Discover Public Health

RESEARCH Open Access

Population-representative inference for primary and secondary outcomes in extended case-cohort designs

Theresa Wimberley[1,2,3*], Oleguer Plana-Ripoll[1,4], Carsten B. Pedersen[1,2,3,5], Marianne G. Pedersen[1,2,3], Julie W. Dreier[1,2], Jakob Christensen[6,7], Henriette Thisted Horsdal[1,2,8], Aske Astrup[1,2], Wesley K. Thompson[9], Dorte Helenius[3,10], John J. McGrath[1,11,12], Naomi R. Wray[13,14], Liselotte Vogdrup Petersen[1,3], Bjarni Vilhjálmsson[1,15,16], Preben B. Mortensen[1,2,3] and Esben Agerbo[1,2,3]

*Correspondence:
Theresa Wimberley
tw.ncrr@au.dk

Full list of author information is available at the end of the article

Abstract

Background The case-cohort design is useful for obtaining population-representative inference with smaller samples, but the performance when dealing with secondary outcomes and extended follow-up remains unclear. We compare the case-cohort design with the full-cohort design across various outcomes, including additional case-groups and extended follow-up, illustrated using the iPSYCH study.

Methods From Danish nationwide registers, we identified the full-population cohort of all individuals born 1981–2008 (n = 1,657,449) and their clinical diagnoses until 2021. The iPSYCH case-cohort sample (n = 141,265) includes specific case-groups with specific psychiatric disorders diagnosed 1994–2015 (n = 93,608) and a random-population subcohort (n = 50,615). We applied inverse probability weights to estimate person-years at risk, age-specific incidence rates, absolute risks, and incidence rate ratios in various scenarios for primary outcomes (affective disorder, bipolar disorder, schizophrenia, autism, and attention-deficit hyperactivity disorder) and secondary outcomes (epilepsy, anxiety, migraine, asthma, diabetes, injury, traumatic brain injury, substance use disorder, and death).

Results Weighted estimates based on the iPSYCH sample aligned with those in the full cohort for both primary and secondary outcomes. For example, weighted absolute risks by age 40 were analogous to full-cohort estimates for both affective disorder (7.9% [7.5–8.3] versus 8.0% [7.9–8.1]) and epilepsy (2.3% [2.1–2.5] versus 2.3% [2.2–2.3]). Similarly, weighted incidence rates and incidence rate ratios mimicked full-cohort estimates.

Conclusion The extended case-cohort design yields valid estimates of age-specific incidence rates, absolute risks, person-years at risk, and incidence rate ratios for primary and secondary outcomes, even with multiple case-groups and extended follow-up.

Keywords Case-cohort design, Inverse probability weighting, Secondary outcomes, Extended follow-up, Register-based research

1 Introduction

A case-cohort study includes all primary outcome cases and a randomly selected subcohort from the target population. The case-cohort design enables unbiased inferences of disease occurrence, when population size or follow-up length make full cohort inference infeasible [1]. The strengths of the randomly selected subcohort have been highlighted in a recent review [2], which also describes methods to upweight individuals in the case-cohort sample to reflect the full cohort and obtain population-representative inference. However, secondary outcomes (outcomes not initially selected as cases) may also be analysed in the case-cohort design. Different inverse probability weights have been suggested for secondary outcomes in the case-cohort design [2–4], but methods are less established as for case–control studies [5–7]. Still, it remains unclear whether extending complex case-cohort designs to investigate secondary outcomes or prolonging the follow-up period provides valid estimates with inverse probability weighting. Therefore, improving the understanding of the case-cohort design for extended applications is essential to guide and improve future epidemiological studies relying on case-cohort designs.

One of the world's largest case-cohort samples is the Lundbeck Foundation Initiative for Integrative Psychiatric Research (iPSYCH) initiated in 2012, encompassing five psychiatric case-groups and a population-based subcohort. The iPSYCH study, that is nested within Danish nationwide registers and in 2015 extended in a second phase to include 141,265 individuals, offers an ideal setting to address these issues [8, 9]. Several outcomes have been studied in the iPSYCH case-cohort sample, including primary outcomes, i.e. disorders in the psychiatric case-groups [10–14], and secondary outcomes such as other psychiatric or non-psychiatric disorders [13], treatment-related outcomes [15, 16], school grades [17], suicidal behaviour [18], and socioeconomic trajectories [19]. However, numerous other outcomes and exposures can be linked via Danish nationwide registers [20].

We compared up-weighted numbers of persons, person-years at risk, age-specific incidence rates, absolute risks, and incidence rate ratios to full-cohort estimates for various extended applications: 1) including all iPSYCH cases vs. outcome of interest or subcohort only; 2) secondary outcomes vs. primary outcomes; and 3) extending follow-up from beyond the original case identification period in 2015 to the end of 2021. To achieve these aims, we applied inverse probability weights based on the two-phase case-cohort sampling procedure, using iPSYCH as an example.

2 Methods

2.1 Data sources

The present study used information from continuously updated nationwide registers: the Danish Civil Registration System [21], the Danish Central Psychiatric Research Register [22], and the Danish National Patient Register [23]. Data were linked through the unique personal identification number assigned to all individuals residing in Denmark [21].

2.2 Study samples

The iPSYCH sample was formed by the union of two case-cohort samples selected from the general population. The first case-cohort sample (previously termed the iPSYCH2012 case-cohort sample) was selected in 2012 from a source population of

singletons born in Denmark May 1, 1981 to December 31, 2008, with known mothers, and who were alive and residing in Denmark at their first birthday. This case-cohort sample included a 2.037% randomly selected subcohort and cases diagnosed by the end of 2012 with one or more of the following disorders: affective disorder (including bipolar disorder), schizophrenia, autism spectrum disorder (autism), and attention-deficit/hyperactivity disorder [8]. The second case-cohort sample was selected in 2015 from an augmented source population—*the full cohort*—born until December 31, 2008. This case-cohort sample included a 1.267% randomly selected subcohort and all cases by the end of 2015 from extended case-groups: affective disorder, schizophrenia spectrum disorder, autism, attention-deficit/hyperactivity disorder, and postpartum psychiatric disorder. The augmented iPSYCH sample included the combined subcohort and all cases, and has previously been referred to as the iPSYCH2015 case-cohort sample [9]. We refer to the combined sample in short as *the iPSYCH sample*.

