard regression methods:

$$P_{diag} = \beta_0 + \beta_1 * birthyear + \beta_2 * diagyear + \beta_3 * age + \beta_4 * age * diagyear$$

However, since regression methods are not suitable even without inclusion of the interaction term, it is not practical to evaluate regressions that include it.

APC Methods and Limitations

APC applies to many fields of study, such as medicine, criminology, economics and sociology. There is a significant amount of journal literature on the topic of APC analysis; one of the earliest papers on APC is by Mason et al. (1973), which suggests some possible solutions to the inherent problems of APC analysis. Glenn (1976) states what he posits is the futility of separating age, period and cohort effects. A fundamental problem is that age, period and cohort are collinear and are confounded by the fact that cohort + age = period, or equivalently for this analysis, birth year + age = diagnosis year. Rodgers (1982) showed that the methods proposed in Mason (1973) in general do not produce reliable results. In particular, he showed that imposing a constraint on the problem to make the terms estimable, “in fact it is exquisitely precise and has

effects that are multiplied so that even a slight inconsistency between the constraint and reality, or small measurement errors, can have very large effects on estimates." Numerous authors have proposed methods that can provide at least partial solutions under certain circumstances. An excellent summary of the methods from the APC literature is the book Age-Period-Cohort Models (O'Brien, 2015) which provides a clear explanation of the problems with using statistical methods to disentangle age, period and cohort effects, and explains some methods that may provide useful results in certain cases, with some limitations, and gives a worked-out example in criminology.

One fundamental problem with APC statistical analysis is what is often referred as the lack of identifiability. That is, there is at least one more variable than there are independent equations, therefore there is an infinite number of possible solutions that satisfy the constraints. Regression methods may generate estimates for the parameters sought but those estimates may contain large errors, possibly without warning from the regression software. In some cases, a regression may fail to converge.

One recommended method of solution which appears to apply to the present analysis is to estimate the age factor based on data other than the primary dataset, apply constraints on the age factor based on this estimate, and set those constraints in the regression model that estimates the birthyear and diagyear effects. The O'Brien book (2015) gives a detailed example from criminology that establishes a set of possible age factor constraints and generates corresponding curves representing the solutions for period and cohort. The age factor constraints used in this example come from a consensus of professional opinion and other sources. The result in that example is a reasonable looking but somewhat wide range of potential valid results. The APC

literature includes much discussion of the potentially arbitrary nature of constraints chosen and applied to APC regressions for the purpose of enabling unique solutions. A sample of the relevant APC literature includes (Mason et al., 1973; Glenn, 1976; Rodgers, 1982; Holford, 1983; Kupper et al., 1985; Robertson, Gandini & Boyle, 1999; O'Brien, 2000; Winship & Harding, 2008; Keyes et al., 2010; Bell & Jones, 2013; Bell & Jones, 2014 and O'Brien, 2015).

For the present analysis of the effects of birth year and diagnostic year on diagnosis probability, we can apply a set of assumptions that may assist in enabling a solution to the APC problem. Specifically, as noted above we can assume there is no effect on case status due to age, which eliminates one source of confounding by the age factor. Under this assumption, the age factor affects only the age at which a case is diagnosed. It does not affect whether or not an individual is a case, and it does not affect whether or not a case is ever diagnosed, as long as the diagnosis occurs no later than the last age considered in the cumulative incidence calculation. This assumption is in contrast to many of problems in the APC class, where age does affect the probability of being a case, for example cancer.

We do not have an independent estimate of the age factor. As explained below under Estimating the values of AgeFactor given values of birthyear and diagyear coefficients, even if we had an independent estimate of the age factor based on other data, that would not be sufficient to produce an unbiased estimate of the desired coefficients. We estimated the age factor by analyzing the primary dataset and used that estimate to form various constraints in regression analyses to estimate the birthyear and diagyear coefficients. Further we performed a sensitivity analysis with minor variations in the constraints based on the estimated age factor. Regression analysis using synthetic data with known parameters as well as using the primary dataset resulted

in a very wide range of coefficient estimates with extreme sensitivity to minor variations in the age factor constraints. That is, the results generally had very large errors, without an evident method to determine reasonable bounds on the true values of the coefficients. This is consistent with (Rodgers 1982). Based on these results, we conclude that APC-based regression approaches are unlikely to produce meaningful results unless we can first establish an accurate and appropriate estimate of the age factor. We also conclude that the simplifying assumption that autism is either present or predetermined from birth is not sufficient to enable a robust solution using APC methods. That is, while age may not confound birthyear or diagyear, the three terms are collinear and hence there is a lack of identifiability and therefore not a unique solution.

New Approaches of Analysis

In this section we show that it is straightforward to estimate the sum of $\beta 1DY + \beta 1BY$. If we knew the values of AgeFactor() in advance we could estimate the individual beta coefficients, or equivalently BetaFraction, which is the fraction of the total that applies to a specific one of the two coefficients. If we knew the value of BetaFraction in advance we could estimate AgeFactor() correctly from the data. Estimating AgeFactor and BetaFraction simultaneously is not possible as the analysis becomes degenerate.

Estimating birthyear and diagyear coefficients given AgeFactor

The section Data Generating Process Model above provides equation [1], which models the assumed data generating process, and the associated equations modeling the birthyear and diagyear effects as exponential and modeling AgeFactor() as a discrete set of values. Substituting the expressions for the individual terms in the process model equation results in equation [2]:

$$P_{diag} = e^{\beta_{0BY}+\beta_{1BY}*birthyear} * e^{\beta_{0DY}+\beta_{1DY}*diagyear} * AgeFactor_{age} \quad [2]$$

Set $\sum_{age=0}^{10} AgeFactor_{age} = 1$. That is, all subjects who are ever diagnosed by age 10 have a probability = 1 of being diagnosed at one of the ages from 0 to 10, where 10 is chosen as the upper age used for cumulative incidence.

