

1 **Time-to-event estimation of birth year prevalence trends:**
2 **a method to enable investigating the etiology of**
3 **childhood disorders including autism**

4 Alexander G. MacInnis^{1*}

5

6 ¹ Independent researcher, Mountain View, California, United States of America

7 * Corresponding author

8 Email: a.macinnis@alumni.stanford.edu (AGM)

9 **Abstract**

10 Measures of incidence are essential for investigating etiology. For congenital diseases
11 and disorders of early childhood, birth year cohort prevalence serves the purpose of
12 incidence. There is uncertainty and controversy regarding the birth prevalence trend of
13 childhood disorders such as autism and intellectual disability because changing
14 diagnostic factors can affect the rate and timing of diagnosis and confound the true
15 prevalence trend. The etiology of many developmental disorders is unknown, and it is
16 important to investigate. This paper presents a novel method, Time-to-Event Prevalence
17 Estimation (TTEPE), to accurately estimate the time trend in birth prevalence of
18 childhood disorders correctly adjusted for changing diagnostic factors. There is no
19 known existing method that meets this need. TTEPE is based on established time-to-
20 event (survival) analysis techniques. Input data are rates of initial diagnosis for each
21 birth year cohort by age or, equivalently, diagnostic year. Diagnostic factors form
22 diagnostic pressure, i.e., the probability of diagnosing cases, which is a function of
23 diagnostic year. Changes in diagnostic criteria may also change the effective
24 prevalence at known times. A discrete survival model predicts the rate of initial
25 diagnoses as a function of birth year, diagnostic year, and age. Diagnosable symptoms
26 may develop with age, affecting the age of diagnosis, so TTEPE incorporates eligibility
27 for diagnosis. Parameter estimation forms a non-linear regression using general-
28 purpose optimization software. A simulation study validates the method and shows that
29 it produces accurate estimates of the parameters describing the trends in birth
30 prevalence and diagnostic pressure. The paper states the assumptions underlying the

Time-to-Event Prevalence Estimation

31 analysis and explores optional additional analyses and potential deviations from
32 assumptions. TTEPE is a robust method for estimating trends in true case birth
33 prevalence controlled for diagnostic factors and changes in diagnostic criteria under
34 certain specified assumptions.

35 **Keywords**

36 Birth year prevalence; birth year cohort; incidence; time-to-event; survival analysis;
37 diagnostic factors; diagnostic pressure; non-linear regression; autism; intellectual
38 disability; childhood disorders; developmental disorders

39 **Introduction**

40 In epidemiology, incidence - the rate of new cases - is a fundamentally important metric
41 for estimating causal associations of time-varying risk factors with rates of a disorder
42 [1,2]. Incidence is different from prevalence, which is the proportion of a defined
43 population with the disorder at a defined time. For some disorders including congenital
44 diseases and developmental disorders such as autism and intellectual disability, birth
45 year cohort prevalence is used instead of incidence [2,3] because disorder incidence is
46 indistinguishable from birth prevalence, and diagnosis may occur later if at all.
47 Sometimes “incidence” is used to mean the rate of incident diagnoses rather than true
48 incidence, which refers to a disorder. While this usage is understandable because
49 observations inherently represent diagnosis and identification, the difference can be

Time-to-Event Prevalence Estimation

50 critically important when there is uncertainty about the rate and timing of diagnosing or
51 identifying cases. Serious developmental disorders such as autism and intellectual
52 disability are important topics of study because they cause a significant reduction in
53 quality of life across the lifespan [4] for the individuals and their families, and they affect
54 large numbers of people.

55

56 Studies of time trends in birth year prevalence or incidence are subject to biases
57 resulting from changes in diagnostic factors and diagnostic criteria [2]. Investigators can
58 estimate incidence and birth prevalence directly from data on diagnoses. But concerns
59 that diagnostic factors and diagnostic criteria may have affected the data can lead to a
60 lack of confidence in the validity of direct estimates. Diagnostic factors are those that
61 influence the probability of diagnosing or identifying cases. Examples include
62 awareness, outreach efforts, screening, diagnostic practice, diagnostic criteria, social
63 factors, policies, and financial incentives for diagnosis. Changes in diagnostic criteria
64 can also have the additional effect of changing the effective birth prevalence by
65 including or excluding as cases some portion of the population compared to prior
66 criteria, with the changes occurring when the new criteria take effect.

67

68 Consider, for example, this question. If reported birth prevalence increased over time,
69 was this caused by changes in actual birth prevalence, the probability of diagnosing
70 cases, diagnostic criteria affecting prevalence, or some combination of these factors? It

Time-to-Event Prevalence Estimation

71 is challenging to disentangle these effects, and there is no known existing method
72 capable of doing so correctly.

73 **Literature Review**

74 There are many studies on the prevalence of developmental disorders, yet very few of
75 them directly address birth year prevalence trends and very few address methods of
76 adjustment for diagnostic factors. The series of reports from the US Centers for Disease
77 Control and Prevention's (CDC) Autism and Developmental Disabilities Monitoring
78 Network (ADDM) [5-13] estimate the prevalence of autism among children who were
79 eight years old at each even-numbered year 2000 through 2016. Each report describes
80 the prevalence of a single year birth cohort, subject to rounding, born eight years before
81 the respective study year. The set of reports represents the trend in birth year
82 prevalence, but the reports describe the findings as simply "prevalence," and do not
83 discuss birth year prevalence or similar names. The ADDM reports suggest that the
84 observed increases in (birth year) prevalence may result from various factors, including
85 changing composition of study sites and geographic coverage, improved awareness,
86 and changes in diagnostic practice and availability of services. However, they do not
87 suggest methods to quantify such effects or to adjust for them. Croen [14] examined
88 birth year prevalence trends in autism and mental retardation in California for birth years
89 1987 to 1994. They concluded that the data and methods available were insufficient to
90 determine how much of the observed increase reflected an increase in true birth
91 prevalence. Hansen [15] recommends using the cumulative incidence of diagnoses of
92 childhood psychiatric disorders for each 1-year birth cohort as a measure of risk. They

Time-to-Event Prevalence Estimation

93 do not suggest an analytical method to estimate or adjust for the effects of diagnostic
94 factors. Nevison [16] presents California Department of Developmental Services data
95 showing a sharp rise in birth year prevalence of autism over several decades but does
96 not discuss methods to adjust for the effects of diagnostic factors.

97

98 Elsabbagh [17] states that investigating time trends in prevalence or incidence requires
99 holding diagnostic factors such as case definition and case ascertainment “under strict
100 control over time,” but does not suggest a method for doing so. Campbell [18] reviews
101 prevalence estimates and describes an ongoing controversy about them. They
102 emphasize the distinction between prevalence and incidence but do not mention birth
103 year prevalence. They summarized the CDC ADDM estimates and stated that one
104 cannot infer incidence from the ADDM prevalence estimates. However, they did not
105 mention that the ADDM estimates are birth year prevalence, which serves the purpose
106 of incidence for childhood disorders. Campbell indicates that analyses should control for
107 certain diagnostic factors, but they do not suggest a method for doing so. Baxter [4]
108 examined prevalence and incidence but did not mention birth year prevalence and did
109 not indicate whether their use of “incidence” refers to the incident diagnoses or
110 incidence of the disorder. Later sections of this paper show why the distinction is crucial.
111 Baxter adjusted for covariates that they assumed introduced bias, including the use of
112 dichotomous variables representing the most recent diagnostic criteria. Such variables
113 inherently represent the time each set of criteria took effect. However, Schisterman [19]
114 shows that controlling for variables on a causal path from the input (time, in this case) to

Time-to-Event Prevalence Estimation

115 the outcome (prevalence or incident diagnoses) constitutes inappropriate adjustment
116 and biases the estimate of the primary effect (i.e., of time on prevalence or incident
117 diagnoses) towards zero. Similarly, Rothman [2] states that controlling for intermediate
118 variables typically causes a bias towards finding no effect.

119

120 Keyes [20] used age-period-cohort analysis to attempt to disentangle the effects of birth
121 year (cohort) from diagnostic year (period) and concluded that period effects best
122 explain observed California data. They also argued that period effects represent
123 diagnostic factors, without noting that birth year prevalence is inherently a cohort effect.

