20 Feb 2026

RE07: An open-source implementation of a claims-based algorithm to identify pregnancy episodes

Dhirishiya P, Onkar Kshirsagar

Disclaimer

The views expressed in this presentation are those of the authors and do not necessarily reflect the views or positions of GSK.

GSK

Agenda

Enhancing Our Understanding of Product Safety in Pregnancy Through Real-World Data

Evidence Gap

Understanding current limitations in pregnancy safety evidence

Pregnancy Algorithm

How pregnancy episodes are identified in real-world data

Use Cases

Practical applications of the algorithm

Open Sourcing the Pregnancy Algorithm in R

Making pregnancy detection accessible for broader research use

GSK

The Patient Conundrum

“I was on this medication for the past 5 years. My doctor will not give me an answer if I should continue taking it.”

“Looking on the internet, there is so much conflicting information but no official source. Is it safe to take it?”

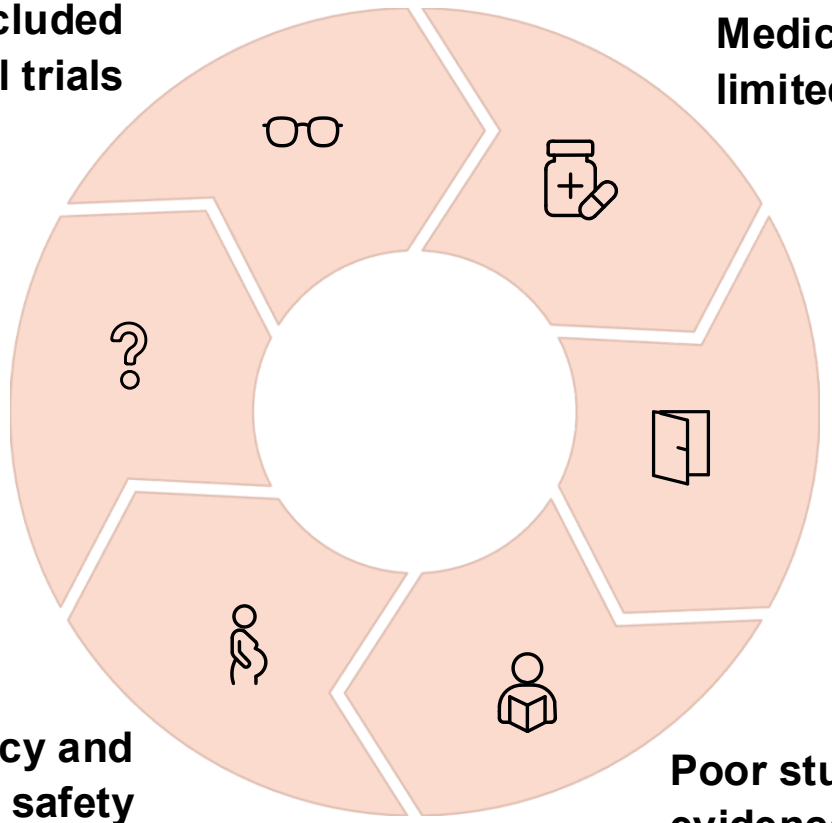
“Maybe I should put up with my condition until the baby is born. I cannot risk it so I will discontinue this treatment.”