The four main samples used in this study are: 1) the full cohort; 2) the subcohort only; 3) the subcohort and outcome; and 4) the subcohort and all case-groups simultaneously (Fig. 1).

2.3 Inclusion probabilities and weights

Inclusion probabilities and weights are presented in Table 1. The inclusion probability (0.03278) for subcohort members born 1981–2005 was calculated based on the initial selection probabilities for the two iPSYCH subcohorts (2.037% and 1.267%), both covering the same birth years, whereas the inclusion probability (1.267%) for subcohort members born 2006–2008 equals the initial selection probability of the second phase sample, sampled from the full cohort including birth years 2006–2008. Furthermore, we present an overall average inclusion probability of 3.054% for the subcohort, inspired by a recent study [24].

Inverse probability of sampling weights assigned at start of follow-up were inspired by Kalbfleisch and Lawless [25]: a case was assigned a weight of one and non-case subcohort members were assigned a weight equal to the inverse inclusion probability, i.e. $30.51 = 1/0.03278$ for non-cases born 1981–2005, and $78.93 = 1/0.01267$ for non-cases born 2006–2008. The overall weight based on the average inclusion probability (*average weight*), is $1/0.03054 = 32.74$. A detailed derivation of the inclusion probabilities and weights is given in supplementary notes 1–2.

2.4 Primary and secondary outcomes

Primary outcomes refer to disorders within the case-groups that were fully selected into the case-cohort design by 2015. The primary outcomes of interest were: affective disorder (main example), schizophrenia, schizophrenia spectrum disorder, bipolar disorder, autism spectrum disorder (autism), and attention-deficit hyperactivity disorder. The secondary outcomes of interest were: epilepsy (main example), anxiety, migraine, asthma, type 1 diabetes, injury, traumatic brain injury, substance use disorder, and death. Diagnoses in the full cohort and nested in the entire iPSYCH sample were identified from the hospital registers until December 31, 2015. Updated diagnoses beyond 2015 were identified in the existing iPSYCH sample from registers updated until December 31, 2021. Diagnostic codes and age cutoffs are provided in Supplementary Table 1.

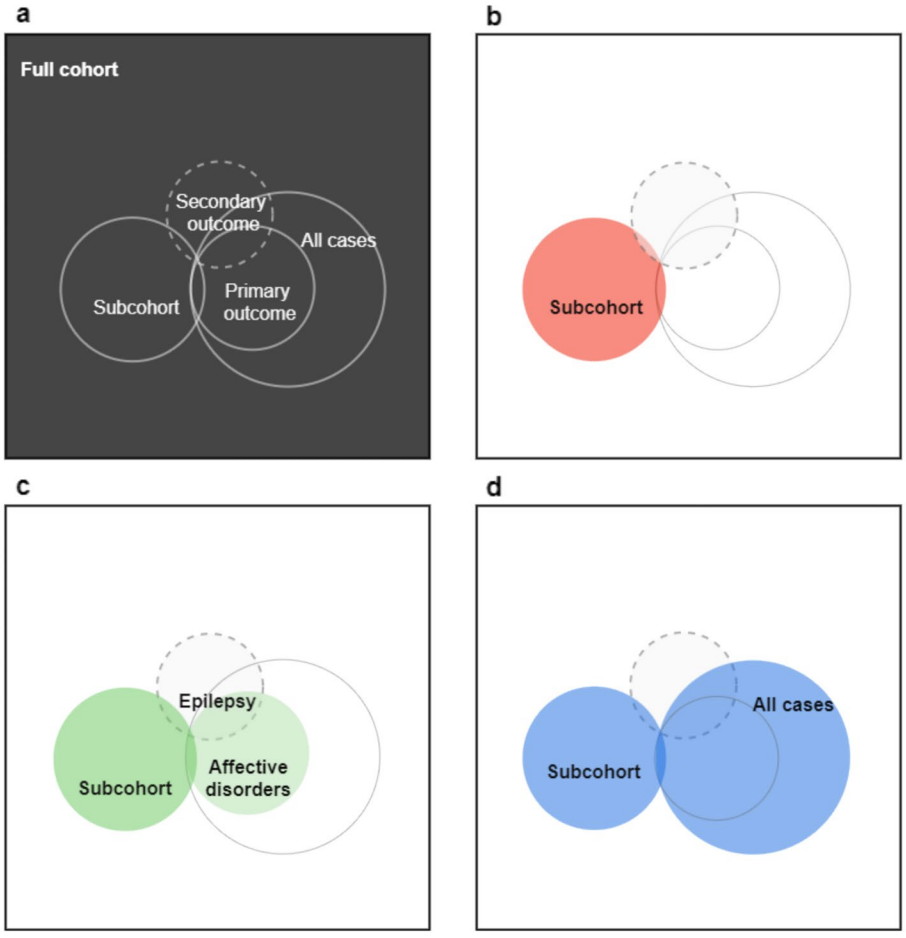

Fig. 1 Venn diagrams visualizing the four study samples. **a** The full cohort (black) refers to the full population of singletons born in Denmark May 1, 1981 to December 31, 2008, with known mothers, and who were alive and residing in Denmark at their first birthday; **b** The subcohort (red) refers to the combined subcohort included in the iPSYCH sample; **c** The subcohort and outcome; the primary outcome of interest (here including affective disorder, green), or the secondary outcome of interest (here epilepsy) nested in the iPSYCH2015 sample; **d** The entire iPSYCH sample including the subcohort and all cases, where all cases refer to individuals belonging to at least one of five psychiatric case groups. All sets are nested within the full cohort at time of the sampling. The combined subcohort and all cases constitutes 3.1% and 5.6% (blue circles), respectively, of the full cohort (black square) totalling 1,657,449 individuals, with an overlap of 2958 cases included in the subcohort

