PLOS ONE

RESEARCH ARTICLE

Temporal Events Detector for Pregnancy Care (TED-PC): A rule-based algorithm to infer gestational age and delivery date from electronic health records of pregnant women with and without COVID-19

Tianchu Lyu[1], Chen Liang[1]*, Jihong Liu[2], Berry Campbell[3], Peiyin Hung[1], Yi-Wen Shih[1], Nadia Ghumman[1], Xiaoming Li[4], on behalf of the National COVID Cohort Collaborative Consortium[¶]

1 Department of Health Services Policy and Management, Arnold School of Public Health, University of South Carolina, Columbia, South Carolina, United States of America, **2** Department of Epidemiology & Biostatistics, Arnold School of Public Health, University of South Carolina, Columbia, South Carolina, United States of America, **3** Department of Obstetrics and Gynecology, School of Medicine, University of South Carolina, Columbia, South Carolina, United States of America, **4** Department of Health Promotion Education and Behaviors, Arnold School of Public Health, University of South Carolina, Columbia, South Carolina, United States of America

¶ Membership of the National COVID Cohort Collaborative Consortium is provided in the Acknowledgments.
* cliang@mailbox.sc.edu

Citation: Lyu T, Liang C, Liu J, Campbell B, Hung P, Shih Y-W, et al. (2022) Temporal Events Detector for Pregnancy Care (TED-PC): A rule-based algorithm to infer gestational age and delivery date from electronic health records of pregnant women with and without COVID-19. PLoS ONE 17(10): e0276923. https://doi.org/10.1371/journal.pone.0276923

Editor: Dong Keon Yon, Kyung Hee University School of Medicine, REPUBLIC OF KOREA

Received: July 22, 2022

Accepted: October 16, 2022

Published: October 31, 2022

Data Availability Statement: Data set used to reach the conclusions drawn in the manuscript is made available for access and reproducing the study upon review and approval of a data use request by National Covid Cohort Collaborative committee. The data set is accessible at https://unite.nih.gov/workspace/compass/view/ri.compass.main.folder.8d143683-2093-4992-a736-ccf3cf8f9117.

Abstract

Objective

Identifying the time of SARS-CoV-2 viral infection relative to specific gestational weeks is critical for delineating the role of viral infection timing in adverse pregnancy outcomes. However, this task is difficult when it comes to Electronic Health Records (EHR). In combating the COVID-19 pandemic for maternal health, we sought to develop and validate a clinical information extraction algorithm to detect the time of clinical events relative to gestational weeks.

Materials and methods

We used EHR from the National COVID Cohort Collaborative (N3C), in which the EHR are normalized by the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM). We performed EHR phenotyping, resulting in 270,897 pregnant women (June 1st, 2018 to May 31st, 2021). We developed a rule-based algorithm and performed a multi-level evaluation to test content validity and clinical validity, and extreme length of gestation (<150 or >300).

Results

The algorithm identified 296,194 pregnancies (16,659 COVID-19, 174,744 without COVID-19) in 270,897 pregnant women. For inferring gestational age, 95% cases (n = 40) have moderate-high accuracy (Cohen's Kappa = 0.62); 100% cases (n = 40) have moderate-high

Funding: Research reported in this publication was supported by the National Institute of Allergy And Infectious Diseases of the National Institutes of Health under Award Numbers 3R01AI127203-05S2 and R21AI170171. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

granularity of temporal information (Cohen's Kappa = 1). For inferring delivery dates, the accuracy is 100% (Cohen's Kappa = 1). The accuracy of gestational age detection for the extreme length of gestation is 93.3% (Cohen's Kappa = 1). Mothers with COVID-19 showed higher prevalence in obesity or overweight (35.1% vs. 29.5%), diabetes (17.8% vs. 17.0%), chronic obstructive pulmonary disease (0.2% vs. 0.1%), respiratory distress syndrome or acute respiratory failure (1.8% vs. 0.2%).

Discussion

We explored the characteristics of pregnant women by different gestational weeks of SARS-CoV-2 infection with our algorithm. TED-PC is the first to infer the exact gestational week linked with every clinical event from EHR and detect the timing of SARS-CoV-2 infection in pregnant women.

Conclusion

The algorithm shows excellent clinical validity in inferring gestational age and delivery dates, which supports multiple EHR cohorts on N3C studying the impact of COVID-19 on pregnancy.

Introduction

Recent findings suggested Coronavirus Disease 2019 (COVID-19) to be associated with an increased risk for adverse pregnancy outcomes and neonatal complications [1, 2]. However, there has been limited knowledge pertaining to the timing of SARS-CoV-2 infection during the pregnancy (e.g., infection in a specific trimester, gestational age [GA], or labor and delivery) and its association with pregnant women's real-time clinical presentation [3]. Timing of viral infection is important because fetuses are more vulnerable to maternal complications and/or viral infection during certain gestational stages [4, 5]. Nevertheless, detecting the timing of viral infection poses substantial challenges to the quality of data and clinical information extraction methodology.

Pregnancy care consists of antenatal, labor and delivery, and postpartum care [6]. Because pregnancy care spans up to 21 months, it involves exceedingly rich clinical data. A complete episode of pregnancy care often involves multiple encounters with multiple health providers, clinical sites, and diverse clinical information systems, meaning that a vast number of clinical events are generated at an unprecedented granularity and data quality. Over decades, comprehensive clinical data for pregnancy care have not been widely available until recent advances in the use of normalized multi-system electronic health records (EHR), such as Observational Health Data Sciences and Informatics (OHDSI)'s Atlas [7], National Institutes of Health (NIH)'s All of Us, and NIH's National COVID Cohort Collaborative (N3C) [8]. which have provided growing Real-World Evidence (RWE) to support pregnancy research during the COVID-19 pandemic [9–12].

One of the unique characteristics of EHR in pregnancy care is the complex temporal relations of clinical events. To better understand the impacts of SARS-CoV-2 infection or the COVID-19 pandemic on pregnancy health, it is important to know the length of pregnancy and the timing of pregnancy-related complications or events in relation to the time of pregnancy (i.e., GA). Both mothers and fetuses experience crucial physiological changes and clinical complications during pregnancy, which generates a substantial number of clinical events in

the EHR. Accurate identification of temporal relations of clinical events across the entire episode of pregnancy care is a fundamental step for clinical decision-making as well as downstream EHR data mining. GA is also the prerequisite for tracing the timing of SARS-CoV-2 infection and COVID-19 vaccination for pregnant women. Although EHR data have unique advantages in preserving temporal information of these clinical events, clinical information extraction methods tailored for pregnancy care have been scarce [13].

Identifying GA from EHR is challenging because GA is a concept of temporal relativity. When it comes to clinical information extraction, this challenge is manifold. First, controlled vocabularies (e.g., ICD, SNOMED-CT, LOINC, RxNorm) along with the relational EHR database architecture are designed to preserve some temporal information of clinical events, yet the information suffers from a low level of granularity (e.g., LOINC 95656–5: Gestational age <30 weeks, LOINC 49085–4: First and Second trimester integrated maternal screen panel), unreliable data entries (e.g., laboratory test results sometimes have delayed time stamps), and incomplete data (e.g., laboratory test results sometime are missing due to many laboratory results are photocopies). Second, approximately 80% of the EHR data consist of unstructured data (i.e., clinical notes), with which a considerable amount of temporal information is in the form of free text that cannot be directly used for quantitative analysis [14]. Third, EHR data for pregnant women are distinct from most other medical specialties in that pregnancy care has scheduled routine visits that are involved with antenatal care, labor and delivery hospitalization, and postpartum care. Incomplete and/or inconsistent data are common because patients are often engaged with different providers, health care systems, and clinical visits with chief complaints not related to pregnancy but would generate data relevant to pregnancy [13]. For example, antenatal care and other medical care during the pregnancy could be at locations different from labor and delivery hospitals. Missing critical information from one clinical site would require researchers to infer such information using data from other sites or other visits. Health records of individual visits span the inpatient and various outpatient visits but may not contain explicit and consistent temporal information (e.g., last menstrual period [LMP] [15], estimated date of delivery [DOD], and GA).

Current biomedical informatics methods for extracting and inferring temporal relations of clinical events include rule-based methods, machine learning, natural language processing (NLP), ontology-based methods, and temporal reasoning [14, 16–23]. Most of these methods utilized unstructured clinical notes in combination with structured EHR data, which is comprehensive for generic temporal information extraction. However, methods focusing on temporal events among pregnant women's EHR are limited. Among studies that extract or infer DOD and GA, LMP and imaging/lab test results are commonly used data; chart review is a commonly used method [17, 20, 22, 23]. Using LMP data requires the providers to accurately document LMP in the EHR, yet in the real world, many EHR datasets have a lot of missing values in LMP. The use of ultrasound test results or other laboratory test results requires the EHR to comprehensively document both laboratory orders and testing results, yet testing results are often missing in real-world EHR datasets. Additionally, using laboratory data alone for inferring GA may not be accurate due to the individual physiological variation among pregnant women. A recent study utilized comprehensive ICD codes of diagnoses and procedures to infer delivery dates [24]. While the study focused on full-term pregnancy with comprehensive medical records during the labor and delivery hospitalization, comprehensive methods remain needed for early-stage pregnancy (e.g., extreme preterm, very preterm) and those who have part of the pregnancy care data and conflicting data documented in EHR. Particularly, there are no published methods for extracting temporal relations of clinical events for pregnant women with COVID-19. This is a critical knowledge gap because temporal relations of clinical events are suggestive of the exact time of viral infection, acute phase of COVID-19, and vaccinations.

