# Time to Event Estimation of Birth Year Prevalence Trends

Alexander MacInnis
Independent researcher
Mountain View, California
Email: a.macinnis@alumni.stanford.edu

## Abstract

Accurate estimates of incidence over time are essential for investigating etiology. For diseases and disorders of early childhood, including autism and intellectual disability, birth year cohort prevalence serves as incidence. There is substantial uncertainty regarding the birth prevalence trend of autism, which is a disorder that results in a lifetime of reduced quality of life and is apparently increasingly prevalent world-wide. This paper presents a novel method to accurately estimate the true birth prevalence of autism and other childhood disorders.

Estimates based on summarizing available data may have substantial errors if the proportion of cases counted in the data changed over time because of changes in diagnostic factors. Controlling directly for diagnostic factors may lead to erroneous estimates. We have not found any existing method for robust estimation of birth prevalence while accounting for changing diagnostic factors. A novel method, Time-to-Event Prevalence Estimation (TTEPE), addresses this need. TTEPE extends established time-to-event analysis techniques. Input data are rates of initial diagnosis by diagnostic year and age for each birth cohort. Diagnostic factors form diagnostic pressure, which is the probability of diagnosing cases. Diagnostic pressure is a function of diagnostic year. Changes in diagnostic criteria may also change the effective prevalence at known times. Non-linear regression estimates parameters using a discrete survival model of the rate of initial diagnoses as a function of birth year, diagnostic year, and age. Diagnosable symptoms may develop with age, affecting eligibility for diagnosis. The paper states the assumptions underlying the analysis. A simulation study validates the TTEPE method, producing accurate estimates of parameters describing trends in birth prevalence and diagnostic pressure. The paper explores optional additional analyses and deviations from assumptions. TTEPE is a robust method for estimating trends in true case birth prevalence controlled for diagnostic factors and changes in diagnostic criteria under certain specified assumptions.

## Keywords

Birth year prevalence, autism, time-to-event, survival analysis, diagnostic factors, non-linear regression

# Introduction

In epidemiology, incidence - the rate of new cases, or newly identified cases - is the preferred measure of rates of occurrence for understanding the causes of a disease or, more generally, disorder[1]. Incidence is in contrast to prevalence, which is the proportion of a population with the disorder at a defined time. Time trend studies of incidence are useful for exploring associations between time-varying potentially causal exposures and rates of a disorder[2]. For some disorders including congenital diseases and developmental disorders such as autism and intellectual disability, epidemiologists use birth year cohort prevalence instead of incidence[2,3] because case status is determined either by birth or shortly afterwards while diagnosis or case identification may occur later. Serious developmental disorders such as autism result in a reduction in quality of life across the lifespan[4]. Accurate estimates of time trends in true birth prevalence can facilitate investigation of both hypothesized risk factors and non-causal explanatory factors.

Studies of time trends in incidence or birth year prevalence are subject to biases resulting from changes in diagnostic factors and diagnostic criteria[2]. We define diagnostic factors as those that influence the probability of diagnosing or identifying cases. Examples include awareness, outreach efforts, screening, diagnostic practice, diagnostic criteria, social factors, policies, and financial incentives for diagnosis. Changes in diagnostic criteria can also have the additional effect of changing the effective birth prevalence by including or excluding as cases some portion of the population compared to prior criteria. Such changes in prevalence occur when the new criteria take effect.

For example, consider this question. If reported birth prevalence increased over time, was this caused by changes in actual birth prevalence, the probability of diagnosing cases, diagnostic criteria affecting prevalence, or some combination of these factors? It is challenging to disentangle these effects and we have not found an existing method capable of doing so correctly. If estimated or assumed trends of birth prevalence are incorrect, assumptions about causal factors may be incorrect, which could result in misdirected research priorities. Uncertainty about birth prevalence trends can persist for many years in the absence of a robust estimation methodology[5].

This paper presents a novel statistical method called time-to-event prevalence estimation (TTEPE). It uses time-to-event survival analysis to estimate the trend in true birth year prevalence, correctly adjusted for the set of diagnostic factors and criteria. An essential element of the method involves modeling the age distribution of initial diagnoses.

TTEPE is particularly applicable to disorders where case status is established by birth or by a known age, and diagnosable symptoms are present by some consistent age. It

is also useful where cases develop diagnosable symptoms gradually over a range of ages.

Data sources suitable for TTEPE analysis provide rates of initial diagnoses by age or equivalently, diagnostic year, for each birth cohort. The rate is the count of initial diagnoses divided by the population of the cohort at the respective age. TTEPE relies on these principles: for each birth cohort, the number of cases at risk of initial diagnosis decreases as cases are diagnosed, and diagnostic year is the time when diagnostic factors affect the probability of diagnoses of cases exhibiting diagnosable symptoms. TTEPE introduces the principle of eligibility for diagnosis, meaning that cases have diagnosable symptoms, where eligibility is a function of age.

## Development of a New Method

Comparison of prevalence estimates across multiple studies generally is not suitable for informing tends in birth prevalence nor incidence[2]. Different prevalence estimates may use different mixes of birth years and ages, as well as numerous other possible differences between prevalence studies[6]. Many combinations of trends in birth prevalence and diagnostic factors could potentially explain observed prevalence trends. One reason is that age ($A$), diagnostic year ($DY$), and birth year ($BY$) are exactly collinear, $DY = BY + A$, subject to rounding, which can lead to unidentified estimates. Another reason is that the age distribution of diagnoses can be different for different solutions of $BY$ and $DY$, as shown below. It can be challenging to estimate the age distribution correctly given the collinearity problem.

Another option is to analyze the cumulative incidence, to a consistent age, of diagnoses in each birth cohort. That approach comes closer to estimating the trend in true birth prevalence, but results are still ambiguous. Here too, many combinations of trends in birth prevalence and diagnostic factors can produce similar trends in cumulative incidence.

### Ambiguity in Estimation

How should one interpret a dataset that produces any one of the cumulative incidence curves illustrated in Figure 1? The figure represents synthetic data; some real-world data may be similar. Observed data might produce a curve resembling any one of the curves in the figure. An exponential curve with a coefficient of 0.1 fits all three plotted lines reasonably well. Does this represent a true increase in birth prevalence with a coefficient of 0.1? Does it result from an exponential increase in the effects of diagnostic factors, with no increase in birth prevalence? Perhaps a combination of both? The three similar cumulative incidence curves represent examples of quite different possible explanations. How to distinguish which one is correct?

One might consider controlling for diagnostic factors over time, e.g., via regression, but that is not sufficient to distinguish between alternative explanations. Diagnostic factors

are a function of time, and birth year prevalence is also a function of time. Controlling directly for diagnostic factors constitutes inappropriate overadjustment. Diagnostic factors form an intermediate variable on a causal path from time to outcome. Controlling for such variables typically produces a "null-biased estimate of the total causal effect," as Schisterman explains[7]. Similarly, Rothman[2] states that controlling for intermediate variables "bias[es] effect estimates, typically, though not necessarily, toward the null (i.e., toward finding no effect)." Schisterman[7] recommends "clearly stating a causal question to be addressed, depicting the possible data generating mechanisms using causal diagrams, and measuring indicated confounders." The TTEPE method direclty addresses these issues.