GSK

The Pregnancy Evidence Gap: A Vicious Cycle

Pregnant women are excluded from clinical trials

Medication use in pregnancy is limited



Medication use in pregnancy remains limited due to unclear safety

Post-marketing studies rely on weak baseline data

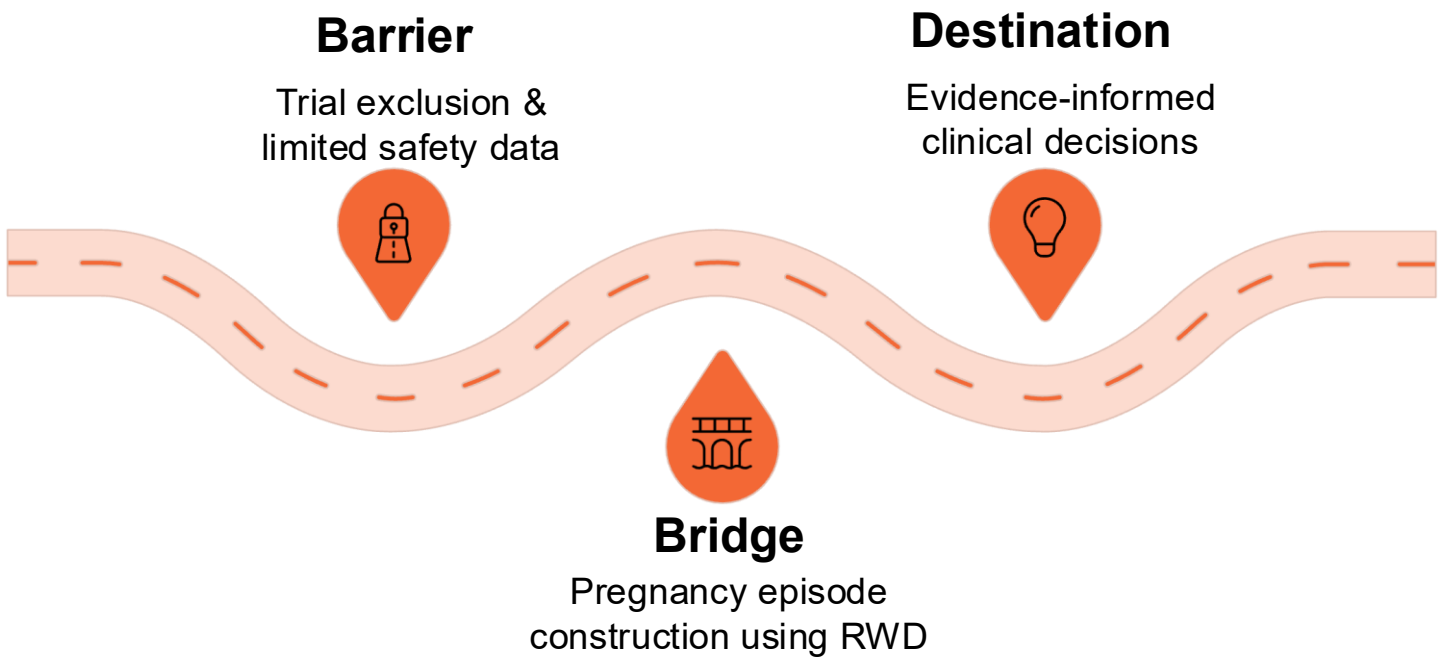
Continued gaps in pregnancy and lactation safety

Poor study design yields limited evidence

Over 90% of medications lack sufficient human pregnancy safety data at the time of approval (Contemporary Clinical Trials, 2022; CDC Medicine and Pregnancy overview).

