The Estimation of Gestational Age at Birth in Database Studies

Short Title: Estimating Gestational Age

Maria Eberg MSc[1], Robert W. Platt PhD[2,3], Kristian B. Filion PhD[1,2,4]

[1] Center for Clinical Epidemiology, Lady Davis Institute, Jewish General Hospital, Montreal, QC

[2] Department of Epidemiology, Biostatistics, and Occupational Health, McGill University, Montreal, QC

[3] Department of Pediatrics, McGill University, Montreal, QC

[4] Department of Medicine, McGill University, Montreal, QC

Word Counts:
Abstract: 248
Main text: 3,997
Total word count: 6,809

Address for Correspondence:

Kristian B. Filion, PhD FAHA
Assistant Professor
Departments of Medicine and of Epidemiology, Biostatistics, and Occupational Health
Jewish General Hospital/McGill University
3755 Cote Ste-Catherine Road, Suite H416.1
Montreal, Quebec, Canada

Telephone: (514) 340-8222 Ext. 8394
Fax: (514) 340-7564
Email: kristian.filion@mcgill.ca

ACKNOWLEDGEMENTS

This study was supported by an operating grant from the Canadian Institutes of Health Research (CIHR; grant number MOP 126171). Dr. Platt holds a Chercheur-National (National Scholar) Award from the Fonds de Recherche en Santé du Québec (Quebec Foundation for Health Research), and is the Albert Boehringer I Chair in Pharmacoepidemiology at McGill University. Dr. Filion holds a New Investigator award from the CIHR.

DISCLOSURES

Dr. Platt reports personal fees from Pfizer, Novartis, Amgen, and AbbVie unrelated to this work. The other authors have no relationships to disclose.

ABSTRACT

Background: Studies on the safety of prenatal medication use require valid estimation of the pregnancy duration. However, gestational age is often incompletely recorded in administrative and clinical databases. Our objective was to compare different approaches to estimating the pregnancy duration.

Methods: Using data from the Clinical Practice Research Datalink and Hospital Episode Statistics, we examined four approaches to estimating missing gestational age: 1) generalized estimating equations for longitudinal data; 2) multiple imputation; 3) estimation based on fetal birth weight and sex; and 4) conventional approaches that assigned a fixed value (39 weeks for all or 39 weeks for full term and 35 weeks for preterm). The gestational age recorded in Hospital Episode Statistics was considered the gold standard. We conducted a simulation study comparing the described approaches in terms of estimated bias and mean square error.

Results: A total of 25,929 infants from 22,774 mothers were included in our "gold standard" cohort. The smallest average absolute bias was observed for the generalized estimating equation that included birth weight, while the largest absolute bias occurred when assigning 39 weeks gestation to all those with missing values. The smallest mean square errors were detected with generalized estimating equations while multiple imputation had the highest mean square errors.

Conclusions: The use of generalized estimating equations resulted in the most accurate estimation of missing gestational age when birth weight information was available. In the absence of birth weight, assignment of fixed gestational age based on term/preterm status may be the optimal approach.

INTRODUCTION

Studies on the safety of medication use during pregnancy, which often involve large administrative or clinical databases, require valid estimation of the pregnancy duration. However, such databases often have inaccurate or incomplete information regarding gestational age at delivery, which may lead to exposure misclassification. The degree of bias depends on the exposure itself and the direction of exposure misclassification.[1,2] As the beginning of pregnancy cannot be easily identified in such databases, investigators must develop different strategies to specify the pregnancy period.

In a review of methods used to estimate the duration of pregnancy in health care databases, Margulis et al.[3] categorized the most common approaches into 5 groups: 1) assigning a uniform duration of pregnancy; 2) estimation based on preterm delivery codes or other pregnancy codes; 3) methods based on the timing of prenatal care; 4) methods based on birth weight; 5) a combination of methods 2 and 3. All of these approaches share a common factor – they are deterministic approaches to estimate pregnancy duration. While these approaches are valid for the majority of pregnancies, they may be suboptimal as they do not maximize the use of demographic and clinical information that is routinely found in administrative and clinical databases.

The objective of this study was to use probabilistic, model-based approaches, including longitudinal models with generalized estimating equations and multiple imputation, to estimate the duration of pregnancy using maternal demographic and clinical information and to contrast the performance of these probabilistic approaches with common deterministic methods, such as assigning a uniform duration of pregnancy, estimating duration of pregnancy using preterm delivery codes and other pregnancy codes, and methods based on birth weight.[3,4]

METHODS

Data source

We conducted a population-based cohort study of patients with a recorded delivery in the Clinical Practice Research Datalink. This database includes demographic characteristics, clinical diagnoses, and prescriptions issued, as well as clinical information such as lifestyle variables (e.g., smoking status, height, weight, alcohol use), clinical measures (e.g., blood pressure readings), and laboratory test results. The Clinical Practice Research Datalink has been validated extensively.[5,6] For approximately 58% of patients[7], Clinical Practice Research Datalink data can also be linked to other National Health Service data sources, including Hospital Episode Statistics data, which contain full hospitalization records.

This study was approved by the research ethics board of the Jewish General Hospital and by the Independent Scientific Advisory Committee of the Clinical Practice Research Datalink (protocol 13_040ARMn).

Study population

We identified women with a delivery recorded in the Clinical Practice Research Datalink between January 1997 and March 2012. Since there may be multiple records related to the same pregnancy, we assumed that any codes in the 259 days (37 weeks) before the delivery date were related to the same pregnancy for term deliveries. For preterm deliveries, a 24-week window was used as this was considered the limit of viability.[8] Clinical Practice Research Datalink additional files were used to obtain the records of gestational age at delivery, birth weight, and baby sex.

We then identified hospitalizations for delivery in Hospital Episode Statistics (ICD-10 codes O80.X-O84.X, O60.X, Z37.X-Z39.X). Potential duplicate records related to the same

delivery were removed as described above. We then linked each delivery record to Hospital Episode Statistics maternity data to obtain additional information on the fetus and gestational age.

Using the datasets created from the Clinical Practice Research Datalink and Hospital Episode Statistics, we defined a cohort of deliveries with recorded gestational age. From this cohort, we created our "gold standard" cohort, a sub-cohort that was restricted to deliveries in which the Clinical Practice Research Datalink date of delivery occurred during a Hospital Episode Statistics hospitalization and with the same length of gestation (calibration ±1week) in both data sources. Birth weight information was included where available; when discrepancies were present between data sources, Hospital Episode Statistics records were preferred. Inclusion was restricted to patients with ≥645 days of observation time in the Clinical Practice Research Datalink prior to delivery to ensure that all patients had ≥1 year of recorded pre-pregnancy history. We further restricted inclusion to mothers aged 14 to 45 years at the time of delivery. Finally, we excluded records with birth weights that were ≥4 standard deviations from the mean birth weight at gestational age[9] as such values were not considered biologically plausible.

Single imputation of gestational age using generalized estimating equations

We used four approaches to impute gestational age, estimated as completed weeks of gestation at the time of delivery, when it was missing in Hospital Episode Statistics.

In the first, we performed single imputation using generalized estimating equations to impute gestational age using demographic and clinical characteristics as independent variables. The method was introduced by Liang and Zeger and is used to estimate regression model parameters for correlated data.[10] With many women having more than one pregnancy recorded in the Clinical Practice Research Datalink, it is reasonable to assume some correlation between

pregnancies. The main advantage of generalized estimating equations lies in the consistent and unbiased estimation of parameters, even when the correlation structure is misspecified. An autoregressive working correlation structure was assumed.

Using covariates with an increasing level of information, we evaluated the accuracy of three different generalized estimating equations. We were interested in the accuracy of the gestational age imputation and not in the interpretation of effect of each covariate. Therefore, after examining a range of dependent variables to satisfy model assumptions, we used the following transformation of gestational age: $log(max_{gw}+1-gw)$, where max_{gw} denotes the maximum gestational age at delivery observed in the study population and *gw* is the number of completed weeks of gestation at delivery.