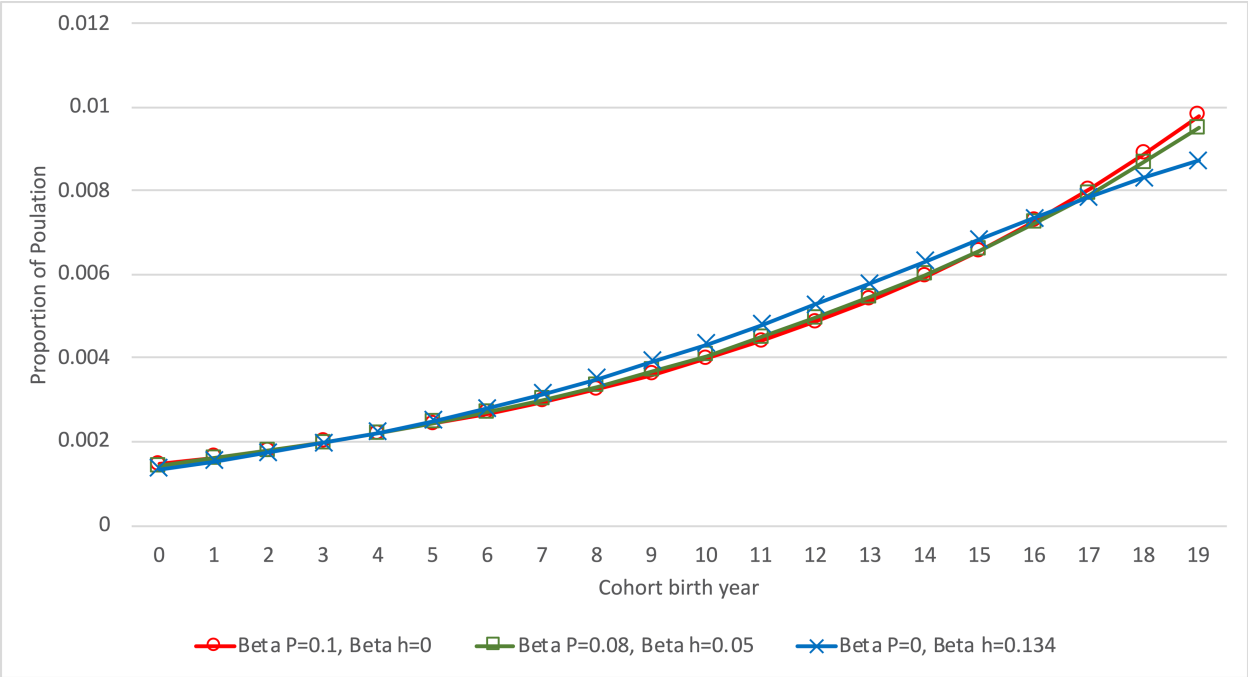


Figure 1 Example of cumulative incidence under three models

Figure 1 illustrates a hypothetical example of cumulative incidence to age ten over 20 consecutive cohorts. The legend lists the parameter value pairs for the three cases. $\beta_P$ is the exponential coefficient of birth year prevalence, and $\beta_h$ is the exponential coefficient of the effect of diagnostic factors by diagnostic year. The data generation process producing these data uses a survival process as detailed below. The variable $h$ represents diagnostic pressure, which is equivalent to hazard, as explained in the section Time to Event Analysis Model. While the three cumulative incidence curves appear similar, the age distributions of diagnoses are evidently different, as Figure 2 shows. The remainder of this paper explains how modeling the age distribution enables robust and unambiguous estimation of the coefficients for birth prevalence and diagnostic factors.

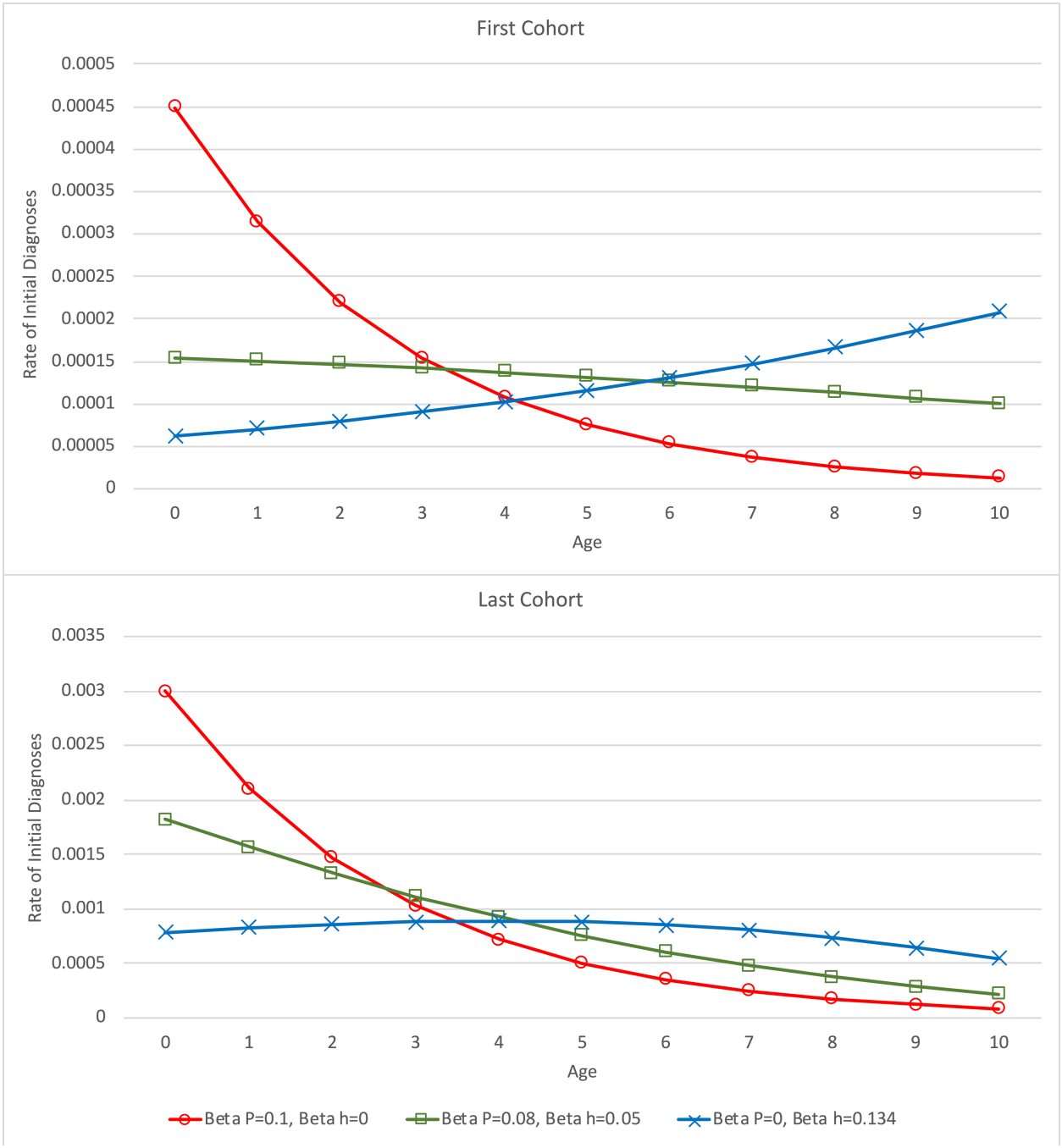


Figure 2 Distribution of diagnoses in the first and last cohorts under three models

## Significance of Birth Year and Diagnostic Year

Diagnostic factors only affect the diagnosis of cases when those cases exhibit diagnosable symptoms, referred to as being eligible for diagnosis. We define diagnostic pressure as the probability of diagnosing eligible undiagnosed cases. Diagnostic

pressure is an effect of the combination of all diagnostic factors. The introduction gives examples.

Diagnostic pressure is equivalent to the hazard $h$ in time-to-event or survival analysis. For each case, the information resulting from diagnostic pressure consists of the time of initial diagnosis, that is, the diagnostic year. Diagnostic pressure has no observable effect before diagnosis of each case, and none after the initial diagnosis since TTEPE considers only initial diagnoses. Hence, the effect of diagnostic pressure on the input data is a function of diagnostic year.

