

A*STAR – Data4Life Session 2

Cohort Definition and Generation with ATLAS 3.0 in D2E

1 October 2026

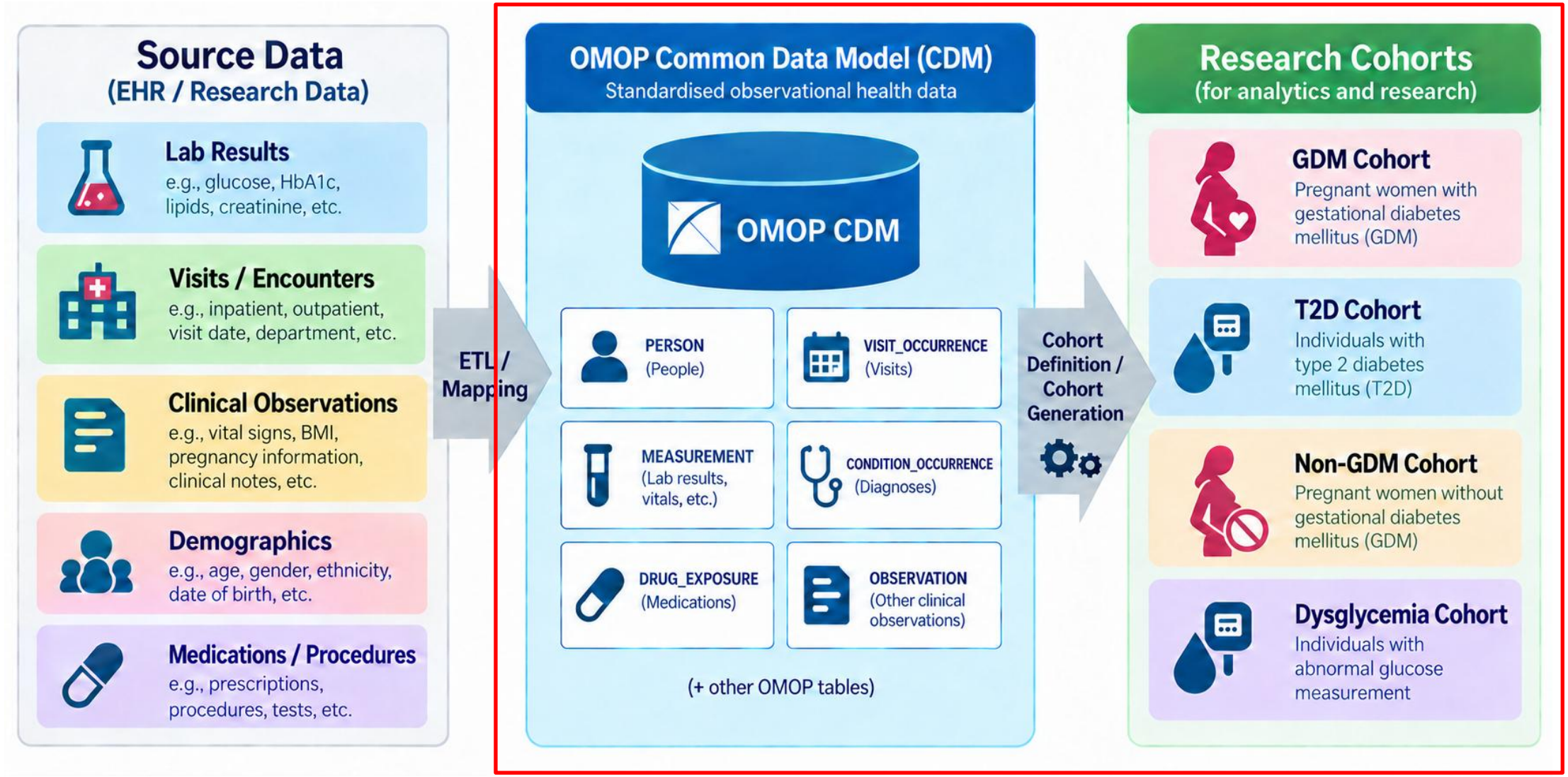
A*STAR:
- Mukkesh P. Kumar
- Cindy Ho
- Erwin Tantoso
- Jing Yang Tan
- Lynnette Tan

Data4Life:
- Tim Walz
- Peter Hoffmann

Today's Objective

- Understand concept sets, cohort definition and cohort generation
- Translate a clinical phenotype into an explicit computable phenotype
- Use the GUSTO GDM/T2D example to connect the research question to ATLAS 3.0 workflow
  - Building Concept sets, Cohort Definition, Cohort Generation and Incidence Rate Analysis
- Demonstration of D2E / ATLAS