When calculating or plotting P(diag) versus either birthyear or diagyear, the value of that independent variable is fixed for each value of the dependent variable. Age, birthyear and diagyear are inter-related with only two independent variables. So in the function of birthyear, replace diagyear with birthyear + age, and in the function of diagyear, replace birthyear with diagyear - age. For either choice of independent variable, the resulting output value for a given value of the independent variable consists of the sum over all values of age using the substituted value of the non-selected independent variable.

P(diag) by birthyear:

$$P_{diag}(birthyear) \quad [3]$$
$$= e^{\beta_{0BY}+\beta_{1BY}*birthyear} * \sum_{age=0}^{10} \left(e^{\beta_{0DY}} * e^{\beta_{1DY}*(birthyear+age)} * AgeFactor_{age}\right)$$

In equation [3], the first term is the probability of being a case as a function of birthyear, and the summed term is the probability of being diagnosed by age 10 given that one is a case.

Since $\sum_{age=0}^{10} AgeFactor_{age} = 1$ we can move the terms $e^{\beta_{0DY}}$ and $e^{\beta_{1DY}*birthyear}$ out of the sum and rearrange as:

$$P_{diag}(birthyear) \quad [4]$$

$$= e^{\beta_{0BY}+\beta_{0DY}} * e^{(\beta_{1BY}+\beta_{1DY})*birthyear} * \sum_{age=0}^{10} \left(e^{\beta_{1DY}*age} * AgeFactor_{age}\right)$$

Similarly for P(diag) by diagyear:

$$P_{diag}(diagyear) \quad [5]$$

$$= e^{\beta_{0BY}+\beta_{0DY}} * e^{(\beta_{1BY}+\beta_{1DY})*diagyear} * \sum_{age=0}^{10} \left(e^{-\beta_{1BY}*age} * AgeFactor_{age}\right)$$

We then convert equations [4] and [5] to linear form by taking the natural log:

F(birthyear):

$$\ln\left(P_{diag}(birthyear)\right) \quad [6]$$

$$= (\beta_{0BY} + \beta_{0DY}) + (\beta_{1BY} + \beta_{1DY}) * birthyear + \ln\left(\sum_{age=0}^{10} \left(e^{\beta_{1DY}*age} * AgeFactor_{age}\right)\right)$$

F(diagyear):

$$\ln\left(P_{diag}(diagyear)\right) \quad [7]$$

$$= (\beta_{0BY} + \beta_{0DY}) + (\beta_{1BY} + \beta_{1DY}) * diagyear + \ln\left(\sum_{age=0}^{10} \left(e^{-\beta_{1BY}*age} * AgeFactor_{age}\right)\right)$$

If we knew the correct value of AgeFactor() we could use this pair of expressions [6] and [7] to find the values of β1BY (birthyear coefficient) and β1DY (diagyear coefficient). The variables birthyear and diagyear in equations [6] and [7] can be replaced with the variable year, where year means birthyear in one case and diagyear in the other. In both equations, the coefficient for year is the sum of the coefficients for birthyear and diagyear, β1BY+β1DY.

Taking the difference between equations [6] and [7], the common intercept β0BY + β0DY and the year term both cancel out and the remaining difference is:

Difference in intercepts = [8]

$$\ln\left(\sum_{age=0}^{10}\left(AgeFactor_{age} * e^{\beta_{1DY}*age}\right)\right) - \ln\left(\sum_{age=0}^{10}\left(AgeFactor_{age} * e^{-\beta_{1BY}*age}\right)\right)$$

To analyze data, find the probability of diagnosis for each value of the independent variable (birthyear or diagyear) where each such probability is the sum of probabilities summed over all values of age, and take the log of each of these values. The range of values of age used for the analysis is based on the greatest year used for cumulative incidence. The range of years naturally differs for analysis by birthyear vs. diagyear due to years for which birthyear and/or diagyear data are available based on the values of age. It is important to select the range of years such that all values of age are represented for each value of birthyear or diagyear. For example, where both birthyear and diagyear data are available from years 1980 through 2015 and the ages 0 to 10 are used for cumulative incidence, analysis by birthyear uses years 1980 through 2005, since birthyear 2005 corresponds to diagyear 2015 which is the last year available. Similarly, analysis by diagyear uses years 1990 through 2015 since 1990 is the first year for which age 10 data are available for birth year 1980.

Using the values of ln(p(diag()) summed over ages vs. year for each of the separate analyses by birthyear and diagyear, perform linear regression vs. year and find the slope and intercept for each. The slopes should be identical but may in practice differ slightly. The value of the slope is the sum of the betas β1BY + β1DY. The difference in the intercepts from these two regressions

is the value that eq. [8] should match given the correct values of AgeFactor(), β1BY and β1DY. In order to obtain the most accurate value of the difference in intercepts from the regression, it is important to offset the year values used in the regression such that year 0 is the median value of the years used in the sums by birthyear and diagyear. This is because the slopes (i.e., regression coefficients) of the sets of sums by birthyear and diagyear may differ, and if the years used in the analysis are significantly different from 0, the difference in slopes translates into an undesired additional difference in intercepts, which biases the results of the subsequent analysis.

Given the value of the sum of betas β1BY+β1DY, we can specify a single control variable BetaFraction that specifies the value of β1BY as a percentage of the sum, thereby specifying the values of both β1BY and β1DY.

Given an assumed value of the discrete set AgeFactor(age), a search over values of BetaFraction using the measured sum β1BY+β1DY and the individual values of β1BY and β1DY based on BetaFraction in equation [8] finds the value of BetaFraction that results in the smallest difference between the resulting value and the actual difference in intercepts of the two linear regressions in the analysis described above. If the values of AgeFactor() are correct, the result should closely approximate the true values of β1BY and β1DY. Experiments with synthetic data show that this is indeed the case; this method produces reasonably accurate estimates of β1BY and β1DY. This method depends on AgeFactor() having a non-zero value at more than one value of age; when there is only one non-zero value the expression in equation [8] is degenerate and does not provide information regarding the best value of BetaFraction.

This method of finding β1BY and β1DY depends on having an accurate estimate of AgeFactor(). Unfortunately that presents a fundamental problem, as explained below.