To identify temporal relations of clinical events imperative for pregnant women with COVID-19, we developed a rule-based clinical information extraction algorithm, namely Temporal Events Detector for Pregnancy Care (TED-PC), which infers GA and DOD using both structured EHR data and annotated clinical notes. The algorithm is designed to capture temporal information to be used for inferring GA and DOD, respectively, so that the complete temporal relations in a pregnancy episode can be replicated and the timing of SARS-CoV-2 infection (in weeks) can be detected. This design is anticipated to be effective for pregnant women with regular labor and delivery hospitalization, without complete hospitalization records, and those who have pre-term delivery, miscarriage, early-stage pregnancy and termination, and multiple births. This algorithm is designed for EHR that are normalized by Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) [8, 25] and is implemented on the N3C enclave. EHR on N3C have 1) individual-level data linked among multiple health systems nationwide and 2) normalized procedures, laboratory tests/results, and annotated clinical notes, which enable the reasoning of GA and DOD for patients with missing and conflicting data. These unique characteristics of OMOP CDM-normalized EHR are critical for detecting the temporal information for pregnant women with COVID-19. The performance and clinical validity of TED-PC were tested systematically on the N3C platform. Presently, this algorithm is used as a critical clinical information extraction tool to identify comprehensive temporal relations of clinical events from multiple COVID-19 pregnant women cohorts on N3C [26–28].

Materials and methods

Data sources

We used the N3C database (level 3, containing dates of clinical events and zip codes), a multicenter clinical data repository that contains de-identified EHR data of individuals with COVID-19 blended with controls (i.e., non-COVID-19) [8]. N3C currently has EHR and medical claims data from more than 73 healthcare systems and institutes across 50 states. The EHR data are normalized using the OMOP CDM [8, 25]. In order to find the full clinical course of each pregnancy, the study cohort included women who met the following conditions: (1) have at least one childbirth between June 1st, 2018 and May 31st, 2021, (2) be aged between 15 to 49 years old at the DOD, and (3) have at least one GA-related record during the pregnancy.

Because the N3C database is normalized by OMOP CDM, we utilized the following resources for EHR phenotyping. The ATHENA vocabulary repository is used for retrieving OMOP CDM concept IDs and phenotyping patients with GA and childbirth-related records. The Algorithms section details the design and the procedures for using the ATHENA.

Algorithms

To retrieve the full spectrum of each pregnancy in the EHR, it is crucial to identify the start date (i.e., pregnancy start) and the end date (i.e., childbirth delivery) of the pregnancy. The start date can be estimated by the GA-related records that indicate the GA (e.g., in weeks, in a range of weeks, or a particular trimester) and the date of the record. The end date can be estimated by the identification of the DOD. Because some pregnant women's EHR data only have either GA-related records or childbirth delivery records, we first estimated GA and DOD, respectively, which resulted in a cohort of pregnant women with estimated GA, denoted as the GA cohort, and a cohort of pregnant women with estimated DOD, denoted as the DOD cohort. Then we estimated the start date and end date of the pregnancy by consolidating the temporal information from both cohorts (Fig 1).

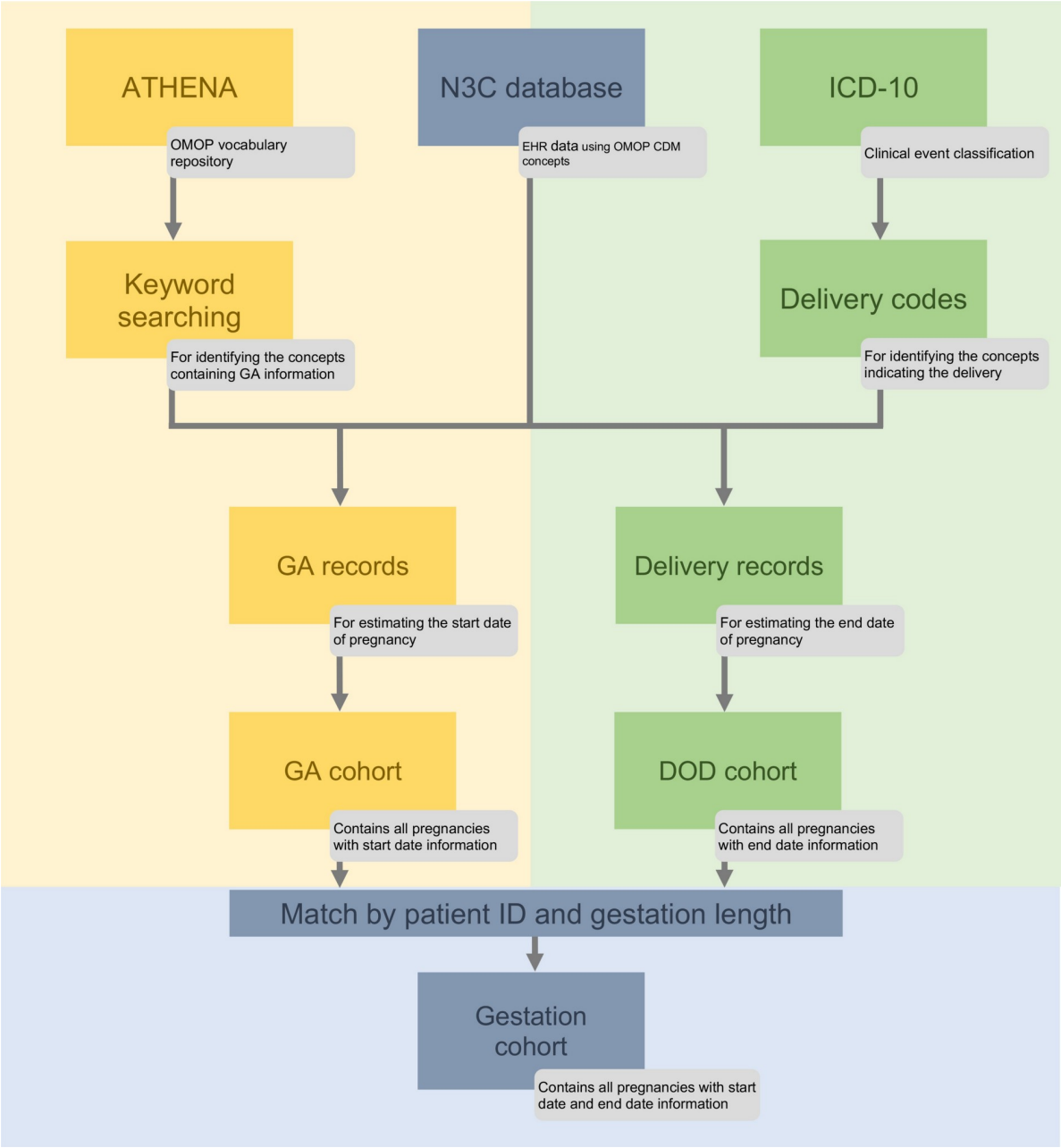

Fig 1. Pipeline of TED-PC for extracting information and building cohorts.

https://doi.org/10.1371/journal.pone.0276923.g001

Gestational age cohort (GA cohort). *Phenotyping*. The purpose of this step is to find OMOP CDM concepts that can be used to retrieve GA-related information from EHR data. For phenotyping the GA cohort, we used a keyword search strategy followed by a review of retrieved OMOP CDM concepts. Using ATHENA, a set of keywords were reviewed and determined: "trimester", "gestation", and "pregnan" (regular expression of "pregnancy" or "pregnant"). These keywords were used in conjunction with three filters in the ATHENA database: (1) "DOMAIN", which included "Condition", "Observation", "Procedure", "Measurement", etc.; (2) "CONCEPT", which included "Standard" and "Non-standard"; (3) "VALIDITY", which included "Valid" and "Invalid". The pseudo-query is:

("trimester" OR "gestation" OR "pregnan") AND (((DOMAIN = "Condition") OR (DOMAIN = "Observation") OR (DOMAIN = "Procedure") OR (DOMAIN = "Measurement")) AND (CONCEPT = "Standard") AND (VALIDITY = "Valid"))

In the review of the returned OMOP CDM concepts, we applied the following three criteria to narrow down the scope step by step to our target concepts: (a) "Whether a record indicates a pregnant patient"; (b) "If yes, whether the record contains GA information of the patient"; (c) "If yes, what is the value of the GA?". Finally, we identified 138 OMOP CDM concepts (See S1 Table). The researcher (TL) who performed the phenotyping was not involved in the phenotyping evaluation.

Rule-based algorithm. We developed a rule-based algorithm to infer GA from EHR data (Fig 2). A critical feature of the algorithm is that we divided all extracted OMOP CDM concepts into four accuracy levels based on their clinical meanings and granularity of the date: high, moderate-high, moderate-low, and low (Table 1) and prioritized the retrieval of GA-related information based on accuracy levels. Table 2 shows the pseudocode for the algorithm.

