Paper RE07

An Open-Source R Implementation of a Validated Claims-Based Algorithm to Identify Pregnancy Episodes

Dhirishiya P, GSK, Bangalore, India
Onkar Kshirsagar, GSK, Bangalore, India

ABSTRACT

Data on medicine and vaccine safety during pregnancy are typically generated post-launch yet post-marketing studies often begin without a clear understanding of medication uptake or characteristics of exposed pregnant individuals. This limits the evidence available to guide prescribing decisions and design robust studies. To support standardized and transparent evidence generation, we translated a validated pregnancy identification algorithm implemented in SAS to an open-source R package, “pregfindr”. The algorithm uses diagnosis and procedure codes in IBM MarketScan® (now Merative) claims data to identify pregnancy episodes, estimate gestational timing, assign trimesters, and classify pregnancy outcomes, including pregnancy losses. We compared pregfindr’s intermediate and episode outputs with the SAS outputs on the same Marketscan extracts. The algorithm has supported regulatory discussions by providing data on the number of pregnant women potentially exposed to a medication for product labelling. It enables safety outcomes assessment during pregnancy, both overall and stratified by timing of exposure during gestation, supporting benefit-risk evaluations and optimized study design. This paper describes the motivation, implementation and practical considerations of the open-source R package and discusses its role in enabling consistent real world evidence generation in pregnancy research.

INTRODUCTION

Pregnant women represent a population for whom evidence on medication and vaccine safety is often limited at the time of product approval. Randomized clinical trials rarely include pregnant women, making real world data sources such as healthcare claims and electronic health records critical for post marketing safety evaluation. However, claims data-based pregnancy research is challenging due to incomplete gestational age information, variability in coding practices, and inconsistent capture of pregnancy outcomes.

Several algorithms have been proposed to identify pregnancies and gestational timing in claims data, but differences in implementation, coding systems, and analytic assumptions can lead to inconsistent results across studies. These differences may affect pregnancy episode counts, estimated gestational timing, trimester assignment, and classification of pregnancy outcomes, particularly when conflicting or incomplete codes are present.

To address this, a standardized pregnancy identification algorithm was developed and evaluated using MarketScan® Commercial and Medicaid data. While analytically robust, the original implementation relied on proprietary SAS code, limiting transparency, reproducibility and reuse.

To improve accessibility and reproducibility, we translated the existing SAS based pregnancy algorithm into R and released it as an open-source package. This aligns with broader efforts to promote transparent, reproducible real-world analytics and support consistent application of established pregnancy algorithms across research teams and studies. In this paper we walk through the algorithm workflow, explain the implementation choices in pregfindr, and describe how we checked it matched the SAS reference.

BACKGROUND AND EVIDENCE GAP

Post-marketing pregnancy safety studies often require early estimates of exposure prevalence, timing of exposure by trimester, and pregnancy outcomes. Without a standardized approach to pregnancy identification, studies may rely on ad hoc definitions that are difficult to compare, reproduce or validate. Inconsistent pregnancy episode construction can affect estimates of exposure windows, outcome incidence, and ultimately study conclusions.

The original pregnancy algorithm addressed many of these challenges by combining diagnosis, procedure, and gestational age codes with clinically informed hierarchies to resolve conflicting information. Validation against published literature demonstrated that the algorithm produced clinically plausible pregnancy outcome distributions and medication exposure estimates consistent with established population-level evidence. However, broader adoption was constrained by the lack of openly available implementation.

OVERVIEW OF THE PREGNANCY ALGORITHM

ADDRESSING INCONSISTENCIES IN PREGNANCY-RELATED CLAIMS RECORDS

A key challenge in pregnancy research using claims data is the presence of multiple and sometimes conflicting indicators of pregnancy timing and outcomes recorded across healthcare encounters. Claims data are generated for billing rather than research purposes, and pregnancy-related codes may appear in different care settings, reflect varying levels of specificity, or represent follow-up care rather than the underlying pregnancy event.

The pregnancy algorithm addresses this challenge by applying predefined hierarchical rules to resolve conflicts in pregnancy timing and outcome information. Rather than relying on a single code or encounter, the algorithm evaluates pregnancy-related records occurring on the same service date and applies ordered decision rules to assign a single trimester classification and outcome. This approach ensures that each pregnancy episode is represented by a coherent and clinically plausible set of attributes, even when the underlying claims data contain overlapping or inconsistent signals.

For pregnancy timing, the algorithm prioritizes direct gestational age information when available and applies outcome- or trimester-based assumptions only when such information is missing. Temporal guardrails are applied to ensure biologically plausible episode construction, such as enforcing minimum separation between distinct pregnancy episodes and preventing postpartum or follow-up care records occurring shortly after a pregnancy outcome from initiating a new episode. These rules help distinguish true new pregnancies from care related to a completed pregnancy and avoid false episode creation driven by administrative coding practices. By explicitly encoding these decisions, the algorithm provides a reproducible and transparent framework for pregnancy episode construction in administrative claims data.