Figure 4 illustrates the log probabilities of cumulative incidence to age 10 by birthyear and by diagyear from the CA-DDS dataset. As indicated in the accompanying text, the slopes of the regression lines for these data are similar but not identical (0.130 and 0.111 respectively) with an average of 0.1206. If the data exactly fit the exponential model, the slopes of these regression lines would be essentially identical and the slope would be equal to the sum of the exponential coefficients for birth year and diagnostic year. According to the data generating model, this difference in intercepts is a function of the BetaFraction and AgeFactor() and could be used to solve betas if we knew the value of AgeFactor().

Figure 5 illustrates the process of estimating BetaFraction from the data, given an assumed value of AgeFactor(). This example uses synthetic data where the true value of BetaFraction is 1.0. This figure also illustrates the very high sensitivity of the BetaFraction estimate to variations in the assumed value of AgeFactor() that is used in the analysis, and the tendency of the estimated value of BetaFraction to follow the value of BetaFraction that is implicit in the choice of method used to estimate AgeFactor(). The three lines in the figure correspond to 3 different assumptions behind the choice of AgeFactor(). For the line labeled AgeFactor BetaFraction=0, AgeFactor is estimated from the data using the implicit assumption that the BetaFraction value is equal to 0. That is, we estimated AgeFactor() by summing the probabilities over values of diagyear for each value of age, without adjusting the results for the correct value of β1BY, which implicitly assumes that β1BY=0, (i.e., BetaFraction =0). The resulting BetaFraction estimate is approximately -0.1, which is close to the value 0 implicitly assumed in the AgeFactor() estimate,

while the true value is 1.0. If we knew the correct value of β1DY we could adjust for it as shown in the following section. Similarly, where the AgeFactor() estimate implicitly assumes that BetaFraction is 0.5, the estimated value of BetaFraction resulting from the analysis is approximately 0.5, and where AgeFactor estimate implicitly assumes that BetaFraction = 1, the estimated value of BetaFraction is approximately 1, which in this example is the correct value.

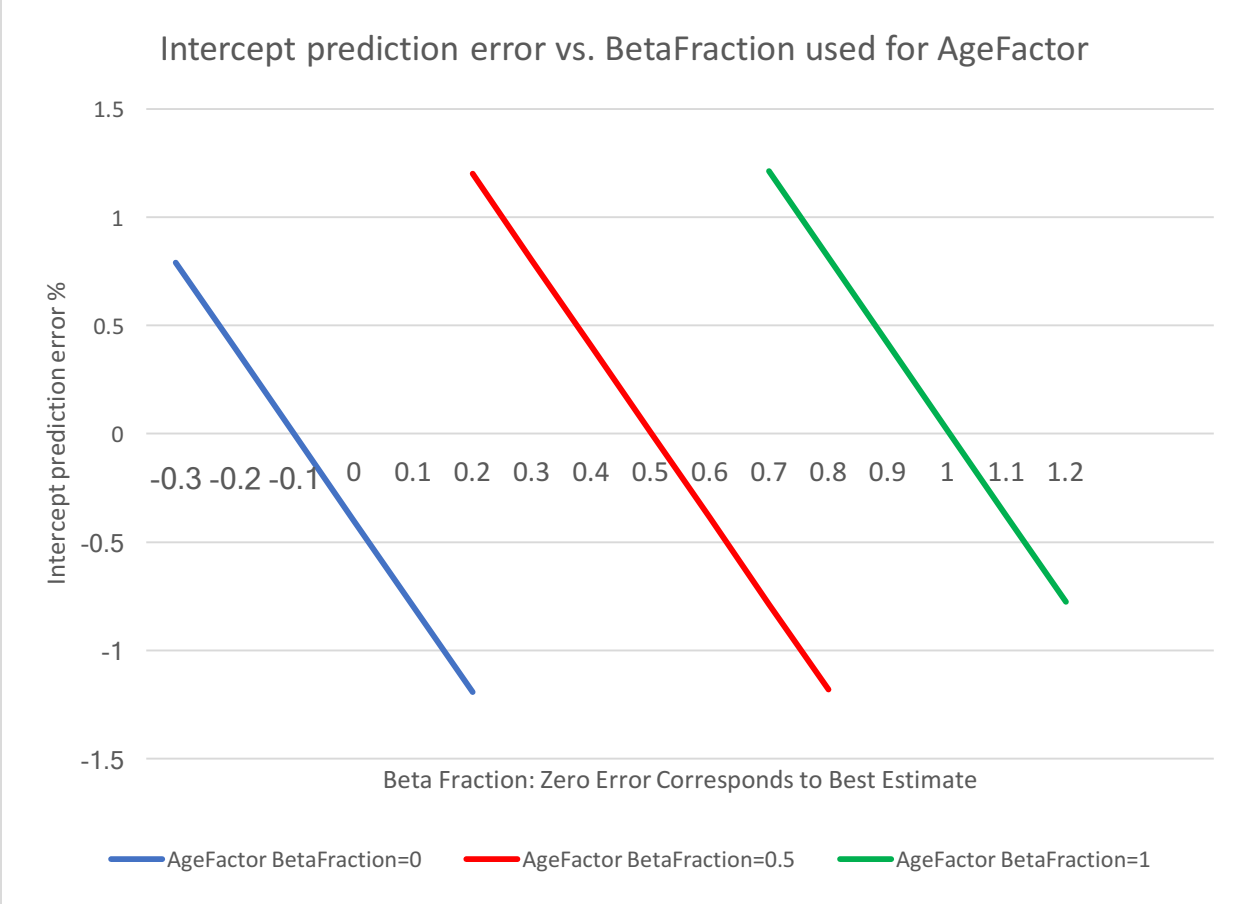

Figure 5 Effect of Choice of Age Factor on Beta Fraction Estimate

Estimating the values of AgeFactor given values of birthyear and diagyear coefficients

We can create two expressions for AgeFactor(age) as functions of sums calculated from the data, the sum of betas calculated from the data as described above, BetaFraction and constants that normalize the sum of the values of AgeFactor() to be equal to 1. We can set these expressions to be equal. If we knew the values of the betas we could find the values of AgeFactor(). However in general we do not have that information before knowing the values of AgeFactor().