Childbirth delivery cohort (DOD cohort). *Phenotyping*. For phenotyping the DOD cohort, we started with a list of CDC-recommended ICD, DRG, and CPT codes used for childbirth delivery and followed by exploring the relevant OMOP CDM concepts using the semantic relationships of concepts on ATHENA. First, we used a set of ICD-10, DRG, and CPT codes suggestive of childbirth delivery (see S2 Table) [29]. These codes were used to retrieve corresponding OMOP CDM concepts in ATHENA in which the resulting CDM concepts were then used to identify the childbirth delivery records in the EHR. Second, since these codes may not comprehensively capture all the OMOP CDM concepts indicating childbirth delivery, we explored the semantic relationships of the OMOP CDM concepts retrieved by these codes and supplemented them with the newly identified concepts [30]. The final concept set contained 105 OMOP CDM standard concepts (See S3 Table). Researchers (TL and YS) who performed the phenotyping and were not involved in the phenotyping evaluation.

Rule-based algorithm. Upon manual chart review of the EHR data, we found the OMOP CDM concepts with the domain of procedure have the highest accuracy with respect to determining the DOD, followed by domains of condition and then observation. Thus, we developed a rule-based algorithm to approximate the true DOD by prioritizing the OMOP CDM 'procedure' domain over the 'condition' domain over the 'observation' domain (Fig 3). Table 3 shows the pseudocode for the algorithm.

All data manipulation, phenotyping, and algorithms were implemented using SQL, R, Python, and PySpark on the N3C platform. Source programming codes are available at N3C, project "[RP-2B9622] Assessing and predicting the clinical outcomes of pregnant women with COVID-19 using machine learning approach."

Evaluation

We performed a multi-level evaluation to test the validity of the algorithm as well as inter-rater reliability. To test the content validity of the OMOP CDM concepts resulting from the phenotyping, two researchers (CL and NG) independently reviewed the concept IDs and their semantic meanings and properties on the ATHENA and dichotomously rated the relevance of all concept IDs. Inter-rater reliability was measured by Cohen's Kappa. Disagreements were discussed and resolved together with a senior OB/GYN physician (BC).

To test the clinical validity of the algorithm for inferring GA, we randomly selected 30 patients from the final cohort, which resulted in 40 distinct pregnancies, including multiple gestations. Their comprehensive medical records on GA, excluding laboratory data, were extracted from the EHR. We calculated the start date of pregnancy by subtracting the GA in

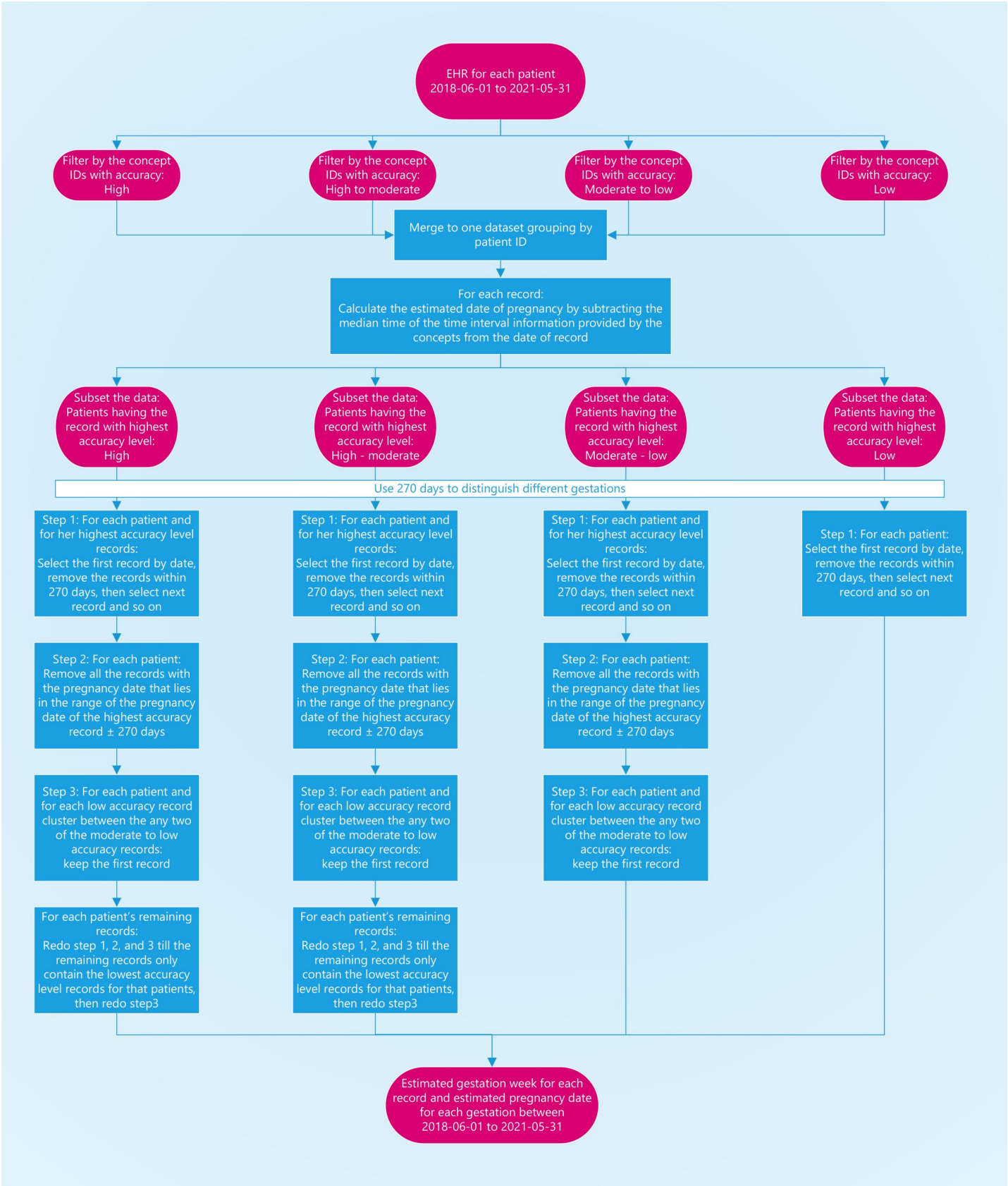
EHR for each patient
2018-06-01 to 2021-05-31
Filter by the concept IDs with accuracy: High
Filter by the concept IDs with accuracy: High to moderate
Filter by the concept IDs with accuracy: Moderate to low
Filter by the concept IDs with accuracy: Low
Merge to one dataset grouping by patient ID
For each record:
Calculate the estimated date of pregnancy by subtracting the median time of the time interval information provided by the concepts from the date of record
Subset the data: Patients having the record with highest accuracy level: High
Subset the data: Patients having the record with highest accuracy level: High - moderate
Subset the data: Patients having the record with highest accuracy level: Moderate - low
Subset the data: Patients having the record with highest accuracy level: Low
Use 270 days to distinguish different gestations
Step 1: For each patient and for her highest accuracy level records:
Select the first record by date, remove the records within 270 days, then select next record and so on
Step 1: For each patient and for her highest accuracy level records:
Select the first record by date, remove the records within 270 days, then select next record and so on
Step 1: For each patient and for her highest accuracy level records:
Select the first record by date, remove the records within 270 days, then select next record and so on
Step 1: For each patient:
Select the first record by date, remove the records within 270 days, then select next record and so on
Step 2: For each patient:
Remove all the records with the pregnancy date that lies in the range of the pregnancy date of the highest accuracy record ± 270 days
Step 2: For each patient:
Remove all the records with the pregnancy date that lies in the range of the pregnancy date of the highest accuracy record ± 270 days
Step 2: For each patient:
Remove all the records with the pregnancy date that lies in the range of the pregnancy date of the highest accuracy record ± 270 days
Step 3: For each patient and for each low accuracy record cluster between the any two of the moderate to low accuracy records:
keep the first record
Step 3: For each patient and for each low accuracy record cluster between the any two of the moderate to low accuracy records:
keep the first record
Step 3: For each patient and for each low accuracy record cluster between the any two of the moderate to low accuracy records:
keep the first record
For each patient's remaining records:
Redo step 1, 2, and 3 till the remaining records only contain the lowest accuracy level records for that patients, then redo step3
For each patient's remaining records:
Redo step 1, 2, and 3 till the remaining records only contain the lowest accuracy level records for that patients, then redo step3
Estimated gestation week for each record and estimated pregnancy date for each gestation between
2018-06-01 to 2021-05-31

Fig 2. Flowchart for estimating gestational age and pregnancy dates.

https://doi.org/10.1371/journal.pone.0276923.g002

weeks from the event date for each record. Two clinical experts (CL and NG) independently reviewed the retrieved records and rated them based on two metrics: accuracy (high/moderate/low) and granularity (high/moderate/low). Accuracy is concerned with the level that the selected OMOP CDM concepts can accurately indicate the GA. For example, GA-related records are typically documented during antenatal care visits. Multiple GA-related records, even when documented at different dates, can suggest a consistent GA. The algorithm-selected concept is most accurate if it is among these records. When there are other GA-related records suggesting a GA different from the algorithm-selected record, the accuracy would not be high. Granularity refers to the extent that the algorithm-selected concept can indicate a specific gestational week. For example, the "gestation period, 38 weeks" has a high granularity level whereas "third-trimester pregnancy" has a low granularity level.