As body mass index and smoking information was missing for part of the cohort, our first generalized estimating equation analysis used only the covariates with complete information: singleton vs multiple gestation, preterm birth, stillbirth, parity >0, a history of spontaneous abortion, complications during pregnancy (pre-eclampsia, other hypertensive disorders during pregnancy, gestational diabetes), a history of complications during previous pregnancies, and maternal information (age at delivery, alcohol-related disorders). The model in our second generalized estimating equation analysis included smoking and body mass index as well as all covariates used in the first model. Due to the exclusion of women with missing smoking and body mass index data, this analysis involved a smaller cohort than the first. Finally, to evaluate the effect of including birth weight in our imputation modeling, restricted cubic splines with three knots were used to model birth weight. With the inclusion of smoking, body mass index, and birth weight, this analysis involved the smallest cohort of the three generalized estimating equation approaches.

To validate the regression results, we defined a separate cohort of deliveries recorded in the Clinical Practice Research Datalink between April 2012 and March 2015. We will refer to this cohort as the external cohort. This external cohort was defined following the same selection steps outlined above. For external validation, we took 1,000 bootstrap samples from the gold standard cohort. We then fitted generalized estimating equations to each selected sample and imputed gestational age at delivery for the external cohort using the regression coefficients derived from the model fit.

Gestational age estimation based on multiple imputation

In our second approach, missing gestational age was estimated using multiple imputation. Multiple imputation provides unbiased estimates under relatively weak assumptions. Multiple imputation proceeds by generating *m* complete datasets where missing values are imputed or filled. Each dataset is analysed separately, and the estimates are then pooled using Rubin's rules.[11] The method can be applied to large datasets with complex patterns of missingness among covariates and uses only complete data quantities with very simple rules of combination. Multiple imputation performs well in situations when data are missing completely at random or missing at random. We used the chained equations method of multiple imputation.

Multiple imputations were performed via predictive mean matching;[12] 10 imputed datasets were created at each time. Predictive mean matching is preferred to regression imputation as it can preserve non-linear relations, even if the structural part of the imputation model is incorrect. When a variable has missing records, predictive mean matching generates predicted values for all observations, including both those with and those without missing data. For each observation with missing data, it identifies a set of records with non-missing data whose predicted values are close

to the predicted value for the observation with missing data. From the subset of those close cases, it samples a record at random and assigns its observed value to substitute for the missing value.

Further inferences were subsequently derived using the standard formulae. For the multiple imputation procedure, we used all the covariates listed in Table 1 (with the exception of baby sex) and evaluated the addition of fetal birth weight to the imputation process. Therefore, multiple imputation method 1 used the same covariates as generalized estimating equation model 2; multiple imputation method 2 used the same covariates as generalized estimating equation model 3.

Single imputation of gestational age using birth weight information

Delivery records were divided into groups based on baby sex and percentile of birth weight. Baby sex information was obtained from the Clinical Practice Research Datalink additional clinical files. Due to low number of deliveries with both missing sex and non-missing birth weight, such records were grouped based on birth weight decile. Using the sex-specific distribution of weight for gestational age at birth (derived from deliveries with complete data), we estimated the missing gestational age by using the median length of gestation from those of similar sex and birth weight.[13]

Conventional approach

In the last approach, we used the conventional method of assigning a fixed gestational age to all deliveries with missing values.[3] Two different approaches were used. In the first, we assigned 39 weeks to all deliveries with missing values. Despite the method being rarely used in current

research, we included it for historical purposes. In the second approach, we assigned 39 weeks to term deliveries and 35 weeks to preterm deliveries.

Simulation study

We conducted a simulation study to compare the described approaches to gestational age estimation, varying the rate of missingness and the set of included variables. We calculated that at least 208 simulations were required to produce an estimate of gestational age within 0.25 weeks of the true gestational age at a 95% confidence level.[14] Therefore, for each method, 250 simulation runs were performed. We considered analyses where, at each iteration, 30% or 50% of the gestational ages were randomly selected and removed. We refer to the data with randomly removed gestational age observations as the validation dataset or cohort. As the data with non-removed gestational age were used to estimate the generalized estimating equation coefficients, we refer to these data as the derivation dataset. Generalized estimating equations used the transformed outcome; all other methods used completed weeks of gestation.

For each method, we calculated the average bias, average absolute bias, average squared prediction error, and the proportion of pregnancies with the estimated completed weeks of gestation within 0, ±1, ±2, ±3, and ±4 weeks of the gold-standard gestational age. We further examined the estimation for term and preterm deliveries separately. All analyses were performed with R version 3.0.2 (R Core Team, R Foundation for Statistical Computing, Vienna, Austria).

Case study: Use of antidepressants

To assess the accuracy of each approach, we assessed prenatal exposure to antidepressants for each delivery. We chose these medications as they are commonly used during pregnancy and

could be associated with an increased risk of autism spectrum disorders and other adverse fetal outcomes.[15] For demonstration purposes, we randomly removed gestational age information for 50% of deliveries and estimated the gestational age using the methods described above. With the true exposure status determined by the Hospital Episode Statistics-defined gestational age, we calculated the sensitivity and specificity of exposure status at anytime during pregnancy and in second or third trimesters of pregnancy for each approach.

RESULTS

A total of 25,929 infants from 22,774 mothers were included in the "gold standard" cohort with a recorded gestational age in both the Clinical Practice Research Datalink and Hospital Episode Statistics (Figure 1). The cohort included 2,877 women >1 delivery. The median gestational age at birth was 40 completed weeks (Table 1, interquartile range: 38, 40). There were 1,484 pregnancies (5.7%) with the recorded gestational age <37 weeks, of which 1,095 had diagnostic codes indicating preterm birth, 93 were multiple gestations, and 31 were stillbirths. The median gestational age at birth for these deliveries was 35 completed weeks (interquartile range: 33.5, 36).

The average bias and prediction error decreased with the use of increasing information in the estimation procedure (Table 2). The smallest average bias (underestimation by 0.002 weeks when 30% of data were set to missing) was observed with multiple imputation, while the generalized estimating equation that included maternal smoking and body mass index information led to the largest average bias (underestimation by 0.554 weeks when 30% of data were set to missing). The average absolute bias ranged from slightly under 1 week (generalized estimating equation model 3) to approximately 1.5 weeks (multiple imputation without birth weight information). The smallest mean squared prediction errors were detected with regression models, while multiple imputation had the highest variability in imputed values. Overall, the generalized estimating equation that included birth weight as a predictor provided the best results, with more than 75% of the estimated gestational ages falling within one week of the true value. However, on average, estimation based on birth weight provided comparable results to generalized estimating equation model 3 without excluding as many records. Importantly, the rate of missing

information did not have an important impact on the accuracy of the examined estimation procedures (eTable 1).

The differences in methods' performance can be assessed in more detail through subgroup analyses of preterm and term deliveries. For preterm deliveries, the assignment of a gestational age of 39 weeks produced results that were more extreme than for any other method. Generalized estimating equation model 3 provided the best estimates for preterm deliveries. For term deliveries, this approach performed the best as well. However, estimation based on birth weight provided equivalent results. Regardless of the approach, the estimation of gestational age for preterm deliveries produced higher prediction errors than for term deliveries (Table 2 and eTable 1).

We identified 1,007 deliveries (3.9%) with ≥1 prescription for an antidepressant at any time during pregnancy (Table 3). For 593 deliveries (2.3%), exposure occurred after 12 weeks of gestation. All methods performed comparably in determining exposure at any time during pregnancy (range of sensitivities: 99.4% to 100.0%; all specificities: 100%). Generalized estimating equation model 3 had one of highest sensitivities and specificities (97.9% and 100.0%, respectively). However, the study sample size had to be reduced (from 25,929 to 17,889 pregnancies) to use this approach because it included body mass index, maternal smoking, and fetal birth weight, which were not available for all pregnancies. On average, multiple imputation method 2 provided the highest sensitivity and specificity without reducing the sample size. All approaches had similar sensitivities (range: 97.5% to 98%) and specificities (range: 99.9% to 100.0%) when assessing antidepressant use in the second or third trimesters (Table 3).

Results of the external validation of the generalized estimating equations (Table 4 and eTables 2 and 3) were comparable to the estimates from the internal validation using the "gold standard" cohort.

We conducted two sensitivity analyses to examine the robustness of our results. In the first sensitivity analysis, we corrected the indicator variable for preterm delivery based on the recorded gestational age (eTables 4 and 5). There were 36 deliveries with a recorded gestational age of ≥37 weeks that had a preterm delivery diagnostic code, and 425 deliveries with a recorded gestational age <37 weeks and no recorded preterm code. After assigning the correct value to the preterm indicator, we repeated the simulations to assess the impact on our results. The bias and prediction errors decreased for all methods except for the conventional approach of assigning 39 weeks gestation to all deliveries.