2.5 Follow-up time

Individuals were followed from age one until first emigration from Denmark, death or last follow-up—whichever came first. Last follow-up was considered at two points in calendar time; 1) December 31, 2015 (iPSYCH selection date); or 2) December 31, 2021 (most recent available information), allowing identification of new cases beyond 2015. The various lengths of follow-ups are illustrated in the Lexis diagram in Supplementary Fig. 1.

2.6 Exposures

To evaluate the population-representativeness of the iPSYCH sample when estimating disease association measures, we included autism as a candidate time-dependent exposure (based on one of the primary outcomes) to estimate incidence rate ratios of affective disorder (primary outcome) and epilepsy (secondary outcome), respectively.

Table 1 Counts, inclusion probabilities, and weights, and upweighted counts for subcohorts and case-cohort samples

Cohort	Criteria	N	Inclusion probability[a]	Weight[b]	Upweighted N
Full cohort	Born in DK 1981–2008, known mother, residing in DK at age 1	N_2 = 1,657,449			
Subcohort					
First phase selection	Random sample of iPSYCH2012 source population[a], born 1981–2005	n_1 = 30,000	$\frac{n_1}{N_1} = 0.02037$	49.09	1,472,762
Second phase selection	Random sample of extended source population (full cohort), born 1981–2008	n_2 = 21,000	$\frac{n_2}{N_2} = 0.01267$	78.93	1,657,449
Born 1981–2005	In combined subcohort, born 1981–2005	48,227	$1 - \left(1 - \frac{n_1}{N_1}\right)\left(1 - \frac{n_2}{N_2}\right) = 0.03278$	30.51	1,471,148.3
Born 2006–2008	In combined subcohort, born 2006–2008	2388	$\frac{n_2}{N_2} = 0.01267$	78.93	188,475.6
All	All in combined subcohort[c]	$n_1 + n_2$ -385=50,615	$\frac{\left(1 - \left(1 - \frac{n_1}{N_1}\right)\left(1 - \frac{n_2}{N_2}\right)\right)}{\frac{N_1}{N_2} + \frac{n_2}{N_2}\frac{N_2 - N_1}{N_2}} = 0.03054$	32.74	1,657,288.6[d]
Subcohort and outcome					
Subcohort and primary outcome	Subcohort and diagnosis of affective disorder (n = 40,482) by 2015	89,800	As above[a]	As above[b]	1,660,541.4
Subcohort and secondary outcome	Subcohort and individuals with epilepsy (n = 4,463) by end of 2015 nested within iPSYCH cases and subcohort	54,258	As above[a]	As above[b]	1,659,629.5
Entire iPSYCH sample					
All cases	Included in the full cohort and in at least one of the case-groups[e] by end of 2015	93,608	1	1	93,608
Non-cases in subcohort	Subcohort members not included in any of the case-groups	47,657	As above[a]	As above[a]	1,567,066.9
Subcohort and all cases	All case groups and subcohort	141,265	As above[a]	As above[b]	1,660,674.9

[a]Individual inclusion probabilities depending on case status and birth cohorts; Cases all have an inclusion probability of 1; non-cases have a birth cohort specific inclusion probability (or the weighted average inclusion probability). Here, the inclusion probability for non-cases in birth cohort 1981–2005 depends on the probability of being selected in any of the two phases, subtracting the probability of being selected in both phases; the inclusion probability for non-cases in birth cohort 2006–2008 equals the second phase selection probability. N_1 = 1,472,762 refers to the source population of the first phase selection (iPSYCH2012) nested within the full cohort, born May 1, 1981 to December 31, 2005

[b]Individual weights are calculated as the inverse of the birth cohort specific inclusion probabilities, if not otherwise specified. All cases are all assigned a weight of 1, regardless of birth year and whether they are also included in the subcohort. Note that, individuals with a secondary outcome are only considered a case if they belong to the iPSYCH case-groups

[c]A number of 385 were included by random in both first and second phase selection

[d]The upweighted N of the subcohort is here shown for the average weight, i.e. based on the average inclusion probability. The upweighted N based on the average weight and the entire iPSYCH sample was 1,654,042.7 (in comparison to 1660 674.9 using birth-cohort-specific weights, last row of table)

[e]Affective disorder, autism spectrum disorder, attention-deficit hyperactivity disorder, schizophrenia spectrum disorder, or postpartum psychiatric disorder

2.7 Statistical analyses

First, we compared overall counts, person-years at risk, and incidence rates from the weighted iPSYCH samples to those from the full cohort. Second, Generalised additive modelling (GAM) with Poisson regression [27] was used to estimate incidence rates for primary outcomes and secondary outcomes, using age as the time scale [26]. The maximum numbers of parameters (k) used for the modelling set to 7. Third, we used the Kaplan–Meier estimator to obtain absolute risks by age 18, 25, 30, and 40 (the maximum age achievable by the end of follow-upon December 31, 2021) and plotted smoothed absolute risks using GAM-models with the maximum numbers of parameters set to 10. Last, we estimated incidence rate ratios using a Cox regression model with a primary outcome as a time-varying exposure to assess associations with another primary outcome and secondary outcome, respectively. Estimates were accompanied by 95% confidence, based on robust variance estimation for the weighted case-cohort samples. Inverse probability of sampling weights were implemented in all models for the case-cohort samples, and estimates were compared to full-cohort estimates.