To test the clinical validity of the algorithm for inferring DOD, we randomly selected 30 gestations from the final cohort. Their records consisting of procedures, conditions, observations, and measurements were extracted from the EHR within ± 14 days of estimated DOD. Two clinical experts (CL and NG) independently reviewed the charts and labeled whether the DOD was correctly inferred by the algorithm.

Despite the average gestation being around 280 days, this estimation varies among individuals. To represent rare cases such as preterm birth, post-term birth, and early-stage termination, we also performed extreme value analysis, in which two clinical experts (TL and CL) performed chart reviews for 30 randomly selected samples with <150 or >300 days of gestation.

Characteristics of pregnant women with and without COVID-19

Using TED-PC, we performed descriptive analyses to explore maternal demographics and underlying conditions (See S4 Table for OMOP CDM concepts) for pregnant women with (cases) and without COVID-19 (controls) which are characterized by temporal information of the gestational weeks when SARS-CoV-2 infection was identified.

Table 1. Concepts categorization by accuracy level.

Accuracy level	Time interval	Definition	Example concept IDs
High	1 week	The concept name specifies the value of GA in weeks (e.g., Gestation period, 15 weeks).	4337360: Gestation period, 1 week
			4097608: Gestation period, 18 weeks
			444098: Gestation period, 40 weeks
Moderate—high	2–5 weeks	The concept name does not specify the value of GA in weeks but specifies the range of GA in weeks which is larger than 1 week and smaller than 6 weeks (e.g., Gestation 9–13 weeks)	4181468: Gestation 9–13 weeks
			44791171: 9–13 weeks gestational age
			45757118: Spontaneous onset of labor between 37 and 39 weeks gestation with planned cesarean section
Moderate—low	6–10 weeks	The concept name does not specify the value of GA in weeks but specifies the range of GA in weeks which is larger than 5 weeks and smaller than 11 weeks (e.g., Gestation 14–20 weeks)	4180111: Third trimester pregnancy less than 36 weeks
			4178165: Gestation 14–20 weeks
			44791170: 14–20 weeks gestational age
Low	11–13 weeks	The concept name does not specify the value of GA in weeks but specifies the trimester (e.g., first trimester)	3657563: First trimester bleeding
			4239938: First trimester pregnancy
			4112238: Third trimester

GA: gestational age.

https://doi.org/10.1371/journal.pone.0276923.t001

Table 2. The pseudocode for the algorithm: Estimating the gestational age.

	Algorithm: GA estimation for each gestation
1:	**procedure** GA ESTIMATION
2:	**Input:** GA-related clinical events, the date of GA-related clinical events
3:	**Output:** Estimated GA for each gestation
4:	pregnancy date <- calculated by the date of GA-related clinical events
5:	**if** the time range indicated by the GA-related clinical events = = 1 week **then**
6:	accuracy <- 1
7:	**else if** 2 weeks < = the time range indicated by the GA-related clinical events < = 5 weeks **then**
8:	accuracy <- 2
9:	**else if** 6 weeks < = the time range indicated by the GA-related clinical events < = 10 weeks **then**
10:	accuracy <- 3
11:	**else if** 11 weeks < = the time range indicated by the GA-related clinical events < = 13 weeks **then**
12:	accuracy <- 4
13:	**endif**
14:	**for** each patient
15:	Sort the GA-related clinical events by event date chronologically
16:	**repeat**
17:	Select the first pregnancy date with the highest accuracy*
18:	**for** the GA-related clinical events in (±270 days of the selected pregnancy date)
19:	Remove
20:	**endfor**
21:	Estimated GA <- the first GA-related clinical event with the highest accuracy*
22:	**until** the last GA-related clinical event
23:	**repeat**
24:	**if** remaining GA-related clinical events exist **then**
25:	**repeat**
26:	Select the first pregnancy date with the highest accuracy*
27:	**for** the GA-related clinical events in (±270 days of the selected pregnancy date)
28:	Remove
29:	**endfor**
30:	Estimated GA <- the first GA-related clinical event with the highest accuracy*
31:	**until** the GA-related clinical event
32:	**else** stop
33:	**endif**
34:	**until** the GA-related clinical event
35:	**endfor**
36:	**endprocedure**

GA: gestational age.

https://doi.org/10.1371/journal.pone.0276923.t002

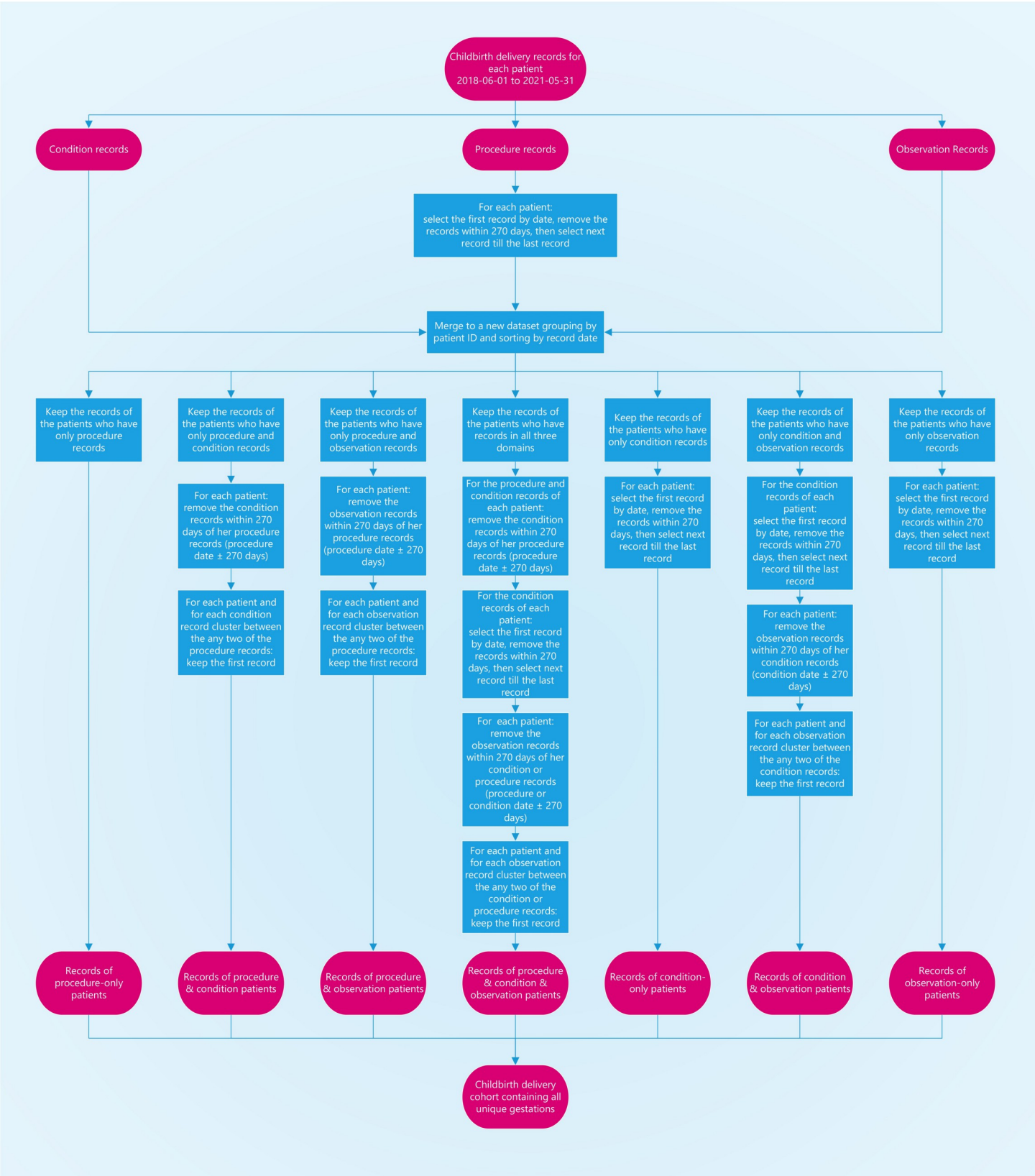

Fig 3. Flowchart for estimating dates of delivery.

https://doi.org/10.1371/journal.pone.0276923.g003

Table 3. The pseudocode for the algorithm: Estimating the date of delivery.

```
Algorithm: DOD estimation for each gestation
procedure DOD ESTIMATION
  Input: childbirth delivery clinical events, the domain and the date of DOD-related clinical events
  Output: Estimated DOD for each gestation
  Event DOD <- the date of childbirth delivery clinical events
  if domain = = "procedure" then
    accuracy <- 1
  else if domain = = "condition" then
    accuracy <- 2
  else if domain = = "observation" then
    accuracy <- 3
  endif
  for each patient
    Sort the childbirth delivery clinical events by event DOD in reversed chronological order
    repeat
      Select the first event DOD with the highest accuracy*
      for the childbirth delivery clinical event in (±270 days of the selected event DOD)
        Remove
      endfor
      Estimated DOD <- the first event DOD with the highest accuracy*
    until the last childbirth delivery clinical event
    repeat
      if remaining childbirth delivery clinical events exist then
        repeat
          Select the first event DOD with the highest accuracy*
          for the childbirth delivery clinical events in (±270 days of the selected event DOD)
            Remove
          endfor
          Estimated DOD <- the first event DOD with the highest accuracy*
          until the last childbirth delivery clinical event
        else stop
        endif
      until the last childbirth delivery clinical event
    endfor
endprocedure
```

DOD: date of delivery.

https://doi.org/10.1371/journal.pone.0276923.t003

Results

Identified OMOP CDM concepts

We identified 2,773 OMOP CDM concepts from the ATHENA vocabulary repository, of which 2,370 indicated pregnancy. Among the concepts relating to pregnancy, 336 have GA-related information. We excluded 189 concepts that either indicated the inaccurate time range broader than one trimester (13 weeks) (e.g., concept ID 21493940: US for pregnancy in the

second or third trimester) or did not have corresponding records in the N3C database (e.g., concept ID 3025286: Gestational age estimated from foot length on US by Mercer 1987 method). Totally, 138 concepts contained useful gestational week information with a time range from one week to one trimester. Within the selected 138 concepts, 42 (30.4%) were in high accuracy, 9 (6.5%) were in moderate-high accuracy, 5 (3.6%) were in moderate-low accuracy, and 82 (59.4%) were in low accuracy.

