Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Development and validation of claims-based algorithms to ascertain gestational age in multiple gestations

Oyoung Kim
Yeouido St. Mary's Hospital, College of Medicine, The Catholic University of Korea
Yuki Gen
St. Vincent's Hospital, College of Medicine, The Catholic University of Korea
Younju Lee
PH-DATA Co., Ltd
Jisu Kim
PH-DATA Co., Ltd
Hyun Sun Ko
Seoul St. Mary's Hospital, College of Medicine, The Catholic University of Korea
In Yang Park
Seoul St. Mary's Hospital, College of Medicine, The Catholic University of Korea
Subeen Hong
unihsy@catholic.ac.kr

Seoul St. Mary's Hospital, College of Medicine, The Catholic University of Korea

Research Article

Keywords: International Classification of Diseases, Algorithms, National Health Programs, Pregnancy, Multiple

Posted Date: July 31st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10134401/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

RESEARCH

Development and validation of claims-based algorithms to ascertain gestational age in multiple gestations

Oyoung Kim[1], Yuki Gen[2], Younju Lee[3], Jisu Kim[3], Hyun Sun Ko[4], In Yang Park[4], Subeen Hong[4]*

[1]*Department of Obstetrics and Gynecology, Yeouido St. Mary's Hospital, College of Medicine, The Catholic University of Korea, Seoul, Republic of Korea*

[2]*Department of Obstetrics and Gynecology, St. Vincent's Hospital, College of Medicine, The Catholic University of Korea, Seoul, Republic of Korea*

[3]*PH-DATA Co., Ltd., Seoul, Republic of Korea*

[4]*Department of Obstetrics and Gynecology, Seoul St. Mary's Hospital, College of Medicine, The Catholic University of Korea, Seoul, Republic of Korea*

Correspondence:

Subeen Hong, MD, PhD

Department of Obstetrics and Gynecology, Seoul St. Mary's Hospital, College of Medicine, The Catholic University of Korea, 222, Banpo-daero, Seocho-gu, Seoul, 06591, Republic of Korea

Tel: 82-2-2588-6169; E-mail: unihsy@naver.com

Abstract

Background: Accurate gestational age (GA) assignment is fundamental to pregnancy epidemiologic research but is not directly available in administrative claims databases. Existing claims-based algorithms were developed primarily for singleton pregnancies; and have not been adequately evaluated for multiple gestations, which differ substantially in obstetric characteristics and healthcare utilization patterns. We aimed to develop and validate claims-based GA estimation algorithms for multiple gestations.

Methods: We identified multiple pregnancies with delivery-related claim codes in the National Health Insurance Service (NHIS) database between 2015 and 2024. Reference-standard GAs were obtained through linkage with the National Health Screening Program for Infants and Children (NHSPIC). Cases with a neonatal birth date outside the maternal hospitalization period or a reference GA of <24 or >40 weeks were excluded. The dataset was randomly divided into development and test sets. Algorithm 1 used preterm birth-related ICD-10 codes, Algorithm 2 used prenatal ultrasonography timing, Algorithm 3 used prenatal screening tests timing, and Algorithm 4 combined these approaches hierarchically. Algorithm performances were evaluated using mean absolute errors (MAE), accuracy within ±7 and ±14 days, and missing rate.

Results: A total of 26,714 eligible multiple gestation pregnancies were included in the final cohort. Algorithm 1, based on preterm birth-related diagnostic codes, showed the lowest estimation accuracy (MAE 10.3 days; accuracy within ±7 or ±14 days; of 51.4% and 79.8%, respectively). Algorithm 2, based on first-trimester targeted sonography, showed the highest estimation accuracy (MAE 5.59 days; accuracy within ±7 or ±14 days, of 81.8% and 91.3%, respectively) but had a missing rate of 27.5%. Algorithm 3 had an MAE of 6.42 days, an accuracy within ±7 or ±14 days of 77.4% and 89.2%, respectively, and a missing rate of 9.6%. Algorithm 4, the hierarchical algorithm, achieved complete GA assignment with no assignment failure while maintaining high estimation performance (MAE 6.14 days; accuracy within ±7 or ±14 days of 79.1% and 89.9%, respectively).

Conclusions: A hierarchical framework that sequentially applied ultrasonography, prenatal screening tests, and diagnostic codes achieved complete GA assignment without missing values while maintaining high estimation performance. This framework provides a potential means of estimating GA in future claims-based studies on multiple gestation pregnancies.

Keywords: International Classification of Diseases, Algorithms, National Health Programs, Pregnancy, Multiple

Background

With the increasing availability of nationwide administrative healthcare databases, claims-based pregnancy research has enabled large-scale epidemiologic investigations of maternal and neonatal outcomes, medication safety, and pregnancy-related health conditions [1, 2]. In pregnancy epidemiology, accurate gestational age (GA) assignment is important for appropriately classifying pregnancy timing, fetal and neonatal outcomes, and time-varying exposures. Inaccurate GA estimation may lead to the misclassification of outcomes and the timing of exposures, thereby biasing epidemiologic risk estimates [3, 4]. However, because administrative claims databases are intended primarily for reimbursement rather than detailed clinical characterization, directly recorded GA information is not usually available in claims data and may only be obtained for selected pregnancies through data linkage [5, 6]. Consequently, to estimate gestational timing, GA is often estimated using indirect indicators, such as diagnostic codes, procedure codes, and the timing of prenatal examinations. Although several studies have developed and validated claims-based GA assignment algorithms, most were conducted in overall pregnancy populations, and thus, their applicability to multiple gestations remains uncertain [5-7]. Multiple gestations differ substantially from singleton pregnancies in terms of preterm birth rates, delivery timing, antenatal surveillance,

and healthcare utilization patterns. These differences may affect coding patterns and healthcare utilization captured in administrative databases and potentially limit the performance of algorithms developed using singleton-dominant populations [8-10]. Therefore, validated GA assignment methods are needed specifically for multiple gestations.