In the second sensitivity analysis, we assessed the accuracy of an approach that combined regression modeling with estimation based on birth weight with the assignment of gestational age based on preterm status (eTable 6). Using generalized estimating equations, we estimated the missing gestational age for records with complete predictor information; otherwise, we assigned 35 or 39 weeks of gestation to pregnancies based on their term/preterm delivery status. We applied a similar approach to those records with missing birth weight. In comparison with other methods using the full cohort, the combination of generalized estimating equations with uniform assignment based on term/preterm status showed superiority over the conventional methods and multiple imputation. This hybrid method performed better on average but lacked precision in the subgroup of preterm deliveries.

DISCUSSION

In this population-based study, we investigated various methods of estimating missing gestational age. Using the maternal and fetal information found in the Clinical Practice Research Datalink and Hospital Episode Statistics, we evaluated methods based on generalized estimating equations, multiple imputation, estimation based on birth weight, and the conventional approach of assigning gestational age independent of maternal characteristics (either 39 weeks for all or 39 weeks for term deliveries and 35 weeks for pre-term deliveries). Using demographic and clinical characteristics, lifestyle information, and reproductive history information, we were able to accurately predict gestational age at delivery within a week of the true value for over 75% of deliveries. Precision for preterm deliveries was lower than for term deliveries with all studied methods, which is expected as, by definition, the potential range of completed weeks of gestation was 37 to 43 for term deliveries and 22 to 36 for preterm deliveries. We also assessed the accuracy of each method of identifying use of antidepressants at any time during pregnancy and during the second or third trimesters, finding that all methods performed similarly well.

Generalized estimating equations provided good estimates of gestational age at birth among preterm and term deliveries. The generalized estimating equation with birth weight as a predictor performed better than the other approaches, especially for preterm deliveries. While multiple imputations provided the least biased estimates on average, the method lacked precision. Among the standard methods examined, assigning 39 weeks to all deliveries was the least accurate, with the highest average squared prediction error of 30.0 weeks in case of preterm deliveries. Assigning 39 weeks gestation to term deliveries and 35 weeks to preterm deliveries resulted in relatively accurate estimates and even outperformed multiple imputation. Estimation based on birth weight performed well on average and in the subgroup of term deliveries.

Gestational age is the most important predictor of perinatal and pediatric outcomes,[16,17] and it is essential that investigators maximize the use of this information in perinatal database studies and pharmacoepidemiologic studies of pregnant women. Our results suggest that the use of a regression model that includes birth weight information is optimal for the estimation of gestational age, particularly in situations where preterm deliveries are of interest and an incorrectly estimated gestational age could have important implications on study results. In studies involving the use of administrative data where no gestational age information is available, caution should be used. In such cases, the naïve approach of assigning 35 and 39 weeks for pre-term and term deliveries is likely the only acceptable solution.

The use of various approaches to estimating gestational age has been examined previously.[3,18-20] However, previous research relied mostly on deterministic methods of estimation: uniform assignment of pregnancy duration or extensive computer algorithms. The former does not reflect the heterogeneity of pregnancy durations observed in a real-world setting, while the latter may not be generalizable to other databases. Some studies claim that the assignment of 35 weeks for pre-term deliveries and 39 weeks for term deliveries is the optimal approach.[3] While we found that this approach was acceptable, our generalized estimating equation approach that relied on maternal and fetal characteristics was more accurate. In addition, some have suggested two-step procedures involving the use of regression in a database that contains recorded gestational age to estimate coefficients and then applying these coefficients to other data sources that do not contain gestational age information.[3] However, the regression models were primarily based on screening tests performed during pregnancy and did not account for maternal and fetal information that could be obtained from administrative data. This proposed approach relied on assumptions that may not be met given the differences across databases in data structure and recording practices. In our

models, we decided to use covariates independent of the timing of prenatal care, including lifestyle variables, medical variables, and delivery details that could be identified in most databases.

This study has several strengths. First, we constructed a "gold standard" cohort in which the gestational age was verified from two data sources (Clinical Practice Research Datalink and Hospital Episode Statistics). Second, one of the major advantages of the Clinical Practice Research Datalink is its large size.[7] With minimal restrictions applied to the study population and the use of population-based data that are representative of the UK population, the results of this analysis are likely to be generalizable to the whole population. Third, we conducted several sensitivity analyses and repeated our analyses using different levels of missingness, and results were consistent across all analyses.

Our study also has some potential limitations. First, we compared model-based approaches to only three of the potential algorithms that could be used to estimate gestational age. For example, another algorithm that is commonly used to estimate gestational age is based on the use of screening test claims.[3] However, Hardy et al.[21] found that most perinatal tests were poorly recorded in the Clinical Practice Research Datalink. Consequently, we did not investigate this approach in the present study. Second, it is important to note the different assumptions about missingness required for the generalized estimating equation and multiple imputation approaches. Multiple imputation provides unbiased estimates assuming data are missing at random, meaning that the propensity for a data point to be missing is not related to the missing observations but is related to some of the observed data. Generalized estimating equations are valid only if data are missing completely at random, i.e., there are no variables, missing or observed, that affect the probability of a data point being missing. By design, in our study, the missingness mechanism was ignorable, and the missing completely at random assumption was thus reasonable. In other settings, one

should explore the missingness mechanism by modeling the probability of gestational age being missing. Third, the identified number of premature deliveries was lower than the UK national average.[22] It is possible that the low proportion can be explained by restricting inclusion to English practices. Therefore, some differences relative to the UK as a whole are possible. Fourth, we assessed the accuracy of all methods in assessing antidepressant use during pregnancy and during the second or third trimesters. While all approaches performed similarly well, the generalizability of these findings to other exposures is unclear. Finally, the applicability of these results to other data sources is unclear. However, demographic characteristics and diagnoses that we used for gestational age estimation (lifestyle variables, comorbidities) are found in most administrative or clinical databases. While smoking status and body mass index are not readily available in administrative databases, diagnostic codes for tobacco dependence and obesity typically are. In addition, national birth cohorts[23-25] usually contain birth details such as birth weight and baby sex. Therefore, our method could be applicable to such cohorts where pregnancy duration is missing for some records.

In conclusion, our results suggest that estimation methods utilizing fetal birth weight information provide the most accurate estimates of gestational age at birth. If birth weight information is available, the use of a generalized estimating equation that includes fetal birth weight as a predictor is suggested as a means to estimate missing gestational age. Otherwise, the conventional approach of assigning a gestational age may be preferred. However, if the conventional approach is used, it is important to differentiate between term and preterm deliveries, as assigning 39 weeks gestation to all deliveries results in substantial bias.

REFERENCES

1. Grzeskowiak LE, Gilbert AL, Morrison JL. Exposed or not exposed? Exploring exposure classification in studies using administrative data to investigate outcomes following medication use during pregnancy. *Eur J Clin Pharmacol* 2012;**68**(5):459-67.
2. Raebel MA, Ellis JL, Andrade SE. Evaluation of gestational age and admission date assumptions used to determine prenatal drug exposure from administrative data. *Pharmacoepidemiol Drug Saf* 2005;**14**(12):829-36.
3. Margulis AV, Setoguchi S, Mittleman MA, Glynn RJ, Dormuth CR, Hernandez-Diaz S. Algorithms to estimate the beginning of pregnancy in administrative databases. *Pharmacoepidemiol Drug Saf* 2013;**22**(1):16-24.
4. Andrade SE, Raebel MA, Morse AN, Davis RL, Chan KA, Finkelstein JA, Fortman KK, McPhillips H, Roblin D, Smith DH, Yood MU, Platt R, J HG. Use of prescription medications with a potential for fetal harm among pregnant women. *Pharmacoepidemiol Drug Saf* 2006;**15**(8):546-54.
5. Herrett E, Thomas SL, Schoonen WM, Smeeth L, Hall AJ. Validation and validity of diagnoses in the General Practice Research Database: a systematic review. *Br.J.Clin.Pharmacol.* 2010;**69**(1):4-14.
6. Khan NF, Harrison SE, Rose PW. Validity of diagnostic coding within the General Practice Research Database: a systematic review. *Br.J.Gen.Pract.* 2010;**60**(572):e128-e136.
7. Herrett E, Gallagher AM, Bhaskaran K, Forbes H, Mathur R, van Staa T, Smeeth L. Data Resource Profile: Clinical Practice Research Datalink (CPRD). *Int J Epidemiol* 2015;**44**(3):827-36.