Algorithm performance

To evaluate phenotyping results, the content validity of the selected concepts was assessed and rated blindfolded by two independent reviewers (CL and NG) who did not participate in the phenotyping. Both reviewers rated all concepts as "valid" (100% agreement).

We evaluated the performance of the GA algorithm in two dimensions: accuracy and granularity. Among the 30 randomly selected pregnant women, eight of them had two gestations and one of them had three gestations during the study time frame. The mean gestation length was 270.15 days with a maximum of 299 days and a minimum of 159 days. Among the 40 pregnancies, one reviewer rated 34 (85.0%) samples as high accuracy, 4 (10.0%) samples as moderate accuracy, and 2 (5.0%) sample as low accuracy. The other reviewer rated 35 (87.5%) samples as high accuracy, 3 (7.5%) samples as moderate accuracy, and 2 (5.0%) samples as low accuracy. The Cohen's Kappa with linear weighting is 0.62, CI = [0.35, 0.90]. For granularity, both reviewers rated the 39 samples as high granularity and one sample as low granularity (100% agreement, unweighted Cohen's Kappa = 1). See Table 4 for the confusion matrix.

For the DOD algorithm, a total of 30 patients' EHR were reviewed independently. Both reviewers rated the 30 samples to be accurate (100% agreement, unweighted Cohen's Kappa = 1)

Extreme value analysis

We randomly selected 30 gestations with a gestation length either smaller than 150 days or greater than 300 days and extracted their EHR. After chart review, 28 of the 30 gestations were extracted with correct GA information, with an accuracy of 93.3%. For the two error cases, the first one was due to the contradiction between the GA records on different dates. For example, there was a record with the concept name "Gestation period, 38 weeks" on date 1, but other records with the same concept on date 2. The second error case was due to the contradiction between the GA records on the same date. For example, on the same date, one record was with the concept name "Gestation period, 36 weeks" and another record was with the concept name "Gestation period, 39 weeks". Among the correctly inferred cases, a certain level of inaccuracy existed. Five gestations only had low accuracy level GA records. Among the extreme case, one of them had only one GA record. Inter-rater reliability is 100% (unweighted Cohen's Kappa = 1).

Table 4. Confusion matrix of the accuracy rating for the performance of the GA algorithm.

		Reviewer2			
		High	**Moderate**	**Low**	**Total**
Reviewer1	High	33	1	0	34
	Moderate	2	1	1	4
	Low	0	1	1	2
	Total	35	3	2	40

https://doi.org/10.1371/journal.pone.0276923.t004

Characteristics of pregnant women with and without COVID-19

Between June 1st, 2018 and May 31st, 2021, a total of 296,194 gestations in 270,897 pregnant women were identified from the N3C database. The mean and median ages are 30.31 and 31, respectively. There were 245,892 women who had one pregnancy during the study time, 24,713 and 292 had two and three pregnancies, respectively. The mean gestational length was 274.14 days. The median was 278 days, with a minimum of 140 days and a maximum of 308 days. N3C data retrieval was completed on 02/12/2022.

Using TED-PC, we identified the timing of SARS-CoV-2 infections in gestational weeks. Fig 4 shows the frequency of infections across gestational weeks. More than half of the infections happened during late pregnancy (between 32 and 41 weeks), which might be related to the increased antenatal visits in late pregnancy. See S5 Table for the selected demographics and underlying conditions of the cohort captured by TED-PC, stratified by trimesters. There were 104,791 and 191,403 gestations before and during the COVID-19 pandemic, respectively, among which there are 16,659 gestations with COVID-19 and 174,744 without COVID-19 peri-pandemic. Age group 30–34 shared the largest proportion across the age groups, followed by the age groups 25–29, 40–44, and 20–24. White people made up the largest percentage of nearly 50% of the total population, followed by Black and Hispanic/Latino races. For mothers who had ever been infected by SARS-CoV-2 before the DOD (before or during the pregnancy), age groups 30–34, 20–24, and 25–29 had the largest percentage. Besides, the

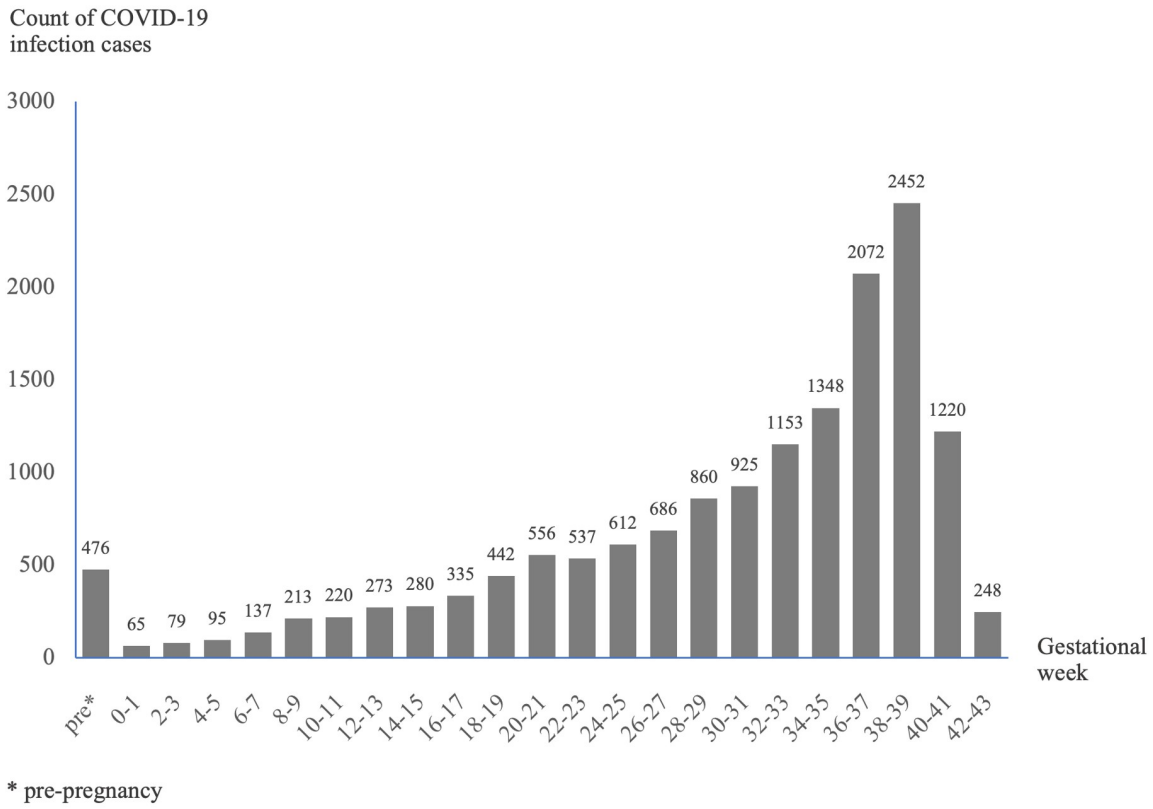

Fig 4. The distribution of the SARS-CoV-2 infection cases by gestational week.

https://doi.org/10.1371/journal.pone.0276923.g004

percentage of White was lower (39.7%) compared with that of pre-pandemic (57.1%). Hispanic/Latino made up the largest proportion across the races in those SARS-CoV-2 infected mothers (31.5%). Compared with the pregnant women without COVID-19, pregnant women with COVID-19 had a higher prevalence in obesity or overweight (35.1% vs. 29.5%), diabetes (17.8% vs. 17.0%), chronic obstructive pulmonary disease (COPD) (0.2% vs. 0.1%), respiratory distress syndrome (ARDS) or acute respiratory failure (ARF) (1.8% vs. 0.2%), myocardial infarction (0.2% vs. 0.1%), and HIV/AIDS (0.6% vs. 0.4%). The characteristics of the proportions shared similar trends when stratified by different trimesters.