Starting from equation [2] we can substitute birthyear+age for diagyear, and re-write the equation as:

$$P_{diag}(birthyear, age) \quad [9]$$
$$= e^{\beta_{0BY}+\beta_{0DY}} * e^{\beta_{1BY}*birthyear} * e^{\beta_{1DY}*(birthyear+age)} * AgeFactor_{age}$$
$$= e^{\beta_{0BY}+\beta_{0DY}} * e^{\beta_{1BY}*birthyear} * e^{\beta_{1DY}*birthyear} * e^{\beta_{1DY}*age} * AgeFactor_{age}$$
$$= e^{\beta_{0BY}+\beta_{0DY}} * e^{(\beta_{1BY}+\beta_{1DY})*birthyear} * e^{\beta_{1DY}*age} * AgeFactor_{age}$$

Substituting diagyear-age for birthyear we can re-write equation [2] as

$$P_{diag}(diagyear, age) \quad [10]$$
$$= e^{\beta_{0BY}+\beta_{0DY}} * e^{\beta_{1BY}*(diagyear-age)} * e^{\beta_{1DY}*diagyear} * AgeFactor_{age}$$
$$= e^{\beta_{0BY}+\beta_{0DY}} * e^{\beta_{1BY}*diagyear} * e^{-\beta_{1BY}*age} * e^{\beta_{1DY}*diagyear} * AgeFactor_{age}$$
$$= e^{\beta_{0BY}+\beta_{0DY}} * e^{(\beta_{1BY}+\beta_{1DY})*diagyear} * e^{-\beta_{1BY}*age} * AgeFactor_{age}$$

To get expressions purely in terms of age, eliminating the variables birthyear and diagyear respectively, re-arrange so AgeFactor(age) is a function of P(diag)(birthyear, age) or

P(diag)(diagyear, age) and sum the probabilities over the applicable values of birthyear or diagyear respectively:

Equation [11]:

$$AgeFactor_{age} = \sum_{birthyear} \frac{P_{diag}(birthyear, age)}{e^{\beta_{0BY}+\beta_{0DY}} * e^{\beta_{1DY}*age} * e^{(\beta_{1BY}+\beta_{1DY})*birthyear}}$$

$$= e^{-\beta_{1DY}*age} * \sum_{birthyear} P_{diag}(birthyear, age) * K_1$$

and

Equation [12]:

$$AgeFxn_{age} = \sum_{birthyear} \frac{P_{diag}(diagyear, age)}{e^{\beta_{0BY}+\beta_{0DY}} * e^{-\beta_{1BY}*age} * e^{(\beta_{1BY}+\beta_{1DY})*diagyear}}$$

$$= e^{\beta_{1BY}*age} * \sum_{diagyear} P_{diag}(diagyear, age) * K_2$$

The value of the constant K1 subsumes the terms in the expression that are constant with respect to age and is selected to force the sum of the values of AgeFactor over the applicable values of age to be equal to 1 when summing over birthyears. Similarly the value of K2 is selected to make the sum of AgeFactor over the applicable ages equal to 1 when summing over values of diagyear. In other words, each discrete value of AgeFactor(age=Age) is found by summing the probabilities of diagnosis over the applicable value of birthyear [11] or diagyear [12] separately for each value of Age and then normalizing the values of AgeFactor such that the sum of the values is equal to 1.

Take the natural log of each of equations [11] and [12]:

$$\ln AgeFactor_{age} = \ln K_1 - \beta_{1DY} * age + \ln\left(\sum_{birthyear} P_{diag}(birthyear, age)\right)$$

$$\ln AgeFactor_{age} = \ln K_2 + \beta_{1BY} * age + \ln\left(\sum_{diagyear} P_{diag}(diagyear, age)\right)$$

Take the sum of those two equations:

$$2 * \ln AgeFactor_{age}$$
$$= \ln K_1 + \ln K_2 + (\beta_{1BY} - \beta_{1DY}) * age$$
$$+ \ln\left(\sum_{birthyear} P_{diag}(birthyear, age)\right) + \ln\left(\sum_{diagyear} P_{diag}(diagyear, age)\right)$$

Re-write in terms of SumBeta and BetaFraction. SumBeta = β1BY+β1DY and β1BY=BetaFraction*SumBeta. Replace ln(K1)+ln(K2) with K3.

$$2 * \ln AgeFactor_{age} \qquad [13]$$
$$= K_3 + (2 * BetaFraction - 1) * SumBeta * age$$
$$+ \ln\left(\sum_{birthyear} P_{diag}(birthyear, age)\right) + \ln\left(\sum_{diagyear} P_{diag}(diagyear, age)\right)$$

If we knew the value of BetaFraction we could solve for AgeFactor and K3, the latter being a nuisance variable.

An important conclusion is that the values of AgeFactor() depend on the value of β1DY or of β1BY according to which of birthyear or diagyear is used for summing the probabilities for each age, and those beta coefficient values are not known in advance when AgeFactor() will be used to estimate the Beta values. In the general case where the values of β1DY and β1BY are both non-zero, the correct or approximate value of the appropriate beta coefficient needs to be used in

equations [11] and [12] to adjust the values resulting from summing the probabilities for each age, before normalizing, in order to produce an accurate estimate of AgeFactor().

In experiments using synthetic data, we know the values of all of the parameters and we can measure the degree to which the estimated values of AgeFactor() match the actual values of AgeFactor(), given the known values of β1DY and β1BY. Experiments produce excellent matches to the actual values of AgeFactor() when analyzing by either birthyear or diagyear. However in those experiments we can specify the correct values of β1DY and β1BY because we know them in advance; clearly that does not apply to real data.