The directed acyclic graph (DAG) in Figure 3 illustrates the causal paths from birth year, diagnostic year, and age to diagnosis. Birth year drives etiologic (causal) factors, which produce the disorder and its symptoms. Diagnostic criteria determine whether each individual's symptoms qualifies them as a case. The criteria may change at specific diagnostic years. Symptoms may vary with age. Diagnostic year drives diagnostic factors, which form diagnostic pressure. Diagnosable symptoms and diagnostic pressure together produce each initial diagnosis.

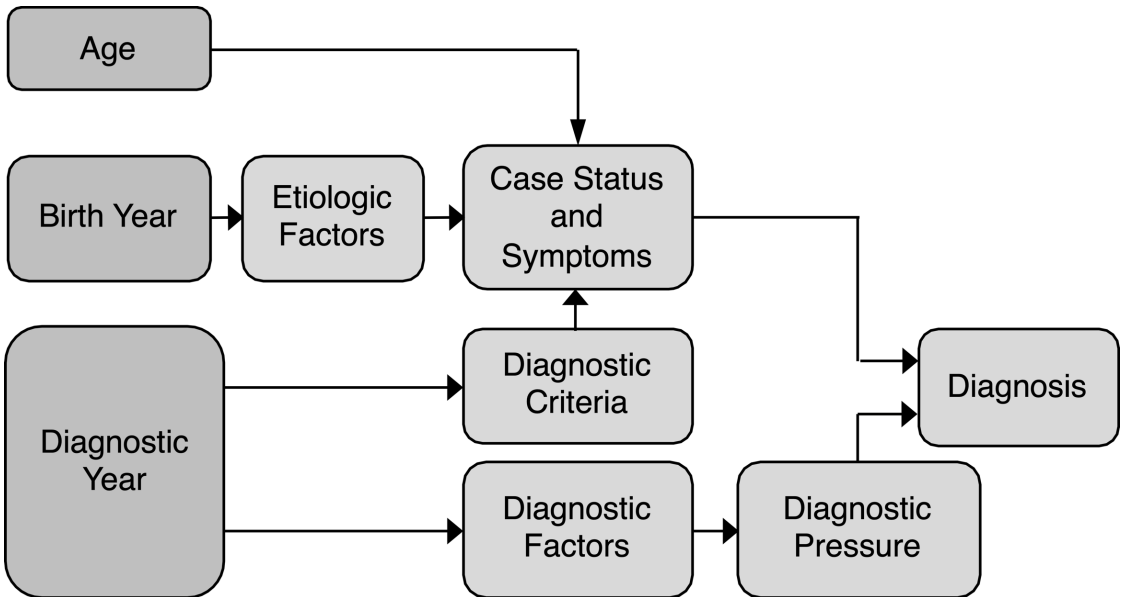


Figure 3 Directed Acyclic Graph Representing Year of Birth, Diagnostic Year and Age

Changes in diagnostic criteria affect the threshold of symptoms that qualify case status. Criteria changes may change the proportion of the cohort classified as cases, i.e., the effective prevalence.

# Time to Event Prevalence Estimation

First, consider the case where the diagnostic criteria do not change the effective prevalence over the interval of interest. The section Changes in Criteria Affecting Prevalence examines the alternative case.

The principle of finding the simplest model that fits the data, sometimes called Occam's Razor, could lead to the conclusion that birth year effects best explain the observed birth year prevalence trends. However, the most parsimonious model is not always the best one[8]. Specifically, diagnostic pressure as a function of diagnostic year might be an essential component for explaining observed diagnosis data.

TTEPE is an extension of established time-to-event methods. TTEPE simultaneously estimates the birth prevalence and diagnostic pressure functions by fitting a model to the data at each age for each cohort. It models the age distribution of rates of initial diagnoses via a survival process as a function of birth year, diagnostic year (birth year plus age), and eligibility.

TTEPE introduces the concept of eligibility. A case is eligible for diagnosis if the individual has diagnosable symptoms, and not eligible if the individual does has not yet developed diagnosable symptoms. For each birth cohort, undiagnosed eligible individuals form the risk set of cases at risk of initial diagnosis. The size of the risk set is $R$. At each age there is a probability of diagnosis of each case in the risk set. This probability is the diagnostic pressure $h$. At each age, newly diagnosed cases are removed from the risk set, and newly eligible cases are added to the risk set. For any given value of $h$, as $R$ decreases or increases, the rate of initial diagnoses $D$ changes accordingly. This process inherently generates the age distribution. The survival function $S$ refers to cases that "survive" diagnosis at each age. If all cases were eligible from birth, $S$ would equal $R$. More generally, however, some cases may initially be ineligible and become eligible as they age, so $R \leq S$. The prevalence $P$ is the proportion of the population that are or become cases. The eligibility factor $E$ is the eligible proportion of $P$; $0 \leq E \leq 1$.

In contrast, in typical survival or time-to-event analysis, including Cox proportional hazards analysis[9], the initial size of the risk set is assumed to have a known value, for instance, an entire population or an entire sample. If the risk set $R$ consisted of the entire population without subtracting diagnosed cases, the estimate $\widehat{h_t} = D_t/R_t$ would be equivalent to the population-based rate of diagnoses $D$ at time $t$. If the condition is rare, it makes little difference whether the risk set is the entire population or the undiagnosed portion.

Population-based rates of initial diagnoses $D$ are observable while the other variables are not. Different values of diagnostic pressure $h$ produce different values of $D$ for a given value of prevalence $P$, as shown in the Illustrative Example section.

## Time to Event Analysis Model

The primary aim is to estimate the temporal trend of birth prevalence $P$ over a range of cohorts, correctly adjusted for diagnostic pressure $h$. Both $P$ and $h$ can vary with time. $P$ is a function of birth year and $h$ is a function of diagnostic year. Estimation of $P$ adjusted for $h$ requires estimating the time-based parameters of both $P$ and $h$ in the time-to-event model and specifying or estimating the eligibility function $E$.

Let $D_{BY,A}$ be the population-based rate of incident diagnoses, where $BY$ is birth year and $A$ is age. The model generates predicted values $\widehat{D_{BY,A}}$. Modeling of proportions rather than counts accommodates changes in the size of the population of each cohort over time, e.g., due to in- and out-migration and deaths. Count values are useful for calculating p-values from chi-square goodness-of-fit measurements. Alternatively, the analysis could model counts directly.

Let $h_{DY}$ be the diagnostic pressure at diagnostic year $DY$. $DY = BY + A$, subject to rounding, so $h_{DY}$ is equivalent to $h_{BY,A}$. Let $P_{BY}$ be the case prevalence of birth year cohort $BY$. Let $R_{BY,A}$ be the discrete risk set function of the population proportion of eligible cases at risk of initial diagnosis at age $A$ for birth year $BY$. TTEPE uses $R$ rather than a discrete survival function $S$ in order to accommodate eligibility changing with age. Let $E_A$ be the discrete eligibility function, the proportion of cases that are eligible at age $A$, bounded by $0 \leq E \leq 1$. At each age $A \geq 1$, $P \times (E_A - E_{A-1})$ is the incremental portion of prevalent cases added to $R$ due to changes in eligibility. For simplicity, assume $E_A$ increases monotonically, i.e., non-decreasing, meaning that cases do not lose eligibility before diagnosis.

Kalbfleisch[10] gives background on general time-to-event theory and equations.

Consider three scenarios, differing by the characteristics of $E_A$. Here we write $h_{DY}$ as $h_{BY,A}$ to clarify the effect of $A$ in $DY = BY + A$.