124 Spiers [21], in a letter regarding Keyes, pointed out that the method used is extremely
125 sensitive to the constraints specified and could as easily have concluded that period,
126 i.e., diagnostic year, effects best explain the data. Spiers also disputed Keyes’
127 interpretation of cohort effects. King [22] implicitly used an age-period-cohort analysis,
128 assuming that period effects are dominant and controlling for birth year. There is
129 extensive literature on the problems with using age-period-cohort analysis to separate
130 the effects of birth year (cohort) from diagnostic year (period). Rodgers [23] states that a
131 constraint of the type used in King “in fact [it] is exquisitely precise and has effects that
132 are multiplied so that even a slight inconsistency between the constraint and reality, or
133 small measurement errors, can have very large effects on estimates.” O’Brien [24], in a
134 book devoted to this topic, states, regarding the relationships of age, period and cohort
135 to the dependent variable, “There is no way to decide except by making an assumption
136 about the relationship between these three variables.” MacInnis [24] showed that

Time-to-Event Prevalence Estimation

137 diagnostic factors are represented by the years of first diagnoses, formulates the
138 problem as one of separating birth year from diagnostic year, and shows that age-
139 period-cohort approaches are not suitable for such analyses. In particular, implicit
140 assumptions to make the model estimable cause the resulting estimates to conform to
141 the assumptions, forming circular logic.

142

143 Campbell [18] and McKenzie [26] both point out that various factors could potentially
144 affect the rate of diagnoses without affecting the true case rate.

145 **Overview**

146 The primary aim of this work is to develop and specify a method to estimate birth
147 prevalence trends, correctly adjusted for trends in the set of diagnostic factors and
148 changes in diagnostic criteria. Armed with such a tool, researchers can quantify the
149 effects of the set of variable causal factors separately from those of the set of diagnostic
150 factors. Where covariates are available, investigators can estimate associations of birth
151 prevalence with a variety of population characteristics that may be causal or
152 explanatory.

153

154 This paper presents a novel statistical method called time-to-event prevalence
155 estimation (TTEPE). It uses time-to-event survival analysis to estimate the trend in true
156 birth year prevalence, correctly adjusted for changes in the set of diagnostic factors and
157 diagnostic criteria. It presents the derivation of the analytical method from first principles
158 and states all the underlying assumptions. A simulation study shows that the method

159 effectively separates and quantifies the birth year trend and the trend in the effects of
160 diagnostic factors, producing accurate estimates.

161 **Method of time-to-event prevalence estimation**

162 **(TTEPE)**

163 **Background**

164 Comparison of prevalence estimates across multiple studies generally is not suitable for
165 informing trends in birth prevalence nor incidence [2]. Different prevalence estimates
166 may use different mixes of birth years and ages, as well as numerous other possible
167 differences between prevalence studies [27]. Many combinations of trends in birth
168 prevalence and diagnostic factors could potentially explain observed prevalence trends.
169 The Introduction section briefly describes the problems with age-period-cohort analyses.
170 Age (A), diagnostic year (DY), and birth year (BY) are exactly collinear, $DY = BY + A$,
171 subject to rounding, which leads to unidentified estimates when using a linear predictor.
172 Another reason is that the age distribution of diagnoses can differ for different solutions
173 of BY and DY , as shown below. It is challenging to estimate the age distribution correctly
174 given the collinearity problem.

175

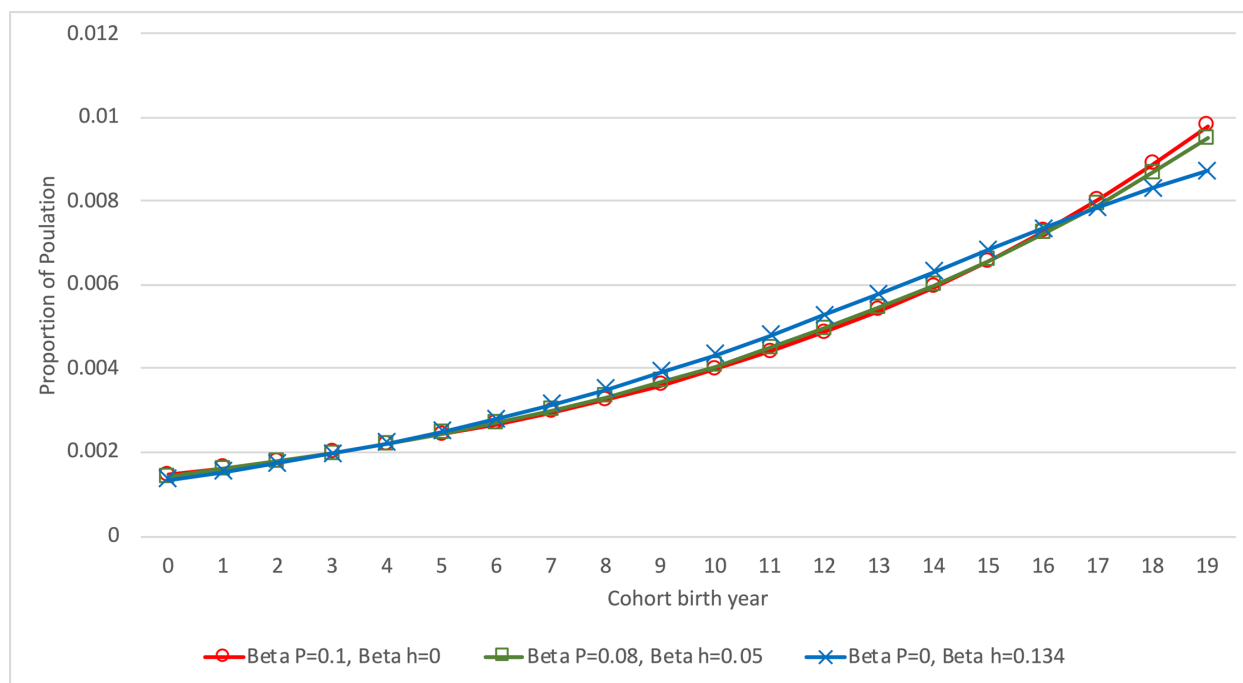
176 Analysis of the cumulative incidence, to a consistent age, of diagnoses in each birth
177 cohort comes closer to estimating the trend in true birth prevalence, but results are still

Time-to-Event Prevalence Estimation

178 ambiguous. Here too, many combinations of trends in birth prevalence and diagnostic
179 factors can produce similar trends in cumulative incidence.