GSK

Bridging the Evidence Gap: Pregnancy Episodes from Real-World Data



Critical Data Gap

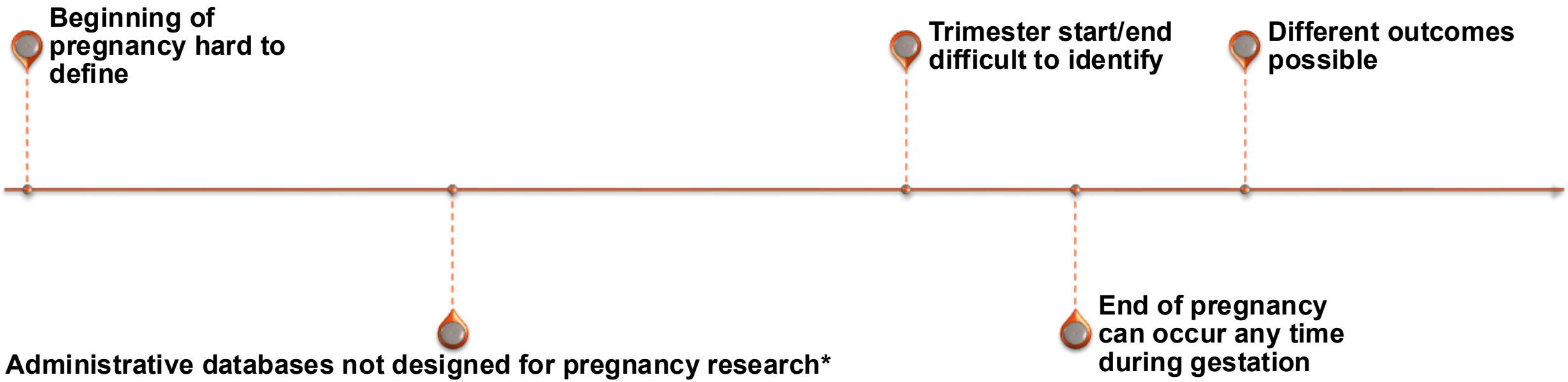
Pregnant patients systematically excluded from clinical trials, creating urgent safety evidence void

Scalable Solution

Structured real-world evidence captures pregnancy outcomes across diverse populations efficiently

GSK

Challenges in Identifying Pregnancies in RWD



Length of pregnancy can vary

Different data-capture across geographies

Different ontologies in use

*Includes challenge: No automated way to address conflicting timing/trimester codes and conflicting outcome codes

Overview of the Pregnancy Algorithm

What the algorithm does

- The algorithm can identify pregnancies, pregnancy length, assign trimesters, and determine outcomes including pregnancy losses based on diagnostic (ICD 9/10) and procedural codes.
- Enables linkage of medication/disease exposures to clinically meaningful pregnancy windows (e.g., by trimester).
- Applied to large US administrative claims databases(MarketScan® Commercial and Medicaid)
- Study populations, including women of childbearing age, are defined based on the specific data release, study period, and age criteria applied in each analysis
- Originally developed in SAS; now implemented in R for open-source access
- Peer-reviewed and published in Birth Defects Research*

*Sumner KM, Ehlinger A, Georgiou ME, Wurst KE. Development and evaluation of standardized pregnancy identification and trimester distribution algorithms in U.S. IBM MarketScan® (now Merative) Commercial and Medicaid data. Birth Defects Res. 2021 Nov 15;113(19):1357-1367 https://doi.org/10.1002/bdr2.1954

What the Pregnancy Algorithm Does

The algorithm structures raw claims data into pregnancy episodes using expert-validated, rule-based hierarchies to maintain accuracy in real-world, noisy data.

1

Identify Pregnancies

Identify women with pregnancy codes relevant to pregnancy and delivery from claims data.

2

Assign Trimester

Assign the trimester that codes correspond to by applying the Trimester Hierarchy to resolve conflicts when multiple codes with conflicting trimester implications.

Example: If a first-trimester ultrasound code is present on the same day as conflicting trimester codes, the trimester is reassigned to first trimester.

3

Determine Pregnancy Outcome

Determine pregnancy outcome for each pregnancy episode and resolve conflicting outcome codes using the Pregnancy Outcome Hierarchy

Example: If spontaneous abortion and stillbirth codes occur on the same day, the episode is classified as spontaneous abortion.

4

Calculate Pregnancy Length

Calculate start and end dates for each pregnancy episode using the Identifying Start and End date Hierarchy.

Example: If gestation week codes are available, pregnancy start date is calculated from the latest week reported.

Assign trimester start and end dates based on the pregnancy start date and standard number of days per trimester.

5

Detect Exposures

Connect maternal exposures to pregnancy windows for analysis. Ensures exposures are anchored to accurate, non-overlapping episodes.