Here is an example that illustrates the sensitivity to the choice of method used to estimate the value of AgeFactor(). Here we use idealized synthetic data from a simple model, where there is no sampling error and the true age factor is specified as 0.25 for ages 3, 4, 5 and 6, and 0 for all other ages. We avoid sampling error by using continuous functions rather than binomial random number generation for the numbers of cases and diagnoses, which is used in the other synthetic datasets.

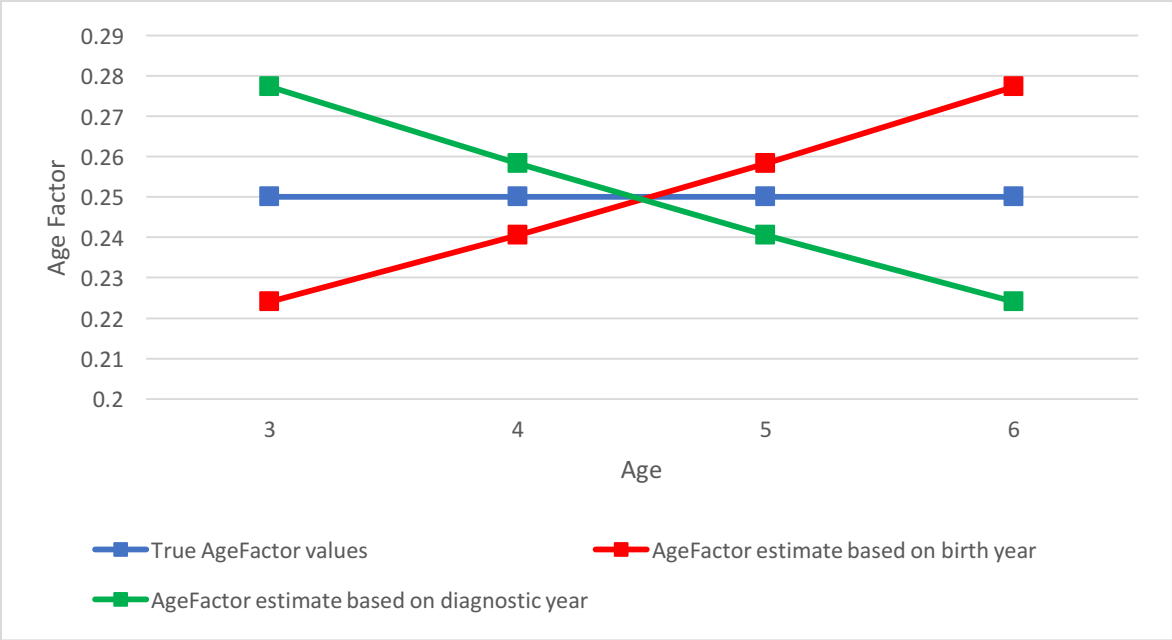

Figure 6 Effect of AgeFactor Estimation Method

Figure 6 illustrates the true age function and the estimates of the age function that would be obtained by summing over birth years in one case, and over diagnostic years in another case, without adjustment for assumed values of β1BY and β1DY. That is, the birth year version implicitly assumes that BetaFraction = 1 (i.e., β1BY = SumBetas) and the diagnostic year version implicitly assumes that BetaFraction = 0 (i.e., β1DY = SumBetas). Summing over birth years without adjustment results in an upward tilt of the AgeFactor estimate, and summing over diagnostic years without adjustment results in a downward tilt of the AgeFactor estimate. This observation helps to visualize the problem with Keyes et al. (2012) as pointed out by Spiers (2013). Keyes et al. estimated the age function from cohort (birth year) data, which would result in an upward tilt of the age factor estimate if the period (diagnostic year) coefficient were positive, and that would result in a bias in the resulting estimates of the birth year and diagnostic year coefficients.

We then used the three resulting values of AgeFactor(), the true value and the estimates based on birthyear and diagnostic year, to estimate the value of BetaFraction. When using the true value of AgeFactor(), in which BetaFraction = 0.5, the estimated value of BetaFraction is exactly 0.5 which is the correct value. When using the unadjusted birth year version of AgeFactor(), the BetaFraction estimate is exactly 1.0, which is the same value implicitly assumed in the estimation of AgeFactor. When using the unadjusted diagnostic year version of AgeFactor(), the BetaFraction estimate is exactly 0, which again is the same value implicitly assumed in the estimation of AgeFactor. The estimate of the beta values depends entirely on the implicit assumptions made in the estimation of the age function, rather than the actual values of the betas, in this idealized case with no sampling error and the data exactly fit the assumed model. This effect helps to explain the results of the following section, which concludes that it is not possible to solve for BetaFraction and AgeFactor() simultaneously.

This effect of the value of BetaFraction implicit in the method used to estimate AgeFactor(), and the resulting effect on the estimate of BetaFraction using the estimated AgeFactor(), may have implications for a broad range of age period cohort problems.

Simultaneous Solution of β1BY, β1DY and AgeFactor

We investigated two potential methods of solving for β1BY, β1DY and AgeFactor() simultaneously. Solving via simultaneous equations based on equation [13] is not possible because there is one less equation (or constraint) than the number of variables to be solved; that is, there is a lack of identifiability. AgeFactor() is a set of values, one for each value of Age used in the analysis. For example, when estimating cumulative incidence to age 10, there are 11 values of age (0 through 10) and hence 11 values of AgeFactor(). K3 and BetaFraction are both

unknowns so there are 13 unknowns. The last two terms in [13] are known as they are calculated from sums of the data being analyzed, and SumBeta is known from analyzing the data as described above. There are 11 simultaneous equations corresponding to equation [13] and the 11 values of age, and one more equation specifying the constraint that the sum of the values of AgeFactor() is equal to 1, so there are 12 equations and 13 unknowns, therefore the solution is not identifiable. There is an infinite number of solutions so neither AgeFactor() nor BetaFraction can be determined.