In this study, we developed and validated claims-based algorithms for GA assingment in multiple gestations using a nationwide population-based healthcare database.

Methods

Data Source

This study used the Korean National Health Insurance Service (NHIS) database, a nationwide administrative healthcare database covering 97% of the Korean population. The database contains healthcare utilization records, including International Classification of Diseases, 10th Revision (ICD-10) diagnostic codes, details of procedures, examinations, and prescription records with corresponding dates, as well as inpatient and outpatient claims data [11]. Algorithms were developed and validated using linked infant health examination records from the National Health Screening Program for Infants and Children (NHSPIC), including date of birth and estimated date of confinement (EDC) and enabled the derivation of the reference-standard GA [12]. Although GAs could be derived from NHSPIC-linked information, such linkage is not routinely available in claims data, and when available, may be limited by live-birth restriction, delayed or restricted linkage, incomplete EDC entries, and potential recall or transcription errors.

Therefore, NHSPIC-derived GA was used as the reference standard for developing and validating claims-based algorithms. The hierachical algorithm (Algorithm 4) was designed to provide a transparent and reproducible approach for future studies using administrative claims databases in which direct GA information is unavailable, incomplete, or cannot be linked.

Study Population and Definition of Delivery Episodes

We identified all multiple gestation pregnancies with delivery-related procedure codes from the NHIS database between January 1, 2015, and December 31, 2024. The delivery-related procedure codes are listed in Supplementary Table 1. Following linkage with NHSPIC records, pregnancies with missing EDC or neonatal birth date information were excluded. We also excluded pregnancies if the neonatal birth date occurred outside the maternal delivery hospitalization period or if the reference GA fell outside the plausible range of 24–40 weeks. A total of 26,714 multiple gestation pregnancies were included in the study. The cohort selection process is summarized in Figure 1.

Derivation of Reference and Claims-Based Gestational Age at Delivery

Reference standard GAs were derived using the EDC and neonatal birth dates recorded in the NHSPIC. Assuming a standard gestational duration of 280 days, the last menstrual period (LMP) was estimated as 280 days before the EDC, and the reference GA was defined as the time between the estimated LMP and the neonatal birth date.

Claims-based algorithms were developed to estimate the LMP using delivery-related diagnoses, prenatal ultrasonography, and prenatal screening tests. Estimated GA was calculated as the interval between the neonatal birth date and the algorithm-derived LMP. Algorithm performance was evaluated by comparing the estimated GA with the reference GA. The derivation and application of representative GA values for claims-based GA assignment are illustrated in Figure 2.

Dataset Splitting and the Derivation of Representative Gestational Age Values

The final analytic cohort was randomly and approximately equally divided into development and test sets. The development set was used for algorithm development and the derivation of representative GA values for claims-based indicators, whereas the independent test set was reserved for algorithm validation.

Algorithm Development

Four claims-based algorithms were developed for GA assignment at delivery. The diagnosis and procedure codes used in this study are provided in Supplementary Table 1.

Algorithm 1: ICD-10 Code–Based Algorithm

Algorithm 1 used preterm birth (PTB)-related diagnostic codes to assign representative GA values. Maternal preterm delivery and neonatal prematurity-related diagnosis codes used in this algorithm are listed in Supplementary Table 1. In the development set, pregnancies with or without PTB-related codes, despite a reference-standard GA of ≥37 weeks or GA <37 weeks respectively, were excluded from the derivation of representative GA values. Representative GA values were then defined as the median reference-standard GA among pregnancies with and without PTB-related diagnostic codes.

Algorithm 2: Sonography-Based Algorithm

Algorithm 2 estimated GAs using the timing of prenatal ultrasonography claims recorded within predefined gestational windows. For each ultrasonography category, the median reference-standard GA at the time of examination was calculated in the development set and assigned as the representative GA for examinations within that category. The evaluated ultrasonography-based approaches included first-trimester sonography, first-trimester targeted sonography, second-/third-trimester sonography, and second-/third-trimester targeted sonography. When multiple examinations were perfomed within the same category, only the earliest examination date was used. Based on estimation performance in the development set, first-trimester targeted sonography was selected for the final sonography-based algorithm.

Algorithm 3: Prenatal Screening–Based Algorithm

Algorithm 3 estimated GA using the timing of prenatal screening claims recorded within predefined gestational windows. The evaluated prenatal screening tests included human chorionic gonadotropin (hCG), maternal serum alpha-fetoprotein (MSAFP), unconjugated estriol (uE3), inhibin A, and the 50-g oral glucose tolerance test (OGTT). For each screening code, a representative GA value was derived from the median reference-standard GA at the time of the examination in the development set. Estimation performance for individual screening tests and their combinations was evaluated

against the reference-standard GA data in the development set. Based on estimation performance, the combination of MSAFP, uE3, and inhibin A was selected for the final prenatal screening–based algorithm.