Based on these considerations, the pregnancy algorithm applies a sequence of hierarchical steps to construct coherent pregnancy episodes from claims data, as described in the below section.

ALGORITHM WORKFLOW

The pregnancy identification algorithm structures raw claims data into discrete pregnancy episodes through a sequence of rule-based steps designed to address the inherent limitations of administrative healthcare data. Claims databases are not collected for pregnancy research purposes and frequently contain incomplete, overlapping, or conflicting information related to gestational timing and pregnancy outcomes. The algorithm applies clinically informed hierarchies to resolve these ambiguities in a consistent and reproducible manner.

Identify Pregnancies:

Pregnancy-related records are identified using diagnosis and procedure codes indicative of pregnancy, delivery, or pregnancy loss. These records serve as anchors for episode construction but may not, on their own, provide sufficient information to define gestational timing.

Assign Trimester:

Trimester assignment is performed by mapping pregnancy-associated codes to gestational periods. When multiple trimester-indicative codes occur on the same service date, a predefined trimester hierarchy is applied to assign a single trimester that reflects the most specific and reliable information. For example, if a first-trimester ultrasound code occurs on the same day as a non-specific trimester indicator, the algorithm assigns the episode to the first trimester, prioritizing the more clinically informative signal.

Determine Pregnancy Outcome:

Pregnancy outcomes are assigned by applying an outcome hierarchy to resolve conflicting outcome codes recorded on the same day. This step ensures that each pregnancy episode is associated with a single outcome classification, even when multiple outcome-related codes are present in the data. For example, if spontaneous abortion and stillbirth codes occur on the same day, the episode is classified as spontaneous abortion. The hierarchy reflects clinically meaningful ordering of outcomes and is consistent with the published methodology.

Calculate Pregnancy Length:

Pregnancy start and end dates are estimated using a predefined hierarchy that combines gestational age codes, trimester assignments, and outcome-specific assumptions. When gestational week codes are available, pregnancy start dates are calculated based on the latest reported gestational age. In the absence of gestational age information, outcome and trimester based assumptions are applied to assign episode boundaries. Trimester start and end dates are subsequently derived from the estimated pregnancy start date using standard trimester durations.

These steps result in non-overlapping, episode-level pregnancy records suitable for downstream analyses.

The steps mirror the logic described in the peer-reviewed publication and are implemented in a modular manner in the R package.

Detailed hierarchy rules for each step can be found here: https://github.com/GSK-Biostatistics/pregfindr/tree/main/vignettes

Figure 1 illustrates the overall structure of the pregnancy identification algorithm, highlighting the hierarchical processing steps used to resolve conflicting timing and outcome information in claims data.