The Journey: From raw data to OMOP CDM to Research Cohorts

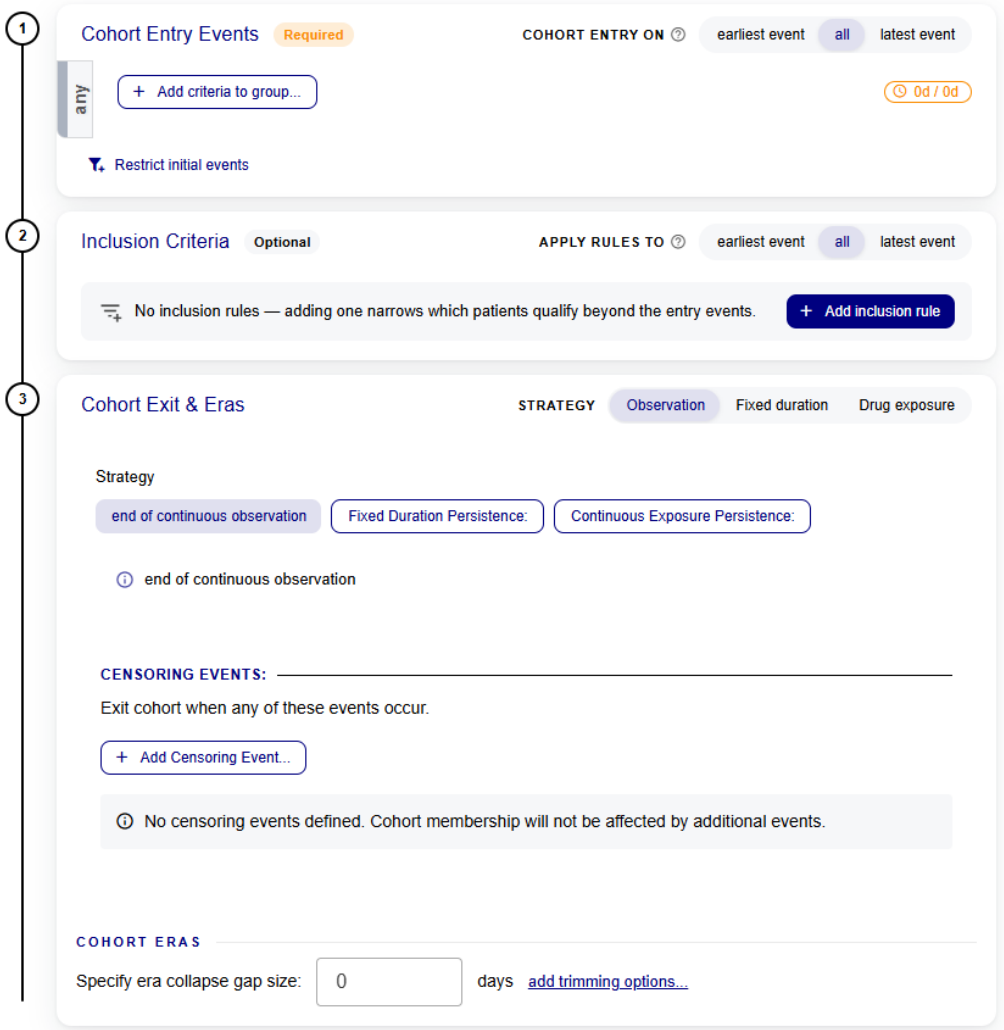


What is a cohort and why do we need it?

- What is a cohort?
  - A cohort is a set of persons who satisfy one or more inclusion criteria for a duration of time.
    - It defines the who, when, and how subjects are included in a cohort
  - Standardized definition
    - Define a specific set of people who meet the inclusion and exclusion criteria over a specific period of time
- Why do we need cohort?
  - Reproducibility
    - Reusable cohort definition ensures the exact same criteria is used across different datasets
  - Distributed network studies
    - Allow standardized analysis code to be executed at each site and hence allows aggregated statistics to be populated identically without pooling sensitive patient-level records

Building blocks of a cohort

- Cohort Entry
  - Specifies the clinical event or criteria that marks the start of a person's time in the cohort.
- Inclusion and Exclusion criteria
  - Rules and restriction criteria applied to filter the population that qualified through the entry event.
- Cohort Exit
  - Defines the exact condition or date when a person's time in the cohort ends