180 **Ambiguity in estimation**

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



190

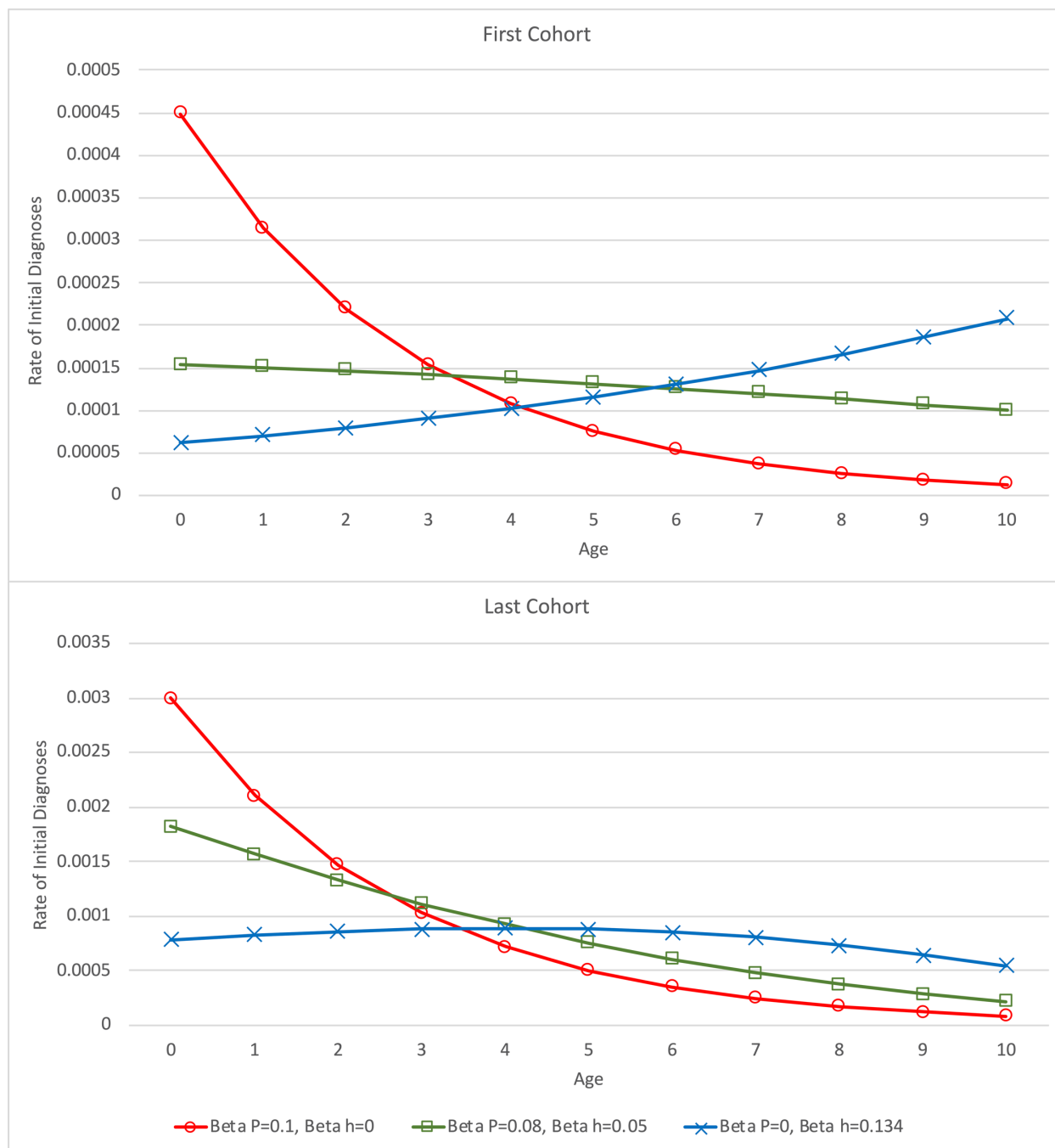
Time-to-Event Prevalence Estimation

191 **Fig 1.** Example of cumulative incidence under three models. β_P is the coefficient for birth
192 prevalence; β_H is the coefficient for the effect of diagnostic factors. Red line with circles
193 represents $\beta_P = 0.1, \beta_H = 0$; green line with squares represents $\beta_P = 0.08, \beta_H = 0.5$; blue line with
194 crosses represents $\beta_P = 0, \beta_H = 0.134$.

195

196 Fig 1 illustrates a hypothetical example of cumulative incidence to age ten over 20
197 consecutive cohorts. The legend lists the parameter value pairs for the three cases. β_P
198 is the exponential coefficient of birth year prevalence, and β_h is the exponential
199 coefficient of the effect of diagnostic factors by diagnostic year. The data generation
200 process producing these data uses a survival process as detailed below. An Excel
201 spreadsheet to generate all plots in this paper is available at OSF [28]. The variable h
202 represents diagnostic pressure, the effect of diagnostic factors, which is equivalent to
203 hazard, as explained in the section Significance of birth year and diagnostic year. While
204 the three cumulative incidence curves appear similar, the age distributions of diagnoses
205 are strikingly different, as Fig 2 shows. The remainder of this paper explains how
206 modeling the age distribution of first diagnoses enables accurate and unambiguous
207 estimation of the coefficients for birth prevalence and diagnostic factors.

Time-to-Event Prevalence Estimation



208

209 **Fig 2.** Distribution of diagnoses in the first and last cohorts under three models. β_P is the
 210 coefficient for birth prevalence; β_H is the coefficient for the effect of diagnostic factors. Red lines
 211 with circles represent $\beta_P = 0.1, \beta_H = 0$; green lines with squares represent $\beta_P = 0.08, \beta_H = 0.05$;
 212 blue lines with crosses represent $\beta_P = 0, \beta_H = 0.134$.

Time-to-Event Prevalence Estimation

213

214 If the disorder's complete cause is in place at birth, then the incidence of the disorder is
215 indistinguishable from birth prevalence. In contrast, diagnosis occurs later, if at all. The
216 incidence of diagnosis represents the combination of birth prevalence and delays and
217 omissions in diagnoses.

218

219 One might consider controlling for diagnostic factors over time, for example, via
220 regression, but that is not sufficient to distinguish between alternative explanations.
221 Diagnostic factors are a function of time, and birth year prevalence is also a function of
222 time. As the Introduction states, controlling directly for diagnostic factors biases the
223 estimates of the main effect, typically towards zero, citing Schisterman [19] and
224 Rothman [2]. Schisterman recommends, "clearly stating a causal question to be
225 addressed, depicting the possible data generating mechanisms using causal diagrams,
226 and measuring indicated confounders." This paper directly addresses these issues.

227 **Significance of birth year and diagnostic year**

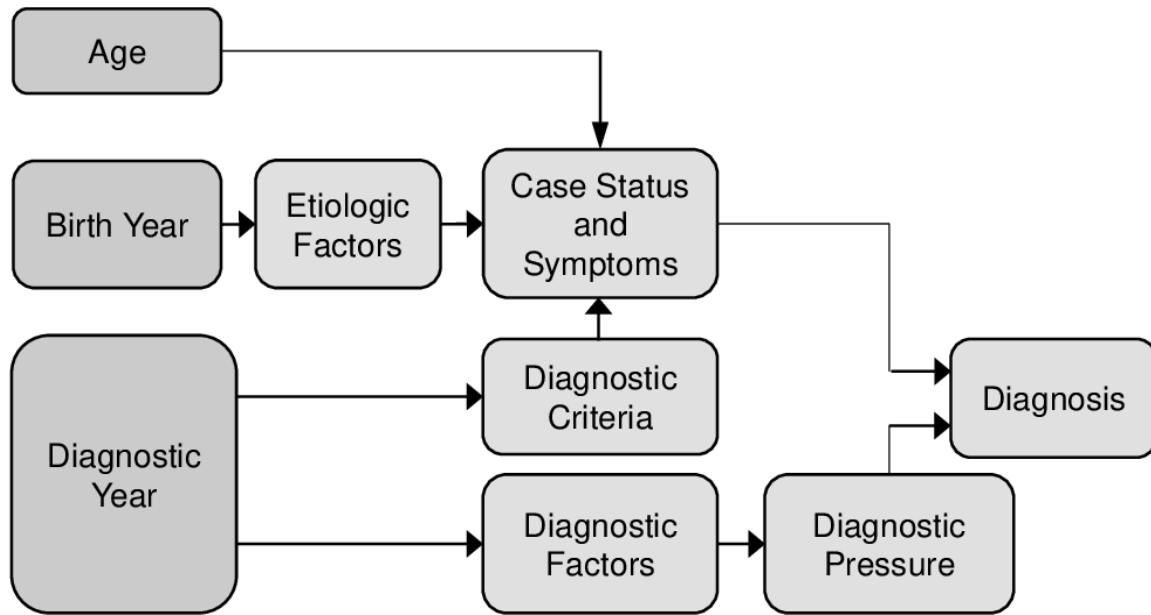
228 Diagnostic factors only affect the diagnosis of cases when those cases exhibit
229 diagnosable symptoms, referred to as being eligible for diagnosis. Diagnostic pressure
230 is the probability of diagnosing eligible undiagnosed cases, and it is an effect of the
231 combination of all diagnostic factors. The Introduction lists examples.

232

Time-to-Event Prevalence Estimation

233 Diagnostic pressure is equivalent to the hazard h in time-to-event or survival analysis.
234 For each case of the disorder, the information resulting from diagnostic pressure
235 consists of the time of initial diagnosis, that is, the diagnostic year. Diagnostic pressure
236 has no observable effect before the diagnosis of each case, and none after the initial
237 diagnosis since TTEPE considers only initial diagnoses. Hence, the effect of diagnostic
238 pressure on the input data is a function of diagnostic year.
239
240 The directed acyclic graph (DAG) in Fig 3 illustrates the causal paths from birth year,
241 diagnostic year, and age to diagnosis. Birth year drives etiologic (causal) factors, which
242 produce the disorder and its symptoms. Diagnostic criteria determine whether each
243 individual's symptoms qualifies them as a case, and criteria may change at specific
244 diagnostic years. Symptoms may vary with age. Diagnostic year drives diagnostic
245 factors, which form diagnostic pressure. Diagnosable symptoms and diagnostic
246 pressure together produce each initial diagnosis.

Time-to-Event Prevalence Estimation



247

248 **Fig 3.** Directed Acyclic Graph Representing Year of Birth, Diagnostic Year and Age

249

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

253 **Development of the TTEPE method**

254 The TTEPE method is based on the DAG of Fig 3 and modeling the age distribution of
255 initial diagnoses. The method avoids the identification problem associated with age-
256 period-cohort analysis, and it avoids the problem of inappropriate adjustment for
257 diagnostic factors.

258

Time-to-Event Prevalence Estimation

259 TTEPE is particularly applicable to disorders where case status is established by birth
260 or by a known age, and diagnosable symptoms are present by some consistent age. It
261 is also useful where cases develop diagnosable symptoms gradually over a range of
262 ages.

263

264 Data sources suitable for TTEPE analysis provide rates of initial diagnoses by age or,
265 equivalently, diagnostic year for each birth cohort. The rate is the count of initial
266 diagnoses divided by the population of the cohort at the respective age.

267

268 TTEPE relies on these principles: for each birth cohort, the number of cases at risk of
269 initial diagnosis decreases as cases are diagnosed, and diagnostic year is the time
270 when diagnostic factors affect the probability of diagnosing cases exhibiting diagnosable
271 symptoms.

272

273 First, consider the case where the diagnostic criteria do not change the effective
274 prevalence over the interval of interest. The section Changes in criteria affecting
275 prevalence examines the alternative.