GSK

What Questions Can This Algorithm Answer?

Safety & Benefit-Risk

- Which pregnancy outcomes occur following exposure to Drug/Vaccine X?
- Do risks vary by trimester or timing of exposure during gestation?

Feasibility & Planning

- How many eligible pregnancies are available in this database for a planned pregnancy safety study?
- Is there sufficient follow-up duration to study rare maternal outcomes?

Labelling & Regulatory

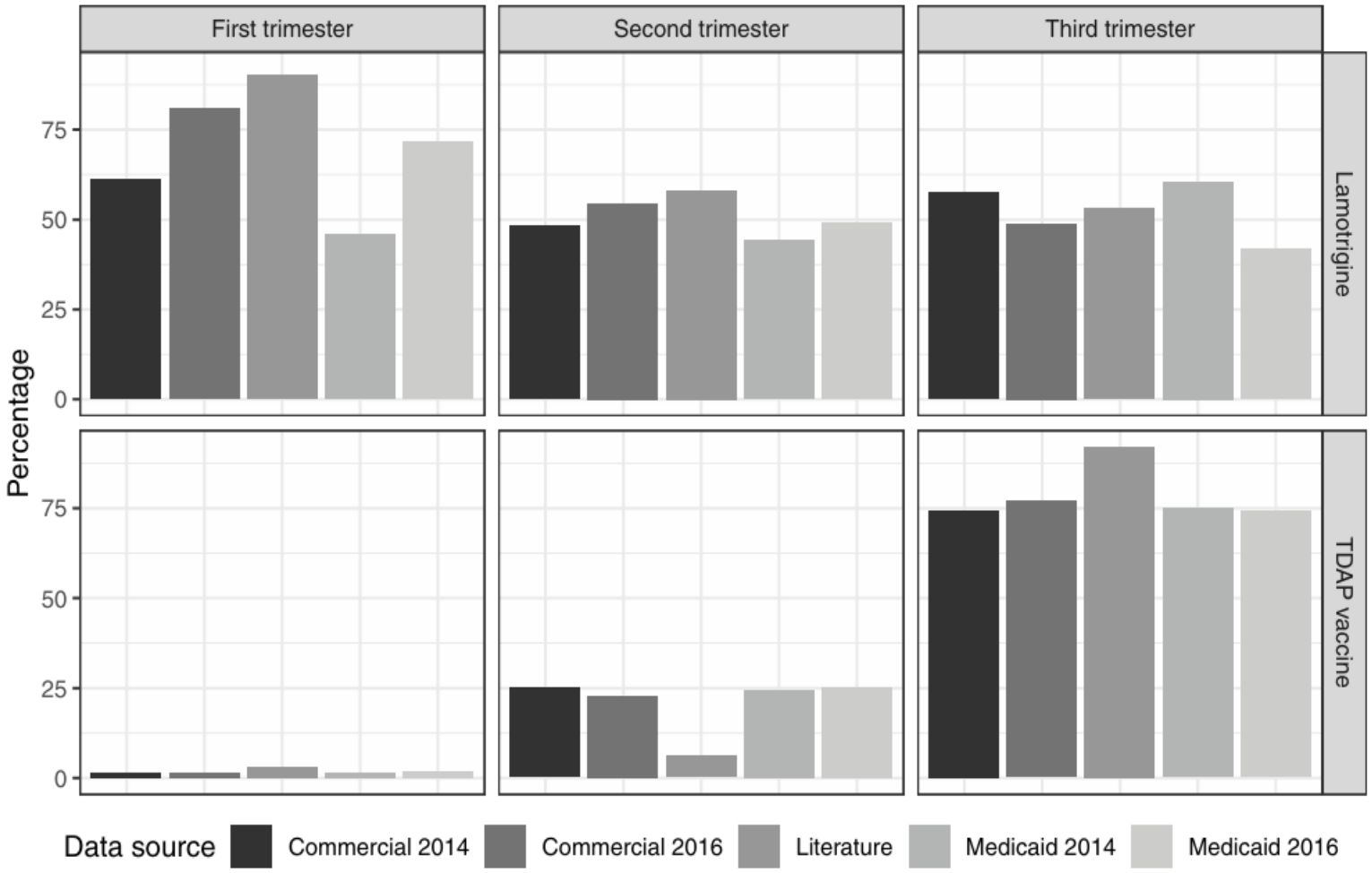
- How many pregnant women have been potentially exposed to Product X since product launch?
- Can we provide evidence to inform product labelling updates or risk minimization measures?