Discussion

Using N3C data, we created the first EHR-based cohort of SARS-CoV-2-infected pregnant women in the US with complete temporal information of clinical events spanning the gestation length, which supports urgently needed COVID-19 research for pregnant women. Our algorithm is among the first that detects exact gestational week of viral infection, early-state pregnancy, preterm birth, early termination, post-term birth, and other adverse clinical events. Because viral infection at different stages of pregnancy is associated with different risks of fetus development and maternal status, TED-PC is generalizable to viral infection aside from COVID-19 as well as adverse events that would affect pregnancy. This algorithm shows the promise to underpin EHR deep phenotyping of pregnancy care as well as machine learning methods, in which both require precise temporal information of clinical events. As a rapid development of the clinical information extraction tool for combating COVID-19, our algorithm is currently supporting several EHR-based cohort studies on the N3C to examine the impact of COVID-19 on pregnant women's real-time clinical inflammatory progression and pregnancy complications.

The accuracy of TED-PC is warranted by a few logic layers. First, compared with previous studies that focused on claims data or required labor efforts, our study took advantage of the OMOP CDM normalized EHR data to categorize normalized concepts into different priority groups [17, 22–24, 31–34]. For example, the GA algorithm prioritized the concepts in the "Procedure" domain over "Condition" domain over "Observation" domain, which logically prevented the algorithm from selecting the records at a higher risk of semantic ambiguity and low granularity. Second, our algorithm categorized the GA-related concepts into different accuracy levels by indicating the time range. This step allows TED-PC to prioritize records with the most accurate information. Third, the use of the 270-day interval in our algorithm enabled us to distinguish different gestations of the same pregnant woman within the time frame. Fourth, the merging and matching process of the GA cohort and the DOD cohort excluded gestations with untrustworthy or missing values, which is common in EHR.

Detection of temporal information for pregnant women with COVID-19 is made available by two major features of OMOP CDM. First, OMOP CDM normalizes multi-system EHR data linked at an individual level. This unique feature enables our algorithm to impute a huge amount of missing temporal values and to resolve conflicts of temporal values among health records from different hospital systems. Second, OMOP CDM includes annotated clinical notes, procedures, and laboratory tests/results, which allows the algorithm to leverage multi-source contextual information for inferring temporal information at an adequate level of granularity.

A few limitations of this study warrant note. First, because several GA-related OMOP CDM concepts do not indicate specific gestational weeks (e.g., "Spontaneous onset of labor between 37 and 39 week gestation with planned cesarean section"), we inferred the gestational weeks using the median time point of the range for these concepts, which may impair the

performance of the algorithm. This impact is mild on the concepts with high or moderate-high accuracy levels since the time range is small, but it could be severe in the concepts with the low accuracy level. Second, our EHR data may not be comprehensive. For example, some examinations or laboratory tests do not have time information, but they are often prescribed to pregnant women during a specific time frame of gestation. Our future direction will aim to improve the performance of TED-PC and test the external validity. From error analysis, data incompleteness and inconsistency remain the major sources of error. Well-designed EHR data imputation methods and a hybrid model of rule-based and machine learning algorithms hold promises for addressing these issues. Although our algorithm is implemented on N3C, it could be potentially repurposed for other OMOP CDM normalized EHR.

Conclusion

We explored and compared the characteristics of pregnant women by different timing of SARS-CoV-2 infection with our newly developed technique: TED-PC, a rule-based algorithm to automatically infer comprehensive temporal information of clinical events from EHR during pregnancy care. The performance of TED-PC is satisfactory as collectively the accuracy and granularity of temporal information are beyond 90%. TED-PC has been implemented on N3C, supporting multiple national EHR cohorts for desperately needed research on the impact of COVID-19 on pregnancy. TED-PC is implemented on N3C data but remains generalizable for OMOP CDM normalized EHR.

Supporting information

S1 Table. OMOP CDM concepts for gestational age-related EHR records.
(DOCX)

S2 Table. ICD, CPT, and DRG codes suggestive of childbirth delivery dates.
(DOCX)

S3 Table. OMOP CDM concepts for delivery date-related EHR records.
(DOCX)

S4 Table. OMOP CDM concepts for underlying conditions.
(DOCX)

S5 Table. Selected demographics and underlying conditions for pregnant women with and without COVID-19 by gestations during pre- and peri-pandemic.
(DOCX)

Acknowledgments

We thank Ms. Ashlee Kim for her support of medical coding. We thank N3C consortium for providing the data and infrastructure. The N3C consortium contact is NCATS_N3C@nih.gov.

N3C attribution

The analyses described in this publication were conducted with data or tools accessed through the NCATS N3C Data Enclave https://covid.cd2h.org and N3C Attribution & Publication Policy v 1.2-2020-08-25b supported by NCATS U24 TR002306. This research was possible because of the patients whose information is included within the data and the organizations (https://ncats.nih.gov/n3c/resources/data-contribution/data-transfer-agreement-signatories)

and scientists who have contributed to the ongoing development of this community resource (https://doi.org/10.1093/jamia/ocaa196).

IRB

The N3C data transfer to NCATS is performed under a Johns Hopkins University Reliance Protocol # IRB00249128 or individual site agreements with NIH. The N3C Data Enclave is managed under the authority of the NIH; information can be found at https://ncats.nih.gov/n3c/resources.

Individual acknowledgments for core contributors

We gratefully acknowledge contributions from the following N3C core teams (leads designated with asterisks):

- CD2H Principal Investigators and N3C Lead Investigators: Melissa A. Haendel*, Christopher G. Chute*, Anita Walden
- NCATS CD2H and N3C Science Officer: Kenneth R. Gersing
- NCATS CD2H and N3C Program Officer: Leonie Misquitta
- NCATS N3C Leadership Team: Joni L. Rutter*, Kenneth R. Gersing*, Penny Wung Burgoon, Samuel Bozzette, Mariam Deacy, Christopher Dillon, Rebecca Erwin-Cohen, Nicole Garbarini, Valery Gordon, Michael G. Kurilla, Emily Carlson Marti, Sam G. Michael, Leonie Misquitta, Lili Portilla, Clare Schmitt, Meredith Temple-O'Connor
- Workstream, subgroup and administrative leaders: Melissa A. Haendel*, Tellen D. Bennett, Christopher G. Chute, David A. Eichmann, Justin Guinney, Warren A. Kibbe, Hongfang Liu, Philip R.O. Payne, Emily R. Pfaff, Peter N. Robinson, Joel H. Saltz, Heidi Spratt, Justin Starren, Christine Suver, Adam B. Wilcox, Andrew E. Williams, Chunlei Wu
- Key liaisons at data partner sites
- Regulatory staff at data partner sites
- Individuals at the sites who are responsible for creating the datasets and submitting data to N3C
- Data Ingest and Harmonization Team: Christopher G. Chute*, Emily R. Pfaff*, Davera Gabriel, Stephanie S. Hong, Kristin Kostka, Harold P. Lehmann, Richard A. Moffitt, Michele Morris, Matvey B. Palchuk, Xiaohan Tanner Zhang, Richard L. Zhu
- Phenotype Team (Individuals who create the scripts that the sites use to submit their data, based on the COVID and Long COVID definitions): Emily R. Pfaff*, Benjamin Amor, Mark M. Bissell, Marshall Clark, Andrew T. Girvin, Stephanie S. Hong, Kristin Kostka, Adam M. Lee, Robert T. Miller, Michele Morris, Matvey B. Palchuk, Kellie M. Walters
- N3C Community Project Management and Operations Team: Anita Walden*, Will Cooper, Patricia A. Francis, Rafael Fuentes, Alexis Graves, Julie A. McMurry, Andrew J. Neumann, Shawn T. O'Neil, Usman Sheikh, Elizabeth Zampino
- Analytics Team (Individuals who build the Enclave infrastructure, help create codesets, variables, and help Domain Teams and project teams with their datasets): Benjamin Amor*, Mark M. Bissell, Katie Rebecca Bradwell, Andrew T. Girvin, Amin Manna, Nabeel Qureshi

• Publication Committee Team: Mary Morrison Saltz*, Christine Suver*, Christopher G. Chute, Melissa A. Haendel, Julie A. McMurry, Andréa M. Volz, Anita Walden, Carolyn Bramante, Jeremy Richard Harper, Wenndy Hernandez, Farrukh M Koraishy, Federico Mariona, Amit Saha, Satyanarayana Vedula

Individual researchers who developed OMOP concept sets that have been used in this study: Daniel Meza, Arti Patel, Richard Zhu, Alfred Anzalone, Arti Patel, and Benjamin Amor.

We acknowledge support from many grants. In addition, access to N3C Data Enclave resources does not imply endorsement of the research project and/or results by NIH or NCATS.