All epidemiological measures were estimated across different scenarios to assess the impact of study population (Fig. 1a–d), outcome (primary outcome or secondary outcome) and length of follow-up (2015 versus 2021). Analyses based on the iPSYCH sample used a robust variance sandwich estimator [3] (Supplementary note 3).

All statistical analyses were performed in R version 4.3.2. Figure 1 was created using the free online diagram software draw.io, and remaining figures were created using ggplot in R version 4.3.2. Programming codes are publicly available for downloading through https://github.com/mtwb24/iPSYCH-design-project.

2.8 Supplementary analyses

First, we estimated age-specific incidence rates and absolute risks for other disorders than affective disorder and epilepsy, i.e. all primary and secondary outcomes of interest. Second, we evaluated the population-representativeness of the weighted incidence rate ratios for other exposures, using traumatic brain injury (secondary outcome) as an example of a time-varying exposure and sex as a time-invariant exposure. Third, we repeated main analyses for absolute risks and incidence rate ratios with the average weight assigned to all non-case subcohort members. Last, we presented unweighted absolute risks and incidence rate ratios to demonstrate the importance of using inverse probability weights.

3 Results

3.1 Study populations and follow-up time

The full cohort included 1,657,449 individuals followed from age 1 until death, emigration, or end of 2015 (or end of 2021). Numbers of primary and secondary outcomes in the different iPSYCH samples are shown in Table 1 and Supplementary Table 1. In the entire iPSYCH sample (n = 141,265) with follow-up extended to end of 2021, we identified 43,519 cases with affective disorder and 5,039 individuals with epilepsy. The person-years of follow-up, weighted by individual weights in the iPSYCH samples, were nearly identical to the corresponding person-years observed in the full cohort across all scenarios (Table 1 and Supplementary Table 2).

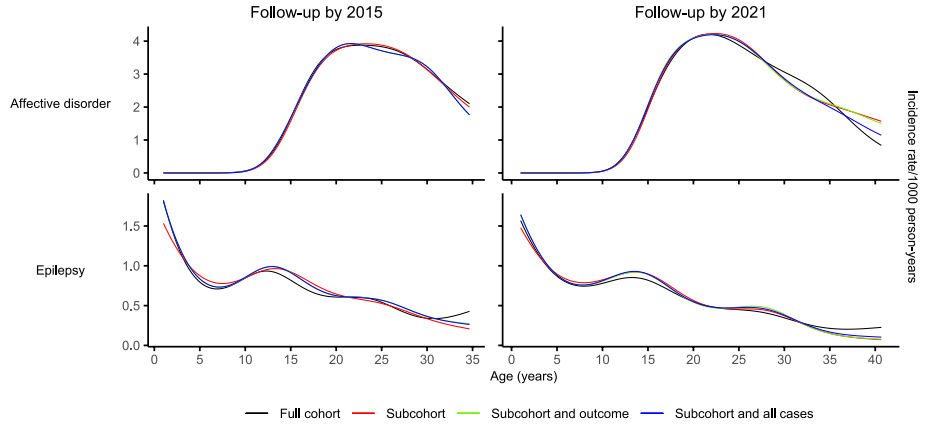

Fig. 2 Incidence rates by age and follow-up period for affective disorder (primary outcome) and epilepsy (secondary outcome) in the full cohort, iPSYCH subcohort and weighted case-cohort samples. Incidence rates were fitted from a poisson generalised additive model with the maximum numbers of parameters (k) used for the modelling set to 7. Follow-up to 2015 (left panel) and 2021 (right panel)

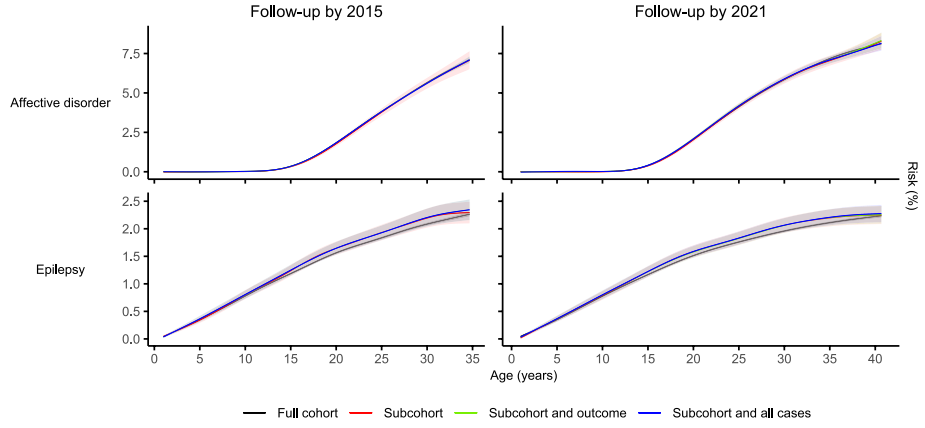

Fig. 3 Absolute risk as a function of age and by follow-up period for affective disorder (primary outcome) and epilepsy (secondary outcome) in the full cohort, iPSYCH subcohort and weighted case-cohort samples. Curves are accompanied by 95% confidence bands, based on robust variance estimation for the case-cohort samples. Curves were smoothed using GAM-models with the maximum numbers of parameters set to 10. Follow-up to 2015 (left panel) and 2021 (right panel)

Incidence rates from the weighted Poisson regression model as a function of age were virtually identical to corresponding estimates in the subcohort and full cohort (Fig. 2). Furthermore, the smoothed curves were also nearly inseparable. Overall, incidence rates for the entire iPSYCH sample were almost identical with full cohort estimates for both primary and secondary outcomes (affective disorder 1.68 per 1000 person-years vs. 1.67 in full cohort; epilepsy, 0.75 vs. 0.72 in full cohort). Incidence rates for all samples and follow-up are shown in Supplementary Table 2. Similarly, the weighted absolute risk curves were almost identical to the analogous full cohort and subcohort curves in all scenarios (Fig. 3). At age 40 years, the risk of affective disorder was estimated to 7.9% (95% CI 7.5–8.3) (vs. full cohort: 8.0% [7.9–8.1]) and risk of epilepsy was estimated to 2.3% (2.1–2.5) (vs. full cohort (2.3% [2.2–2.3]). Absolute risks at ages 18, 25, 30 and 40 for all scenarios can be found in Supplementary Table 3.