276

277 The principle of finding the simplest model that fits the data, sometimes called Occam's
278 Razor, could lead to the conclusion that birth year effects best explain the observed
279 birth year prevalence trends. However, the most parsimonious model is not always the

Time-to-Event Prevalence Estimation

280 best one [29]. Specifically, diagnostic pressure as a function of diagnostic year could
281 potentially be an essential component for explaining observed diagnosis data.

282

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

288

289 TTEPE introduces the concept of eligibility. A case is eligible for diagnosis if the
290 individual has diagnosable symptoms and not eligible if the individual has not yet
291 developed diagnosable symptoms. For each birth cohort, undiagnosed eligible
292 individuals form the risk set of cases at risk of initial diagnosis. The size of the risk set is
293 denoted R . At each age there is some probability of diagnosis of each case in the risk
294 set. This probability is the diagnostic pressure h . At each age, newly diagnosed cases
295 are removed from the risk set, and newly eligible cases are added to the risk set. For
296 any given value of h , as R decreases or increases, the rate of initial diagnoses D
297 changes accordingly. This process generates the modeled age distribution. The survival
298 function S refers to cases that “survive” diagnosis at each age. If all cases were eligible
299 from birth, S would equal R . More generally, however, some cases may initially be
300 ineligible and become eligible as they age, so $R \leq S$. The prevalence P is the proportion

Time-to-Event Prevalence Estimation

301 of the population that are or become cases. The eligibility factor E is the eligible
302 proportion of P , $0 \leq E \leq 1$.

303

304 In contrast, in typical survival or time-to-event analysis, including Cox proportional
305 hazards analysis [30], the initial size of the risk set is assumed to have a known value,
306 for instance, an entire population or an entire sample. If the risk set R consisted of the
307 entire population without subtracting diagnosed cases, the estimate $\hat{h}_t = D_t/R_t$ would
308 be equivalent to the population-based rate of diagnoses D at time t . If the disorder is
309 rare, it makes little difference whether the risk set is the entire population or the
310 undiagnosed portion.

311

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

315 **Time-to-event analysis model**

316 The analysis model enables estimation of the temporal trend of birth prevalence P over
317 a range of cohorts, correctly adjusted for diagnostic pressure h . Both P and h can vary
318 with time. P is a function of birth year, and h is a function of diagnostic year. Estimation
319 of P adjusted for h requires estimating the time-based parameters of both P and h in
320 the time-to-event model and specifying or estimating the eligibility function E .

321

Time-to-Event Prevalence Estimation

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

328

329 Let h_{DY} be the diagnostic pressure at diagnostic year DY . $DY = BY + A$, subject to
330 rounding, so h_{DY} is equivalent to $h_{BY,A}$. Let P_{BY} be the case prevalence of birth year
331 cohort BY . Let $R_{BY,A}$ be the discrete risk set function of the population proportion of
332 eligible cases at risk of initial diagnosis at age A for birth year BY . TTEPE uses R rather
333 than a discrete survival function S to accommodate eligibility changing with age. Let E_A
334 be the discrete eligibility function, the proportion of cases that are eligible at age A ,
335 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
336 prevalent cases added to R due to changes in eligibility. For simplicity, assume E_A
337 increases monotonically, i.e., non-decreasing, meaning that cases do not lose eligibility
338 before diagnosis.

339

340 Kalbfleisch [31] gives background on general time-to-event theory and equations.

341

342 Consider three scenarios, differing by the characteristics of E_A . Here we write h_{DY} as
343 $h_{BY,A}$ to clarify the effect of A in $DY = BY + A$.

Time-to-Event Prevalence Estimation

344

345 Scenario: constant $E_A = 1$. All cases are eligible from birth, so $E_A = 1$ for all values of A .

346 This scenario is equivalent to standard time-to-event models that do not consider

347 eligibility. For $A \geq 1$, $E_A - E_{A-1} = 0$. For the first year of age, $A = 0$, $R_{BY,0} = P_{BY}E_0 = P_{BY}$

348 and $D_{BY,0} = R_{BY,0}h_{BY,0} = P_{BY}h_{BY,0}$.

349 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

350 $D_{BY,1} = R_{BY,1}h_{BY,1} = P_{BY}(1 - h_{BY,0})h_{BY,1}$.

351 For $A = 2$, $R_{BY,2} = R_{BY,1} - D_{BY,1} = P_{BY}(1 - h_{BY,0}) - P_{BY}(1 - h_{BY,0})h_{BY,1} =$

352 $P_{BY}(1 - h_{BY,0})(1 - h_{BY,1})$ and $D_{BY,2} = R_{BY,2}h_{BY,2} = P_{BY}(1 - h_{BY,0})(1 - h_{BY,1})h_{BY,2}$.

353 Similarly, for $A = 3$,

354 $R_{BY,3} = P_{BY}(1 - h_{BY,0})(1 - h_{BY,1})(1 - h_{BY,2})$ and

355 $D_{BY,3} = P_{BY}(1 - h_{BY,0})(1 - h_{BY,1})(1 - h_{BY,2})h_{BY,3}$.

356

357 Generally, for $A \geq 1$,

358
$$R_{BY,A} = P_{BY} \prod_{a=0}^{A-1} (1 - h_{BY,a})$$

359 and

360
$$D_{BY,A} = P_{BY} \prod_{a=0}^{A-1} (1 - h_{BY,a}) h_{BY,A} \quad (1)$$

361

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

Time-to-Event Prevalence Estimation

363
$$S_{BY,A} = P_{BY} - \sum_{a=0}^{A-1} D_{BY,a} \quad (2)$$

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

365

366 Scenario: Increasing E_A . $E_0 < 1$ and E_A increases monotonically with A . For $A = 0$,

367 $R_{BY,0} = E_0 P_{BY}$ and $D_{BY,0} = E_0 P_{BY} h_{BY,0}$. For each $A \geq 1$, $R_{BY,A} = R_{BY,A-1} - D_{BY,A-1} +$

368 $(E_A - E_{A-1})P_{BY}$. The incremental increase of E_A causes an incremental increase in $R_{BY,A}$.

369 Then,

370
$$D_{BY,A} = R_{BY,A} h_{BY,A} = (R_{BY,A-1} - D_{BY,A-1}) h_{BY,A} + (E_A - E_{A-1}) P_{BY} h_{BY,A} \quad (3)$$

371 Equation (3) can be useful as a procedural definition. We can write equivalent

372 expressions for $R_{BY,A}$ and $D_{BY,A}$ as sums of expressions similar to equation (1), where

373 each summed expression describes the portion of P_{BY} that becomes eligible at each

374 age according to E_A . For $A \geq 1$,

375

376
$$R_{BY,A} = \sum_{a=0}^{A-1} (E_a - E_{a-1}) P_{BY} \prod_{b=a}^{A-1} (1 - h_{BY,b})$$

377

378
$$D_{BY,A} = \sum_{a=0}^{A-1} (E_a - E_{a-1}) P_{BY} \prod_{b=a}^{A-1} (1 - h_{BY,b}) h_{BY,A} \quad (4)$$

379

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

381

Time-to-Event Prevalence Estimation

382 Scenario: Plateau E_A . E_A increases from $E_0 < 1$ and plateaus at $E_A = 1$ for $A \geq AE$,
 383 where AE is the age of complete eligibility, $AE < M$, and M is the maximum age
 384 included in the analysis. Equation (3) applies, noting that for $A > AE$, $(E_A - E_{A-1}) = 0$.
 385 Equivalently, combine equation (2) with the fact that $E_{AE} = 1$ to obtain $R_{AE} = S_{AE} = P_{BY} -$
 386 $\sum_{a=0}^{AE-1} D_{BY,a}$, so

$$387 \quad D_{BY,AE} = R_{BY,AE} h_{BY,AE} = S_{BY,AE} h_{BY,AE} = (P_{BY} - \sum_{a=0}^{AE-1} D_{BY,a}) h_{BY,AE} \quad (5)$$

388 and for $A > AE$,

$$389 \quad R_{BY,A} = S_{BY,A} = (P_{BY} - \sum_{a=0}^{AE-1} D_{BY,a}) \prod_{b=AE}^{A-1} (1 - h_{BY,b})$$

390

$$391 \quad D_{BY,A} = (P_{BY} - \sum_{a=0}^{AE-1} D_{BY,a}) \prod_{b=AE}^{A-1} (1 - h_{BY,b}) h_{BY,A} \quad (6)$$

392

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

399 **Prevalence, cumulative incidence and censoring**

400 The case prevalence in each cohort is the cumulative incidence of initial diagnoses
401 through the last age of follow-up plus the censored portion. This assumes that any
402 difference in competing risks between cases and non-cases in the age range analyzed
403 is small enough to be ignored. This assumption is consistent with Hansen [15]. If the
404 rate of deaths of cases before initial diagnosis exceeds that of the entire population of
405 the cohort at the same ages, that excess would constitute a competing risk and would
406 reduce the estimated prevalence accordingly.