Reusability Across Products

- Can we turn this into a standard pregnancy module that multiple products, studies and teams can reuse?
- Can we plug the algorithm into other RWD workflows (feasibility tools, safety analyses, dashboards, etc.)?

GSK

Example From Published Evaluation: Trimester-Specific Exposure Patterns



The algorithm aligns medication and vaccination records with pregnancy episodes to estimate exposure by trimester.

Lamotrigine (maintenance medication): exposure is observed across all trimesters and is relatively similar by trimester, consistent with ongoing treatment through pregnancy.

TDAP (pregnancy-recommended vaccine): exposure is concentrated later in pregnancy, with a clear peak in the third trimester, reflecting guideline-recommended timing.

Sumner et al., Birth Defects Res 2021. Exposure to lamotrigine and TDAP by trimester among pregnant women in MarketScan Commercial and Medicaid databases, compared with Sentinel estimates.

GSK

Open Sourcing the Pregnancy Algorithm in R

What We Released and Why It Matters

Why Open Source?

The original SAS implementation was proprietary and not externally shareable. The R version addresses this limitation by providing a fully accessible, transparent solution that the broader research community can use and contribute to.

What We Released

- A modular, well-documented R package: pregfindr
- Rule-based logic for pregnancy episode identification, trimester classification, and outcome assignment
- Example workflows in package vignettes (GitHub/docs)

Key Benefits

- Ensures consistency and transparency in real-world pregnancy research
- Expands access across industry and academia
- Enables faster, lower-cost enhancements and broader adoption
- Supports regulatory and public health efforts with validated methodology

The R implementation produced equivalent results to the SAS version for pregnancy episode counts, outcome distributions, and episode timing, with stepwise validation across hierarchy stages using the same MarketScan extracts.

GSK

Architecture of pregfindr

Input Layer

Raw claims data from MarketScan (Commercial or Medicaid) combined with external codelists (e.g., outcomes, gestational age, trimester)

Algorithm Core

Three hierarchical functions resolve conflicts:
`apply_trimester_hierarchy()` - resolves conflicting trimester codes
`apply_outcome_hierarchy()` - resolves outcome conflicts
`apply_pregnancy_start_end_hierarchy()` - estimates episode dates using hierarchy logic.

Structured Output

Clean, patient-level pregnancy episodes for downstream analysis.

Flexible design: Supports custom inputs including age ranges, timeframes, and configurable data storage backends. Clean separation between data extraction, processing logic, and episode construction ensures maintainability and reproducibility.

GSK

How to Use the pregfindr Package

Step 1: Load extracted claims data

- Patient ID
- Diagnosis/Procedure Codes
- Date of service
- Primary diagnosis indicator
- Claims File Type (Inpatient / Outpatient)
- Pregnancy-related codes (from curated codelists):
  - Trimester indicators
  - Pregnancy outcomes
  - Secondary outcomes

Step 2: apply_trimester_hierarchy()

- Creates clean trimester code using predefined trimester hierarchy

Step 3: apply_output_hierarchy()

- Assigns single pregnancy outcome per episode

Step 4: apply_pregnancy_start_end_hierarchy()

- Patient ID
- Pregnancy Start & End Dates
- Pregnancy length
- Trimester Start & End Dates
- Outcome

Detailed guidance, input specifications, variable naming conventions and hierarchy logic are available in pregfindr/vignettes/ on GitHub.