We also investigated a brute force maximum likelihood estimation (MLE) approach to solving BetaFraction and AgeFactor() simultaneously. That approach is based on the method described above in the section Estimating birthyear and diagyear coefficients given AgeFactor, extended to specify the appropriate value of AgeFactor() for each value of BetaFraction in the search, rather than a fixed value of AgeFactor(). In this way, the values of β1BY, β1DY and AgeFactor() are searched simultaneously based on the single control variable BetaFraction. However, experimental evidence using synthetic data shows that the predicted value of the difference in intercepts is nearly identical for all values of BetaFraction, and the error between the predicted and actual values of the difference intercepts is very small for all values of BetaFraction. This directly implies that the effect of choosing the values of AgeFactor() based on assumed values of β1BY and β1DY cancels out the effects of the same choices of β1BY and β1DY on the overall expression, giving an accurate prediction value regardless of the choice of BetaFraction. This result is consistent with the lack of identification found via analysis of the set of simultaneous equations above; there is an infinite set of solutions, and no one solution can be identified as being correct.

Age Factor: Significance and Estimation

Establishing an accurate set of values for the age factor is essential for producing unbiased, reasonably accurate estimates of the period (diagnostic year) and cohort (birth year) coefficients, as explained in previous sections. In some analysis problems, such as the present case, we may not care about the values of the age factor but we need them anyway in order to estimate the coefficients we do care about. The age factor estimate needs to be accurate in order to obtain suitable results, due to the high sensitivity to errors in the age factor.

The problem is further compounded by the fundamental problems with obtaining an accurate estimate of the age factor. As explained above, any estimate of the age factor based on data inherently depends on assumptions of the coefficients for birth year and/or diagnostic year. An estimate of the age factor based on analysis by birth year implicitly assumes that the diagnostic year coefficient is zero, unless we use an explicit assumption of that coefficient to adjust the age factor estimate. Similarly, an estimate of age factor based on diagnostic year analysis implicitly assumes that the birth year coefficient is zero, unless we explicitly adjust using an assumed value of the birth year coefficient. In general, in APC analyses we do not have the values of the birth year and diagnostic year coefficients and hence we cannot create unbiased estimates of the age factor. If we assume a value for one or the other of these coefficients for use in estimating the age factor, that assumed value has a large impact on the coefficient value that results from estimation using the age factor, which can lead to meaningless results. This problem exists even if the data used for generating the age factor estimate is separate and distinct from the data used for the primary analysis; it does not help to have an independent data source to estimate the age factor, contradicting the advice in O'Brien (2015).

Conclusions

The primary conclusion of this study is that it is not possible to generate unbiased estimates of the separate birth year and diagnostic year coefficients using the conventional regression, restricted regression using APC methods, or the novel analytical methods introduced here. The problem presents a lack of identifiability due to an insufficient number of constraints compared to the number of variables that need to be solved for in order to find the values of the two coefficients of interest. An approach from the APC literature to solve this problem involves estimating the age factor from data other than the primary dataset. However this approach generally produces unreliable results for two reasons. First, any estimate of the age factor involves implicitly assuming that either the birth year or diagnostic year coefficient is equal to zero, or else explicitly adjusting using an assumed value of at least one of the coefficients being sought, and hence the age factor estimate is biased in a way that is directly related to the true unknown values of these coefficients. Second, regression using restrictions based on an estimate of the age factor, or direct algebraic solutions such as presented here, are highly sensitive to minor deviations in the restrictions.

The statistical analysis of incidence data does not provide evidence for or against specific hypotheses of true case increase or non-etiological factors.

We showed a simple way to determine the value of the sum of the birth year and diagnostic year coefficients conditional on the assumptions behind the specified data generating process model. This sum is the coefficient of the time factor log-linear analysis by either birth year or diagnostic year of sums over the ages included in the analysis. Since we can readily find the sum of the two coefficients, we can specify both of the individual coefficient values using a single fraction term,

however we cannot reliably estimate the value of the fraction term, and hence we cannot separate the birth year and diagnostic year effects.

Analysis shows that in the CA-DDS dataset the sum of the coefficients for birth year and diagnostic year is 0.1206, corresponding to an increase of 12.82% per year.

Discussion

The question of the true values of the coefficients for birth year and diagnostic year in autism incident diagnoses is a very important one, and one that remains unanswered. These coefficients represent the full set of time varying causal (etiologic) factors and the full set of time varying non-etiologic factors respectively. The birth year coefficient also predicts the future case load of adults with autism who will need services, in many cases very significant and expensive services, which are usually paid for by either the federal or state government. As of the time of this thesis there have been no published studies showing unbiased estimates of these coefficients. While we were not able to obtain the results we sought using the methods described here, there may be opportunities for other approaches to find the answers.

We developed and explained a new analytical method to estimate the birth year and diagnostic year coefficients given values of the age factor, or to estimate the age factor given values of the birth year and diagnostic year coefficients.

One of many open questions related to autism is whether there is an epidemic. According to the CDC, an epidemic "refers to an increase, often sudden, in the number of cases of a disease above what is normally expected in that population in that area." (CDC, n.d., principles of epidemiology). The expected number of cases is generally the baseline or endemic value. In the

case of autism it is not completely clear what the baseline prevalence or incidence values are, however the baseline should logically be based on data representing a time period from the past. Clearly the measured prevalence and diagnostic incidence have increased dramatically since approximately 1980, implying the presence of an epidemic. If one chooses a definition of epidemic that relies only on the true case prevalence, as opposed to diagnosed case prevalence, and postulates that the entire measured increase might be due entirely to diagnostic and other non-etiological factors, then one might conclude that it is currently unknown whether or not there is an epidemic because the science does not yet have reliable estimates of the individual birth year and diagnostic year coefficients. However, there is currently no evidence that the diagnostic year effect is 100% of the sum of the birth year and diagnostic year effects, and there is no evidence that the set of all non-etiologic factors explains all of the measured increase.