407

408 In all three scenarios of E_A , we can express P as a function of S and the cumulative
409 incidence $CI = \sum_{a=0}^{A-1} D_{BY,a}$ for $A > 0$, by rearranging equation (2) as $P = S_A + CI_{A-1}$.

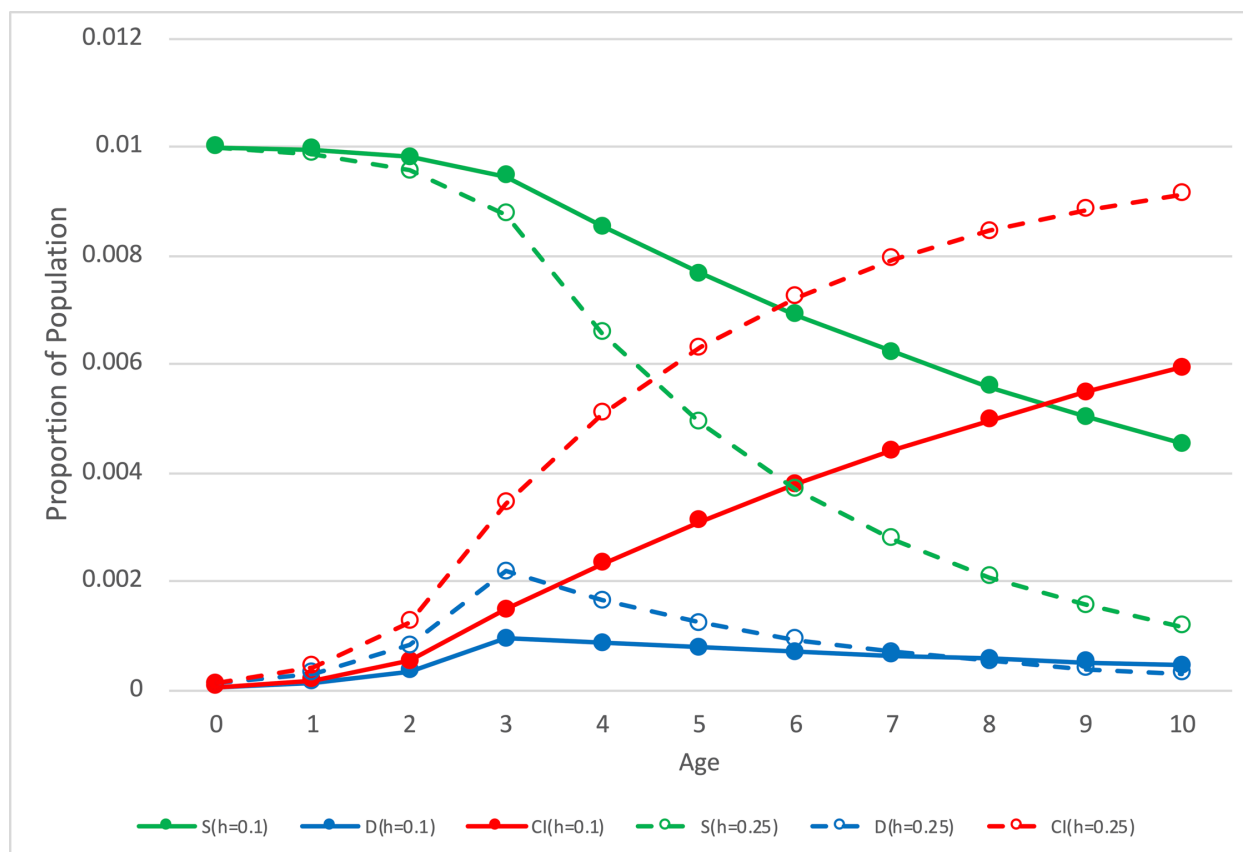
410 Assuming that eligibility at the last age of follow-up $E_M = 1$, $S_M = R_M$. Then, $P = R_M +$
411 CI_{M-1} and $D_M = R_M h_M$. The censored proportion is $S_{M+1} = S_M - D_M$, which is equivalent
412 to $S_{M+1} = R_M - R_M h_M = R_M(1 - h_M)$. After estimating the model parameters, the
413 estimated censored proportion is $\widehat{S}_{M+1} = \widehat{R}_M(1 - \widehat{h}_M)$.

414 **Illustrative example**

415 Fig 4 illustrates an example according to the plateau E_A scenario showing the
416 relationships between prevalence, diagnosis rates, the survival function, and cumulative
417 incidence CI with two different values of diagnostic pressure $h = 0.1$ and $h = 0.25$ and
418 prevalence $P = 0.01$. In this example, $E = 1$ for $A \geq AE = 3$ and h takes on one of two
419 constant values. The value of h determines the shapes of these functions vs. age. This

Time-to-Event Prevalence Estimation

420 example shows constant values of h purely for clarity, not as an assumption nor a
 421 limitation of TTEPE.



422

423 **Fig 4.** Example of a survival process for two values of diagnostic pressure h .

424 The green lines S denote survival, the blue lines D denote the rate of diagnoses, and the red
 425 lines CI denote cumulative incidence. The solid lines represent $h=0.1$, and the dotted lines
 426 represent $h=0.25$.

427

428 As cases are diagnosed, S decreases and CI increases. R is not shown; $R = S$ for $A \geq$

429 $AE = 3$. Only D is observable.

430 **Assumptions**

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

433

434 1. The eligibility function E_A under consistent diagnostic criteria is consistent across
435 cohorts.

436 2. The diagnostic pressure applies equally to all eligible undiagnosed cases at any
437 given diagnostic year.

438 3. The case prevalence under consistent diagnostic criteria within each cohort is
439 constant over the range of ages included in the analysis.

440 4. Case status is binary according to the applicable diagnostic criteria.

441 5. The discrete-time interval (e.g., one year) is small enough that the error
442 introduced by treating the variable values as constant within each interval is
443 negligible.

444 6. No false positives.

445 7. Data represent truly initial diagnoses.

446 8. Any difference in competing risks between cases and non-cases in the age range
447 analyzed is small enough to be ignored.

448

449 The assumption of a consistent eligibility function means that cases develop

450 diagnosable symptoms as a function of age, and that function is the same for all cohorts

451 under consistent diagnostic criteria. The section Changes in criteria affecting prevalence
452 discusses a separate effect that might make the eligibility function appear inconsistent.

453 **Estimating parameters**

454 TTEPE performs a non-linear regression that estimates the parameters of a model of
455 $D_{BY,A}$ using general-purpose optimization software. The model is based on equations (1)
456 through (6) selected based on the eligibility scenario. The model produces estimates
457 $\widehat{D}_{BY,A}$ from the parameters and independent variables, and the software finds the
458 parameter values that minimize a cost function $\text{cost}(D, \widehat{D})$. One suitable implementation
459 of optimization software in the Python language is the `curve_fit()` function in the SciPy
460 package (`scipy.optimize.curve_fit` in SciPy v1.5.2). Its cost function is $(D - \widehat{D})^2$, so it
461 minimizes the sum of squared errors. Python software to perform this regression and
462 the simulations described below is available at OSF [28].

463
464 Investigators should choose which model equation to use based on knowledge or
465 estimates of the eligibility function E_A . The constant E_A scenario and equation (1)
466 assume that all cases are eligible from birth, which may not be valid for some disorders.
467 The validity of the assumption that all cases are eligible by a known age AE , i.e., the
468 plateau E_A scenario and equation (6), may be supported by either external evidence,
469 e.g., the definition of the disorder, or estimation of E_A . The least restrictive approach of
470 the increasing E_A scenario uses equation (4) to estimate E_A . Non-parametric estimates
471 \widehat{E}_A can inform a choice of a parametric form of E_A . The value of E_A at the maximum age
472 studied M should be set to 1 to ensure the estimates are identifiable. If $E_A = 1$ for all $A \geq$

Time-to-Event Prevalence Estimation

473 AE , that fact and the value of AE should be apparent from estimates \widehat{E}_A , and the plateau
474 E_A scenario applies.

475

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

480

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

485

486 Suppose the population proportion of cases represented in the data is unknown for all
487 cohorts. In that case, estimates of the absolute prevalence, or the intercept, may be
488 underestimated by an unknown scale factor. If that proportion is known for at least one
489 cohort, we can use it to calibrate the intercept. Proportional changes in prevalence
490 between cohorts are unaffected by underestimation of the intercept. If the population
491 proportion of cases included in the sample changes over time, that change reflects
492 changing diagnostic factors, and the estimated parameters of h automatically represent
493 such changes.