<u>Scenario: constant $E_A = 1$.</u> All cases are eligible from birth, so $E_A = 1$ for all values of $A$. This scenario is equivalent to standard time-to-event models that do not consider eligibility. For $A \geq 1$, $E_A - E_{A-1} = 0$. For the first year of age, $A = 0$, $R_{BY,0} = P_{BY}E_0 = P_{BY}$ and
$D_{BY,0} = R_{BY,0}h_{BY,0} = P_{BY}h_{BY,0}$.
For $A = 1$, $R_{BY,1} = P_{BY} - D_{BY,0} = P_{BY} - P_{BY}h_{BY,0} = P_{BY}(1 - h_{BY,0})$ and
$D_{BY,1} = R_{BY,1}h_{BY,1} = P_{BY}(1 - h_{BY,0})h_{BY,1}$.
For $A = 2$, $R_{BY,2} = R_{BY,1} - D_{BY,1} = P_{BY}\left(1 - h_{BY,0}\right) - P_{BY}\left(1 - h_{BY,0}\right)h_{BY,1} =$
$P_{BY}\left(1 - h_{BY,0}\right)(1 - h_{BY,1})$ and $D_{BY,2} = R_{BY,2}h_{BY,2} = P_{BY}(1 - h_{BY,0})(1 - H_{BY,1})h_{BY,2}$.
Similarly, for $A = 3$,
$R_{BY,3} = P_{BY}(1 - h_{BY,0})(1 - h_{BY,1})(1 - h_{BY,2})$ and
$D_{BY,3} = P_{BY}(1 - h_{BY,0})(1 - h_{BY,1})(1 - h_{BY,2})h_{BY,3}$.

Generally, for $A \geq 1$,

$$R_{BY,A} = P_{BY} \prod_{a=0}^{A-1} (1 - h_{BY,a})$$

and

$$D_{BY,A} = P_{BY} \prod_{a=0}^{A-1} \left(1 - h_{BY,a}\right) h_{BY,A} \tag{1}$$

In all three scenarios in this paper, the survival function is:

$$S_{BY,A} = P_{BY} - \sum_{a=0}^{A-1} D_{BY,a} \tag{2}$$

The summation term is the cumulative incidence of initial diagnoses through age $A - 1$.

<u>Scenario: Increasing $E_A$.</u> $E_0 < 1$ and $E_A$ increases monotonically with $A$. For $A = 0$, $R_{BY,0} = E_0 P_{BY}$ and $D_{BY,0} = E_0 P_{BY} h_{BY,0}$. For each $A \geq 1$, $R_{BY,A} = R_{BY,A-1} - D_{BY,A-1} + (E_A - E_{A-1}) P_{BY}$. The incremental increase of $E_A$ causes an incremental increase in $R_{BY,A}$. Then,

$$D_{BY,A} = R_{BY,A} h_{BY,A} = \left(R_{BY,A-1} - D_{BY,A-1}\right) h_{BY,A} + (E_A - E_{A-1}) P_{BY} h_{BY,A} \tag{3}$$

Equation (3) can be useful as a procedural definition. We can write equivalent expressions for $R_{BY,A}$ and $D_{BY,A}$ as sums of expressions similar to equation (1), where each summed expression describes the portion of $P_{BY}$ that becomes eligible at each age according to $E_A$. For $A \geq 1$,

$$R_{BY,A} = \sum_{a=0}^{A-1} (E_a - E_{a-1}) P_{BY} \prod_{b=a}^{A-1} (1 - h_{BY,b})$$

$$D_{BY,A} = \sum_{a=0}^{A-1} (E_a - E_{a-1}) P_{BY} \prod_{b=a}^{A-1} \left(1 - h_{BY,b}\right) h_{BY,A} \tag{4}$$

where $E_{-1}$ is defined to be 0. $E_A$ can be defined parametrically or non-parametrically.

<u>Scenario: Plateau $E_A$.</u> $E_A$ increases from $E_0 < 1$ and plateaus at $E_A = 1$ for $A \geq AE$, where $AE$ is the age of complete eligibility, $AE < M$, and $M$ is the maximum age included in the analysis. Equation (3) applies, noting that for $A > AE$, $(E_A - E_{A-1}) = 0$. Equivalently, combine equation (2) with the fact that $E_{AE} = 1$ to obtain $R_{AE} = S_{AE} = P_{BY} - \sum_{a=0}^{AE-1} D_{BY,a}$, so

$$D_{BY,AE} = R_{BY,AE}h_{BY,AE} = S_{BY,AE}h_{BY,AE} = (P_{BY} - \sum_{a=0}^{AE-1} D_{BY,a})\, h_{BY,AE} \tag{5}$$

and for $A > AE$,

$$R_{BY,A} = S_{BY,A} = (P_{BY} - \sum_{a=0}^{AE-1} D_{BY,a}) \prod_{b=AE}^{A-1} (1 - h_{BY,b})$$

$$D_{BY,A} = (P_{BY} - \sum_{a=0}^{AE-1} D_{BY,a}) \prod_{b=AE}^{A-1} \left(1 - h_{BY,b}\right) h_{BY,A} \tag{6}$$

The scenario of increasing $E_A$ is a general formulation and may not be needed in practice. The plateau $E_A$ scenario may be appropriate when external information, such as the definition of the disorder's case status, indicates the value of $AE$, or when investigators specify $AE$ based on estimates of $E_A$ found using Equation (4). Equations (5) and (6) do not model $E_A$ nor $D_{BY,A}$ for $A < AE$. Rather, they use the empirical values of $D_{BY,A}$ for $A < AE$.

### Prevalence, Cumulative Incidence and Censoring

The case prevalence in each cohort is the cumulative incidence of initial diagnoses through the last age of follow-up plus the censored portion, assuming that any difference in competing risks between cases and non-cases in the age range analyzed is small enough to be ignored. If the rate of deaths of cases before initial diagnosis exceeds that of the entire population of the cohort at the same ages, that excess would constitute a competing risk and would reduce the estimated prevalence accordingly.

In all three scenarios of $E_A$, we can express $P$ as a function of $S$ and the cumulative incidence $CI = \sum_{a=0}^{A-1} D_{BY,a}$ for $A > 0$, re-arranging equation (2) as $P = S_A + CI_{A-1}$. Assuming that eligibility at the last age of follow-up $E_M = 1$, $S_M = R_M$. Then, $P = R_M + CI_{M-1}$ and $D_M = R_M h_M$. The censored proportion is $S_{M+1} = S_M - D_M$, which is equivalent to $S_{M+1} = R_M - R_M h_M = R_M(1 - h_M)$. After estimating the model parameters, the estimated censored proportion is $\widehat{S_{M+1}} = \widehat{R_M}(1 - \widehat{h_M})$.

### Illustrative Example

Figure 4 illustrates an example according to the plateau $E_A$ scenario showing the relationships between prevalence, diagnosis rates, the survival function, and cumulative incidence $CI$ with two different values of diagnostic pressure $h$ = 0.1 and $h$ = 0.25 and prevalence $P$ = 0.01. In this example, $E = 1$ for $A \geq AE = 3$ and $h$ takes on one of two constant values. The value of $h$ determines the shapes of these functions vs. age. This example shows constant values of $h$ purely for clarity, not as an assumption nor a limitation of TTEPE.