Incidence rate ratios estimated in weighted Cox regression models for the association between autism (primary outcome used as a time-dependent exposure) and affective

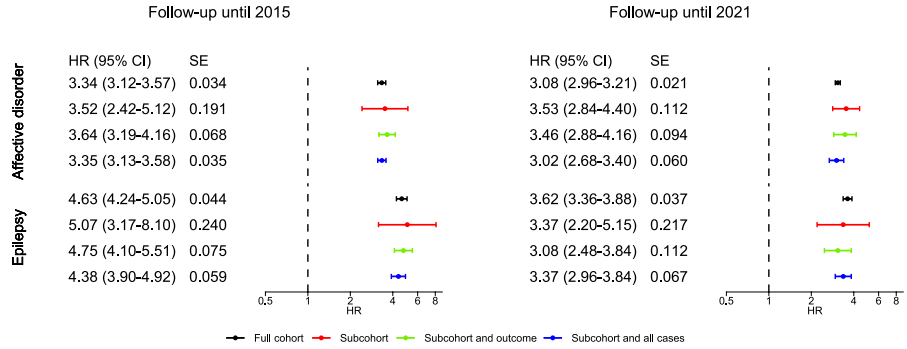

Fig. 4 Incidence rate ratios of affective disorder and epilepsy according to a history of autism in the full cohort, iPSYCH subcohort and case-cohort samples. Hazard ratios with 95% CIs were estimated from a crude Cox regression model with autism spectrum disorder (primary outcome) as a time-dependent exposure and affective disorder and epilepsy as outcomes. Robust standard errors were used in the weighted models. Follow-up to 2015 (left panel) and 2021 (right panel). Abbreviations: CI: Confidence interval, HR: Hazard ratio, SE: standard error

disorder and epilepsy were similar in size to the full cohort and subcohort (Fig. 4). For affective disorder with follow-up until the end of 2021, the incidence rate ratio in the entire iPSYCH sample was 3.02 (95% CI 2.68–3.40) (vs. full cohort: 3.08 (2.96–3.21), and for epilepsy the incidence rate ratio was 3.37 (2.96–3.84) (vs. full cohort: 3.62 (3.36–3.88)). In the iPSYCH samples, standard errors were expectedly largest for the smallest sample, i.e. the subcohort (e.g. the standard error for epilepsy by 2015 was 0.240) and analogous robust standard errors were smallest for the sample (except full cohort) including the subcohort plus all cases (standard error = 0.059). By extending follow-up to the end of 2021, standard errors expectedly slightly decreased for the subcohort (from 0.240 to 0.217), but slightly increased in the weighted samples (from 0.059 to 0.067).

3.2 Supplementary analyses

Repeating the main analyses for other disorders showed that weighted incidence rates and absolute risks matched those of the full cohort for all primary and secondary outcomes (Supplementary Figs. 2–5). Minor deviations from the full cohort were observed in incidence rates for some childhood disorders (autism and attention-deficit hyperactivity disorder) when restricted to the subcohort only or at the extremes of the age spectrum. Incidence rate ratios were comparable across samples for both traumatic brain injury as time-varying exposure and sex as time-fixed exposure, and for all other primary and secondary outcomes (Supplementary Fig. 6–9). The robust standard errors increased after extending follow-up when a primary outcome was studied as the outcome, whereas they decreased when a secondary outcome was studied as the outcome. When average weights were used, all absolute risks and incidence rate ratios closely resembled the analogous measures based on the entire cohort. (Supplementary Figs. 10–11). When the weights were ignored, the absolute risk at age 30 years was between 2 and 12 times larger than the analogous risk based on the full cohort (Supplementary Fig. 12). Similarly, incidence rate ratios were blatantly biased in unweighted analyses due to the underrepresented unexposed person-time at risk (Supplementary Fig. 13).

4 Discussion

We used nationwide registry data to show that inverse probability weights produce valid and population-representative inference in a complex case-cohort sample like iPSYCH. Our study demonstrated that the inverse probability weights performed impeccably in all scenarios and extensions of the case-cohort design by effectively upweighting individuals and person-years to match those of the full cohort. Thus, all epidemiological measures of disease occurrence based on inverse probability weights closely resembled those from the full cohort across all scenarios, including sample ascertainment, exposures, follow-up duration, and primary and secondary outcomes. Unsurprisingly, absolute risks based on *unweighted* data were hugely overestimated, particularly for primary but also for secondary outcomes. Similarly, *unweighted* incidence rate ratios were profoundly biased, underscoring the necessity of using weights to obtain population-representative inference based on case-cohort samples.

Our approach extends the inverse-probability weights method to accommodate extended follow-up and complex sampling. Additionally, we introduce a novel method to generate population denominators for person-years at risk and population size based on highly ascertained case-cohort samples. Despite the relatively smaller size of the second birth cohort (2006–2008) and large weights, where a subcohort member represents nearly 80 individuals in the full cohort, our proposed weights performed flawlessly across all epidemiological measures, even for rare outcomes such as epilepsy. Our analyses demonstrate that inference should include all primary case-groups and the subcohort, even when focusing on a single primary outcome. Most previous iPSYCH studies have excluded other case-groups, resulting in inefficient data use—a point well-established in the statistical literature [28].