Data partners with released data

The following institutions whose data is released or pending:

Available

Advocate Health Care Network—UL1TR002389: The Institute for Translational Medicine (ITM) • Boston University Medical Campus—UL1TR001430: Boston University Clinical and Translational Science Institute • Brown University—U54GM115677: Advance Clinical Translational Research (Advance-CTR) • Carilion Clinic—UL1TR003015: iTHRIV Integrated Translational health Research Institute of Virginia • Charleston Area Medical Center—U54GM104942: West Virginia Clinical and Translational Science Institute (WVCTSI) • Children's Hospital Colorado—UL1TR002535: Colorado Clinical and Translational Sciences Institute • Columbia University Irving Medical Center—UL1TR001873: Irving Institute for Clinical and Translational Research • Duke University—UL1TR002553: Duke Clinical and Translational Science Institute • George Washington Children's Research Institute—UL1TR001876: Clinical and Translational Science Institute at Children's National (CTSA-CN) • George Washington University—UL1TR001876: Clinical and Translational Science Institute at Children's National (CTSA-CN) • Indiana University School of Medicine—UL1TR002529: Indiana Clinical and Translational Science Institute • Johns Hopkins University—UL1TR003098: Johns Hopkins Institute for Clinical and Translational Research • Loyola Medicine—Loyola University Medical Center • Loyola University Medical Center—UL1TR002389: The Institute for Translational Medicine (ITM) • Maine Medical Center—U54GM115516: Northern New England Clinical & Translational Research (NNE-CTR) Network • Massachusetts General Brigham—UL1TR002541: Harvard Catalyst • Mayo Clinic Rochester—UL1TR002377: Mayo Clinic Center for Clinical and Translational Science (CCaTS) • Medical University of South Carolina—UL1TR001450: South Carolina Clinical & Translational Research Institute (SCTR) • Montefiore Medical Center—UL1TR002556: Institute for Clinical and Translational Research at Einstein and Montefiore • Nemours—U54GM104941: Delaware CTR ACCEL Program • NorthShore University HealthSystem—UL1TR002389: The Institute for Translational Medicine (ITM) • Northwestern University at Chicago—UL1TR001422: Northwestern University Clinical and Translational Science Institute (NUCATS) • OCHIN—INV-018455: Bill and Melinda Gates Foundation grant to Sage Bionetworks • Oregon Health & Science University—UL1TR002369: Oregon Clinical and Translational Research Institute • Penn State Health Milton S. Hershey Medical Center—UL1TR002014: Penn State Clinical and Translational Science Institute • Rush University Medical Center—UL1TR002389: The Institute for Translational Medicine (ITM) • Rutgers, The State University of New Jersey—UL1TR003017: New Jersey Alliance for Clinical and Translational Science • Stony Brook University—U24TR002306 • The Ohio State University—UL1TR002733: Center for Clinical and Translational Science • The State University of New

York at Buffalo—UL1TR001412: Clinical and Translational Science Institute • The University of Chicago—UL1TR002389: The Institute for Translational Medicine (ITM) • The University of Iowa—UL1TR002537: Institute for Clinical and Translational Science • The University of Miami Leonard M. Miller School of Medicine—UL1TR002736: University of Miami Clinical and Translational Science Institute • The University of Michigan at Ann Arbor—UL1TR002240: Michigan Institute for Clinical and Health Research • The University of Texas Health Science Center at Houston—UL1TR003167: Center for Clinical and Translational Sciences (CCTS) • The University of Texas Medical Branch at Galveston—UL1TR001439: The Institute for Translational Sciences • The University of Utah—UL1TR002538: Uhealth Center for Clinical and Translational Science • Tufts Medical Center—UL1TR002544: Tufts Clinical and Translational Science Institute • Tulane University—UL1TR003096: Center for Clinical and Translational Science • University Medical Center New Orleans—U54GM104940: Louisiana Clinical and Translational Science (LA CaTS) Center • University of Alabama at Birmingham—UL1TR003096: Center for Clinical and Translational Science • University of Arkansas for Medical Sciences—UL1TR003107: UAMS Translational Research Institute • University of Cincinnati—UL1TR001425: Center for Clinical and Translational Science and Training • University of Colorado Denver, Anschutz Medical Campus—UL1TR002535: Colorado Clinical and Translational Sciences Institute • University of Illinois at Chicago—UL1TR002003: UIC Center for Clinical and Translational Science • University of Kansas Medical Center—UL1TR002366: Frontiers: University of Kansas Clinical and Translational Science Institute • University of Kentucky—UL1TR001998: UK Center for Clinical and Translational Science • University of Massachusetts Medical School Worcester—UL1TR001453: The UMass Center for Clinical and Translational Science (UMCCTS) • University of Minnesota—UL1TR002494: Clinical and Translational Science Institute • University of Mississippi Medical Center—U54GM115428: Mississippi Center for Clinical and Translational Research (CCTR) • University of Nebraska Medical Center—U54GM115458: Great Plains IDeA-Clinical & Translational Research • University of North Carolina at Chapel Hill—UL1TR002489: North Carolina Translational and Clinical Science Institute • University of Oklahoma Health Sciences Center—U54GM104938: Oklahoma Clinical and Translational Science Institute (OCTSI) • University of Rochester—UL1TR002001: UR Clinical & Translational Science Institute • University of Southern California—UL1TR001855: The Southern California Clinical and Translational Science Institute (SC CTSI) • University of Vermont—U54GM115516: Northern New England Clinical & Translational Research (NNE-CTR) Network • University of Virginia—UL1TR003015: iTHRIV Integrated Translational health Research Institute of Virginia • University of Washington—UL1TR002319: Institute of Translational Health Sciences • University of Wisconsin-Madison—UL1TR002373: UW Institute for Clinical and Translational Research • Vanderbilt University Medical Center—UL1TR002243: Vanderbilt Institute for Clinical and Translational Research • Virginia Commonwealth University—UL1TR002649: C. Kenneth and Dianne Wright Center for Clinical and Translational Research • Wake Forest University Health Sciences—UL1TR001420: Wake Forest Clinical and Translational Science Institute • Washington University in St. Louis—UL1TR002345: Institute of Clinical and Translational Sciences • Weill Medical College of Cornell University—UL1TR002384: Weill Cornell Medicine Clinical and Translational Science Center • West Virginia University—U54GM104942: West Virginia Clinical and Translational Science Institute (WVCTSI).

Submitted

Icahn School of Medicine at Mount Sinai—UL1TR001433: ConduITS Institute for Translational Sciences • The University of Texas Health Science Center at Tyler—UL1TR003167:

Center for Clinical and Translational Sciences (CCTS) • University of California, Davis—UL1TR001860: UCDavis Health Clinical and Translational Science Center • University of California, Irvine—UL1TR001414: The UC Irvine Institute for Clinical and Translational Science (ICTS) • University of California, Los Angeles—UL1TR001881: UCLA Clinical Translational Science Institute • University of California, San Diego—UL1TR001442: Altman Clinical and Translational Research Institute • University of California, San Francisco—UL1TR001872: UCSF Clinical and Translational Science Institute.

Pending

Arkansas Children's Hospital—UL1TR003107: UAMS Translational Research Institute • Baylor College of Medicine—None (Voluntary) • Children's Hospital of Philadelphia—UL1TR001878: Institute for Translational Medicine and Therapeutics • Cincinnati Children's Hospital Medical Center—UL1TR001425: Center for Clinical and Translational Science and Training • Emory University—UL1TR002378: Georgia Clinical and Translational Science Alliance • HonorHealth—None (Voluntary) • Loyola University Chicago—UL1TR002389: The Institute for Translational Medicine (ITM) • Medical College of Wisconsin—UL1TR001436: Clinical and Translational Science Institute of Southeast Wisconsin • MedStar Health Research Institute—UL1TR001409: The Georgetown-Howard Universities Center for Clinical and Translational Science (GHUCCTS) • MetroHealth—None (Voluntary) • Montana State University—U54GM115371: American Indian/Alaska Native CTR • NYU Langone Medical Center—UL1TR001445: Langone Health's Clinical and Translational Science Institute • Ochsner Medical Center—U54GM104940: Louisiana Clinical and Translational Science (LA CaTS) Center • Regenstrief Institute—UL1TR002529: Indiana Clinical and Translational Science Institute • Sanford Research—None (Voluntary) • Stanford University—UL1TR003142: Spectrum: The Stanford Center for Clinical and Translational Research and Education • The Rockefeller University—UL1TR001866: Center for Clinical and Translational Science • The Scripps Research Institute—UL1TR002550: Scripps Research Translational Institute • University of Florida—UL1TR001427: UF Clinical and Translational Science Institute • University of New Mexico Health Sciences Center—UL1TR001449: University of New Mexico Clinical and Translational Science Center • University of Texas Health Science Center at San Antonio—UL1TR002645: Institute for Integration of Medicine and Science • Yale New Haven Hospital—UL1TR001863: Yale Center for Clinical Investigation.

International Committee of Medical Journal Editors (ICMJE) statement

Authorship was determined using ICMJE recommendations.

Author Contributions

Conceptualization: Tianchu Lyu, Chen Liang.

Data curation: Tianchu Lyu, Chen Liang, Yi-Wen Shih.

Formal analysis: Tianchu Lyu, Chen Liang, Berry Campbell.

Funding acquisition: Chen Liang, Jihong Liu, Berry Campbell, Peiyin Hung, Xiaoming Li.

Investigation: Chen Liang.

Methodology: Chen Liang.

Project administration: Chen Liang, Jihong Liu, Xiaoming Li.

Resources: Chen Liang, Berry Campbell.

Software: Chen Liang.

Supervision: Chen Liang, Jihong Liu.

Validation: Chen Liang, Berry Campbell, Nadia Ghumman.

Visualization: Chen Liang.

Writing – original draft: Tianchu Lyu, Chen Liang.

Writing – review & editing: Tianchu Lyu, Chen Liang, Jihong Liu, Berry Campbell, Peiyin Hung, Yi-Wen Shih, Nadia Ghumman, Xiaoming Li.