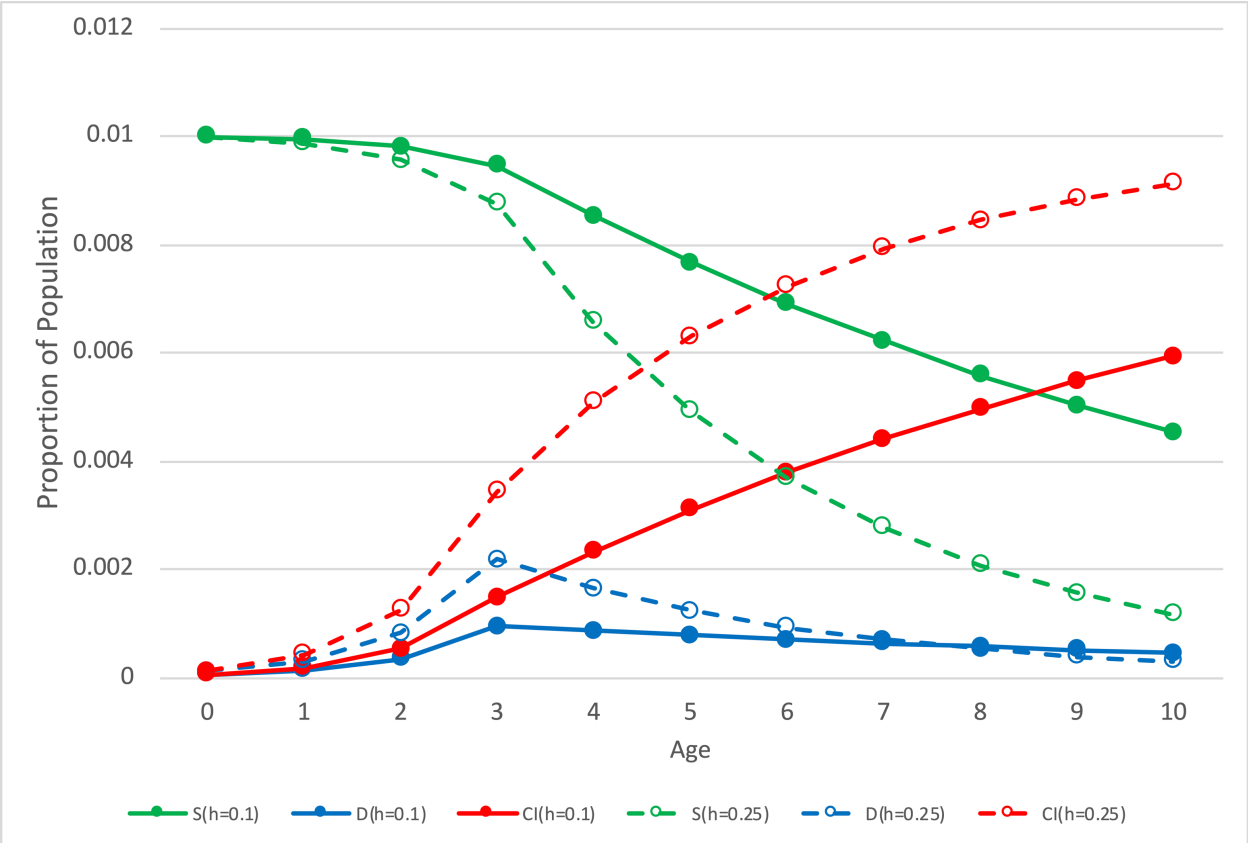


Figure 4 Example of Survival Process for Two Values of Diagnostic Pressure $h$.
S: Survival, P: Prevalence, D: Diagnoses

As cases are diagnosed, $S$ decreases and $CI$ increases. $R$ is not shown; $R = S$ for $A \geq AE = 3$. Only $D$ is observable.

## Estimating Parameters

Several baseline assumptions enable TTEPE analysis. Some assumptions may be relaxed, as discussed below.

1. The eligibility function $E_A$ under consistent diagnostic criteria is consistent across cohorts.
2. The diagnostic pressure applies equally to all eligible undiagnosed cases at any given diagnostic year.
3. The case prevalence under consistent diagnostic criteria within each cohort is constant over the range of ages included in the analysis.
4. Case status is binary according to the applicable diagnostic criteria.
5. The discrete time interval (e.g., one year) is small enough that the error introduced by treating the variable values as constant within each interval is negligible.
6. No false positives.

7. Data represent truly initial diagnoses.
8. Any difference in competing risks between cases and non-cases in the age range analyzed is small enough to be ignored.

The assumption of a consistent eligibility function means that cases develop diagnosable symptoms as a function of age and that function is the same for all cohorts under consistent diagnostic criteria. The section Changes in Criteria Affecting Prevalence discusses a separate effect that might make the eligibility function appear to be inconsistent.

TTEPE performs a non-linear regression that estimates the parameters of a model of $D_{BY,A}$ using general optimization software. The model is based on Equations (1) through (6) selected based on the eligibility scenario. The model produces estimates $\widehat{D_{BY,A}}$ from the parameters and independent variables, and the software finds the parameter values that minimize a cost function cost($D, \widehat{D}$). One suitable implementation of optimization software in the Python language is the curve_fit() function in the SciPy package. Its cost function is $(D - \widehat{D})^2$, so it minimizes the sum of squared errors. Python software to perform this regression and the simulations described below is available at OSF[11].

Investigators should choose which model equation to use based on knowledge or estimates of the eligibility function $E_A$. The constant $E_A$ scenario and Equation (1) assume that all cases are eligible from birth, which may not be valid for some disorders. The validity of the assumption that all cases are eligible by a known age $AE$, i.e., the plateau $E_A$ scenario and equation (6), may be supported by either external evidence, e.g., the definition of the disorder, or estimation of $E_A$. The least restrictive approach of the increasing $E_A$ scenario uses equation (4) to estimate $E_A$. Non-parametric estimates $\widehat{E_A}$ can inform a choice of a parametric form of $E_A$. The value of $E_A$ at the maximum age studied $M$ should be set to 1 to ensure the estimates are identifiable. If $E_A = 1$ for all $A \geq AE$, that fact and the value of $AE$ should be apparent from estimates $\widehat{E_A}$, and the plateau $E_A$ scenario applies.

Investigators should choose forms of $P_{BY}$ and $h_{BY,A}$ appropriate to the dataset. Linear, first-order exponential, second-order exponential or non-parametric models may be appropriate. Graphical and numeric model fit combined with degrees of freedom can guide the optimum choice of a well-fitting parsimonious model.

TTEPE preferably estimates $P$ and $h$ simultaneously over a series of cohorts, utilizing data points from all cohorts, thereby enabling robust estimation and flexible model specification. Alternatively, under some conditions, it may be possible to estimate $h$ in a single cohort and estimate $P$ of that cohort based on $\hat{h}$.

If the population proportion of cases represented in the data is unknown for all cohorts, estimates of the absolute prevalence, or the intercept, may be underestimated by an unknown scale factor. If that proportion is known for at least one cohort, we can use it to

calibrate the intercept. Proportional changes in prevalence between cohorts are unaffected by underestimation of the intercept. If the population proportion of cases included in the sample changes over time, that change reflects changing diagnostic factors. The estimated parameters of $h$ automatically represent such changes.

## Changes in Criteria Affecting Prevalence

Changes in diagnostic criteria could potentially affect rates of initial diagnoses by changing the effective prevalence. This mechanism is distinct from diagnostic pressure. Changes in criteria may change the effective prevalence within a cohort, without affecting symptoms or etiology, by including or excluding as cases some portion of the cohort population compared to prior criteria. A criteria change may change the effective prevalence of any cohort where the birth year is greater than or equal to the year the change took effect. For birth years before the year of criteria change, a change in criteria that changes the effective prevalence causes an increase or decrease in the size of the risk set $R$ starting at the diagnostic year the change took effect. Generally, diagnostic criteria should be given in published documents, such that changes in criteria correspond to effective dates of new or revised specifications.

Let $\{CF_{cy}\}$ be the set of criteria factors that induce a multiplicative effect on effective prevalence due to criteria changes that occurred at criteria years $\{cy\}$ after the first $DY$ included in the study. $P_{BY}$ is the prevalence of cohort $BY$ before the effect of any of $\{CF_{cy}\}$. For each cohort $BY$, the effective prevalence $EP_{BY,A}$ at age A is

$$EP_{BY,A} = P_{BY} \prod_{cy \leq (BY+A)} CF_{cy} \tag{7}$$

where $CF_0$, the value in effect before the first $DY$ in the study, equals 1. The combination of $P_{BY}$ and the effects of all $\{CF_{cy} | cy \leq BY + A\}$ determines the final effective prevalence of each cohort.