Inference for secondary outcomes, including age-specific incidences, absolute risks, and incidence rate ratios, was comparable between weighted samples and the full cohort. Thus, our findings indicate that secondary outcomes can be effectively studied using the sampled, as has also been demonstrated in previous simulation studies [29]. In our study, we found comparable results for all epidemiological measures for both epilepsy and all other secondary outcomes, including very rare outcomes such as death in the iPSYCH cohort, though incidence rates were closer to subcohort estimates than full cohort estimates. Robust variance estimation yielded the highest precision for primary outcomes, which is unsurprising given that the number of cases in the case-cohort sample equals that in the full cohort.

When extending the follow-up, weighted case-cohort analyses closely reflected full cohort estimates for both primary and secondary outcomes. Intriguingly, extended follow-up led to a precision loss of the weighted incidence rate ratios for the primary outcomes. This may be explained by primary outcomes being fully sampled by 2015 compared to secondary outcomes being identified in the selected sample. With the population-based subcohort already selected, there is a great potential for further extensions with additional primary or secondary outcomes.

4.1 Strengths and limitations

A key strength of this study is the ability to link nationwide registers, allowing comparison between case-cohort estimates from the iPSYCH design and full cohort estimates, and continuously updated information for extended follow-up. The iPSYCH sample is

a unique genetic data source and includes over 90,000 cases with specific and fully sampled psychiatric disorders, and a population-representative subcohort; thus eliminating any selection bias [30]. In contrast, the UK biobank is subject to participation bias with more selected with much lower age-adjusted prevalence of health outcomes and lower mortality rates compared to the general UK population [31]. The iPSYCH case-cohort study enables genetic epidemiologic research in relation to various register-based long-term secondary outcomes, as demonstrated in this study.

Our study had several limitations. First, our analyses relied on proposed time-independent weights without comparison to other types of weighting schemes. Our time-independent inverse probability weights were derived in the spirit of Kalbfleisch and Lawless [25]. A weight of 1 was assigned to all cases, and a weight equivalent to the sampling fraction was assigned to non-cases (see supplementary notes 1–2). Our approach aligns with previous suggestions [32], also for the stratified case-cohort design [33], and for the study of secondary outcomes with suggested augmentations to the inverse probability of sampling weighted estimator [29]. Kalbfleisch and Lawless weights have previously been compared to time-dependent weighting schemes (Borgan II weights [34], Barlow [32], and Prentice [1]). Simulation studies found minuscule differences in bias and precision when comparing weighted and unweighted full-cohort Cox regression models [3, 35]. Furthermore, our analyses with time-independent weights closely matched those from the full cohort in various scenarios, indicating minimal benefit from time-dependent weights at the cost of increased analytical complexity. Second, the absolute risk curves are slightly overestimated as we did not account for competing risks due to death [36]. However, very few individuals died before age 40 which is the maximum age in our study. Additionally, no statistical method currently exists for deriving a robust variance estimator for case-cohort studies with competing risks. Third, not all our scenarios were relevant for all birth cohorts, e.g. later birth cohorts (born 2006–2008) will not be at risk of later-onset disorders such as affective disorder; thus the relevant birth cohort should be considered when implementing the birth cohort or average weights, as reflected in our sensitivity analysis using average weights based on all birth cohort years. Nevertheless, in this study we mimicked the iPSYCH design, thereby maintaining consistency across samples and outcomes. Furthermore, the robustness of results across outcomes of different prevalence, age of onset, and diagnostic validity suggests that our results may likely generalize to other secondary outcomes.

Finally, our weights may not be optimal if additional restrictions are applied after the initial case-cohort selection. For example, genetic case-cohort data such as iPSYCH are often restricted due to genotyping errors [37, 38] and confounding from population stratification. The latter restriction may, however, be less common in the future owing to the growing aspiration to avoid ancestry-based restrictions [39–41].

5 Conclusion

Our study provides a proof-of-concept that inverse probability weighting performs well across various extensions of the iPSYCH case-cohort design, including secondary phenotypes. Weighted estimates of disease occurrence and association matched those of the full cohort, even for secondary outcomes, samples with multiple case-groups, and extended follow-up. In contrast, unweighted estimates were severely biased. Our study

demonstrates that inverse probability weighting is a viable method for generalizing epidemiological findings from extended case-cohort studies.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12982-025-00888-w.

Additional file 1.

Author contributions

EA, TW, and OPR conceptualized and designed the study. EA obtained access to the data. TW managed the data and conducted the statistical analyses. EA and TW coordinated the project and drafted all versions of the manuscript. TW, OPR, CBP, MGP, JWD, JC, HTH, AA, WKT, DH, JJM, NRW, LVP, BV, PBM, and EA contributed to the interpretation of the results, critically revised the manuscript, approved the final version for publication, and agreed to be accountable for all aspects of the work. PBM secured funding for the study.

Funding

This work was supported by the Danish National Research Foundation, via a Niels Bohr Professorship to John McGrath, the Novo Nordisk Foundation (NNF16OC0019126 and NNF22OC0075033) and the Lundbeck Foundation (R400-2022–1205). LVP were supported by the Novo Nordic Foundation (grant no NNF23OC0085941). The iPSYCH initiative is funded by the Lundbeck Foundation (grant nos. R102-A9118 and R155-2014–1724).

Data availability

This study uses Danish register data from the Social Science Health Genetics project (PI: E. Agerbo). Data access was obtained through secured servers on Statistics Denmark and approved by the Danish Data Protection Agency, the Danish Health Data Authority, and Statistics Denmark. Individual-level data cannot be shared due to national guidelines, but data access can be granted upon application to the above-mentioned authorities.