8. El-Metwally D, Vohr B, Tucker R. Survival and neonatal morbidity at the limits of viability in the mid 1990s: 22 to 25 weeks. *J Pediatr* 2000;**137**(5):616-22.

9. Kramer MS, Platt RW, Wen SW, Joseph KS, Allen A, Abrahamowicz M, Blondel B, Breart G. A new and improved population-based Canadian reference for birth weight for gestational age. *Pediatrics* 2001;**108**(2):E35.

10. Diggle PJ, Heagerty PK, Liang KY, Zeger SL. *Analysis of Longitudinal Data.* 2nd edition ed. New York, NY: Oxford University Press, 2002.

11. Rubin DB. *Multiple Imputation for Nonresponse in Surveys.* New York, NY: Wiley, 1987.

12. Heitjan DF, Little RJA. Multiple Imputation for the Fatal Accident Reporting System. *Journal of the Royal Statistical Society. Series C(Applied Statistics)* 1991;**40**(1):13-29.

13. Piper JM, Mitchel EF, Jr., Ray WA. Presumptive eligibility for pregnant Medicaid enrollees: its effects on prenatal care and perinatal outcome. *Am J Public Health* 1994;**84**(10):1626-30.

14. Burton A, Altman DG, Royston P, Holder RL. The design of simulation studies in medical statistics. *Stat Med* 2006;**25**(24):4279-92.

15. Rai D, Lee BK, Dalman C, Golding J, Lewis G, Magnusson C. Parental depression, maternal antidepressant use during pregnancy, and risk of autism spectrum disorders: population based case-control study. *BMJ* 2013;**346**:f2059.

16. Callaghan WM, MacDorman MF, Rasmussen SA, Qin C, Lackritz EM. The contribution of preterm birth to infant mortality rates in the United States. *Pediatrics* 2006;**118**(4):1566-73.

17. Sowards KA. What is the leading cause of infant mortality? A note on the interpretation of official statistics. *Am J Public Health* 1999;**89**(11):1752-4.

18. Margulis AV, Palmsten K, Andrade SE, Charlton RA, Hardy JR, Cooper WO, Hernandez-Diaz S. Beginning and duration of pregnancy in automated health care databases: review of estimation methods and validation results. *Pharmacoepidemiol Drug Saf* 2015;**24**(4):335-42.

19. Cea-Soriano L, Garcia Rodriguez LA, Fernandez Cantero O, Hernandez-Diaz S. Challenges of using primary care electronic medical records in the UK to study medications in pregnancy. *Pharmacoepidemiol Drug Saf* 2013;**22**(9):977-85.

20. Toh S, Mitchell AA, Werler MM, Hernandez-Diaz S. Sensitivity and specificity of computerized algorithms to classify gestational periods in the absence of information on date of conception. *Am J Epidemiol* 2008;**167**(6):633-40.

21. Hardy JR, Holford TR, Hall GC, Bracken MB. Strategies for identifying pregnancies in the automated medical records of the General Practice Research Database. *Pharmacoepidemiol.Drug Saf* 2004;**13**(11):749-759.

22. National Health Service. Premature labour and birth. http://www.nhs.uk/conditions/pregnancy-and-baby/pages/premature-early-labour.aspx Accessed 2016-07-21, 2016.

23. Micali N, Stemann Larsen P, Strandberg-Larsen K, Nybo Andersen AM. Size at birth and preterm birth in women with lifetime eating disorders: a prospective population-based study. *Bjog* 2016;**123**(8):1301-10.

24. Vandentorren S, Bois C, Pirus C, Sarter H, Salines G, Leridon H. Rationales, design and recruitment for the Elfe longitudinal study. *BMC Pediatr* 2009;**9**:58.

25. Rappazzo KM, Warren JL, Meyer RE, Herring AH, Sanders AP, Brownstein NC, Luben TJ. Maternal residential exposure to agricultural pesticides and birth defects in a 2003 to 2005 North Carolina birth cohort. *Birth Defects Res A Clin Mol Teratol* 2016;**106**(4):240-9.

FIGURE LEGEND

Figure 1 Flow diagram describing cohort construction

Table 1. Baseline demographic and clinical characteristics of the cohort.

Characteristic [a]	Deliveries (N = 25,929)
Demographic and lifestyle information	
Age (years)	
Mean (SD)	30.3 (5.88)
Median (IQR)	31 (26, 35)
Body mass index, n (%)	
< 18.5 kg/m^2	836 (3.2)
18.5-25 kg/m^2	11,280 (43.5)
25-30 kg/m^2	5,026 (19.4)
≥ 30 kg/m^2	3,157 (12.2)
Missing	5,630 (21.7)
Smoking status, n (%)	
Never	12,018 (46.3)
Ever	11,070 (42.7)
Missing	2,841 (11.0)
Alcohol abuse, n (%)	1,215 (4.7)
Delivery information	
Gestational age at delivery (weeks)	
Mean (SD)	39.3 (1.84)
Median (IQR)	40 (38, 40)
Fetal birth weight (grams)	
Mean (SD)	3,401.8 (547.59)
Median (IQR)	3,415 (3080, 3746)
Missing, n (%)	714 (2.8)
Baby sex, n (%)	
Male	13,027 (50.2)
Female	12,548 (48.4)
Missing	354 (1.4)
Number of previous pregnancies, n (%)	
0	12,334 (47.6)
1	9,922 (38.3)
2	2,870 (11.1)
3	617 (2.4)
≥4	186 (0.7)
Number of previous miscarriages, n (%)	
0	21,163 (81.6)
1	3,828 (14.8)
2	721 (2.8)
3	162 (0.6)
≥4	55 (0.2)
Preterm delivery, n (%) [b]	1,095 (4.2)
Multiple gestations, n (%)	200 (0.8)
Stillbirth, n (%)	46 (0.2)

Pregnancy complications, n (%)	
History of gestational diabetes	142 (0.5)
Gestational diabetes during current pregnancy	456 (1.8)
History of hypertensive disorders during pregnancy	212 (0.8)
Hypertensive disorders during current pregnancy	153 (0.6)

Abbreviations: SD: standard deviation; IQR: interquartile range.

[a] Alcohol abuse, parity, miscarriages and history of pregnancy complications were assessed before the beginning of current pregnancy. Smoking status was assessed in the past 5 years of medical history. Body mass index was estimated using the latest available weight measurement up to the end of the first trimester (first 12 weeks) of the current pregnancy. Current pregnancy conditions were assessed in the second and third trimesters of the current pregnancy (after 12 weeks of gestation).

[b] In sensitivity analyses, the indicator was corrected for 461 delivery records according to the recorded gestational age.

Table 2. Validation of estimated gestational age at birth against gestational age recorded in the "gold-standard" cohort (50% of gestational age records missing).

	GEE			Multiple Imputation		Estimation Based on Birth Weight	Conventional Methods	
Statistic	Model 1[a] (n=25,929)	Model 2[b] (n=18,405)	Model 3[c] (n=17,889)	MI 1[d] (n=25,929)	MI 2[e] (n=25,929)	Method 1[f] (n=25,215)	Method 1[g] (n=25,929)	Method 2[h] (n=25,929)
				All deliveries				
Bias[i]	-0.52	-0.55	-0.34	-0.02	-0.002	0.06	-0.27	-0.44
Absolute bias	1.19	1.16	0.96	1.50	1.34	1.01	1.31	1.18
Mean squared prediction error[j]	2.47	2.29	1.61	4.25	3.33	1.87	3.45	2.52
Correct estimation	23.6%	23.9%	30%	22.1%	24.5%	30.8%	23.0%	24%
± 1 week	44.4%	44.7%	48%	36.9%	39.5%	45.9%	43.0%	45.1%
± 2 weeks	26.4%	26%	18.6%	23.2%	22.4%	17.9%	25.0%	25.1%
± 3 weeks	4%	4%	3%	11.2%	9.1%	4%	5.9%	4%
± 4 weeks	1%	0.7%	0%	4%	3%	1%	1%	1%
± 5 weeks	0%	0%	0%	1%	1%	0%	0%	0%
				Preterm deliveries				
Bias[i]	-0.33	-0.24	-0.47	-0.06	-0.01	2.03	4.76	0.76
Absolute bias	2.14	1.79	1.20	2.67	2.01	2.49	4.77	1.74
Mean squared prediction error[j]	9.04	6.93	2.81	15.61	10.02	8.42	30.03	7.99
				Term deliveries				
Bias[i]	-0.53	-0.56	-0.33	-0.02	-0.002	-0.02	-0.49	-0.49
Absolute bias	1.14	1.14	0.96	1.45	1.31	0.95	1.15	1.15
Mean squared prediction error[j]	2.18	2.11	1.56	3.75	3.04	1.60	2.28	2.28