Algorithm 4: Hierarchical Algorithm

Algorithm 4 was developed as a hierarchical approach to minimize missing values while maintaining estimation accuracy. The missing rate was defined as the proportion of pregnancies for which GA could not be estiaated because of unavailable claims-based indicators. Claims-based indicators were sequentially applied according to their performance in the development set, with the best-performing algorithm given priority. The sonography-based algorithm was applied first, followed by the prenatal screening–based algorithm and, finally, the ICD-10 code–based algorithm when no preceding indicator was available.

Outcome Measures

The primary outcome was the accuracy of GA estimation against the reference GA. Mean absolute error (MAE) was defined as the mean absolute difference between estimated and reference-standard GA vales. As performed in previous validation studies, estimation accuracy was evaluated based on whether the estimated GA differed from the reference-standard GA by within ±7 or ±14 days [13, 14]. The missing rate was defined as the proportion of pregnancies for which GA could not be estimated because of unavailable claims-based indicators.

Statistical Analysis

Esitimated GA distributions were summarized using means, standard deviation (SD), median, and interquartile range (IQR). Differences between algorithm-derived MAEs were assessed using the Friedman test, followed by Bonferroni-adjusted Wilcoxon signed-rank tests for pairwise comparisons. MAE comparisons across Algorithms 1–3 were restricted to pregnancies with available estimates ; Algorithm 4 was excluded because it hierarchically integrated Algorithms 1-3.

Results

Study population and construction of the reference dataset

A total of 46,883 multiple gestation pregnancies with delivery-related procedure codes were identified in the NHIS database, and these included 46,066 twin pregnancies, 792 triplet pregnancies, and 25 quadruplet pregnancies. To construct the reference standard GA dataset, maternal records were linked to NHSPIC records. After linkage, 28,054 pregnancies had available reference standard GA information; Pregnancies with a missing estimated delivery date (n = 18,791) or neonatal birth date (n = 38) were excluded. In addition, 1,340 pregnancies were excluded for failing to meet study eligibility criteria. These included 989 pregnancies without a neonatal birth date recorded within the maternal hospitalization period and 351 pregnancies with a reference GA outside the 24–40 week range. The final analytic cohort consisted of 26,714 pregnancies, and included 26,064 twin pregnancies and 650 higher-order multiple pregnancies (Figure 1). The cohort was randomly divided into development and test datasets (13,358 and 13,356 pregnancies, respectively).

Mean GA at delivery was 35 weeks and 5 days (SD, 2 weeks and 2 days) in the overall cohort. Mean GA was 35 weeks and 5 days (SD, 2 weeks and 2 days) for twin pregnancies and 33 weeks and 5 days (SD, 2 weeks and 3 days) for higher-order multiple pregnancies. (Table 1).

Derivation of Representative Gestational Age Values and Algorithm Selection

Representative GA values and the performance of candidate approaches evaluated using the development dataset are summarized in Table 2.

For Algorithm 1, 2,943 pregnancies (22.0%) with discordant PTB-related diagnostic codes and reference standard GA were excluded from the derivation of representative GA values. For pregnancies with PTB-related diagnostic codes, the median GA was 35 weeks and 2 days, compared with 37 weeks and 3 days for those without PTB-related diagnostic codes.

For Algorithm 2, representative GA values were 5 weeks and 3 days for first-trimester sonography, 12 weeks and 1 day for first-trimester targeted sonography, 16 weeks for second-/third-trimester sonography, and 21 weeks and 3 days for second-/third-trimester targeted sonography. For the

ultrasonography-based approaches evaluated, first-trimester targeted sonography had the best overall performance in the development set and was therefore selected for Algorithm 2

For Algorithm 3, representative GA values were derived from the median reference-standard GA associated with each prenatal screening test. The representative GA values were 15 weeks and 2 days for human chorionic gonadotropin (hCG), 16 weeks and 1 day for maternal serum alpha-fetoprotein (MSAFP), unconjugated estriol (uE3), and inhibin A, and 24 weeks and 2 days for the 50-g oral glucose tolerance test (OGTT). Of the evaluated prenatal screening tests and test combinations, the combination of MSAFP, uE3, and inhibin A showed the best overall performance in the development set and was therefore selected for Algorithm 3. The performance of the candidate approaches was evaluated separately for twin and higher-order multiple pregnancies, as shown in Supplementary Table 2.

Validation of Gestational Age Assignment Algorithms

Representative GA values derived in the development set were applied to the independent test set for validation; algorithm performance is summarized in Table 3.

Algorithm 1, based on PTB-related diagnostic codes, showed the lowest estimation performance, with an MAE of 10.3 days, accuracy within ±7 days of 51.4%, and accuracy within ±14 days of 79.8%; No missing values were observed.

Algorithm 2, based on first-trimester targeted sonography, showed the best overall estimation performance, with an MAE of 5.59 days, accuracy within ±7 days of 81.8%, and accuracy within ±14 days of 91.3%. However, this algorithm had the highest missing rate (27.5%).

Algorithm 3, based on prenatal screening tests, had an MAE of 6.42 days, accuracy within ±7 days of 77.4%, accuracy within ±14 days of 89.2%, and a missing rate of 9.6%.

Analyses were restricted to pregnancies with available GA estimates (n = 8,777) to compare algorithm estimation errors. MAE differences of Algorithms 1–3 were statistically significant

(Friedman test, $P < 0.0001$), and all pairwise comparisons remained significant after the Bonferroni correction (Supplementary Table 4).