494 **Changes in criteria affecting prevalence**

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

506

507 Let $\{CF_{cy}\}$ be the set of criteria factors that induce a multiplicative effect on effective
508 prevalence due to criteria changes that occurred at criteria years $\{cy\}$ after the first
509 DY included in the study. P_{BY} is the prevalence of cohort BY before the effect of any of
510 $\{CF_{cy}\}$. For each cohort BY , the effective prevalence $EP_{BY,A}$ at age A is

511
$$EP_{BY,A} = P_{BY} \prod_{cy \leq (BY+A)} CF_{cy} \quad (7)$$

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

Time-to-Event Prevalence Estimation

515

516 For a given BY and increasing A , $BY + A$ crossing any cy causes a step-change in the
517 effective prevalence EP . Using a general formulation of eligibility E_A , per the increasing
518 E_A scenario and equation (3), and for clarity substituting $BY + A$ for DY , we obtain the
519 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 $A \geq 1$,

520
$$R_{BY,A} = R_{BY,A-1} - D_{BY,A-1} + E_A(EP_{BY,A} - EP_{BY,A-1}) + (E_A - E_{A-1})EP_{BY,A}$$

521 and

522
$$D_{BY,A} = [R_{BY,A-1} - D_{BY,A-1} + E_A(EP_{BY,A} - EP_{BY,A-1}) + (E_A - E_{A-1})EP_{BY,A}]h_{DY} \quad (8)$$

523 The term $EP_{BY,A} - EP_{BY,A-1}$ represents the change in the effective prevalence EP when
524 $BY + A$ crosses one of $\{cy\}$. As each CF_{cy} takes effect at $cy = DY = BY + A$, the newly
525 effective CF_{cy} changes $EP_{BY,A}$ and R in all BY cohorts where cy corresponds to an age A
526 in the range of ages studied. These changes in R affect the rates of initial diagnoses D .
527 For cohorts born after cy , CF_{cy} applies to all ages.

528

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

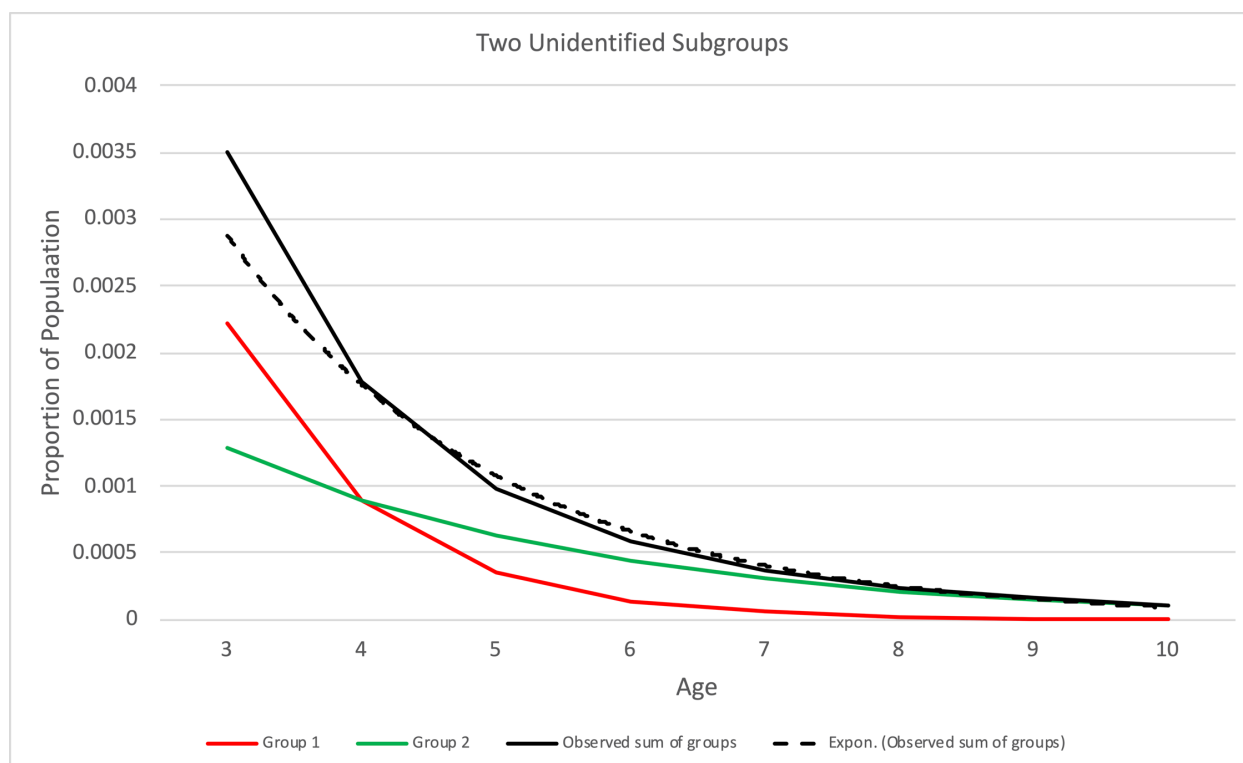
533

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

536 **Potential violations of assumptions**

537 Suppose a dataset represents a non-homogeneous set of cases with different effective
538 values of diagnostic pressure h applying to different unidentified subgroups at the same
539 DY . That would violate the assumption that the diagnostic pressure applies equally to all
540 eligible undiagnosed cases at any given DY . Cases may have differing degrees of
541 symptom severity, and more severe symptoms may result in earlier diagnosis [32],
542 implying greater diagnostic pressure. Fig 5 illustrates this situation. The figure illustrates
543 constant values of h purely for clarity, not as an assumption nor a limitation. If data
544 represent unidentified subgroups with differential diagnostic pressure, the distribution of
545 diagnoses is a sum of distributions with different values of h . Such a sum of distributions
546 with different values of h may impair fit with a model that assumes homogeneous h . For
547 data where $E = 1$ for $A \geq AE$ (plateau E_A), adjusting the assumed value of AE used in
548 estimation, called AE^* , may mitigate such errors, as shown in the Simulation study
549 section. Stratified estimation using subgroup data, if available, can avoid the issue of
550 unidentified non-homogenous subgroups.

Time-to-Event Prevalence Estimation



551

552 **Fig 5.** Example where observed diagnosis rates represent two unidentified subgroups with
553 different values of diagnostic pressure. The red and green lines represent rates of diagnosis of
554 the two subgroups. The solid black line shows the aggregate diagnosis rates. The dotted line
555 shows the exponential fit to the aggregate diagnosis rates. The age of eligibility $AE = 3$ in this
556 example.

557

558 Any imbalance of case prevalence between in-migration and out-migration to and from
559 the region defining the population over the study period would violate the assumption of
560 constant prevalence within each cohort.

561

562 If some in-migrating cases were diagnosed before in-migration and their subsequent re-
563 diagnoses in the study region were labeled as initial diagnoses, that would violate the

Time-to-Event Prevalence Estimation

564 assumption of truly first diagnoses. Such an effect would be most evident at greater
565 ages after the diagnosis of most cases. Bounding the maximum age studied M to a
566 modest value, sufficient to capture most initial diagnoses, can minimize any resulting
567 bias.

568
569 Apart from subgroups with non-homogeneous h , it is theoretically possible for h to have
570 different effective values for cases of different ages with the same symptom severity at
571 the same DY . Such an effect would represent an age bias in diagnostic pressure,
572 independent of symptom severity. If there is a reason to suspect such an age bias,
573 investigators can add an age parameter to h in the model and estimate its parameters.

574
575 For datasets where diagnosis follows best practices using gold-standard criteria, the
576 lack of false positives may be a fair assumption. It would be difficult to discover any
577 false positives in that case. Where diagnosis uses a less precise process, some false
578 positives might occur. For example, diagnosticians might tend to produce a positive
579 diagnosis of individuals who do not meet formal diagnostic criteria, perhaps under
580 pressure from the patient or parents, or to facilitate services for the individual. In
581 scenarios where the rate of false positives is significant, the age distribution relative to
582 the rates of true diagnoses may be important. If false positives are uniformly distributed
583 over the age range studied, they would cause a constant additive offset to the rates of
584 diagnoses. True case diagnoses should be more common at younger ages, and less
585 common as the risk pool is depleted, so false positives may be relatively more evident

586 at older ages. If false positives are more common at older ages, their effect may be
587 even more obvious.

588 **Model fit**

589 To ensure robust conclusions, investigators should test the model fit to ascertain both
590 model correctness and parameter estimation accuracy. The model fits well if summary
591 measures of the error are small and individual point errors are unsystematic and small
592 [33]. One can examine the fit both graphically and numerically. Plots of $D_{BY,A}$ vs. $\widehat{D}_{BY,A}$ at
593 all ages for individual cohorts can illuminate any issues with fit, which might occur at
594 only some cohorts or ages. Visualization of the model vs. data can expose aspects of
595 the data that might not fit well in a model with few parameters, possibly suggesting a
596 higher-order model or semi-parametric specifications.