Abbreviations: GEE: generalized estimating equations; MI: multiple imputation.
[a] Model 1: Generalized estimating equations with autoregressive working correlation matrix, adjusted for maternal age at delivery; maternal history of alcohol consumption; indicator of singleton vs multiple gestation; presence of the preterm birth or still birth indicator; previous parity; history of miscarriages; complications during current pregnancy (pre-eclampsia, other hypertensive disorders during pregnancy, gestational diabetes); and complications during previous pregnancies.
[b] Model 2: Model 1 + body mass index, maternal history of smoking.
[c] Model 3: Model 2 + fetal birth weight.
[d] MI 1: Multiple imputation with predictive mean matching and 10 imputed datasets, birth weight information was not used.
[e] MI 2: Multiple imputation with predictive mean matching and 10 imputed datasets, birth weight information was used.
[f] Birth Weight Method 1: Gestational age set to the median of gestational age for respective birth weight percentile and baby sex. Median gestational age was used for respective birth weight decile when baby sex was missing.
[g] Conventional Method 1: Gestational age of 39 weeks assumed for all deliveries.
[h] Conventional Method 2: Gestational age of 39 weeks assumed for deliveries with no indication of preterm delivery, gestational age of 35 weeks assumed for deliveries with indication of preterm delivery.
[i] Mean bias over 250 simulations; in each iteration, calculated as predicted value minus the observed.
[j] Average prediction error over 250 simulations; in each iteration, calculated as the difference of the observed and predicted values squared.

Table 3. Classification of prenatal exposure status to antidepressants based on estimated gestational age versus gestational age obtained from the "gold standard" cohort [a].

	True exposure in 2nd and 3rd trimesters		True exposure anytime during pregnancy	
Estimated exposure	Specificity	Sensitivity	Specificity	Sensitivity
GEE Model 1[b]	100.0	97.6	100.0	99.4
GEE Model 2[c]	100.0	97.5	100.0	100.0
GEE Model 3[d]	100.0	97.9	100.0	100.0
MI 1[e]	99.9	98.0	100.0	100.0
MI 2[f]	100.0	98.0	100.0	100.0
Birth Weight Method 1[g]	99.9	97.9	100.0	99.8
Conventional Method 1[h]	100.0	97.6	100.0	99.8
Conventional Method 2[i]	100.0	97.6	100.0	99.8

[a] Restricted to deliveries with missing gestational age.

Abbreviations: GEE: generalized estimating equation; MI: multiple imputation.

[b] GEE Model 1: Generalized estimating equations with autoregressive working correlation matrix, adjusted for maternal age at delivery; maternal history of alcohol consumption; indicator of singleton vs multiple gestation; presence of the preterm birth or still birth indicator; previous parity; history of miscarriages; complications during current pregnancy (pre-eclampsia, other hypertensive disorders during pregnancy, gestational diabetes); and complications during previous pregnancies.

[c] GEE Model 2: Model 1 + body mass index, maternal history of smoking.

[d] GEE Model 3: Model 2 + fetal birth weight.

[e] MI 1: Multiple imputation with predictive mean matching and 10 imputed datasets, birth weight information was not used.

[f] MI 2: Multiple imputation with predictive mean matching and 10 imputed datasets, birth weight information was used.

[g] Birth Weight Method 1: Gestational age set to the median of gestational age for respective birth weight percentile and baby sex.

[h] Conventional Method 1: Gestational age of 39 weeks assumed for all deliveries.

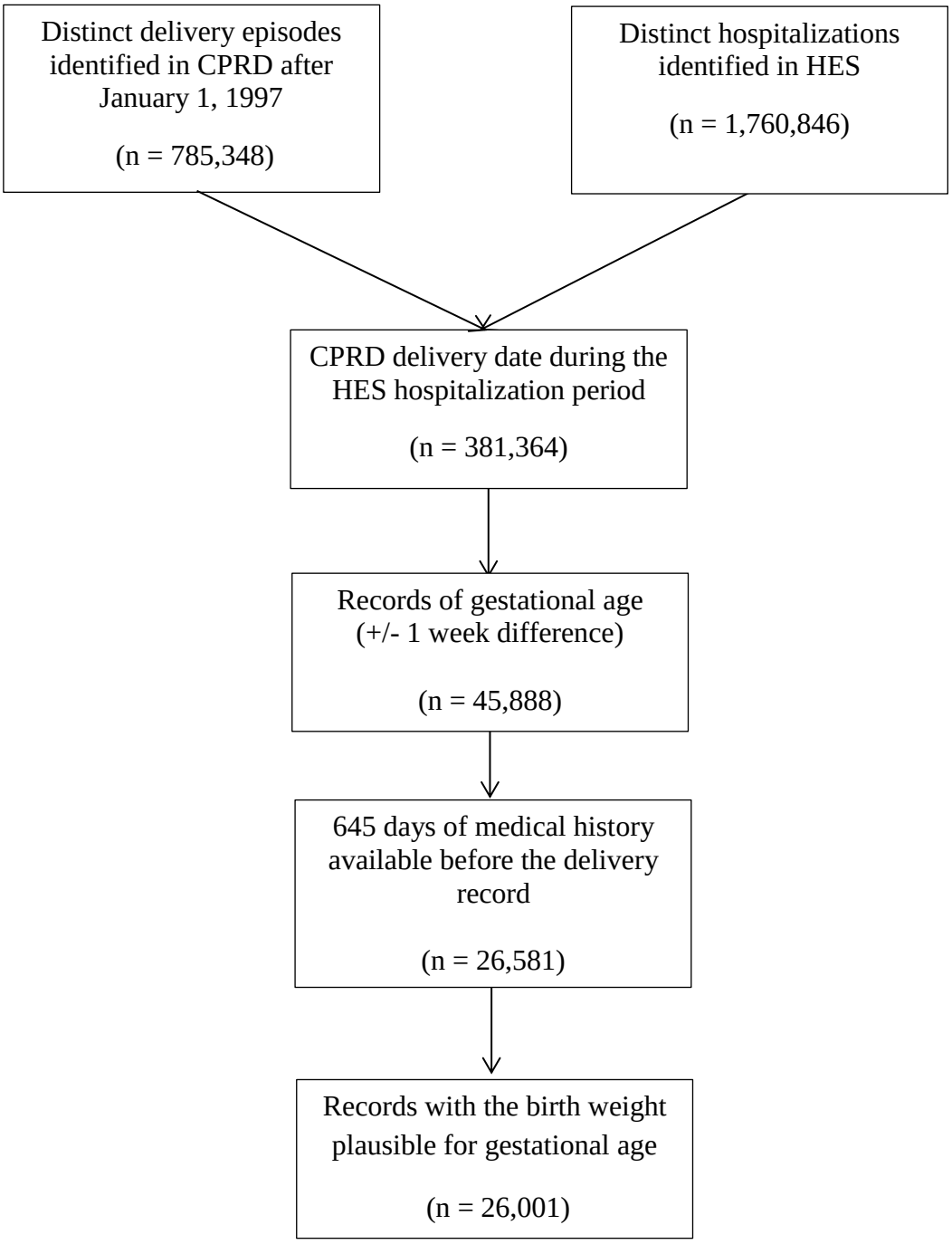
[i] Conventional Method 2: Gestational age of 39 weeks assumed for deliveries with no indication of preterm delivery, gestational age of 35 weeks assumed for deliveries with indication of preterm delivery.