Algorithm 4, the hierarchical algorithm combining Algorithms 1–3, achieved complete GA assignment without missing values. In the hierarchical approach, Algorithm 4 attempted the component algorithms in a fixed order: Algorithm 2 first, followed by Algorithm 3, and finally Algorithm 1 if needed. Using this sequence, Algorithm 2 provided the GA estimate in 72.5% of pregnancies, Algorithm 3 in 24.7%, and Algorithm 1 in the remaining 2.9%. Algorithm 4 had an MAE of 6.14 days, accuracy within ±7 days of 79.1%, and accuracy within ±14 days of 89.9%. While its estimation performance was slightly lower than that of Algorithm 2, Algorithm 4 had a performance comparable to Algorithm 3 and eliminated missing estimates. The validation performance of the candidate approaches was also evaluated separately for twin and higher-order multiple pregnancies, as shown in Supplementary Table 3.

Discussion

In this nationwide claims-based study of multiple gestation pregnancies, we developed and validated four GA assignment algorithms using nationwide linked administrative healthcare data. Algorithm 2, based on first-trimester targeted sonography, had the highest estimation accuracy, whereas Algorithm 4, a hierarchical algorithm, achieved complete GA assignment without missing values and a high overall performance. These findings provide a validated approach for GA assignment in multiple gestation pregnancies and may support future population-based research using administrative healthcare databases.

Given the increasing prevalence of multiple gestations associated with assisted reproductive technologies, including in vitro fertilization, accurate GA assignment in this population is becoming increasingly important in pregnancy epidemiologic research [15-17].

Multiple gestations are associated with earlier delivery timing and substantially higher preterm birth rates than singleton pregnancies [8, 10, 18]. Representative GA values derived from PTB-related

diagnostic code classifications in our multiple gestation cohort were shifted toward earlier GAs compared with those reported in previous claims-based studies. This difference was primarily driven by pregnancies without PTB-related diagnostic codes, for which the representative GA was 37 weeks and 3 days in our cohort, compared with 39 weeks reported by Margulis et al. and 38 weeks and 6 days in a recent NHIS-based study [5, 19]. These findings suggest that applying representative GA values derived from singleton-dominant populations to multiple gestations may introduce systematic gestational timing misclassifications, which highlights the need for population-specific validation of claims-based GA assignment approaches for multiple gestation pregnancies.

Of the evaluated approaches, Algorithm 2, based on first-trimester targeted ultrasonography, showed the highest estimation accuracy. Similar findings have been reported in previous claims-based GA assignment studies [19]. The superior performance of this approach likely reflects the highly standardized timing of first-trimester targeted ultrasonography, which is performed within a narrowly defined GA range for nuchal translucency measurement as part of first-trimester aneuploidy screening, making it a more precise indicator of gestational timing than other ultrasonographic examinations.

Algorithm 3, based on prenatal screening tests, also demonstrated favorable performance, particularly with the combined use of MSAFP, uE3, and inhibin A. Although OGTT is recommended within a defined GA range, its timing is influenced by maternal risk factors and clinical circumstances, reducing its specificity as a GA marker [20]. In contrast, MSAFP, uE3, and inhibin A are closely linked to aneuploidy screening, which must be performed within specific GA ranges for accurate interpretation, making these markers more specific indicators of gestational timing. In contrast, despite being a component of the standard quad screening test, hCG had a lower estimation accuracy than the other markers, which may have been because hCG is frequently measured for clinical indications beyond aneuploidy screening, such as early pregnancy viability assessment and ectopic pregnancy evaluation, reducing its specificity as a GA indicator.

Algorithm 1, which relied solely on PTB-related diagnostic codes, had a lower estimation accuracy than the other algorithms. Unlike procedure- or test-based indicators, diagnostic codes reflect broad clinical categories rather than specific gestational milestones and therefore provide less precise information regarding gestational timing.

Algorithm 4, the hierarchical algorithm that integrated all available claims-based information, achieved a favorable balance between estimation accuracy and data completeness. Although its estimation accuracy was slightly lower than that of Algorithm 2 alone, it eliminated missing estimates while maintaining high overall performance. Because the exclusion of pregnancies with missing GA information may reduce statistical power and introduce selection bias, complete GA assignment is an important consideration in large administrative database studies. In this regard, the hierarchical algorithm offers a pragmatic approach that balances estimation accuracy with data completeness.

Strengths and Limitations

The use of linked nationwide healthcare claims and infant screening databases, which enabled the development and validation of GA assignment algorithms against reference standard GA information, is a major strength of this study. In addition, more than 26,000 multiple gestation pregnancies were included, and to the best of our knowledge, this represents one of the largest cohorts specifically assembled for the development and validation of claims-based GA assignment algorithms for multiple gestations.

Most previous claims-based GA assignment studies were conducted in overall pregnancy populations and did not specifically focus on multiple gestation pregnancies. By addressing this underrepresented population, our study helps fill an important evidence gap in claims-based pregnancy research.

Another important strength is that the present study evaluated three complementary categories of claims-based indicators; prenatal ultrasonography, prenatal screening tests, and diagnostic coding

patterns, allowing direct comparison of different GA assignment strategies within the same population. The hierarchical framework that integrated these indicators, according to estimation accuracy and data availability, achieved a favorable balance between estimation accuracy and data completeness. By prioritizing first-trimester targeted ultrasonography when available and subsequently incorporating prenatal screening and diagnostic code information, the framework allowed clinically precise gestational markers to be preferentially utilized while preserving population coverage through secondary claims-based indicators.