597
598 If the model uses an assumed age of complete eligibility AE^* that differs from the true
599 value of AE represented by the data, model fit may be impaired, particularly if $AE^* < AE$.
600 As the Simulation study section shows, setting $AE^* < AE$ can result in estimation errors.
601 Setting $AE^* > AE$ tends not to impair model fit and may improve it in the case of non-
602 homogeneous subgroups; see Fig 5. The presence of non-homogeneous subgroups
603 may be evident from examining model fit.

604
605 The chi-square test statistic, applied to the overall model, individual cohorts, and single
606 ages across cohorts, is a numerical approach to assess absolute model fit. The p-value

607 associated with the chi-square statistic utilizes observed and expected count values, not
608 proportions. The p-value incorporates the effect of the number of parameters in the
609 model via the degrees of freedom.

610 **Simulation study**

611 We tested the TTEPE method via a simulation study where we know the ground truth of
612 all parameters, following the recommendations in Morris [34]. There are six pairs of
613 values of β_h and β_p , each ranging from 0 to 0.1 in steps of 0.02, and each pair sums to
614 0.1. In one parameter set, the prevalence increases as $e^{0.1 \times BY}$ and diagnostic pressure
615 is constant; in another parameter set, the prevalence is constant and diagnostic
616 pressure increases as $e^{0.1 \times DY}$; and the other four parameter sets represent various rates
617 of change of both variables. In all cases, $P = 0.01$ at the final BY , $h = 0.25$ at the final
618 DY , $AE = 3$, $M = 10$, and there are 20 successive cohorts. These simulations assume the
619 investigators know the correct value $AE^* = 3$ from either knowledge of the disorder or
620 estimation of E_A . The study synthesized each data model as real-valued proportions
621 without sampling and with binomial random sample generation of incident diagnoses.
622 For sampling, the population of each cohort is a constant 500,000. Monte Carlo
623 simulation of parameter estimation bias and model standard error (SE) used 1000
624 iterations of random data generation for each set of parameters. Parameter estimation
625 is as described above, implemented using the Python SciPy `curve_fit()` function.
626

Time-to-Event Prevalence Estimation

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

632

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

True Parameters		Bias \widehat{P}	Bias $\widehat{\beta}_p$	Bias \widehat{h}	Bias $\widehat{\beta}_h$
		at final BY		at final DY	
β_p	β_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

635 $P = 0.01$ at the final BY , $h = 0.25$ at the final DY , $AE^* = AE = 3$, $M = 10$, and there are
 636 20 successive cohorts.

637 β_p, β_h are coefficients for prevalence and hazard, respectively.

638 P , prevalence; BY , birth year; DY , diagnostic year.

639

Time-to-Event Prevalence Estimation

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

645

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

True		\widehat{P} at final <i>BY</i>		$\widehat{\beta}_P$		\widehat{h} at final <i>DY</i>		$\widehat{\beta}_h$	
Parameters									
β_P	β_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

648 Population of each cohort = 500,000. $P=0.01$ at the final *BY*, $h=0.25$ at the final

649 *DY*, $AE^*=AE=3$, $M=10$, and there are 20 successive cohorts.

650 β_P, β_h are coefficients for prevalence and hazard, respectively.

651 P , prevalence; *BY*, birth year; *DY*, diagnostic year.

652

Time-to-Event Prevalence Estimation

653 Table 3 gives results where estimation uses an assumed value AE^* that, in some cases,
654 does not match the true value of $AE = 3$ represented by the data. Synthesis uses one
655 homogenous group with consistent h at each value of DY . Estimation using $AE^* = 2$
656 results in substantial estimation errors and model misfit that is obvious from plots of
657 data vs. model (not shown). Estimation using $AE^* = 3$, $AE^* = 4$, or $AE^* = 5$ produces
658 accurate results, with slightly more error where $AE^* = 5$. Plots show that the model fits
659 well in all three cases (not shown). The choice of AE^* is not critical as long as $AE^* \geq AE$.
660 These data use real-valued proportions to avoid confusing model mismatch with
661 sampling effects.
662

Time-to-Event Prevalence Estimation

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

AE^* used in estimation	Bias \hat{P} at final BY	Bias $\hat{\beta}_P$	Bias \hat{h} at final DY	Bias $\hat{\beta}_h$
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

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

669
 670 Table 4 shows results with an intentional mismatch between estimation assuming one
 671 homogeneous group and data representing two subgroups with different values of h ,
 672 illustrated in Fig 5. Note the visible error of the exponential fit to the data at age = 3 and
 673 a good fit for age > 3. In this synthetic dataset, the two subgroups are of equal size, and
 674 the true value of h in one group is twice that of the other. This information is not known
 675 to the estimation, and the data do not indicate subgroup size nor membership. In the
 676 worst case, estimation uses $AE^* = AE = 3$, and the $\hat{\beta}_P$ bias is 0.001, which is 1% of the
 677 actual value of 0.1. This error is due to the subgroups having different hazards, which

Time-to-Event Prevalence Estimation

678 are not accounted for in the estimation. When using $AE^* = 4$ or $AE^* = 5$, the $\widehat{\beta}_P$ bias
679 becomes 6×10^{-4} or less, and model fit is improved (not shown).

680

681 **Table 4. Comparison of the effect of the choice of assumed AE^* vs. true value of**
682 **$AE = 3$, with two unidentified subgroups with different hazards, mismatched to**
683 **analysis.**

AE^* used in estimation	Bias \widehat{P} at final BY	Bias $\widehat{\beta}_P$	Bias \widehat{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

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

689 Discussion

690 Readers may suspect that the estimates are unidentified, i.e., not unique, due to
691 possible interaction between age, diagnostic year, and birth year, such that estimates
692 may be biased even if the model fit is excellent. While that concern is appropriate for
693 analytical methods that assume an age distribution, ignore it, or estimate it

Time-to-Event Prevalence Estimation

694 inappropriately, including age-period-cohort methods, TTEPE avoids that problem and
695 produces uniquely identified estimates. TTEPE models the age distribution of initial
696 diagnoses as a non-linear function of birth year, diagnostic year, and other variables.

697

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

700

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

706

707 TTEPE provides accurate estimates of prevalence parameters with a strong power to
708 detect small differences. The Monte Carlo simulation study in Table 2 shows a
709 magnitude of bias of the prevalence coefficient $\widehat{\beta}_p$ not exceeding 6.5×10^{-5} or 0.0065%
710 per year. The model SE of $\widehat{\beta}_p$ ranges from 0.0012 to 0.0017, where the true β_p ranges
711 from 0 to 0.1. Using the largest observed SE and $1.96 \times SE$ as a threshold for 95%
712 confidence intervals, the method can detect differences in β_p of 0.0033, i.e., 0.33% per
713 year. Investigators can expect similar performance for real-world datasets that meet the
714 baseline assumptions and have characteristics comparable to the simulated data. The
715 population size and prevalence affect the SE. Note that in the simulation study, there

Time-to-Event Prevalence Estimation

716 are 20 cohorts and 11 ages (0 through 10), so there are 220 data points. Each data
717 point is an independent binomial random sample. The analysis estimates four
718 parameters that define the curves that fit the data. The large number of independent
719 data points and the small number of model parameters help to produce the small bias
720 and model SE. If the population of each cohort or the prevalence was substantially
721 smaller, or if the number of parameters was greater, we would expect the bias and SE
722 to be larger. These could occur with small geographic regions, very rare disorders, or
723 higher-order or semi-parametric models, respectively.

724
725 TTEPE is useful for answering some important questions, such as the actual trend in
726 case prevalence over multiple birth cohorts of disorders such as autism
727 and intellectual disability, as described in Elsabbagh [17] and McKenzie [26]
728 respectively. Accurate trend estimates can inform investigation into etiology. Where
729 datasets include appropriate covariates, stratified analysis can estimate the
730 relationships between various population characteristics and trends in true case
731 prevalence and diagnostic factors. Example covariates include sex, race, ethnicity,
732 socio-economic status, geographic region, parental education, environmental exposure,
733 genetic profile, and other potential factors of interest.

734
735 It may be feasible to extend TTEPE to disorders where the time scale starts at some
736 event other than birth. For example, the time origin might be the time of completion of a
737 sufficient cause, and various outcomes may serve as events of interest. It is important

Time-to-Event Prevalence Estimation

738 to ensure that the eligibility function with respect to the time origin is consistent across
739 cohorts.

740

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

745 **Acknowledgments**

746 The author thanks Dr. Lu Tian for his expert advice on survival analysis methods; Dr.
747 Lorene Nelson and Dr. Kristin Sainani for guidance on my thesis which was the genesis
748 of this project and for comments on this paper; Dr. Michael Sigman and Dr. Larry Tang
749 for their thoughtful reviews of the paper; and the Stanford Biomedical Data Science
750 team for their project reviews and insightful comments.