For a given $BY$ and increasing $A$, $BY + A$ crossing any $cy$ causes a step change in the effective prevalence $EP$. Using a general formulation of eligibility $E_A$, per the increasing $E_A$ scenario and Equation (3), and for clarity substituting $BY + A$ for $DY$, we obtain the following. For $A = 0$, $R_{BY,0} = E_0 EP_{BY,0}$ and $D_{BY,0} = E_0 EP_{BY,0} h_{BY,0}$. For $A \geq 1$,

$$R_{BY,A} = R_{BY,A-1} - D_{BY,A-1} + E_A(EP_{BY,A} - EP_{BY,A-1}) + (E_A - E_{A-1})EP_{BY,A}$$

and

$$D_{BY,A} = [R_{BY,A-1} - D_{BY,A-1} + E_A(EP_{BY,A} - EP_{BY,A-1}) + (E_A - E_{A-1})EP_{BY,A}]h_{DY} \tag{8}$$

The term $EP_{BY,A} - EP_{BY,A-1}$ represents the change in the effective prevalence $EP$ when $BY + A$ crosses one of $\{cy\}$. As each $CF_{cy}$ takes effect at $cy = DY = BY + A$, the newly effective $CF_{cy}$ changes $EP_{BY,A}$ and $R$ in all $BY$ cohorts where $cy$ corresponds to an age $A$ in the range of ages studied. These changes in $R$ affect the rates of initial diagnoses $D$. For cohorts born after $cy$, $CF_{cy}$ applies to all ages.

The parameters of $P_{BY}$ quantify the birth year prevalence controlled for diagnostic criteria changes, which are represented by $\{CF_{cy}\}$. In other words, $P_{BY}$ is the cohort prevalence that would have occurred if the initial criteria had applied over all diagnostic years included in the study.

To estimate the parameters, use a software model of Equation (8) with optimization software, as described in the previous section.

## Potential Violations of Assumptions

If a dataset represents a non-homogeneous set of cases with different effective values of diagnostic pressure $h$ applying to different unidentified subgroups at the same $DY$, that violates the assumption that the diagnostic pressure applies equally to all eligible undiagnosed cases at any given $DY$. Cases may have differing degrees of symptom severity, and more severe symptoms may result in earlier diagnosis[12], implying greater diagnostic pressure. Figure 5 illustrates this situation. The figure illustrates constant values of $h$ purely for clarity, not as an assumption nor a limitation. Stratified estimation using subgroup data, if available, can avoid the violation of the assumption of homogeneous diagnostic pressure. If data represent unidentified subgroups with differential diagnostic pressure, the distribution of diagnoses is a sum of distributions with different values of $h$. Such a sum of distributions with different values of $h$ may impair fit with a model that assumes homogeneous $h$. For data where $E = 1$ for $A \geq AE$ (plateau $E_A$), adjusting the assumed value of $AE$ used in estimation, called $AE^*$, may mitigate such errors, as shown in the Simulation Study section.

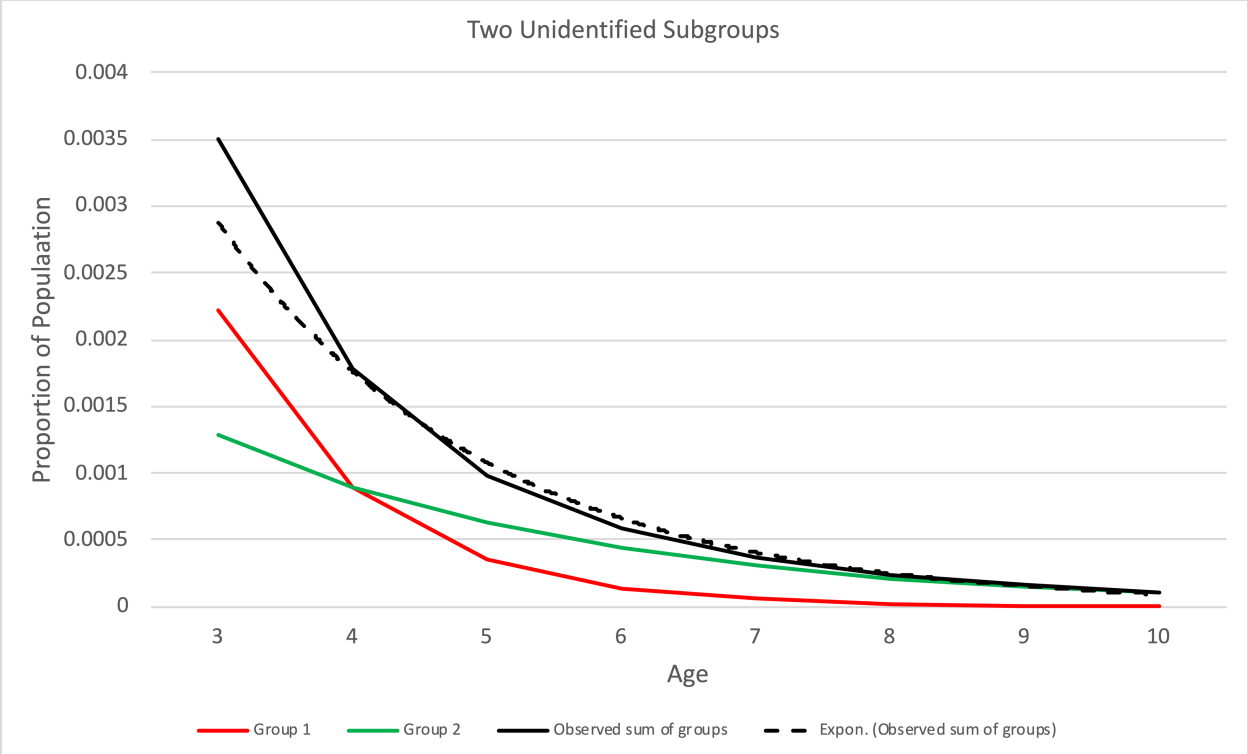


Figure 5 Example where observed diagnosis rates represent two unidentified subgroups with different values of diagnostic pressure

Any imbalance of case prevalence between in-migration and out-migration to and from the region defining the population over the study period would violate the assumption of constant prevalence within each cohort. If some in-migrating cases were diagnosed before in-migration and their subsequent re-diagnoses in the study region were labeled as initial diagnoses, that would violate the assumption of truly first diagnoses. Such an effect would be most evident at greater ages after the diagnosis of most cases. Bounding the maximum age studied $M$ to a modest value, sufficient to capture most initial diagnoses, can minimize any resulting bias.

Apart from subgroups with non-homogeneous $h$, it is theoretically possible for $h$ to have different effective values for cases of different ages with the same symptom severity at the same $DY$. Such an effect would represent an age bias in diagnostic pressure, independent of symptom severity. If there is a reason to suspect such a bias, investigators can add an age parameter to $h$ in the model and estimate its parameters.

For datasets where diagnosis follows best practices using gold-standard criteria, the lack of false positives may be a fair assumption, and it would be difficult discover any false positives. Where diagnosis uses a less precise process, some false positives might occur. For example, diagnosticians might tend to produce a positive diagnosis of individuals who do not meet case diagnostic criteria, perhaps under pressure from the patient or parents, or to facilitate services for the individual. In scenarios where the rate

of false positives is significant, the age distribution relative to the rates of true diagnoses may be important. If false positives are uniformly distributed over the age range studied, the result would be a pure additive effect to the rates of diagnoses. True case diagnoses should be more common at younger ages and less common as the risk pool is depleted, so false positives may be relatively more evident at older ages. If the age distribution of false positives is weighted toward older ages, the effect may be further pronounced.

## Model Fit

To ensure robust conclusions, investigators should test the model fit to ascertain both model correctness and parameter estimate accuracy. The model fits well if summary measures of the error are small and individual point errors are unsystematic and small[13]. One can examine the fit both graphically and numerically. Plots of $D_{BY,A}$ vs. $\widehat{D_{BY,A}}$ at all ages for individual cohorts can illuminate any issues with fit, which might occur at only some cohorts or ages. Visualization of the model vs. data can expose aspects of the data that might not fit well in a model with few parameters, which may suggest a model with more parameters or semi-parametric specifications.