An additional advantage of the proposed framework is its clinical interpretability. Unlike more complex machine learning approaches, which may improve estimation accuracy but often function as black boxes and are difficult to implement or validate across different healthcare databases, the present algorithms rely on straightforward clinical events that are easy to identify and interpret, including prenatal ultrasonography, prenatal screening tests, and diagnostic codes. [21]. Consequently, each GA estimate can be directly linked to a specific healthcare utilization record, facilitating transparency, reproducibility, and assessment of potential sources of bias and misclassification. Furthermore, because the proposed framework relies on routinely collected prenatal care indicators and reflects the general temporal structure of prenatal care rather than features unique to a specific healthcare system, this clinically interpretable approach may be adaptable to other administrative healthcare databases following local validation and calibration.

Despite these strengths, several limitations should be considered. First, the algorithms were developed using the Korean NHIS system and thus may partially reflect characteristics specific to Korean prenatal care utilization and claims coding practices. In particular, the relatively high utilization of prenatal care services in South Korea may have increased the availability of ultrasonography and prenatal screening claims used in the present algorithms. Therefore, external validation is warranted before applying these algorithms in other international healthcare systems.

Second, reference-standard GA information was available only for pregnancies successfully linked to infant screening records. Consequently, the study population may not fully represent all multiple gestation pregnancies captured in the claims database.

Third, as with all administrative claims databases, coding inaccuracies, delayed claim submissions, institutional variation in coding practices, and residual misclassification of claims-based indicators may have affected algorithm performance.

Conclusions

Finally, the present study focused primarily on GA assignmnet at delivery and did not evaluate longitudinal GA reconstruction across pregnancy. In addition, because validation relied on linkage to infant screening records, the proposed algorithms were evaluated primarily for live births and may therefore not be directly generalizable to pregnancies ending in miscarriage or stillbirth.

In conclusion, we developed and validated claims-based GA assigment algorithms for multiple gestation pregnancies using nationwide linked administrative healthcare data. Among the evaluated approaches, first-trimester targeted ultrasonography demonstrated the highest estimation accuracy, whereas the hierarchical algorithm achieved a favorable balance between estimation accuracy and data completeness by enabling complete GA assignment without missing values. These findings highlight the importance of population-specific validation of GA assignment methods in multiple gestation pregnancies and provide a practical framework for future claims-based research in this population.

List of abbreviations

GA: Gestational age

PTB : Preterm birth

MAE : Mean absolute errors

EDC: Estimated date of confinement

LMP: Last menstrual period

Declarations

Ethics approval and consent to participate

This study was approved by the Institutional Review Board of Seoul St. Mary's Hospital, The Catholic University of Korea (No. KC24ZISI0746). The requirement for informed consent was waived by the Institutional Review Board of Seoul St. Mary's Hospital, The Catholic University of Korea, because this study used de-identified administrative claims data. All methods were performed in accordance with the relevant guidelines and regulations, including the Declaration of Helsinki.

Consent for publication

Not applicable

Availability of data and materials

The datasets generated and/or analyzed during the current study are not publicly available because they were obtained from the National Health Insurance Service of Korea under restricted access. Data

may be accessed from the National Health Insurance Service of Korea after appropriate approval and permission.

Competing interests

The authors report no conflicts of interest.

Funding

This research was supported by a grant of Patient-Centered Clinical Research Coordinating Center (PACEN) funded by the Ministry of Health and Welfare, Republic of Korea (grant number: RS-2025-25410838)

Authors' contributions

OK: Writing – original draft, methodology, investigation, formal analysis, and visualization

YL: Data curation and investigation

JK: Data curation and investigation

YG: Methodology, Data curation and investigation

HSK: Data curation, and investigation

IYP: Data curation, and investigation

SH: Writing – review and editing, conceptualization, formal analysis, funding acquisition, methodology, supervision, and project administration

Acknowledgements

The authors acknowledge financial support from the Catholic Medical Center Research Foundation made in the program year of 2026

Reference

1. Ailes EC, Simeone RM, Dawson AL, Petersen EE, Gilboa SM. Using insurance claims data to identify and estimate critical periods in pregnancy: An application to antidepressants. Birth Defects Res A Clin Mol Teratol. 2016;106:927–34.
2. Andrade SE, Berard A, Nordeng HME, Wood ME, van Gelder M, Toh S. Administrative Claims Data Versus Augmented Pregnancy Data for the Study of Pharmaceutical Treatments in Pregnancy. Curr Epidemiol Rep. 2017;4:106–16.
3. Savitz DA, Dole N, Herring AH. Methodologic issues in the design and analysis of epidemiologic studies of pregnancy outcome. Stat Methods Med Res. 2006;15:93–102.
4. Matcho A, Ryan P, Fife D, Gifkins D, Knoll C, Friedman A. Inferring pregnancy episodes and outcomes within a network of observational databases. PLoS One. 2018;13:e0192033.
5. 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;22:16–24.
6. Zhu Y, Hampp C, Wang X, Albogami Y, Wei YJ, Brumback BA, et al. Validation of algorithms to estimate gestational age at birth in the Medicaid Analytic eXtract-Quantifying the misclassification of maternal drug exposure during pregnancy. Pharmacoepidemiol Drug Saf. 2020;29:1414–22.
7. Li Q, Andrade SE, Cooper WO, Davis RL, Dublin S, Hammad TA, et al. Validation of an algorithm to estimate gestational age in electronic health plan databases. Pharmacoepidemiol Drug Saf. 2013;22:524–32.
8. American College of Obstetricians and Gynecologists. Multifetal Gestations: Twin, Triplet, and Higher-Order Multifetal Pregnancies: ACOG Practice Bulletin, Number 231. Obstet Gynecol. 2021;137:e145–e62.