Table 4. Bootstrap validation of regression model performance on external data. [a]

	GEE for Longitudinal Data		
	Model 1 [b] (n = 2,640)	Model 2 [c] (n = 2,176)	Model 3 [d] (n = 2,098)
Statistic	Median (95% CI)	Median (95% CI)	Median (95% CI)
	All deliveries		
Bias	-0.57 (-0.59, -0.55)	-0.57 (-0.60, -0.46)	-0.16 (-0.20, -0.14)
Absolute bias	1.10 (1.09, 1.11)	1.07 (1.06, 1.08)	0.87 (0.86, 0.87)
Mean squared prediction error	1.97 (1.95, 1.98)	1.87 (1.83, 1.89)	1.33 (1.31, 1.34)
	Preterm Deliveries		
Bias	-0.40 (-0.71, -0.16)	-0.07 (-0.28, 0.03)	-0.21 (-0.31, -0.14)
Absolute bias	1.61 (1.49, 1.74)	1.24 (1.17, 1.36)	0.89 (0.84, 0.94)
Mean squared prediction error	4.70 (4.46, 4.89)	3.37 (3.17, 3.71)	1.38 (1.25, 1.49)
	Term deliveries		
Bias	-0.57 (-0.59, -0.57)	-0.59 (-0.62, -0.48)	-0.16 (-0.20, -0.14)
Absolute bias	1.07 (1.07, 1.07)	1.07 (1.05, 1.07)	0.87 (0.86, 0.87)
Mean squared prediction error	1.81 (1.81, 1.82)	1.81 (1.77, 1.82)	1.32 (1.31, 1.34)

[a] A total of 1,000 iterations were performed. At each iteration generalized estimating equation models were fitted with the selected bootstrap sample. Estimated model coefficients were used to predict gestational age records from the external cohort.
[b] Model 1: Generalized estimating equations with autoregressive working correlation matrix, adjusted for maternal age at delivery; maternal history of alcohol consumption; indicator of singleton vs multiple gestation; presence of the preterm birth or still birth indicator; previous parity; history of miscarriages; complications during current pregnancy (pre-eclampsia, other hypertensive disorders during pregnancy, gestational diabetes); and complications during previous pregnancies.
[c] Model 2: Model 1 + body mass index, maternal history of smoking.
[d] Model 3: Model 2 + fetal birth weight.

Figure 1.


Distinct delivery episodes identified in CPRD after January 1, 1997
(n = 785,348)
Distinct hospitalizations identified in HES
(n = 1,760,846)
CPRD delivery date during the HES hospitalization period
(n = 381,364)
Records of gestational age (+/- 1 week difference)
(n = 45,888)
645 days of medical history available before the delivery record
(n = 26,581)
Records with the birth weight plausible for gestational age
(n = 26,001)

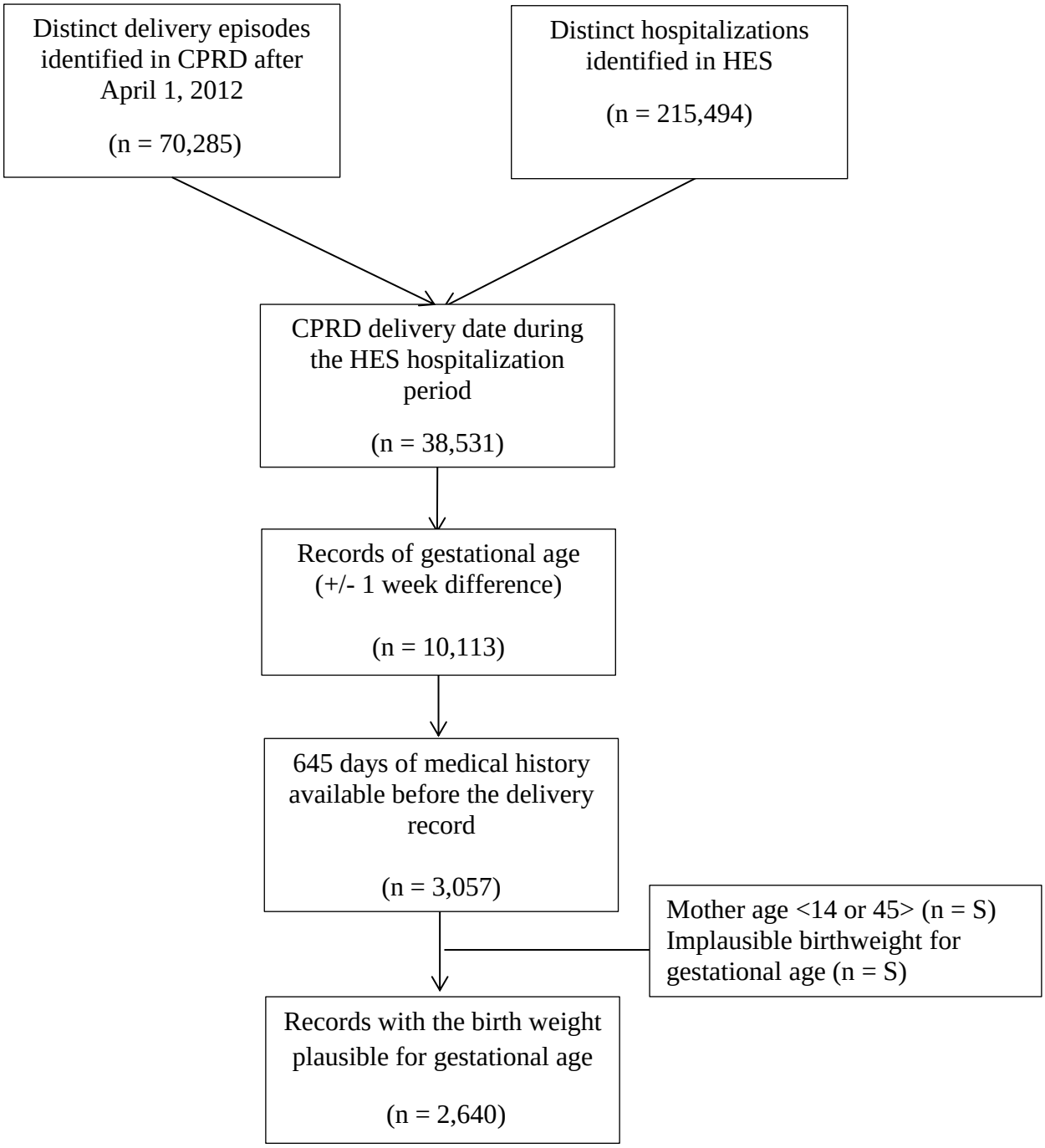
Supplemental Tables

eTable 1. Validation of estimated gestational age at birth against gestational age recorded in the "gold-standard" cohort (50% of gestational age records missing).

	GEE			Multiple Imputation		Estimation Based on Birth Weight	Conventional Methods	
Statistic	Model 1[a] (n=25,929)	Model 2 [b] (n=18,405)	Model 3 [c] (n=17,889)	MI 1 [d] (n=25,929)	MI 2 [e] (n=25,929)	Method 1 [f] (n=25,215)	Method 1 [g] (n=25,929)	Method 2 [h] (n=25,929)
All deliveries								
Bias [i]	-0.53	-0.56	-0.34	-0.02	-0.003	0.06	-0.27	-0.45
Absolute bias	1.18	1.16	0.96	1.50	1.33	1.00	1.31	1.18
Mean squared prediction error [j]	2.47	2.26	1.60	4.25	3.32	1.84	3.45	2.52
Correct estimation								
± 1 week	23.5%	23.8%	30%	22%	24.5%	30.9%	23%	23.9%
± 2 weeks	44.4%	44.8%	48%	36.9%	39.5%	46%	43.1%	45.2%
± 3 weeks	26.5%	26.1%	18.6%	23.2%	22.4%	17.6%	25%	25.2%
± 4 weeks	4%	4%	3%	11.2%	9.2%	4%	5.9%	4%
± 5 weeks	1%	0.5%	0%	4%	3%	1%	1%	1%
Preterm deliveries								
Bias [i]	-0.34	-0.23	-0.46	-0.03	-0.003	2.04	4.76	0.76
Absolute bias	2.14	1.75	1.20	2.66	2.01	2.48	4.77	1.75
Mean squared prediction error [j]	8.99	6.64	2.79	15.56	9.94	8.36	30.10	8.05
Term deliveries								
Bias [i]	-0.53	-0.58	-0.33	-0.02	-0.003	-0.02	-0.49	-0.49
Absolute bias	1.14	1.14	0.96	1.45	1.30	0.94	1.15	1.15
Mean squared prediction error [j]	2.18	2.10	1.56	3.75	3.03	1.57	2.28	2.28