If the model uses an assumed age of complete eligibility $AE^*$ that differs from the true value of $AE$ represented by the data, model fit may be impaired, particularly if $AE^* < AE$. As the Simulation Study section shows, setting $AE^* < AE$ can result in estimation errors, while setting $AE^* > AE$ tends not to impair model fit and may improve it in case of non-homogeneous subgroups. The presence of non-homogeneous subgroups may be evident from examining model fit.

The chi-square test statistic, applied to the overall model, to individual cohorts, and single ages across cohorts, is a numerical approach to assess absolute model fit. The p-value associated with the chi-square statistic utilizes observed and expected count values, not proportions. The p-value incorporates the effect of the number of parameters in the model via the degrees of freedom.

## Simulation Study

We tested the TTEPE method via a simulation study where we know the ground truth of all parameters, following the recommendations in Morris[14]. There are six pairs of values of $\beta_h$ and $\beta_P$, each ranging from 0 to 0.1 in steps of 0.02, and each pair sums to 0.1. That is, in one parameter set the prevalence increases as $e^{0.1\times BY}$ and diagnostic pressure is constant; in another parameter set, the prevalence is constant and diagnostic pressure increases as $e^{0.1\times DY}$; and the other four parameter sets represent various rates of change of both variables. In all cases, $P = 0.01$ at the final $BY$, $h = 0.25$ at the final $DY$, $AE = 3$, $M = 10$, and there are 20 successive cohorts. These simulations assume the correct value $AE^* = 3$ is known from either knowledge of the disorder or estimation of $E_A$. The study synthesized each data model both as real-valued

proportions without sampling and with binomial random sample generation of incident diagnoses, with the population of each cohort set to a constant 500,000. Monte Carlo simulation of parameter estimation bias and model standard error (SE) used 1000 iterations of random data generation for each set of parameters. Parameter estimation is described above, implemented using the Python SciPy curve_fit() function.

Table 1 shows results using real-valued proportions without sampling, isolating the estimation process from random sampling variations. It shows the bias in estimating each of the four model parameters for each of the six combinations of $\beta_h$ and $\beta_P$. The biases are minimal; the greatest bias magnitude in $\widehat{\beta_P}$ occurs with $\beta_P$ = 0 and $\beta_h$ = 0.1 and is on the order of $10^{-10}$.

**Table 1** Simulation results of parameter optimization, real-valued proportions with no sampling.

| True Parameters | | Bias $\hat{P}$ at final $BY$ | Bias $\widehat{\beta_P}$ | Bias $\hat{h}$ at final $DY$ | Bias $\widehat{\beta_h}$ |
|---|---|---|---|---|---|
| $\beta_P$ | $\beta_h$ | | | | |
| 0.1 | 0 | 5.9E-12 | 8.9E-11 | -5.5E-10 | -1.8E-10 |
| 0.08 | 0.02 | 0 | 0 | -2.8E-17 | -1.4E-17 |
| 0.06 | 0.04 | -1.7E-18 | 0 | 1.1E-16 | 1.4E-17 |
| 0.04 | 0.06 | 3.5E-18 | 1.4E-17 | -5.6E-17 | -6.9E-18 |
| 0.02 | 0.08 | 0 | 3.5E-18 | 5.6E-17 | -1.4E-17 |
| 0 | 0.1 | -4.7E-11 | -6.6E-10 | 2.2E-9 | 7.7E-10 |

$P$= 0.01 at the final $BY$, $h$ =0.25 at the final $DY$, $AE^*$=$AE$=3, $M$ = 10, and there are 20 successive cohorts.
$\beta_P, \beta_h$ are coefficients for prevalence and hazard, respectively.
$BY$ is birth year, $DY$ is diagnostic year.

Table 2 gives results from Monte Carlo analysis of the same parameter sets where the data use binomial sampling. It shows the bias and model SE of each parameter for each parameter set. The bias of the primary parameter $\widehat{\beta_P}$ remains small, on the order of $10^{-5}$ or $10^{-6}$. The SE is relevant when considering sampling, and it shows the effect of sampling compared to Table 1.

**Table 2** Simulation results of parameter optimization, Monte Carlo with binomial sampling, 1000 iterations.

| True Parameters | | $\hat{P}$ at final $BY$ | | $\widehat{\beta_P}$ | | $\hat{h}$ at final $DY$ | | $\widehat{\beta_h}$ | |
|---|---|---|---|---|---|---|---|---|---|
| $\beta_P$ | $\beta_h$ | Bias | SE | Bias | SE | Bias | SE | Bias | SE |
| 0.1 | 0 | -2.0E-6 | 1.0E-4 | -2.0E-5 | 0.0013 | 3.3E-5 | 0.0070 | -5.4E-6 | 0.00191 |
| 0.08 | 0.02 | -2.6E-6 | 1.1E-4 | -3.2E-5 | 0.0012 | 1.5E-4 | 0.0072 | 4.4E-5 | 0.00191 |
| 0.06 | 0.04 | 7.8E-6 | 1.2E-4 | 2.7E-5 | 0.0013 | -4.4E-4 | 0.0079 | -1.2E-4 | 0.00207 |
| 0.04 | 0.06 | 6.5E-5 | 0.0015 | 6.5E-5 | 0.0015 | -5.8E-4 | 0.0085 | -1.6E-4 | 0.00224 |
| 0.02 | 0.08 | -2.0E-6 | 1.6E-4 | -9.8E-6 | 0.0016 | 4.4E-4 | 0.0086 | 9.4E-5 | 0.00227 |
| 0 | 0.1 | 4.5E-6 | 1.8E-4 | 7.1E-6 | 0.0017 | 2.0E-4 | 0.0094 | 2.7E-5 | 0.00234 |

Population of each cohort = 500,000. $P$=0.01 at the final $BY$, $h$=0.25 at the final $DY$, $AE^*$=$AE$=3, $M$ = 10, and there are 20 successive cohorts.
$\beta_P, \beta_h$ are coefficients for prevalence and hazard, respectively.
$BY$ is birth year, $DY$ is diagnostic year.

Table 3 gives results where estimation uses an assumed value $AE^*$ that, in some cases, does not match the true value of $AE = 3$ represented by the data. Synthesis uses one homogenous group with consistent $h$ at each value of $DY$. Estimation using $AE^* = 2$ results in substantial estimation errors and model mis-fit that is obvious from plots of data vs. model (not shown). Estimation using $AE^* = 3$, $AE^* = 4$, or $AE^* = 5$ produces accurate results, with slightly more error where $AE^* = 5$. Plots show that the model fits well in all three cases (not shown). The choice of $AE^*$ is not critical as long as $AE^* \geq AE$. These data use real-valued proportions to avoid confusing model mismatch with sampling effects.

**Table 3** Comparison of the effect of the choice of assumed $AE^*$ vs. true value of $AE = 3$, with one homogeneous group of cases.

| $AE^*$ used in estimation | Bias $\hat{P}$ at final $BY$ | Bias $\widehat{\beta_P}$ | Bias $\hat{h}$ at final $DY$ | Bias $\widehat{\beta_h}$ |
|---|---|---|---|---|
| 2 | 0.002 | -0.019 | -0.0096 | 0.036 |
| 3 | 5.9E-12 | 8.9E-11 | -5.5E-10 | -1.8E-10 |
| 4 | -4.4E-12 | -6.6E-11 | 4.8E-10 | 1.64E-10 |
| 5 | 1.5E-11 | 2.2E-10 | -1.85E-9 | -7.18E-10 |

True values: $\beta_P$ = 0.1, $\beta_h$ = 0, $P = 0.01$ at the final $BY$, $h$=0.25 at the final $DY$, $AE = 3$. Maximum age $M$ = 10. There are 20 successive cohorts. Diagnostic pressure is consistent across cases at each $DY$. Simulation uses real values, no sampling.