9. Monden C, Pison G, Smits J. Twin Peaks: more twinning in humans than ever before. Hum Reprod. 2021;36:1666–73.
10. Kim O, Lee HY, Ko HS, Park IY, Hong S. Predictive value of a short cervix in the early third trimester for preterm birth in twin pregnancies. BMC Pregnancy Childbirth. 2026.
11. Seong SC, Kim YY, Khang YH, Heon Park J, Kang HJ, Lee H, et al. Data Resource Profile: The National Health Information Database of the National Health Insurance Service in South Korea. Int J Epidemiol. 2017;46:799–800.
12. Kim JH, Lee JE, Shim SM, Ha EK, Yon DK, Kim OH, et al. Cohort profile: National Investigation of Birth Cohort in Korea study 2008 (NICKs-2008). Clin Exp Pediatr. 2021;64:480–8.
13. Moll K, Wong HL, Fingar K, Hobbi S, Sheng M, Burrell TA, et al. Validating Claims-Based Algorithms Determining Pregnancy Outcomes and Gestational Age Using a Linked Claims-Electronic Medical Record Database. Drug Saf. 2021;44:1151–64.
14. Chiu YH, Huybrechts KF, Zhu Y, Straub L, Bateman BT, Logan R, et al. Internal validation of gestational age estimation algorithms in health-care databases using pregnancies conceived through fertility procedures. Am J Epidemiol. 2024;193:1168–75.
15. Sunderam S, Kissin DM, Zhang Y, Jewett A, Boulet SL, Warner L, et al. Assisted Reproductive Technology Surveillance - United States, 2018. MMWR Surveill Summ. 2022;71:1–19.
16. Choi SH, Park YS, Shim KS, Choi YS, Chang JY, Hahn WH, et al. Recent trends in the incidence of multiple births and its consequences on perinatal problems in Korea. J Korean Med Sci. 2010;25:1191–6.
17. Brown JP, Yland JJ, Williams PL, Huybrechts KF, Hernandez-Diaz S. Accounting for Twins and Other Multiple Births in Perinatal Studies of Live Births Conducted Using Healthcare Administration Data. Epidemiology. 2025;36:165–73.

18. Hong JY, Lee HR, Kim Y, Kim YM, Sung JH, Choi SJ, et al. Changes in the perinatal outcomes of twin pregnancies delivered at a tertiary referral center in Korea during a 24-year period from 1995 to 2018. Obstet Gynecol Sci. 2020;63:251–60.
19. Kim HJ, Cho H, Jung SY, Shin JY, Choe SA, Choi NK. Evaluating and Refining Claims-Based Algorithms for Pregnancy Outcomes and Gestational Age Estimation in Korea. J Korean Med Sci. 2026;41:e138.
20. American College of Obstetricians and Gynecologists. ACOG Practice Bulletin No. 190: Gestational Diabetes Mellitus. Obstet Gynecol. 2018;131:e49–e64.
21. Rudin C. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. Nat Mach Intell. 2019;1:206–15.

Table 1. Gestational age distribution in the reference dataset by plurality

	Overall	Twin	≥ Triplet
Number of pregnancies	26,714	26,064	650
Mean GA at delivery , weeks (SD)	35w5d (2w2d)	35w5d (2w2d)	33w5d (2w3d)
GA at delivery, n (%)			
24–27 weeks	345 (1.3)	330 (1.3)	15 (2.3)
28–31 weeks	1,532 (5.7)	1,437 (5.5)	95 (14.6)
32–36 weeks	18,707 (70.0)	18,191 (69.8)	516 (79.4)
≥37 weeks	6,130 (22.9)	6,106 (23.4)	24 (3.7)

Data are presented as n (%) unless otherwise indicated. Gestational age is presented as mean gestational age and expressed as completed weeks and days.

The ≥ triplet category includes triplet and higher-order multiple pregnancies. Abbreviations: GA, gestational age; SD, standard deviation

Table 2. Representative gestational age values and the performance of claims-based gestational age assignment approaches in the development set

	Representative GA (median)	MAE (days)	Accuracy within ±7 days (%)	Accuracy within ±14 days (%)	Missing rate (%)
ICD-10–based approach		9.67	55.6	81.3	22.0
PTB-related code present	35w2d				
PTB-related code absent	37w3d				
Ultrasonography-based approaches					
First-trimester sonography	5w3d	9.24	59.4	83.0	25.7
First-trimester targeted sonography	12w1d	5.69	81.6	91.1	27.5
Second-/third-trimester sonography	16w0d	12.25	59.9	82.5	19.9
Second-/third-trimester targeted sonography	21w3d	11.60	49.9	72.5	23.9
Mean Across Ultrasonography		10.39	69.5	83.5	19.5
Prenatal screening–based approaches					
hCG	15w2d	42.87	14.0	43.7	8.3
MSAFP	16w1d	6.64	77.2	89.2	12.9
uE3	16w1d	6.42	78.3	89.6	23.2
Inhibin A	16w1d	6.15	78.2	89.9	18.5
50g OGTT	24w2d	48.45	22.1	42.4	1.0
Combined prenatal screening markers (MSAFP, uE3, and Inhibin A)		6.49	77.1	89.1	10.0

Representative GA values were derived from the median GAs observed in the development set. Abbreviations: GA, gestational age; MAE, mean absolute error; PTB, preterm birth; MSAFP, maternal serum alpha-fetoprotein; hCG, human chorionic gonadotropin; uE3, unconjugated estriol; OGTT, oral glucose tolerance test.