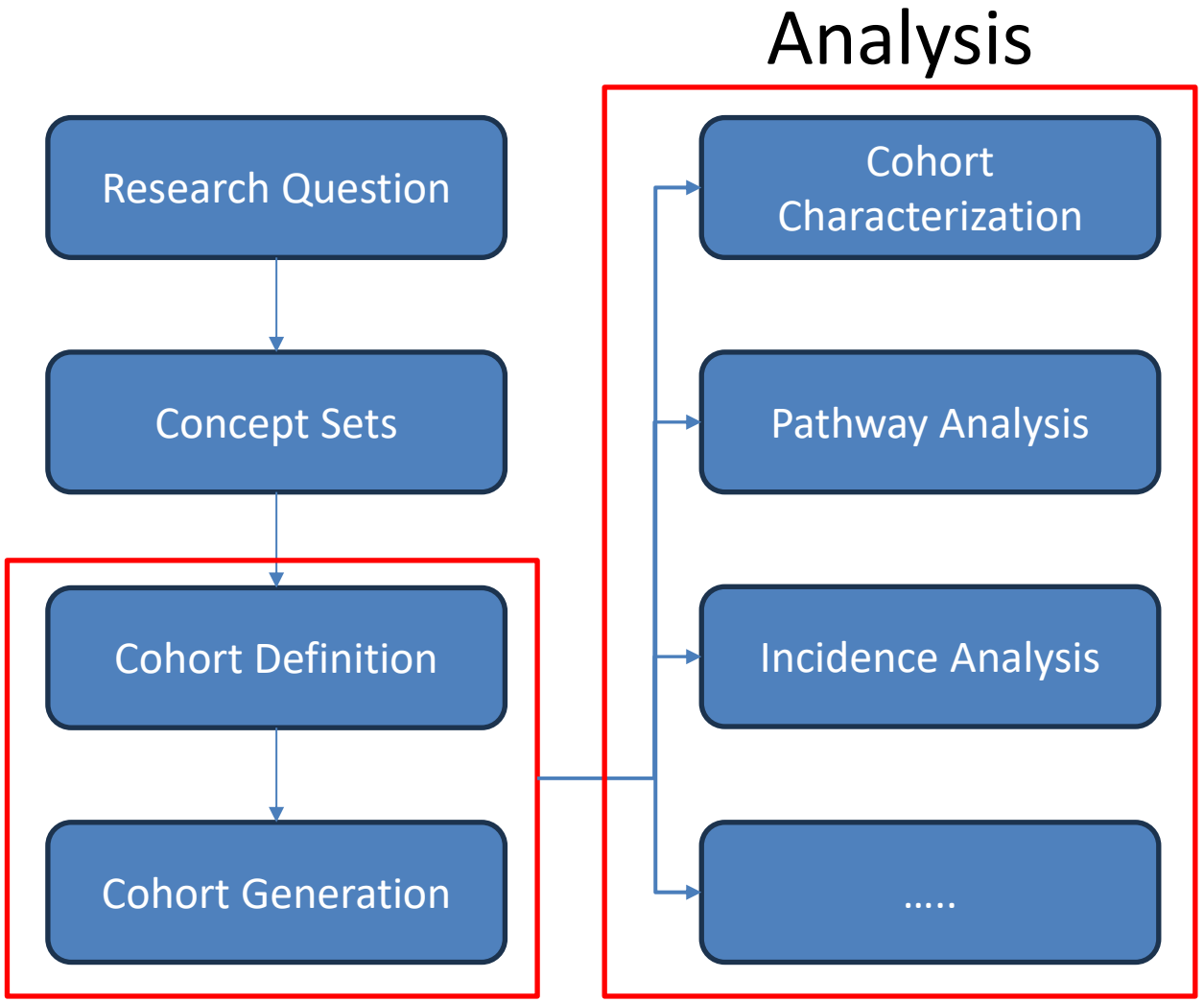
From clinical phenotype to computable phenotype

- Clinical phenotype vs computable phenotype
  - Clinical phenotype is expression of a clinical condition in a patient by clinician (Human Narrative)
  - Computable phenotype is expression of a clinical condition, characteristic, or set of features using a standardized, machine-readable algorithm

Clinical Statement	Needs for Computable Phenotype
Women with Gestational Diabetes Mellitus (GDM)	How is GDM identified? Diagnosis? OGTT? Fasting Glucose? Both?
Abnormal glucose	Which measurement concepts? What thresholds?
During pregnancy	How is pregnancy/delivery identified?
No previous diabetes	Which concepts? What look-back period?
Adults	Age at which date?

Main idea: Every ambiguity in clinical definition must become an explicit rule in the computable phenotype.