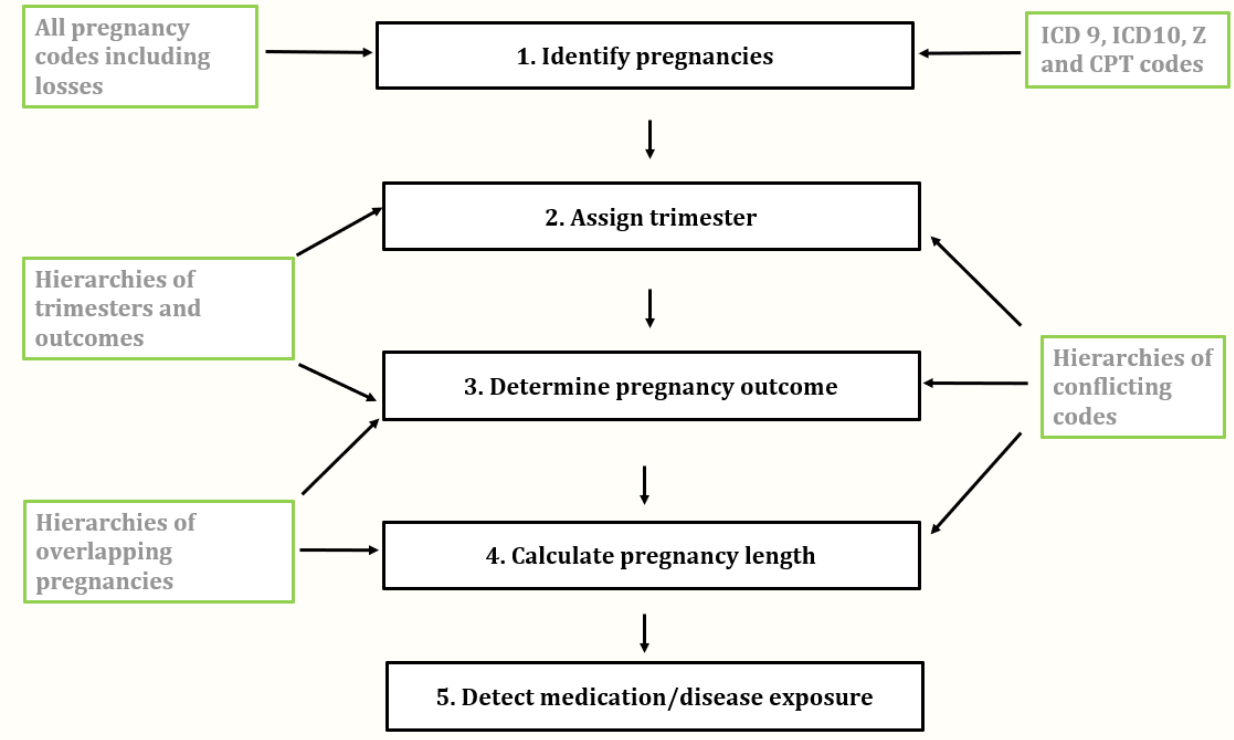

Figure 1. High-level overview of the pregnancy identification algorithm implemented in pregfindr. Raw pregnancy-related claims records are processed through a series of hierarchical steps to resolve trimester indicators, assign pregnancy outcomes, and estimate pregnancy start and end dates, resulting in a clean, episode-level pregnancy dataset suitable for downstream real-world analyses.

DATA SOURCE CONTEXT

Healthcare claims data, such as MarketScan® Commercial and Medicaid databases, provide longitudinal records of healthcare utilization across large and diverse populations. While these data sources do not capture pregnancy information directly, they contain diagnosis and procedure codes that can be leveraged to infer pregnancy episodes, gestational timing, and outcomes. Their scale and longitudinal nature make them particularly valuable for post-marketing pregnancy safety evaluation, feasibility assessments, and exposure characterization, if pregnancy episodes are identified using standardized and validated methods. This approach can be extended to other data sources with appropriate code mapping and validation.

OPEN-SOURCE R IMPLEMENTATION (PREGFINDR)

DESIGN PRINCIPLES

The R implementation was guided by several design considerations:

- Direct translation of the published algorithm logic to ensure alignment with the validated SAS implementation.
- Modular structure, with each major algorithmic step implemented as a separate function.
- Separation of data extraction from processing, allowing users to prepare input data using their preferred data platforms.
- Clear documentation, including function-level help files and worked examples in package vignettes.

PACKAGE STRUCTURE

The pregfindr package provides functions corresponding to each algorithm stage:

- apply_trimester_hierarchy() - resolves conflicting trimester codes
- apply_outcome_hierarchy() - resolves outcome conflicts
- apply_pregnancy_start_end_hierarchy() - estimates episode dates using hierarchy logic.

Input to the package is a claims-level dataset containing patient identifiers, service dates, pregnancy-related diagnosis or procedure codes, and along with any raw trimester or outcome signals indicated by those codes based on curated codelists which are available within the package. The primary output is a pregnancy episode-level dataset containing patient identifiers and estimated pregnancy start and end dates, which can be further used to derive gestational attributes for downstream analyses.

INSTALLATION AND USE

The package is publicly available via GitHub and can be installed using standard R tools for GitHub-hosted packages. Example workflows and detailed guidance on input requirements and hierarchy logic are provided in the package vignettes.

Pregfindr repository: https://github.com/GSK-Biostatistics/pregfindr

VALIDATION AND CONSISTENCY ASSESSMENT

To ensure consistency with the original SAS implementation, outputs generated by the open-source R package were compared against results produced using the proprietary SAS code applied to the same MarketScan® data extracts. Validation focused on key episode-level metrics relevant for pregnancy research, including the number of pregnancies identified, distribution of pregnancy outcomes, and timing of pregnancy episodes.

Validation was performed in a stepwise manner, with intermediate outputs examined at each major hierarchy stage. This approach enabled early identification of discrepancies related to data ordering, handling of missing values, and interpretation of gestational timing rules, which are known sources of variation across analytic implementations.

The R implementation produced equivalent results to the SAS outputs across these metrics, demonstrating that the translation preserved the original algorithm logic. This stepwise validation approach provided confidence that the open-source implementation could be used interchangeably with the validated SAS version for supported data sources.

PRACTICAL CONSIDERATIONS

DATA ENVIRONMENT

The package is designed to operate on extracted claims-level datasets and does not impose requirements on how data are stored or queried upstream. Data extraction may be performed using database technologies such as SQL, Spark, or cloud-based platforms, after which pregfindr functions are applied in R. Input datasets do not require pre-sorting, and missing values are permitted; the algorithm is designed to accommodate incomplete or conflicting signals commonly observed in administrative claims data, provided required identifiers and service dates are available.