751 **References**

- 752 1. Szklo M, Nieto FJ. Epidemiology Beyond the Basics. 1st ed. Burlington (MA): Jones
753 & Bartlett; 2014.
- 754 2. Rothman KJ, Greenland S, Lash TL. Modern Epidemiology. 3rd ed. Philadelphia:
755 Wolters Kluwer; 2008.

Time-to-Event Prevalence Estimation

- 756 3. GBD 2017 Disease and Injury Incidence and Prevalence Collaborators. Global,
757 regional, and national incidence, prevalence, and years lived with disability for 354
758 diseases and injuries for 195 countries and territories, 1990–2017: a systematic
759 analysis for the Global Burden of Disease Study 2017. *Lancet*. 2018;392:1789–858.
760 Suppl 1.
- 761 4. Baxter AJ, Brugha TS, Erskine HE, Scheurer RW, Vos T, Scott JG. The
762 epidemiology and global burden of autism spectrum disorders. *Psychol Med*.
763 2015;45(3):601–613.
- 764 5. Centers for Disease Control. Prevalence of Autism Spectrum Disorders — Autism
765 and Developmental Disabilities Monitoring Network, Six Sites, United States, 2000.
766 *MMWR Surveillance Summaries*. 2007; 56(SS01);1-11. Available from:
767 <https://www.cdc.gov/mmwr/index.html>.
- 768 6. Centers for Disease Control. Prevalence of Autism Spectrum Disorders — Autism
769 and Developmental Disabilities Monitoring Network, Six Sites, United States, 2002.
770 *MMWR Surveillance Summaries*. 2007; 56(SS01);12-28.
- 771 7. Centers for Disease Control. Brief Update: Prevalence of Autism Spectrum
772 Disorders — Autism and Developmental Disabilities Monitoring Network, United
773 States, 2004. *MMWR Surveillance Summaries*. 2009; 58(SS-10);21-24.
- 774 8. Centers for Disease Control. Prevalence of Autism Spectrum Disorders — Autism
775 and Developmental Disabilities Monitoring Network, United States, 2006. *MMWR*
776 *Surveillance Summaries*. 2009; 58(SS-10);1-20.

Time-to-Event Prevalence Estimation

- 777 9. Centers for Disease Control. Prevalence of Autism Spectrum Disorders — Autism
778 and Developmental Disabilities Monitoring Network, 14 Sites, United States, 2008.
779 MMWR Surveillance Summaries. 2012;61(SS-3):1-19.
- 780 10. Centers for Disease Control. Prevalence of Autism Spectrum Disorders — Autism
781 and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2010.
782 MMWR Surveillance Summaries. 2014;63(SS-2):1-21.
- 783 11. Centers for Disease Control. Prevalence of Autism Spectrum Disorders — Autism
784 and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2012.
785 MMWR Surveillance Summaries. 2018;65(13):1-23.
- 786 12. Centers for Disease Control. Prevalence of Autism Spectrum Disorders — Autism
787 and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2014.
788 MMWR Surveillance Summaries. 2018;67(6):1-23.
- 789 13. Centers for Disease Control. Prevalence of Autism Spectrum Disorders — Autism
790 and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2016.
791 MMWR Surveillance Summaries. 2020;69(4):1-12.
- 792 14. Croen LA, Grether JK, Hoogstrate J, Selvin S. The Changing Prevalence of Autism
793 in California. J Autism Dev Disord. 2002;32(3):207-215.
- 794 15. Hansen SN, Overgaard M, Andersen PK, Parner ET. Estimating a population
795 cumulative incidence under calendar time trends. BMC Med Res Methodol.
796 2017;17:7.

Time-to-Event Prevalence Estimation

- 797 16. Nevison C, Blaxill M, Zahorodny W. California Autism Prevalence Trends from 1931
798 to 2014 and Comparison to National ASD data from IDEA and ADDM. *J Autism Dev*
799 *Disord.* 2018; (doi.org/10.1007/s10803-018-3670-2). Suppl S1.
- 800 17. Elsabbagh M, Divan G, Koh Y-J, Kim YS, Kauchali S, Marcin C, et al. Global
801 prevalence of autism and other pervasive developmental disorders. *Autism*
802 *Research.* 2012;5:160–179.
- 803 18. Campbell CA, Davarya S, Elsabbagh M, Madden L, Fombonne E. Prevalence and
804 the Controversy. In: Matson JL, Sturmey P, editors. *International Handbook of*
805 *Autism and Pervasive Developmental Disorders.* New York: Springer; 2011 pp. 25-
806 35.
- 807 19. Schisterman EF, Cole SR, Platt RW. Overadjustment Bias and Unnecessary
808 Adjustment in Epidemiologic Studies. *Epidemiology.* 2009;20(4):488-495.
- 809 20. Keyes KM, Susser E, Cheslack-Postava K, Fountain C, Liu K, Bearman PS. Cohort
810 effects explain the increase in autism diagnosis among children born from 1992 to
811 2003 in California. *Int J Epidemiol.* 2012;41(2):495-503
- 812 21. Spiers N. Cohort effects explain the increase in autism diagnosis among children
813 born from 1992 to 2003 in California [letter]. *Int J Epidemiol.* 2013;42:1520–1521.
- 814 22. King M, Bearman P. Diagnostic change and the increased prevalence of autism. *Int*
815 *J Epidemiol.* 2009; 38:1224–1234.
- 816 23. Rodgers WL. Estimable Functions of Age, Period, and Cohort Effects. *Am Sociol*
817 *Rev.* 1982;47(6):774-787.

Time-to-Event Prevalence Estimation

- 818 24. O'Brien RM. Age-Period-Cohort Models. Boca Raton (FL): CRC Press; 2015.
- 819 25. MacInnis AG. Autism Prevalence Trends by Birth Year and Diagnostic Year:
820 Indicators of Etiologic and Non-Etiologic Factors – an Age Period Cohort Problem
821 [thesis]. Stanford (CA): Stanford University; 2017 DOI:
822 10.13140/RG.2.2.11821.59360. Available from:
823 [https://www.researchgate.net/publication/322724736_Thesis_Autism_Prevalence_Tr](https://www.researchgate.net/publication/322724736_Thesis_Autism_Prevalence_Trends_by_Birth_Year_and_Diagnostic_Year_Indicators_of_Etiologic_and_Non-Etiologic_Factors_-_an_Age_Period_Cohort_Problem)
824 [ends_by_Birth_Year_and_Diagnostic_Year_Indicators_of_Etiologic_and_Non-](https://www.researchgate.net/publication/322724736_Thesis_Autism_Prevalence_Trends_by_Birth_Year_and_Diagnostic_Year_Indicators_of_Etiologic_and_Non-Etiologic_Factors_-_an_Age_Period_Cohort_Problem)
825 [Etiologic_Factors_-_an_Age_Period_Cohort_Problem.](https://www.researchgate.net/publication/322724736_Thesis_Autism_Prevalence_Trends_by_Birth_Year_and_Diagnostic_Year_Indicators_of_Etiologic_and_Non-Etiologic_Factors_-_an_Age_Period_Cohort_Problem)
- 826 26. McKenzie K, Milton M, Smith G, Ouellete-Kuntz H. Systematic Review of the
827 Prevalence and Incidence of Intellectual Disabilities: Current Trends and Issues. Cur
828 Dev Disord Rep. 2016;3:104-115.
- 829 27. Fombonne E. Epidemiology of pervasive developmental disorders. *Pediatr Res*.
830 2009;65(6):591-598.
- 831 28. MacInnis AG. Time-to-event Prevalence Estimation TTEPE [software]. 2020. OSF
832 repository. Available from: <https://doi.org/10.17605/OSF.IO/WPNKU>.
- 833 29. Findley DF. Counterexamples to Parsimony and BIC. *Ann Inst Stat Math*.
834 1991;43(3):505-514.
- 835 30. Cox DR. Regression Models and Life Tables. *J R Stat Soc Series B Stat Methodol*.
836 1972;34(2):187-220.
- 837 31. Kalbfleisch JD, Prentice RL. *The Statistical Analysis of Failure Time Data*. 2nd ed.
838 Hoboken (NJ): Wiley; 2002.

Time-to-Event Prevalence Estimation

- 839 32. Hyman SL, Levy SE, Myers SM, AAP COUNCIL ON CHILDREN WITH
840 DISABILITIES, SECTION ON DEVELOPMENTAL AND BEHAVIORAL
841 PEDIATRICS. Identification, Evaluation, and Management of Children With Autism
842 Spectrum Disorder. *Pediatrics*. 2020;145(1):e20193447. doi: 10.1542/peds.2019-
843 3447.
- 844 33. Hosmer DW, Lemeshow S, Sturdivant RX. *Applied Logistic Regression*. 3rd ed.
845 Hoboken (NJ): Wiley; 2013.
- 846 34. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical
847 methods. *Stat Med*. 2019;38:2074–2102.