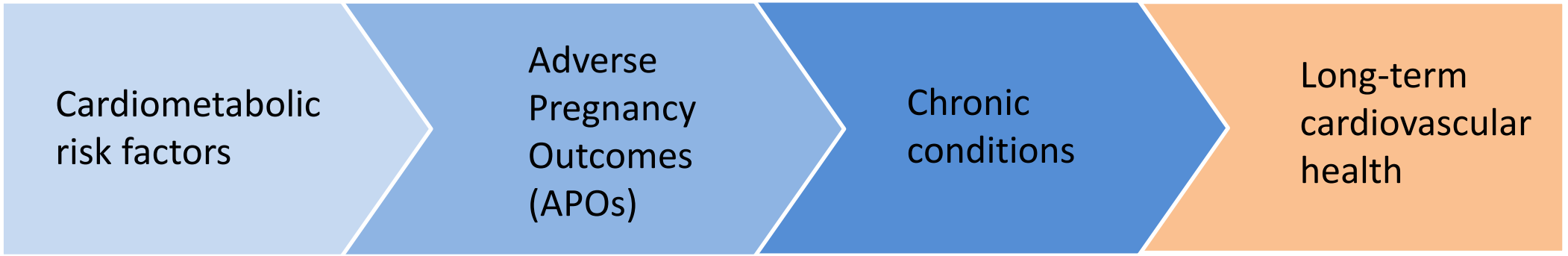
GUSTO: Workflow



- Identify Research Question
- Define concept sets
- Cohort definition
- Cohort generation
- Cohort analysis

Formulating study idea

PREGNANCY — POST - PREGNANCY



e.g. Obesity, hypertension, dyslipidemia, hyperglycemia	e.g. Hypertensive disorders of pregnancy (HDPs), **gestational diabetes mellitus (GDM)**, preterm birth	e.g. Hypertension, **diabetes**, dyslipidemia	Cardiovascular events: e.g. Heart attack, Stroke, Heart failure

There is a knock-on effect from **maternal health during pregnancy** all the way to **long-term cardiovascular health**

For this demo: We focus on the relationship between GDM and T2D

GUSTO: Research Question

Combined analysis of gestational diabetes and maternal weight status from pre-pregnancy through post-delivery in future development of type 2 diabetes

Ling-Wei Chen [#1], Shu E Soh [#1,2], Mya-Thway Tint [1,3], See Ling Loy [4,5], Fabian Yap [5,6], Kok Hian Tan [5,7], Yung Seng Lee [1,2,8], Lynette Pei-Chi Shek [1,2,8], Keith M Godfrey [9], Peter D Gluckman [1,10], Johan G Eriksson [1,3,11,12], Yap-Seng Chong [1,3], Shiao-Yng Chan [13,14]

Affiliations + expand

PMID: 33658531 PMCID: PMC7930020 DOI: 10.1038/s41598-021-82789-x

Women with GDM had a higher risk of later T2D [relative risk (95% CI) 12.07 (4.55, 32.02)] and dysglycemia [3.02 (2.19, 4.16)] compared with non-GDM women

- Research question: Does GDM during pregnancy increase the risk of developing T2D (or dysglycemia) in the years following delivery?

GUSTO: Concept Sets

- Concept set is a reusable, data-agnostic expression that defines a collection of standardized vocabulary concepts (OMOP concept IDs) representing a specific clinical items
- Identify variables and concepts to define computable phenotypes
  - Delivery (Mapped to 35815042 as delivery method)
  - Gestational age (3012266)
  - Fasting glucose in serum or plasma (3018251)
  - OGTT 2 hour (3017538)
  - History of Diabetes or GDM (1340204 (201826), 1340204 (201254), 3175877)
  - Type 2 Diabetes (201826)
- Identify the thresholds to define GDM or dysglycaemia
  - GDM: Fasting Glucose >= 7.0 mmol/L OR OGTT 2 hour >= 7.8 mmol/L
  - Postpartum dysglycemia: Fasting Glucose between 6.1 mmol/L and 6.9 mmol/L OR OGTT 2 hour between 7.8 mmol/L and 11.0 mmol/L.

GUSTO: Cohort Definition

Cohort Building Blocks	GDM Cohort	Non-GDM Cohort	T2D Outcome Cohort	Dysglycemia Outcome Cohort
Cohort Entry (index date)	• Delivery • Age >= 18	• Delivery • Age >= 18	• First occurrence of T2D	• Fasting Glucose between 6.1 – 6.9 mmol/L; OR OGTT 2-hr between 7.8 – 11 mmol/L
Inclusion Criteria	• 24 weeks <= Gestational age <= 28 weeks • Fasting Glucose >= 7.0 mmol/L; OR • OGTT 2-hr >= 7.8 mmol/L	• 24 weeks <= Gestational age <= 28 weeks • Fasting Glucose < 7.0 mmol/L; AND • OGTT 2-hr < 7.8 mmol/L	• Has “delivery” prior to T2D	• Has “delivery” anytime until 90 days before dysglycemia
Exclusion Criteria	• Exclude subjects with prior T1D/T2D/GDM	• Exclude subjects with prior T1D/T2D/GDM	• None	• None
Cohort Exit	• End of observation	• End of observation	• End of observation	• End of observation

Delivery is the index date for the GDM/non-GDM target cohorts.