Strengths

This paper explains logically why the separate effects of birth year and diagnostic year correspond directly to the sets of etiologic and non-etiologic factors, respectively. It makes the case via a DAG, which is inherently subject to debate, so this particular conclusion is not iron-clad. It explores existing and novel approaches to estimating the birth year and diagnostic year effects, their sum and the age factor. It explains a simple way to estimate accurately the sum of the birth year and diagnostic year effects. It explores various ways to attempt to solve the question of the proportions of this sum that are allocated to the individual birth year and diagnostic year effects, and concludes it is not possible to do so using any of the existing and novel methods explored. It provides a potentially valuable conclusion about the relationship between the age factor and the cohort and period coefficients, which shows that the APC

approach of using an independent age factor estimate is less valuable than is indicated in the literature.

Limitations

The primary limitation of this paper is that it fails to find estimates for the values of the birth year and diagnostic year coefficients, which was the primary specific aim. Therefore it cannot draw any new conclusions as to the relative effects of etiologic and non-etiologic factors on the large measured increase in the cumulative incidence of autism diagnoses. As a result, apparently it will be necessary to use significantly different methods in order to estimate these important coefficients. There are other methods described in the APC literature which we have not yet tried for this problem; it is unknown whether they would provide reliable, unbiased results.

Future work

Other approaches to estimating change in autism prevalence by birth may be possible. Under the null hypothesis the case prevalence has not changed over time, only the non-etiologic factors have changed. That implies that the case prevalence was the same for birth year 1980 as it was for birth year 2005, and it was much greater than the measured prevalence currently indicates such that the vast majority of those with autism born in the early years were never diagnosed. For each individual born in any year decades before today who has or had autism, that same person still has autism today unless he or she recovered from autism or died, and there is the possibility that the individual moved out of any given geographic region of study. It should be possible, if complex, to study the true prevalence of autism in specified populations of adults with known ages, and adjust for rates of death, recovery, immigration and emigration.

It may be possible to analyze the relationship between dynamic factors and events that would be expected to cause deviations from the straight -line log-linear trends of cumulative incidence versus birth year or diagnostic year and observed deviations. There may be testable hypotheses regarding associating such known events and changes in potentially explanatory factors with deviations. It may be practical to quantify the effects of these events based on analysis of the associations.

Another potential direction for future research is to perform an in-depth review of the evidence of the effects of hypothesized etiologic and non-etiologic factors, and evaluate the degree to which models based on them can explain the observed statistics.

References

Autism Speaks (2010). What is Causing the Increase in Autism Prevalence? *Autism Speaks Official Blog*, https://autismspeaksblog.wordpress.com/2010/10/22/got-questions-answers-to-your-questions-from-the-autism-speaks%E2%80%99-science-staff-2/

Becerra, T.A., Wilhelm, M., Olsen, J., Cockburn, M. & Ritz, B. (2013). Ambient air pollution and autism in Los Angeles County, California. *Environmental Health Perspectives, 121*(3), 380-386

Bell, A. & Jones, K. (2013). The impossibility of separating age, period and cohort effects. *Social Science & Medicine, 93,* 163-165

Bell, A. & Jones, K. (2014). Don't birth cohorts matter? A commentary and simulation exercise on Reither, Hauser, and Yang's (2009) age-period-cohort study of obesity. *Social Science & Medicine, 101,* 176-180

Blaxill, M.F., Baskin, D.S. & Spitzer, W.O. (2003). Commentary: Blaxill, Baskin, and Spitzer on Croen et al. (2002), the changing prevalence of autism in California. *Journal of Autism and Developmental Disorders, 33*(2), 223-226

California Department of Developmental Services. (2002). *Autism spectrum disorders: Best practice guidelines for screening, diagnosis and assessment.* http://www.dds.ca.gov/Autism/docs/ASD_Best_Practice2002.pdf

CDC. (n.d.) Autism and Developmental Disabilities Monitoring (ADDM) Network. https://www.cdc.gov/ncbddd/autism/addm.html

CDC. (2016). Prevalence and characteristics of autism spectrum disorder among children aged 8 years - Autism and developmental disabilities monitoring network, 11 sites, United States, 2012. *Surveillance Summaries, 65*(3), 1–23

CDC. (n.d.) Principles of epidemiology in public health practice, third Edition. An introduction to applied epidemiology and biostatistics. Lesson 1: Introduction to Epidemiology. Section 11: Epidemic Disease Occurrence. https://www.cdc.gov/ophss/csels/dsepd/ss1978/lesson1/section11.html

Croen, L.A., Grether, J.K., Hoogstrate, J. & Selvin, S. (2002). The changing prevalence of autism in California. *Journal of Autism and Developmental Disorders*, *32*(3), 207-215

Croen, L.A., Grether, J.K., Hoogstrate, J. & Selvin, S. (2003). A Response to Blaxill, Baskin, and Spitzer on Croen *et al.* (2002), "The changing prevalence of autism in California". *Journal of Autism and Developmental Disorders*, *33*, 227 doi:10.1023/A:1022964132203

Croen, L.A., Grether, J.K., Yoshida, C.K., Odouli, R. & Hendrick, V. (2011). Antidepressant use during pregnancy and childhood autism spectrum disorders. *Archives of General Psychiatry*, *68*(11), 1104-1112

Croen, L.A. (2017). Private communication

Elsabbagh, M., Divan, G., Koh, Y-K., Young, S.K., Kauchali, S., Marcin, C. … Fombonne, E. (2012). Global prevalence of autism and other pervasive developmental disorders. *Autism Research*, *5*, 160–179

Fombonne, E. (2009). Epidemiology of pervasive developmental disorders. *Pediatric Research, 65(*6), 591-598

Glenn, N.D. (1976). Cohort analysts' futile quest: Statistical attempts to separate age, period and cohort effects. *American Sociological Review, 41*(5), 900-904

Goldani, A.A., Downs, S.R., Widjaja, F., Lawton, B. & Hendren, R. (2014). Biomarkers in Autism. *Frontiers in Psychiatr*y, 5(100), 1-13

Hallmayer, J., Cleveland, S., Torres, A., Phillips, J., Cohen, B., Torigoe, T., … Risch, N. (2011). Genetic heritability and shared environmental factors among twin pairs with autism. *Archives of General Psychiatry, 68*(11), 1095-1102