Table 3. Validation performance of claims-based gestational age assignment algorithms in the test set

Algorithm	MAE (days)	Accuracy within ±7 days (%)	Accuracy within ±14 days (%)	Missing rate (%)
Algorithm 1 (ICD-10 code–based)	10.3	51.4	79.8	0.0
Algorithm 2 (ultrasonography-based)	5.59	81.8	91.3	27.5
Algorithm 3 (prenatal screening-based)	6.42	77.4	89.2	9.6
Algorithm 4 (Hierarchical algorithm)	6.14	79.1	89.9	0.0

MAE was defined as the mean absolute difference between the estimated and reference GAs. Accuracies within ±7 and ±14 days are presented as proportions of estimates within 7 and 14 days, respectively, of reference GAs. Missing rate was defined as the proportion of pregnancies for which GA could not be estimated by the algorithms. Abbreviations: MAE, mean absolute error; ICD-10, International Classification of Diseases, 10th Revision

Figure Legends

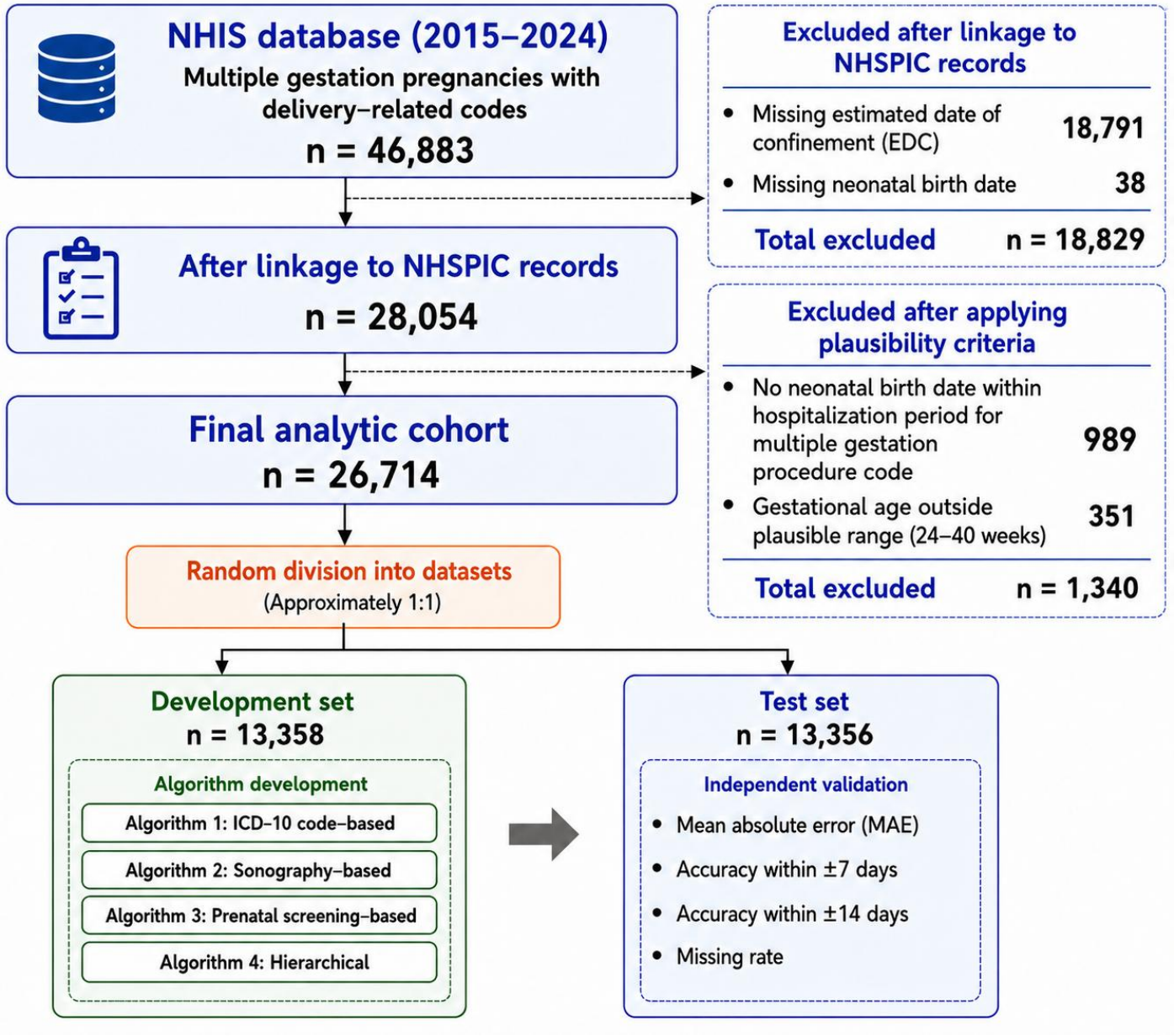


Figure 1. Flow diagram of cohort selection and study design.