Declarations

Ethics approval and consent to participate

This is an observational study. All data are de-identified and not recognizable at an individual level. By Danish law, analysis of such register data does not require ethical review board approval. This research was conducted within the approvals of register-based research granted by authorities, including the Danish Health Data Authority, the Danish Data Protection Agency, and Statistics Denmark. As these data are used without explicit consent, confidentiality of this wealth of data is protected via Danish data protection procedures that are layered on top of the European Union's (EU) GDPR rules and regulations. Nonetheless, we have utilized several approaches to access to the Danish data while ensuring compliance with Danish ethical permissions, Danish law, and EU GDPR.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Author details

[1]The National Centre for Register-Based Research, Department of Public Health, School of Business and Social Sciences, Aarhus University, Fuglesangs Allé 26, 8210 Aarhus, Denmark
[2]Centre for Integrated Register-Based Research (CIRRAU), Aarhus University, Aarhus, Denmark
[3]The Lundbeck Foundation Initiative for Integrative Psychiatric Research (iPSYCH), Aarhus, Denmark
[4]Department of Clinical Epidemiology, Aarhus University and Aarhus University Hospital, Aarhus, Denmark
[5]Hammel Neurorehabilitation Centre and University Research Clinic, Aarhus University, Hammel, Denmark
[6]Department of Neurology, Affiliated Member of the European Reference Network EpiCARE, Aarhus University Hospital, Aarhus, Denmark
[7]Department of Clinical Medicine, Aarhus University, Aarhus, Denmark
[8]BERTHA-Big Data Centre for Environment and Health, Aarhus University, Aarhus, Denmark
[9]Center for Population Neuroscience and Genetics, Laureate Institute for Brain Research, Tulsa, OK, USA
[10]Institute of Biological Psychiatry, Mental Health Services, Copenhagen University Hospital, Roskilde, Denmark
[11]Queensland Centre for Mental Health Research, The Park Centre for Mental Health, Wacol, QLD 4072, Australia
[12]Queensland Brain Institute, The University of Queensland, St Lucia, QLD 4076, Australia
[13]Department of Psychiatry, University of Oxford, Oxford, UK
[14]Institute for Molecular Bioscience, The University of Queensland, St Lucia, Australia
[15]Department of Molecular Biology and Genetics, Bioinformatics Research Centre, Aarhus University, Aarhus, Denmark
[16]Novo Nordisk Foundation Center for Genomic Mechanisms of Diseases, The Broad Institute of MIT and Harvard, Cambridge, MA, USA

Received: 21 February 2025 / Accepted: 13 August 2025
Published online: 08 October 2025