Hansen, S.F., Schendel, D.E. & Parner, E.T. (2014). Explaining the increase in the prevalence of autism spectrum disorders - the proportion attributable to changes in reporting practices. *JAMA Pediatrics,* doi:10.1001/jamapediatrics.2014.1893, E1-E7

Hertz-Picciotto, I. & Delwiche, L. (2009). The Rise in Autism and the Role of Age at Diagnosis. *Epidemiology, 20*(1), 84–90

Holford, T.R. (1983). The estimation of age, period and cohort effects for vital rates. *Biometrics, 39*, 311-324

Kamowski-Shakibai, M.T., Kollia, B. & Magaldi, N. (2017) Autism spectrum disorders and communication disorders: Influences of advanced parental age and use of assisted reproductive technology. *Advances in Neurodevelopmental Disorders, 1*, 21–30

Keyes, K.M., Utz, R.L., Robinson, W. & Li, G. (2010). What is a cohort effect? Comparison of three statistical methods for modeling cohort effects in obesity prevalence in the United States, 1971–2006. *Social Science & Medicine, 70*, 1100–1108

Keyes, K.M., Susser, E., Cheslack-Postava, K., Fountain, C., Liu, K. & Bearman, P.S. (2012) Cohort effects explain the increase in autism diagnosis among children born from 1992 to 2003 in California. *International Journal of Epidemiology, 41*(2), 495-503

Keyes, K.M. & Bearman, P.S. (2013) Reply to Spiers: Cohort effects explain the increase in autism diagnosis among children born from 1992 to 2003 in California. *International Journal of Epidemiology,* 13(42), 1521

King, M. & Bearman, P. (2009). Diagnostic change and the increased prevalence of autism. *International Journal of Epidemiology, 38,* 1224-1234

Kupper, L.L., Janis, J.M., Karmous, A. & Greenberg, B.G. (1985). Statistical age-period-cohort analysis: a review and critique. *Journal of Chronic Diseases, 38*(10), 811-830

Lyall, K., Croen, L., Daniels, J., Fallin, M.D., Ladd-Acosta, C., Lee, B.K., … Newschaffer, C. (2017). The changing epidemiology of autism spectrum disorders. *Annual Review of Public Health, 38*, 81-102

Mandell, D.S. & Palmer, R.F. (2005). Differences among states in the identification of autistic spectrum disorders. *Archives of Pediatric and Adolescent Medicine, 159*, 266–269

Mason, K.A., Mason, W.M., Winsborough, H.H. & Poole, W.K. (1973). Some methodological issues in cohort analysis of archival data. *American Sociological Review, 38*(2), 242-258

Mazumdar, S., Winter, A., Liu K.-Y. & Bearman, P. (2013) Spatial clusters of autism births and diagnoses point to contextual drivers of increased prevalence. *Social Science and Medicine, 95,* 87-96

Ng, M., de Montigny, J.G., Ofner, M. & Do, M.T. (2017). Environmental factors associated with autism spectrum disorder: a scoping review for the years 2003–2013. *Health Promotion and Chronic Disease Prevention in Canada*, *37*(1), 1-23

O'Brien, R.M. (2000). Age period cohort characteristic models. *Social Science Research, 29*, 123–139

O'Brien, R.M. (2015). *Age-period-cohort models: Approaches and analyses with aggregate data.* Boca Raton FL: CRC Press

Polyak, A., Kubina, R.M. & Girirajan, S. (2015) Comorbidity of intellectual disability confounds ascertainment of autism: Implications for genetic diagnosis. *American Journal of Medical Genetics Part B, 168B*, 600–608

Robertson, C. & Boyle, P. (1986). Age period and cohort models: The use of individual records. *Statistics in Medicine, 5*, 527-538

Robertson, C., Gandini, S. & Boyle, P. (1999) Age-period-cohort models: A comparative study of available methodologies. *Journal of Clinical Epidemiology, 52*(6), 569–583

Rodgers ,W.L. (1982). Estimable functions of age, period, and cohort effects. *American Sociological Review, 47*(6), 774-787

Sanders, S.J., Murtha, M.T., Gupta, A.R., Murdoch, J.D., Raubeson, M.J., Willsey, A.J., … State, M.W. (2012). De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature, 485*, 237-241

Shattuck, P.T. (2006). The contribution of diagnostic substitution to the growing administrative prevalence of autism in us special education. *Pediatrics, 117*(4), 1028-1037

Shelton, J.F., Geraghty, E.M., Tancredi, D.J., Delwiche, L.D., Schmidt, R.J., Ritz, B., … Hertz-Picciotto, I. (2014). Neurodevelopmental disorders and prenatal residential proximity to agricultural pesticides: The CHARGE study. *Environmental Health Perspectives, 122*(10)*,* 1103–1109, http://dx.doi.org/10.1289/ehp.1307044

Spiers, N. (2013). Letter: Cohort effects explain the increase in autism diagnosis among children born from 1992 to 2003 in California. *International Journal of Epidemiology, 42*, 1520–1521

Volk, H.E., Lurmann, F., Penfold, B., Hertz-Picciotto, I., McConnell, R. (2013). Traffic-related air pollution, particulate matter, and autism. *JAMA Psychiatry, 70*(1), 71-77

Winship, C. & Harding, D. (2008). A mechanism-based approach to the identification of age–period–cohort models. *Sociological Methods and Research, (36)*3, 362-401

Autism Prevalence Trends by Birth Year and Diagnostic Year:
Indicators of Etiologic and Non-Etiologic Factors – an Age Period Cohort Problem

Alexander G. MacInnis
June 2017

Approved for Submission to the Division of Epidemiology
Department of Health Research and Policy
Stanford University School of Medicine

Epidemiology reader: _____ Date:_____
Lorene Nelson
Associate Professor, Division of Epidemiology
Department of Health Research and Policy

Co-Reader: _____ Date:_____
Kristin Sainani
Associate Professor (Teaching)
Department of Health Research and Policy