- GDM status is defined using gestational age 24-28 weeks plus glucose/OGTT from the same date or testing episode.

GUSTO: Cohort Generation

- Execute the logic defined in the “Cohort Definition” on the data source
  - Identify all qualifying persons/events
  - Apply inclusion/exclusion criteria
  - Determine cohort start and end dates

Review target cohort counts

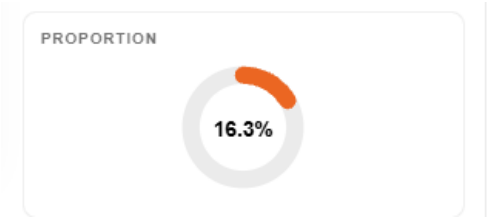
Stages	GDM Cohort	Non-GDM Cohort	T2D Outcome	Dysglycemia Outcome
Initial Eligible Subjects	1194	1194	76	377
Subjects Excluded	1034	436	0	147
Final Cohort Size	160	758	76	230

Cohort generation not only gives us the final number of subjects. It also allows us to examine how subjects are included or excluded at each stage, which is important for validating our phenotype.

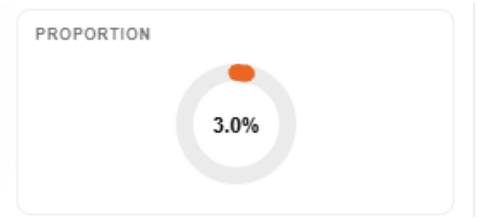
GUSTO: Analysis -> Incidence Rates (T2D)

- Target Cohort
  - GDM Cohort
  - Non-GDM Cohort
- Outcome Cohort
  - Postpartum T2D Outcome Cohort

GDM Cohort



Non-GDM Cohort

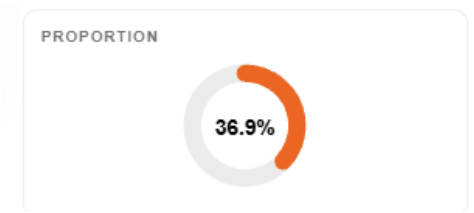


Women with GDM had a ~5.4-fold higher proportion of postpartum T2D than women without GDM (16.3% vs 3.0%).

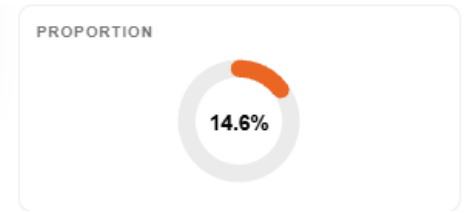
GUSTO: Analysis -> Incidence Rates (Dysglycemia)

- Target Cohort
  - GDM Cohort
  - Non-GDM Cohort
- Outcome Cohort
  - Postpartum Dysglycemia Outcome Cohort

GDM Cohort



Non-GDM Cohort



Postpartum dysglycaemia was more frequent among women with GDM (36.9%) than women without GDM (14.6%), approximately 2.5-fold higher.

Summary: From Research Question to Research Cohort

- Define the research question
- Define concept sets
- Build cohort definition
- Generate cohort and validate by reviewing attrition
- Cohort analysis

Research question → Concept sets → Cohort definition → Cohort generation → Analysis

Data2Evidence Demo

https://github.com/OHDSI/Data2Evidence/


Terima Kasih
THANK YOU
GRACIAS
ARIGATO
SHUKURIA
GOZAIMASHITA
EFCHARISTO
BOLZÏN
MERCI
谢谢
감사합니다
ありがとう
தமிழ்
DANKSCHEEN
BIYAN
SHUKRIA
TASHAKKUR ATU
YAQHANYELAY
SUKSAMA
EKHMET
SPASIBO
DENKAUJA
NENACHALHYA
GRAZIE
MEHRBANI
MAAKE
KOMAPSUMNIDA
JUSPAXAR
PALDIES
MERASTAWHY
GAEJTHO
FAKAAUE
AGUYJE
LAH
ATTO
SIKOMO
MINMONCHAR