Table 4 shows results with an intentional mismatch between estimation assuming one homogeneous group and data representing two subgroups with different values of $h$, illustrated in Figure 5. Note the visible error of the exponential fit to the data at age = 3,

and good fit for age > 3. In this synthetic dataset, the two subgroups are of equal size, and the true value of $h$ in one group is twice that of the other. This information is not known to the estimation, and the data do not indicate subgroup size nor membership. In the worst case, estimation uses $AE^* = AE = 3$ and the $\widehat{\beta_P}$ bias is 0.001, which is 1% of the actual value of 0.1. This error is due to the subgroups having different hazards, which are not accounted for in the estimation. When using $AE^* = 4$ or $AE^* = 5$, the $\widehat{\beta_P}$ bias becomes $6 \times 10^{-4}$ or less, and model fit is evidently improved (not shown).

**Table 4** Comparison of the effect of the choice of assumed $AE^*$ vs. true value of $AE = 3$, with two unidentified subgroups with different hazards, mis-matched to analysis.

| $AE^*$ used in estimation | Bias $\hat{P}$ at final $BY$ | Bias $\widehat{\beta_P}$ | Bias $\hat{h}$ at final $DY$ | Bias $\widehat{\beta_h}$ |
|---|---|---|---|---|
| 3 | -4.3E-4 | 0.001 | 0.0018 | -0.002 |
| 4 | -3.8E-4 | 6.1E-4 | -0.004 | -0.0016 |
| 5 | -3.3E-4 | 3.5E-4 | -0.0097 | -0.0011 |

True values: $\beta_P$ = 0.1, $\beta_h$ = 0, $P = 0.01$ at the final $BY$, $h$ =0.25 at the final $DY$, $AE = 3$. Two equal-sized groups of cases where the diagnostic pressure $h$ of one group is twice that of the other, while the estimation assumes one homogeneous group. Maximum age $M$ = 10. There are 20 successive cohorts. Simulation uses real values, no sampling.

## Discussion

Readers may suspect that the estimates are unidentified, i.e., not unique, due to possible interaction between age, diagnostic year, and birth year. While that concern is appropriate with analytical methods that assume an age distribution, ignore it, or estimate it inappropriately, TTEPE avoids that problem and produces uniquely identified estimates. TTEPE models the age distribution of initial diagnoses directly as a function of birth year, diagnostic year, and other variables.

Keep in mind that the eligibility function $E_A$ is different from the age distribution of diagnoses; $E_A$ is an attribute of the disorder under study.

This paper states the assumptions that underlie TTEPE analysis. The DAG of Figure 3 illustrates the assumed causal paths from birth year, diagnostic year, and age, including the set of time-varying diagnostic factors and the effect of changes in criteria on effective prevalence. The DAG and associated analysis appear to cover all plausible mechanisms to explain observed trends in rates of initial diagnoses.

TTEPE provides accurate estimates of prevalence parameters with a strong power to detect small differences. The Monte Carlo simulation study in Table 2 shows a magnitude of bias of the prevalence coefficient $\widehat{\beta_P}$ of $\leq 6.5 \times 10^{-5}$ or 0.0065% per year. The model SE of $\widehat{\beta_P}$ ranges from 0.0012 to 0.0017, where the true $\beta_P$ ranges from 0 to

0.1. Using the largest observed SE and $1.96 \times SE$ as a threshold for 95% confidence intervals, the method can detect differences in $\beta_P$ of 0.0033, i.e., 0.33% per year. We can expect similar performance for real-world datasets that meet the baseline assumptions and have characteristics comparable to the simulated data. The population size and the prevalence affect the SE. Note that in the simulation study there are 20 cohorts and 11 ages (0 through 10), so there are 220 data points. Each data point is an independent binomial random sample. The analysis estimates four parameters that define the curves that fit the data. The large number of independent data points and the small number of model parameters help explain the small bias and model SE. If the population of each cohort or the prevalence were substantially smaller, we would expect the bias and SE to be larger. These could occur with small geographic regions or very rare disorders, respectively.

TTEPE is useful for answering some fundamental questions, such as the actual trend in case prevalence over multiple birth cohorts of disorders such as autism and intellectual disability, as described in Elsabbagh[15] and McKenzie[16] respectively. Where datasets include appropriate covariates, stratified analysis can estimate the relationships between trends in true case prevalence and diagnostic factors, and population characteristics such as sex, race, ethnicity, socio-economic status, geographic region, parental education, environmental exposure, genetic profile, and other potential factors of interest.

It may be feasible to extend TTEPE to disorders where the time scale starts at some event other than birth. For example, the time origin might be the time of completion of a sufficient cause, and various outcomes may serve as events of interest. It is important to ensure that the eligibility function with respect to the time origin is consistent across cohorts.

Investigators may utilize domain knowledge to inform specialized analyses. For example, they may incorporate knowledge of mortality rates and standardized mortality ratios, rates of recovery from the condition before diagnosis, or the characteristics of migration in and out of the study region.

## References

1. Porta, M, ed. (2008). *A Dictionary of Epidemiology, fifth edition,* pp 70. New York, NY: Oxford University Press
2. Rothman, K, Greenland, S, Lash, T. (2008). *Modern Epidemiology, third edition*, pp 47, 516, and 656. Philadelphia PA: Wolters Kluwer
3. GBD 2017 Disease and Injury Incidence and Prevalence Collaborators. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic

analysis for the Global Burden of Disease Study 2017, Supplementary appendix 1. *Lancet*. 2018; 392: 1789–858

4. Baxter, A, Brugha, A, Erskine, R, Vos, T, Scott, J. The epidemiology and global burden of autism spectrum disorders. *Psychological Medicine.* 2015;45(3):601–613
5. Campbell, C, Davarya, S, Elsabbagh, M, Madden, L, Fombonne, E. (2011) Prevalence and the Controversy. Matson, J, Sturmey, P. (eds) *International Handbook of Autism and Pervasive Developmental Disorders,* pp 24-35 New York, NY: Springer
6. Fombonne, E. Epidemiology of pervasive developmental disorders. *Pediatric Research.* 2009;65(6):591-598
7. Schisterman, E, Cole, S, Platt, R. Overadjustment Bias and Unnecessary Adjustment in Epidemiologic Studies. *Epidemiology*. 2009;20:488-495
8. Findley, D. Counterexamples to Parsimony and BIC. *Ann. Inst. Statist. Math.* 1991;43(3):505-514
9. Cox DR. Regression Models and Life Tables. *J Royal Stat. Soc.* 1972;34(2):187-220
10. Kalbfleisch, J, Prentice, R. (2002). *The Statistical Analysis of Failure Time Data, second edition*, pp 8-9, 36-47. Hoboken NJ: Wiley
11. MacInnis, A. (2020, July 19). Time-to-event Prevalence Estimation TTEPE. https://doi.org/10.17605/OSF.IO/WPNKU
12. Hyman, S, Levy, S, Meyers, S. Later age at diagnosis was associated with reported mild presentation. *Pediatrics*. 2020; 145(1): e20193447
13. Hosmer, D, Lemeshow, S, Sturdivant, R (2013). *Applied Logistic Regression, Third Edition*, pp. 153-157. Hoboken NJ: Wiley
14. Morris, T, White, I, Crowther, M. Using simulation studies to evaluate statistical methods. *Statistics in Medicine*. 2019;38:2074–2102
15. Elsabbagh, M, Divan, G, Koh, Y-K, et al. Global Prevalence of Autism and Other Pervasive Developmental Disorders. *Autism Res.* 2012;5:160–179
16. McKenzie K, Milton M, Smith G, Ouellete-Kuntz H. Systematic Review of the Prevalence and Incidence of Intellectual Disabilities: Current Trends and Issues. *Cur Dev Disord Rep*. 2016;3:104-115