GSK

Design Considerations

Modular Structure

Core hierarchy steps are implemented as independent functions, allowing stepwise validation and re-execution without rerunning the full pipeline, whilst supporting repeated execution on large claims datasets.

Maintainability

The codebase follows widely adopted R conventions, supporting readable workflows and simplifying updates to logic or inputs with minimal refactoring.

Decoupled Processing

Algorithm logic remains independent of data extraction and storage mechanisms, enabling deployment across different environments once inputs adhere to expected structures.

Transparency

The full implementation of the published algorithm is openly available, supporting reproducibility, independent review, and reuse across research communities.

GSK

Ensuring Reproducibility Across Implementations

Reproducibility required precise alignment on missing values, date and week definitions, and consistent ordering of records across all implementations.

Stepwise validation of intermediate hierarchy stages proved essential to confirm equivalence and identify discrepancies early in the pipeline.

Applying complex, rule-based hierarchies to large claims datasets demanded disciplined execution management and careful computational resource planning.

Open-sourcing emphasised the critical need for clear abstractions, comprehensive documentation, and intuitive interfaces to support external validation and broader adoption.

GSK

Implementation scope & known constraints

- Faithful translation, not refactor: The R implementation is a close, line-by-line translation of the validated SAS reference to preserve reproducibility and transparency.
- Dependence on coded healthcare data: Algorithm performance reflects the presence and quality of diagnosis, procedure, and billing codes available in the input claims data. Configuration management and change control are recommended to ensure reproducibility and traceability.
- Incomplete clinical capture: Pregnancies with minimal healthcare utilization, including very early losses, may not be fully captured due to reliance on coded encounters.
- Insurance coverage matters: Continuous enrollment supports more complete pregnancy episode construction; coverage gaps may affect episode timing or identification.
- Portability limits: The method was validated using U.S. MarketScan claims data; application to other databases or coding systems requires local code mapping and re-evaluation.

GSK

What's Next

Extending and Scaling the Open-Source Pregnancy Algorithm

Adaptation to Additional Data Sources

Extend the implementation to claims and healthcare databases across different geographies, exploring adaptation to alternative coding systems and terminology standards where feasible.

Flexible, Cohort-Driven Execution

Enable workflows that commence from predefined cohorts, such as medication-specific or disease-specific populations, rather than requiring full database scans.

Ongoing Alignment with Evolving Standards

Maintain compatibility with updates to coding terminologies, database structures, and data quality frameworks as real-world data sources continue to evolve.

Integration into Analytics Workflows

Support seamless use in feasibility assessments, safety signal evaluation, and downstream analytics, including integration with study planning tools and interactive dashboards.

Continued Collaboration and Validation

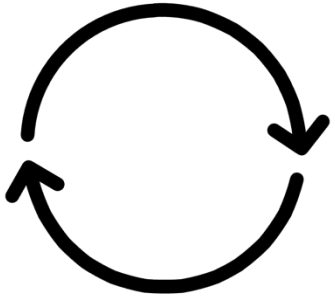
Encourage active collaboration with researchers, data partners, and regulatory bodies to validate and refine the implementation across diverse data sources and use cases.

GSK

Summary

What this work enables

Standardized Identification



Reproducible Results

Open-Source Implementation

GSK

Thank You

Acknowledgements

Federico Concas
Digital Data and Analytics Manager, GSK

Jaspreet Multani
Principal Data Scientist, GSK

Betsy Georgiou
Associate Director, Epidemiology

Keele Wurst
Head, Immunology & Fibrosis Epidemiology

Questions?

We welcome your thoughts and feedback

Repository:
GSK-Biostatistics/pregfindr

Get in Touch

Dhirishiya P
dhirishiya.x.p@gsk.com

Onkar Kshirsagar
onkar.s.kshirsagar@gsk.com