Abbreviations: GEE: generalized estimating equations; MI: multiple imputation.
[a] Model 1: Generalized estimating equations with autoregressive working correlation matrix, adjusted for maternal age at delivery; maternal history of alcohol consumption; indicator of singleton vs multiple gestation; presence of the preterm birth or still birth indicator; previous parity; history of miscarriages; complications during current pregnancy (pre-eclampsia, other hypertensive disorders during pregnancy, gestational diabetes); and complications during previous pregnancies.
[b] Model 2: Model 1 + body mass index, maternal history of smoking.
[c] Model 3: Model 2 + fetal birth weight.
[d] MI 1: Multiple imputation with predictive mean matching and 10 imputed datasets; birth weight information was not used.
[e] MI 2: Multiple imputation with predictive mean matching and 10 imputed datasets; birth weight information was used.
[f] Birth Weight Method 1: Gestational age set to the median of gestational age for respective birth weight percentile and baby sex. Median gestational age was used for respective birth weight decile when baby sex was missing.
[g] Conventional Method 1: Gestational age of 39 weeks assumed for all deliveries.
[h] Conventional Method 2: Gestational age of 39 weeks assumed for deliveries with no indication of preterm delivery, gestational age of 35 weeks assumed for deliveries with indication of preterm delivery.
[i] Mean bias over 250 simulations; in each iteration, calculated as predicted value minus the observed.
[j] Average prediction error over 250 simulations; in each iteration, calculated as the difference of the observed and predicted values squared.

eFigure 1. External validation cohort selection



Abbreviations: S: suppressed data to comply with CPRD privacy restrictions (denotes count <5).

eTable 2. Baseline demographic and clinical characteristics of the external validation cohort.

Characteristic *	Deliveries (N = 2,640)
Demographic and lifestyle information	
Age (years)	
Mean (SD)	29.9 (5.92)
Median (IQR)	30 (26, 34)
Body mass index, n (%)	
< 18.5 kg/m^2	81 (3.1)
18.5-25 kg/m^2	1,036 (39.2)
25-30 kg/m^2	614 (23.3)
≥ 30 kg/m^2	495 (18.8)
Missing	414 (15.7)
Smoking status, n (%)	
Never	1,282 (48.6)
Ever	1,296 (49.1)
Missing	62 (2.3)
Alcohol abuse, n (%)	222 (8.4)
Delivery information	
Gestational age at delivery (weeks)	
Mean (SD)	39.3 (1.71)
Median (IQR)	40 (39, 40)
Fetal birth weight (grams)	
Mean (SD)	3,423.9 (528.48)
Median (IQR)	3,460 (3,116, 3,750)
Missing, n (%)	87 (3.3)
Baby sex, n (%)	
Male	1,325 (50.2)
Female	1,238 (46.9)
Missing	77 (2.9)
Number of previous pregnancies, n (%)	
0	2,458 (93.1)
1	S
2	S
Number of previous miscarriages, n (%)	
0	2,090 (79.2)
1	437 (16.6)
2	85 (3.2)
3	20 (0.8)
≥4	8 (0.3)
Preterm delivery, n (%)	140 (5.3)
Multiple gestations, n (%)	7 (0.3)
Stillbirth, n (%)	S
Pregnancy complications, n (%)	
History of gestational diabetes	29 (1.1)

Gestational diabetes during current pregnancy	70 (2.7)
History of hypertensive disorders during pregnancy	21 (0.8)
Hypertensive disorders during current pregnancy	11 (0.4)

Abbreviations: SD: standard deviation; IQR: interquartile range; S: suppressed data to comply with CPRD privacy restrictions (denotes count <5).

* Alcohol abuse, parity, miscarriages and history of pregnancy complications were assessed before the beginning of current pregnancy. Smoking status was verified in the last 5 years prior to delivery. Body mass index was estimated using the latest available weight measurement up to the end of the first trimester (first 12 weeks) of the current pregnancy. Current pregnancy conditions were assessed in the second and third trimesters of the current pregnancy (after 12 weeks of gestation).

eTable 3. Average parameter estimates for GEE obtained from 1,000 bootstrap samples.

	GEE Coefficient (SE)		
Variable	**Model 1**	**Model 2**	**Model 3**
Intercept	1.41 (0.004)	1.44 (0.01)	2.36 (0.03)
Age, <25 years	0	0	0
Age, 26-30 years	-0.001 (0.005)	0.006 (0.006)	0.01 (0.006)
Age, 31-35 years	0.005 (0.005)	0.01 (0.006)	0.03 (0.006)
Age, 36 + years	0.03 (0.006)	0.03 (0.007)	0.04 (0.007)
History of Alcohol Abuse	0.002 (0.009)	-0.0002 (0.01)	-0.003 (0.009)
History of Smoking	--	0.005 (0.004)	-0.007 (0.004)
Body mass index, < 18.5 kg/m^2	--	0	0
Body mass index, 18.5-25 kg/m^2	--	-0.05 (0.01)	-0.01 (0.010)
Body mass index, 25-30 kg/m^2	--	-0.05 (0.01)	-0.001 (0.01)
Body mass index, ≥ 30 kg/m^2	--	-0.06 (0.01)	0.009 (0.01)
History of Gestational Hypertension	0.05 (0.02)	0.06 (0.02)	0.04 (0.02)
Gestational Hypertension during the current pregnancy	0.14 (0.02)	0.14 (0.03)	0.08 (0.03)
History of Gestational Diabetes	0.04 (0.02)	0.05 (0.03)	0.07 (0.03)
Gestational Diabetes during the current pregnancy	0.18 (0.01)	0.20 (0.01)	0.20 (0.01)
History of Miscarriage	0.01 (0.005)	0.02 (0.006)	0.01 (0.005)
Preterm delivery	0.78 (0.006)	0.76 (0.007)	0.51 (0.007)
Multiple Gestation Pregnancy	0.23 (0.02)	0.24 (0.02)	0.15 (0.02)
Stillbirth	0.24 (0.06)	0.19 (0.08)	0.02 (0.05)
Previous Parity	0.04 (0.004)	0.04 (0.004)	0.06 (0.004)
Birth Weight in grams (Linear Component)	--	--	-0.0003 (0.00001)
Birth Weight in grams (Non-linear Component)	--	--	0.0001 (0.00001)

Abbreviations: SE: standard error

The dependent variable was $log(max_{gw}+1-gw)$, where max_{gw} denotes the maximum gestational age at delivery observed in the study population and *gw* is the number of completed weeks of gestation at delivery.

eTable 4. Validation of estimated gestational age at birth against gestational age recorded in the "gold-standard" cohort, corrected values of preterm indicator (30% of gestational age records missing).

	GEE			Multiple Imputation		Conventional Methods	
Statistic	Model 1[a] (n=25,929)	Model 2 [b] (n=18,405)	Model 3 [c] (n=17,889)	MI 1 [d] (n=25,929)	MI 2 [e] (n=25,929)	Method 1 [f] (n=25,929)	Method 2 [g] (n=25,929)
Bias [h]	-0.58	-0.59	-0.35	-0.02	-0.002	-0.27	-0.50
Absolute bias	1.14	1.13	0.95	1.41	1.26	1.30	1.13
Mean squared prediction error [i]	2.15	2.06	1.51	3.60	2.87	3.44	2.17
Correct estimation	23.8%	24.1%	30.1%	22.9%	25.5%	23.0%	24.0%
± 1 week	44.8%	45.2%	48.4%	37.7%	40.1%	43.0%	46.2%
± 2 weeks	27%	26.3%	18.7%	23.4%	22.4%	25.0%	25.4%
± 3 weeks	3.4%	3.9%	2.2%	11.1%	9.0%	5.9%	3.0%
± 4 weeks	0.4%	0.1%	0%	3.5%	2.0%	1%	0%
± 5 weeks	0%	0%	0%	0.9%	0%	1%	0%
Bias [h]	-0.36	-0.31	-0.45	-0.09	-0.02	4.79	0.79
Absolute bias	2.03	1.76	1.17	2.42	1.81	4.79	1.67
Mean squared prediction error [i]	7.32	6.07	2.65	13.46	8.56	29.83	7.55
Bias [h]	-0.59	-0.60	-0.35	-0.01	-0.001	-0.57	-0.57
Absolute bias	1.09	1.09	0.94	1.35	1.23	1.09	1.09
Mean squared prediction error [i]	1.84	1.85	1.46	3.00	2.52	1.84	1.84

[*] Preterm status indicator was corrected for 461 delivery records based on the gestational age obtained from the "gold standard" cohort. Abbreviations: MI: multiple imputation.