References

1. Prentice RL. A case-cohort design for epidemiologic cohort studies and disease prevention trials. Biometrika. 1986;73(1):1–11.
2. O'Brien KM, Lawrence KG, Keil AP. The case for case-cohort: an applied epidemiologist's guide to reframing case-cohort studies to improve usability and flexibility. Epidemiology. 2022;33(3):354–61. https://doi.org/10.1097/EDE.0000000000001469.
3. Kulathinal S, Karvanen J, Saarela O, Kuulasmaa K. Case-cohort design in practice - experiences from the MORGAM project. Epidemiol Perspect Innov. 2007;4:15. https://doi.org/10.1186/1742-5573-4-15.
4. Saarela O, Kulathinal S, Karvanen J. Secondary analysis under cohort sampling designs using conditional likelihood. J Prob Stat. 2012;2012: 931416.
5. Sofer T, Cornelis MC, Kraft P, Tchetgen Tchetgen EJ. Control function assisted IPW estimation with a secondary outcome in case-control studies. Stat Sin. 2017;27(2):785–804.
6. Tchetgen Tchetgen EJ. A general regression framework for a secondary outcome in case–control studies. Biostatistics. 2014;15(1):117–28.
7. Xing C, McCarthy JM, Dupuis J, et al. Robust analysis of secondary phenotypes in case-control genetic association studies. Stat Med. 2016;35(23):4226–37. https://doi.org/10.1002/sim.6976.
8. Pedersen CB, Bybjerg-Grauholm J, Pedersen MG, et al. The iPSYCH2012 case-cohort sample: new directions for unravelling genetic and environmental architectures of severe mental disorders. Mol Psychiatry. 2018;23(1):6–14. https://doi.org/10.1038/mp.2017.196.
9. Bybjerg-Grauholm J, Pedersen CB, Bækvad-Hansen M, Pedersen MG. The iPSYCH2015 Case-Cohort sample: updated directions for unravelling genetic and environmental architectures of severe mental disorders. medRxiv preprint https://doi.org/10.1101/2020.11.30.20237768. 2020.
10. Agerbo E, Sullivan PF, Vilhjalmsson BJ, et al. Polygenic risk score, parental socioeconomic status, family history of psychiatric disorders, and the risk for schizophrenia: a Danish population-based study and meta-analysis. JAMA Psychiat. 2015;72(7):635–41. https://doi.org/10.1001/jamapsychiatry.2015.0346.
11. Agerbo E, Trabjerg BB, Borglum AD, et al. Risk of early-onset depression associated with polygenic liability, parental psychiatric history, and socioeconomic status. JAMA Psychiat. 2021;78(4):387–97. https://doi.org/10.1001/jamapsychiatry.2020.4172.
12. Schendel D, Munk Laursen T, Albinana C, et al. Evaluating the interrelations between the autism polygenic score and psychiatric family history in risk for autism. Autism Res. 2022;15(1):171–82. https://doi.org/10.1002/aur.2629.
13. Wimberley T, Brikell I, Astrup A, et al. Shared familial risk for type 2 diabetes mellitus and psychiatric disorders: a nationwide multigenerational genetics study. Psychol Med. 2024. https://doi.org/10.1017/S0033291724001053.
14. Munk-Olsen T, Di Florio A, Madsen KB, et al. Postpartum and non-postpartum depression: a population-based matched case-control study comparing polygenic risk scores for severe mental disorders. Transl Psychiatry. 2023;13(1):346. https://doi.org/10.1038/s41398-023-02649-2.
15. Brikell I, Wimberley T, Albinana C, et al. Genetic, clinical, and sociodemographic factors associated with stimulant treatment outcomes in ADHD. Am J Psychiatry. 2021;178(9):854–64. https://doi.org/10.1176/appi.ajp.2020.20121686.
16. Liu X, Trinh NT, Wray NR, et al. Impact of genetic, sociodemographic, and clinical features on antidepressant treatment trajectories in the perinatal period. Eur Neuropsychopharmacol. 2024;81:20–7. https://doi.org/10.1016/j.euroneuro.2024.01.010.
17. Jefsen OH, Holde K, McGrath JJ, et al. Polygenic risk of mental disorders and subject-specific school grades. Biol Psychiatry. 2024;96(3):222–9. https://doi.org/10.1016/j.biopsych.2023.11.020.
18. Erlangsen A, Appadurai V, Wang Y, et al. Genetics of suicide attempts in individuals with and without mental disorders: a population-based genome-wide association study. Mol Psychiatry. 2020;25(10):2410–21. https://doi.org/10.1038/s41380-018-0218-y.
19. Ronda V, Agerbo E, Bleses D, et al. Family disadvantage, gender, and the returns to genetic human capital. Scand J Econ. 2022;124(2):550–78.
20. Erlangsen A, Fedyszyn I. Danish nationwide registers for public health and health-related research. Scand J Public Health. 2015;43(4):333–9. https://doi.org/10.1177/1403494815575193.
21. Pedersen CB. The Danish Civil Registration System. Scand J Public Health. 2011;39(7 Suppl):22–5.
22. Mors O, Perto GP, Mortensen PB. The Danish psychiatric central research register. Scand J Public Health. 2011;39(7 Suppl):54–7.
23. Schmidt M, Schmidt SA, Sandegaard JL, Ehrenstein V, Pedersen L, Sorensen HT. The Danish National Patient Registry: a review of content, data quality, and research potential. Clin Epidemiol. 2015;7:449–90. https://doi.org/10.2147/CLEP.S91125.
24. Parner ET, Andersen PK, Overgaard M. Cumulative risk regression in case-cohort studies using pseudo-observations. Lifetime Data Anal. 2020;26(4):639–58. https://doi.org/10.1007/s10985-020-09492-3.
25. Kalbfleisch JD, Lawless JF. Likelihood analysis of multi-state models for disease incidence and mortality. Stat Med. 1988;7(1–2):149–60. https://doi.org/10.1002/sim.4780070116.
26. Frome EL. The analysis of rates using poisson regression-models. Biometrics. 1983;39(3):665–74. https://doi.org/10.2307/2531094.
27. Carstensen B. Epidemiology with R. Oxford: Oxford University Press; 2021.
28. Borgan Ø, Goldstein L, Langholz B. Methods for the analysis of sampled cohort data in the Cox proportional hazards model. Ann Stat. 1995;23(5):1749–78.
29. Pan Y, Cai J, Kim S, Zhou H. Regression analysis for secondary response variable in a case-cohort study. Biometrics. 2018;74(3):1014–22. https://doi.org/10.1111/biom.12838.
30. Schoeler T, Speed D, Porcu E, Pirastu N, Pingault JB, Kutalik Z. Participation bias in the UK biobank distorts genetic associations and downstream analyses. Nat Hum Behav. 2023;7(7):1216–27. https://doi.org/10.1038/s41562-023-01579-9.
31. Fry A, Littlejohns TJ, Sudlow C, et al. Comparison of sociodemographic and health-related characteristics of UK biobank participants with those of the general population. Am J Epidemiol. 2017;186(9):1026–34. https://doi.org/10.1093/aje/kwx246.
32. Barlow WE. Robust variance estimation for the case-cohort design. Biometrics. 1994;50(4):1064–72.

33. Breslow NE, Lumley T, Ballantyne CM, Chambless LE, Kulich M. Using the whole cohort in the analysis of case-cohort data. Am J Epidemiol. 2009;169(11):1398–405. https://doi.org/10.1093/aje/kwp055.
34. Borgan O, Langholz B. Nonparametric estimation of relative mortality from nested case-control studies. Biometrics. 1993;49(2):593–602.
35. Petersen L, Sorensen TI, Andersen PK. Comparison of case-cohort estimators based on data on premature death of adult adoptees. Stat Med. 2003;22(24):3795–803. https://doi.org/10.1002/sim.1672.
36. Andersen PK, Keiding N. Interpretability and importance of functionals in competing risks and multistate models. Stat Med. 2012;31(11–12):1074–88. https://doi.org/10.1002/sim.4385.
37. Grove J, Ripke S, Als TD, et al. Identification of common genetic risk variants for autism spectrum disorder. Nat Genet. 2019;51(3):431–44. https://doi.org/10.1038/s41588-019-0344-8.
38. Demontis D, Walters RK, Martin J, et al. Discovery of the first genome-wide significant risk loci for attention deficit/hyperactivity disorder. Nat Genet. 2019;51(1):63–75. https://doi.org/10.1038/s41588-018-0269-7.
39. Mester R, Hou K, Ding Y, et al. Impact of cross-ancestry genetic architecture on GWASs in admixed populations. Am J Hum Genet. 2023;110(6):927–39. https://doi.org/10.1016/j.ajhg.2023.05.001.
40. Ding Y, Hou K, Xu Z, et al. Polygenic scoring accuracy varies across the genetic ancestry continuum. Nature. 2023;618(7966):774–81. https://doi.org/10.1038/s41586-023-06079-4.
41. Kachuri L, Chatterjee N, Hirbo J, et al. Principles and methods for transferring polygenic risk scores across global populations. Nat Rev Genet. 2024;25(1):8–25. https://doi.org/10.1038/s41576-023-00637-2.

Publisher's Note