PERFORMANCE AND SCALE

Processing large claims datasets can be computationally intensive, particularly during hierarchy resolution steps. The package is intended to be run in appropriate analytic environments with sufficient memory and compute resources. Intermediate outputs may optionally be written to disk to support stepwise execution, validation, and quality review in large-scale analyses.

SCOPE AND LIMITATIONS

The pregfindr package represents a faithful translation of a validated pregnancy algorithm into an open-source R implementation, prioritizing reproducibility and transparency over refactoring or methodological modification. As with the original algorithm, pregnancy identification relies on recorded healthcare encounters and may not capture very early pregnancy losses or events with incomplete claims. Gestational timing is estimated using available diagnosis and procedure codes and clinically informed assumptions described in the published methodology, and residual uncertainty should be considered in trimester-specific analyses. Continuous insurance enrollment supports more complete pregnancy episode construction, while coverage gaps may affect episode identification or timing. Although the underlying method was validated using U.S. administrative claims data, application to other databases or coding systems requires appropriate local code mapping and re-evaluation. These considerations should be taken into account when interpreting results.

APPLICATIONS OF THE ALGORITHM

The pregnancy identification algorithm supports a range of real-world evidence applications where accurate gestational timing is required.

Safety and risk evaluation:

Pregnancy episodes and trimester assignments enable assessment of maternal or fetal outcomes following exposure to medications or vaccines, both overall and by timing of exposure during gestation. This is particularly relevant for evaluating trimester-specific risk profiles.

Study feasibility and planning:

The algorithm can be used to estimate the number of eligible pregnancies within a database, assess available follow-up time, and evaluate whether sufficient sample size exists to study rare pregnancy-related outcomes.

Regulatory and labeling:

Pregnancy episode identification supports estimation of the number of pregnant individuals potentially exposed to a product since launch and can inform evidence generation for product labeling updates or risk minimization activities.

Across use cases, the algorithm provides a reusable, standardized pregnancy framework, letting teams focus on study-specific analyses, improving efficiency and comparability. The consistent pregnancy definitions also facilitate integration into broader real-world analytics workflows.

DISCUSSION

Releasing the pregnancy algorithm as an open-source R package addresses key barriers to transparency and reuse in pregnancy-focused real-world evidence. The modular design allows users to apply the core algorithmic logic to data extracted from different environments, while the published documentation supports correct and consistent use. Emphasis on documentation and example workflows was critical to ensure that external users could understand and apply the methodology without access to internal development context.

FUTURE DIRECTIONS

Potential extensions of this work include adaptation to additional claims or electronic health record databases, incorporation of alternative coding systems used outside the United States, and integration into feasibility or analytics toolkits. Continued collaboration with external researchers may further refine and validate the approach across diverse data sources.

CONCLUSION

Identifying pregnancy episodes and gestational timing in claims data is a non-trivial task that requires careful handling of incomplete and conflicting information. The pregfindr R package provides an open-source implementation of a validated pregnancy identification algorithm, enabling transparent and reproducible construction of pregnancy episodes from administrative claims data.

By translating the established methodology into a modular R package with clear documentation, this work supports consistent application of pregnancy identification logic across studies and research teams. The availability of an open-source implementation lowers barriers to reuse and facilitates alignment across pregnancy-focused real-world evidence studies.

As real-world data continue to play an important role in post-marketing research, tools such as pregfindr provide a practical foundation for generating consistent evidence on medication and vaccine use during pregnancy.

REFERENCES

Sumner KM, Ehlinger A, Georgiou ME, Wurst KE. Development and evaluation of standardized pregnancy identification and trimester distribution algorithms in U.S. IBM MarketScan® Commercial and Medicaid data. Birth Defects Research. 2021.

pregfindr GitHub repository: https://github.com/GSK-Biostatistics/pregfindr

ACKNOWLEDGMENTS

The authors would like to acknowledge Betsy Georgiou and Keele Wurst for their contributions to the development and clinical interpretation of the pregnancy algorithm, and Jaspreet Multani and Federico Concas for their support in the R implementation and open-source release. This work was conducted as part of real-world evidence activities at GlaxoSmithKline.

CONTACT INFORMATION

Your comments and questions are valued and encouraged.
Contact the author at:
Author Name: Dhirishiya
Company: GSK
Address: Bangalore, India
Email: dhirishiya.x.p@gsk.com

Author Name: Onkar Kshirsagar
Company: GSK
Address: Bangalore, India
Email: onkar.s.kshirsagar@gsk.com