Multiple gestation pregnancies were identified from the NHIS database and linked to NHSPIC records, which were only used to construct the reference-standard GA dataset and not for the implementation of the final claims-based algorithms. After exclusions, the final cohort was randomly divided into development and test datasets for algorithm development and validation. Abbreviations: NHIS, National Health Insurance Service; NHSPIC, National Health Screening Program for Infants and Children; ICD-10, International Classification of Diseases, 10th Revision

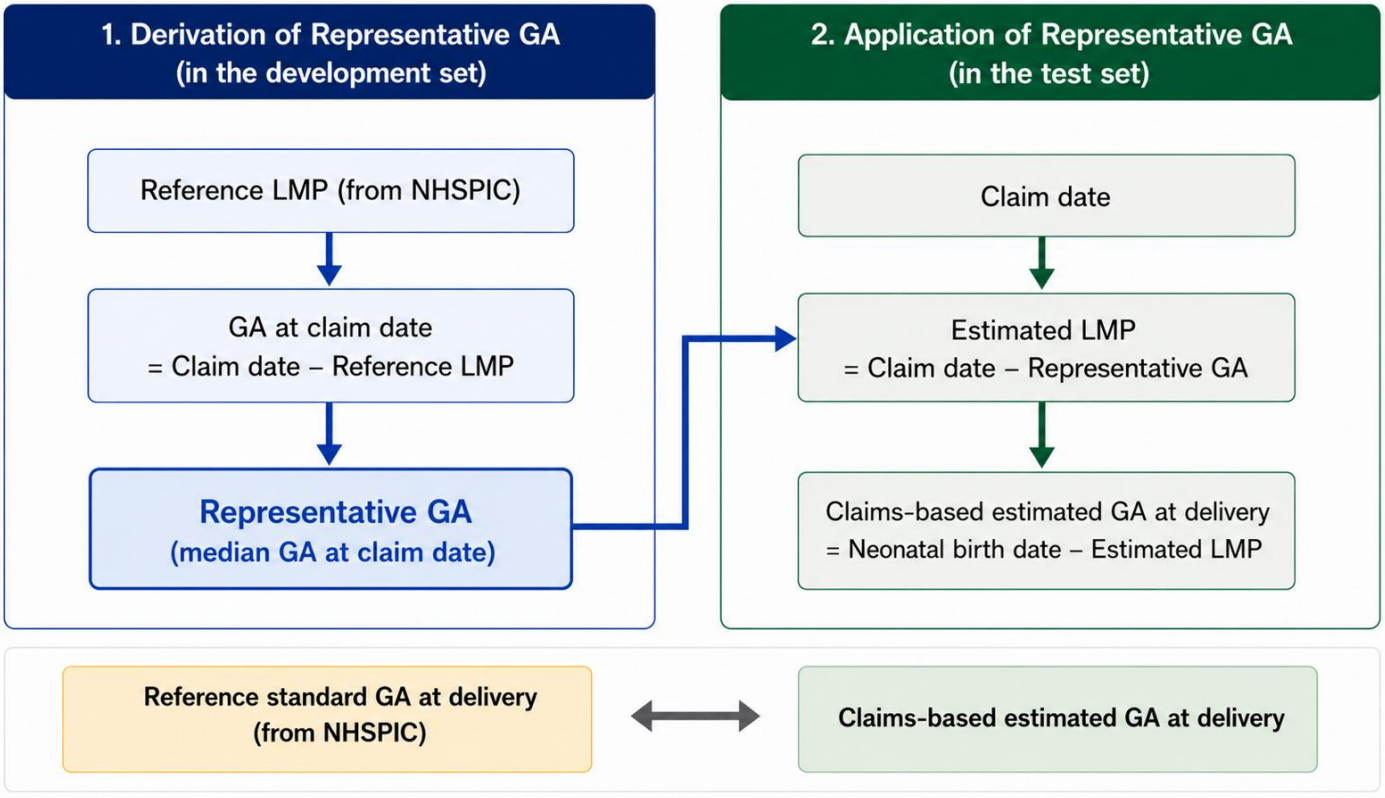


Figure 2. Derivation and application of representative gestational age values for claims-based gestational age assignment

Representative GA values were derived in the development set and subsequently applied to the test set to estimate GA at delivery. Performance was evaluated by comparing claims-based estimated GA at delivery with reference-standard GA at delivery. Abbreviations: NHSPIC, National Health Screening Program for Infants and Children; GA, gestational age; LMP, last menstrual period.

Additional file 1

File name: Supplementary Table_1.docx

Title: Supplementary Table 1. Claims-based codes used for gestational age estimation

Additional file 2

File name: Supplementary Table_2.docx

Title: Supplementary Table 2. Performance of claims-based gestational age assignment approaches in the development set, stratified by multiplicity

Additional file 3

File name: Supplementary Table_3.docx

Title: Supplementary Table 3. Validation performance of claims-based gestational age assignment algorithms in the test set by multiplicity

Additional file 4

File name: Supplementary Table_4.docx

Title: Supplementary Table 4. Comparison of gestational age estimation errors among three assignment algorithm

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryTable1.docx
- SupplementaryTable2.docx
- SupplementaryTable3.docx
- SupplementaryTable4.docx