[a] Model 1: Generalized estimating equations with autoregressive working correlation matrix, adjusted for maternal age at delivery; maternal history of alcohol consumption; indicator of singleton vs multiple gestation; presence of the preterm birth or still birth indicator; previous parity; history of miscarriages; complications during current pregnancy (pre-eclampsia, other hypertensive disorders during pregnancy, gestational diabetes); and complications during previous pregnancies.

[b] Model 2: Model 1 + body mass index, maternal history of smoking.
[c] Model 3: Model 2 + fetal birth weight.
[d] MI 1: Multiple imputation with predictive mean matching and 10 imputed datasets, birth weight information was not used.
[e] MI 2: Multiple imputation with predictive mean matching and 10 imputed datasets, birth weight information was used.
[f] Conventional Method 1: Gestational age of 39 weeks assumed for all deliveries.
[g] Conventional Method 2: Gestational age of 39 weeks assumed for deliveries with no indication of preterm delivery, gestational age of 35 weeks assumed for deliveries with indication of preterm delivery.
[h] Mean bias over 250 simulations, in each iteration calculated as predicted value minus the observed.
[i] Average Prediction error over 250 simulations, in each iteration calculated as the difference of the observed and predicted values squared.

eTable 5. Validation of estimated gestational age at birth against gestational age recorded in the "gold-standard" cohort, corrected values of preterm indicator (50% of gestational age records missing).

	GEE			Multiple Imputation		Conventional Methods	
Statistic	Model 1[a] (n=25,929)	Model 2 [b] (n=18,405)	Model 3 [c] (n=17,889)	MI 1 [d] (n=25,929)	MI 2 [e] (n=25,929)	Method 1 [f] (n=25,929)	Method 2 [g] (n=25,929)
Bias [h]	-0.58	-0.60	-0.35	-0.01	0.002	-0.27	-0.49
Absolute bias	1.15	1.13	0.94	1.41	1.26	1.30	1.13
Mean squared prediction error [i]	2.15	2.06	1.51	3.58	2.86	3.45	2.17
Correct estimation	23.7%	24.0%	30.1%	22.9%	25.5%	23.0%	24.0%
± 1 week	44.8%	45.2%	48.5%	37.6%	40.1%	43.1%	46.2%
± 2 weeks	27.1%	26.5%	18.6%	23.4%	22.4%	25%	25.4%
± 3 weeks	3.4%	3.8%	2.2%	11.1%	9.0%	5.8%	3.1%
± 4 weeks	0.4%	0.2%	0.0%	3.5%	2.0%	1.0%	0.0%
± 5 weeks	0.0%	0.0%	0.0%	0.9%	0.0%	1.0%	0.0%
Bias [h]	-0.36	-0.31	-0.46	-0.05	0.009	4.80	0.80
Absolute bias	2.04	1.76	1.17	2.41	1.79	4.80	1.69
Mean squared prediction error [i]	7.36	6.08	2.63	13.38	8.44	30.07	7.66
Bias [h]	-0.59	-0.61	-0.35	-0.01	0.001	-0.57	-0.57
Absolute bias	1.09	1.10	0.93	1.35	1.23	1.09	1.09
Mean squared prediction error [i]	1.84	1.86	1.46	2.99	2.52	1.84	1.84

[*] Preterm status indicator was corrected for 461 delivery records based on the gestational age obtained from the "gold standard" cohort. Abbreviations: MI: multiple imputation.

[a] Model 1: Generalized estimating equations with autoregressive working correlation matrix, adjusted for maternal age at delivery; maternal history of alcohol consumption; indicator of singleton vs multiple gestation; presence of the preterm birth or still birth indicator;

previous parity; history of miscarriages; complications during current pregnancy (pre-eclampsia, other hypertensive disorders during pregnancy, gestational diabetes); and complications during previous pregnancies.
[b] Model 2: Model 1 + body mass index, maternal history of smoking.
[c] Model 3: Model 2 + fetal birth weight.
[d] MI 1: Multiple imputation with predictive mean matching and 10 imputed datasets, birth weight information was not used.
[e] MI 2: Multiple imputation with predictive mean matching and 10 imputed datasets, birth weight information was used.
[f] Conventional Method 1: Gestational age of 39 weeks assumed for all deliveries.
[g] Conventional Method 2: Gestational age of 39 weeks assumed for deliveries with no indication of preterm delivery, gestational age of 35 weeks assumed for deliveries with indication of preterm delivery.
[h] Mean bias over 250 simulations, in each iteration calculated as predicted value minus the observed.
[i] Average Prediction error over 250 simulations, in each iteration calculated as the difference of the observed and predicted values squared.

eTable 6. Comparison of GEE and estimation based on birth weight using only complete data against a combined approach of assigning a uniform gestational age based on preterm status to records with missing covariate data.

	GEE				Estimation Based on Birthweight	
Statistic	Model 2 [a] (n=18,405)	Model 2 and Conventional method 2 [b] (n=25,929)	Model 3 [c] (n=17,889)	Model 3 and Conventional method 2 [d] (n=25,929)	Method 1[e] (n=25,215)	Method 1 and Conventional Method 2[f] (n=25,929)
Rate of missingness = 30%						
			All deliveries			
Bias [g]	-0.55	-0.47	-0.34	-0.32	0.06	0.07
Absolute bias	1.16	1.18	0.96	1.05	1.01	1.02
Mean squared prediction error [h]	2.29	2.51	1.61	2.09	1.87	1.97
			Preterm deliveries			
Bias [g]	-0.24	0.39	-0.47	0.33	2.03	1.99
Absolute bias	1.79	1.92	1.20	1.60	2.49	2.48
Mean squared prediction error [h]	6.93	8.51	2.81	6.38	8.42	8.85
			Term deliveries			
Bias [g]	-0.57	-0.51	-0.33	-0.35	-0.02	-0.01
Absolute bias	1.14	1.15	0.96	1.03	0.95	0.96
Mean squared prediction error [h]	2.11	2.24	1.56	1.90	1.60	1.67
Rate of missingness = 50%						
			All deliveries			
Bias [g]	-0.56	-0.48	-0.34	-0.32	0.06	0.08
Absolute bias	1.16	1.18	0.96	1.05	1.00	1.01
Mean squared prediction error [h]	2.26	2.50	1.60	2.08	1.84	1.94

			Preterm deliveries			
Bias [g]	-0.23	0.41	-0.46	0.33	2.04	1.99
Absolute bias	1.75	1.91	1.20	1.60	2.48	2.48
Mean squared prediction error [h]	6.64	8.50	2.79	6.40	8.36	8.80
			Term deliveries			
Bias [g]	-0.56	-0.52	-0.33	-0.35	-0.02	-0.01
Absolute bias	1.14	1.15	0.96	1.03	0.94	0.95
Mean squared prediction error [h]	2.10	2.24	1.56	1.89	1.57	1.64

[a] Model 2: Generalized estimating equations with autoregressive working correlation matrix, adjusted for maternal age at delivery; maternal history of alcohol consumption; indicator of singleton vs multiple gestation; presence of the preterm birth or still birth indicator; previous parity; history of miscarriages; complications during current pregnancy (pre-eclampsia, other hypertensive disorders during pregnancy, gestational diabetes); and complications during previous pregnancies; body mass index, and maternal history of smoking.

[b] Model 2 and Conventional method 2: GEE model 2 used for records with non-missing covariate information. For deliveries with missing BMI or smoking, gestational age was imputed with a uniform value of 39 or 35 weeks depending on preterm status.

[c] Model 3: Model 2 + fetal birth weight.

[d] Model 3 and Conventional method 2: GEE model 3 used for records with non-missing covariate information. For deliveries with missing BMI, smoking or birth weight, gestational age was imputed with a uniform value of 39 or 35 weeks depending on preterm status.

[e] Birth Weight Method 1: Gestational age set to the median of gestational age for respective birth weight percentile and baby sex. Median gestational age is used for respective birth weight decile when baby sex is missing.

[f] Birth Weight Method 1 and Conventional Method 2: Gestational age set to the median of gestational age for respective birth weight percentile and baby sex. Median gestational age is used for respective birth weight when baby sex is missing. For deliveries with missing birth weight, gestational age was imputed with a uniform value of 39 or 35 weeks depending on preterm status.

[g] Mean bias over 250 simulations; in each iteration, calculated as predicted value minus the observed.

[h] Average prediction error over 250 simulations; in each iteration, calculated as the difference of the observed and predicted values squared.