References

1. Allotey J, Stallings E, Bonet M, Yap M, Chatterjee S, Kew T, et al. Clinical manifestations, risk factors, and maternal and perinatal outcomes of coronavirus disease 2019 in pregnancy: living systematic review and meta-analysis. BMJ. 2020; 370:m3320. https://doi.org/10.1136/bmj.m3320 PMID: 32873575
2. Zambrano LD, Ellington S, Strid P, Galang RR, Oduyebo T, Tong VT, et al. Update: Characteristics of Symptomatic Women of Reproductive Age with Laboratory-Confirmed SARS-CoV-2 Infection by Pregnancy Status—United States, January 22-October 3, 2020. MMWR Morb Mortal Wkly Rep. 2020; 69 (44):1641–7. https://doi.org/10.15585/mmwr.mm6944e3 PMID: 33151921
3. Piekos SN, Roper RT, Hwang YM, Sorensen T, Price ND, Hood L, et al. The effect of maternal SARS-CoV-2 infection timing on birth outcomes: a retrospective multicentre cohort study. The Lancet Digital Health. 2022; 4(2):e95–e104. https://doi.org/10.1016/S2589-7500(21)00250-8 PMID: 35034863
4. Class QA, Lichtenstein P, Långström N, D'Onofrio BM. Timing of prenatal maternal exposure to severe life events and adverse pregnancy outcomes: a population study of 2.6 million pregnancies. Psychosom Med. 2011; 73(3):234–41. https://doi.org/10.1097/PSY.0b013e31820a62ce PMID: 21321257
5. Racicot K, Mor G. Risks associated with viral infections during pregnancy. The Journal of Clinical Investigation. 2017; 127(5):1591–9. https://doi.org/10.1172/JCI87490 PMID: 28459427
6. MedlinePlus. Pregnancy care [cited 2022 Mar. 02]. Available from: https://medlineplus.gov/ency/article/007214.htm.
7. Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, et al. Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. MEDINFO 2015: eHealth-enabled Health: IOS Press; 2015. p. 574–8. PMID: 26262116
8. Haendel MA, Chute CG, Bennett TD, Eichmann DA, Guinney J, Kibbe WA, et al. The National COVID Cohort Collaborative (N3C): Rationale, design, infrastructure, and deployment. Journal of the American Medical Informatics Association. 2020; 28(3):427–43.
9. Häyrinen K, Saranto K, Nykänen P. Definition, structure, content, use and impacts of electronic health records: A review of the research literature. International Journal of Medical Informatics. 2008; 77 (5):291–304. https://doi.org/10.1016/j.ijmedinf.2007.09.001 PMID: 17951106
10. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. Journal of the American Medical Informatics Association. 2013; 20 (1):144–51. https://doi.org/10.1136/amiajnl-2011-000681 PMID: 22733976
11. Leon LJ, McCarthy FP, Direk K, Gonzalez-Izquierdo A, Prieto-Merino D, Casas JP, et al. Preeclampsia and Cardiovascular Disease in a Large UK Pregnancy Cohort of Linked Electronic Health Records. Circulation. 2019; 140(13):1050–60.
12. Liyanage H, Williams J, Byford R, de Lusignan S. Ontology to identify pregnant women in electronic health records: primary care sentinel network database study. BMJ Health Care Inform. 2019; 26(1): e100013. https://doi.org/10.1136/bmjhci-2019-100013 PMID: 31272998
13. Roser BJ, Rubin SE, Nagarajan N, Wieland DL, Benfield NC. A data extraction algorithm for assessment of contraceptive counseling and provision. American Journal of Obstetrics and Gynecology. 2018; 218(3):333.e1–.e5. https://doi.org/10.1016/j.ajog.2017.11.578 PMID: 29175248
14. Sun W, Rumshisky A, Uzuner O. Temporal reasoning over clinical text: the state of the art. Journal of the American Medical Informatics Association. 2013; 20(5):814–9. https://doi.org/10.1136/amiajnl-2013-001760 PMID: 23676245
15. Spong CY. Defining "Term" Pregnancy: Recommendations From the Defining "Term" Pregnancy Workgroup. JAMA. 2013; 309(23):2445–6. https://doi.org/10.1001/jama.2013.6235 PMID: 23645117

16. Banda JM, Seneviratne M, Hernandez-Boussard T, Shah NH. Advances in Electronic Phenotyping: From Rule-Based Definitions to Machine Learning Models. Annual Review of Biomedical Data Science. 2018; 1(1):53–68. https://doi.org/10.1146/annurev-biodatasci-080917-013315 PMID: 31218278

17. Hornbrook MC, Whitlock EP, Berg CJ, Callaghan WM, Bachman DJ, Gold R, et al. Development of an Algorithm to Identify Pregnancy Episodes in an Integrated Health Care Delivery System. Health Services Research. 2007; 42(2):908–27. https://doi.org/10.1111/j.1475-6773.2006.00635.x PMID: 17362224

18. Kovačević A, Dehghan A, Filannino M, Keane JA, Nenadic G. Combining rules and machine learning for extraction of temporal expressions and events from clinical narratives. Journal of the American Medical Informatics Association. 2013; 20(5):859–66. https://doi.org/10.1136/amiajnl-2013-001625 PMID: 23605114

19. Li F, Du J, He Y, Song H-Y, Madkour M, Rao G, et al. Time event ontology (TEO): to support semantic representation and reasoning of complex temporal relations of clinical events. Journal of the American Medical Informatics Association. 2020; 27(7):1046–56. https://doi.org/10.1093/jamia/ocaa058 PMID: 32626903

20. 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. Pharmacoepidemiology and Drug Safety. 2013; 22(5):524–32. https://doi.org/10.1002/pds.3407 PMID: 23335117

21. Lin Y-K, Chen H, Brown RA. MedTime: A temporal information extraction system for clinical narratives. Journal of Biomedical Informatics. 2013; 46:S20–S8. https://doi.org/10.1016/j.jbi.2013.07.012 PMID: 23911344

22. Margulis AV, Setoguchi S, Mittleman MA, Glynn RJ, Dormuth CR, Hernández-Díaz S. Algorithms to estimate the beginning of pregnancy in administrative databases. Pharmacoepidemiology and Drug Safety. 2013; 22(1):16–24. https://doi.org/10.1002/pds.3284 PMID: 22550030

23. 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(2):e0192033. https://doi.org/10.1371/journal.pone.0192033 PMID: 29389968

24. Canelón SP, Burris HH, Levine LD, Boland MR. Development and evaluation of MADDIE: Method to Acquire Delivery Date Information from Electronic health records. International Journal of Medical Informatics. 2021; 145:104339. https://doi.org/10.1016/j.ijmedinf.2020.104339 PMID: 33232918

25. OHDSI. OMOP Common Data Model 2021 [cited 2022 Mar. 01]. Available from: https://www.ohdsi.org/data-standardization/the-common-data-model/.

26. Liu J, Hung P, Liang C, Zhang J, Qiao S, Campbell BA, et al. Multilevel determinants of racial/ethnic disparities in severe maternal morbidity and mortality in the context of the COVID-19 pandemic in the USA: protocol for a concurrent triangulation, mixed-methods study. BMJ Open. 2022; 12(6):e062294. https://doi.org/10.1136/bmjopen-2022-062294 PMID: 35688597

27. Liang C, Weissman S, Olatosi B, Poon EG, Yarrington ME, Li X. Curating a knowledge base for individuals with coinfection of HIV and SARS-CoV-2: a study protocol of EHR-based data mining and clinical implementation. BMJ Open. 2022; 12(9):e067204. https://doi.org/10.1136/bmjopen-2022-067204 PMID: 36100301

28. Lyu T, Liang C, editors. Predict Pregnancy Outcomes in the COVID-19 Pandemic Using Electronic Health Records and Machine Learning Approach. 2022 IEEE 10th International Conference on Healthcare Informatics (ICHI); 2022 11–14 June 2022.

29. AIM. AIM Data Guide 2020 [cited 2022 Mar. 07]. Available from: https://safehealthcareforeverywoman.org/wp-content/uploads/AIM_DataGuide_Oct2020.pdf.

30. Song YP, Skinner JP, Bynum JMDMPH, Sutherland JP, Wennberg JEMDMPH, Fisher ESMDMPH. Regional Variations in Diagnostic Practices. The New England Journal of Medicine. 2010; 363(1):45–53. https://doi.org/10.1056/NEJMsa0910881 PMID: 20463332

31. Blotière P-O, Weill A, Dalichampt M, Billionnet C, Mezzarobba M, Raguideau F, et al. Development of an algorithm to identify pregnancy episodes and related outcomes in health care claims databases: An application to antiepileptic drug use in 4.9 million pregnant women in France. Pharmacoepidemiology and Drug Safety. 2018; 27(7):763–70. https://doi.org/10.1002/pds.4556 PMID: 29763992

32. MacDonald SC, Cohen JM, Panchaud A, McElrath TF, Huybrechts KF, Hernández-Díaz S. Identifying pregnancies in insurance claims data: Methods and application to retinoid teratogenic surveillance. Pharmacoepidemiology and Drug Safety. 2019; 28(9):1211–21. https://doi.org/10.1002/pds.4794 PMID: 31328328

33. Manson JM, McFarland B, Weiss S. Use of an Automated Database to Evaluate Markers for Early Detection of Pregnancy. American Journal of Epidemiology. 2001; 154(2):180–7. https://doi.org/10.1093/aje/154.2.180 PMID: 11447053

34. Naleway AL, Gold R, Kurosky S, Riedlinger K, Henninger ML, Nordin JD, et al. Identifying pregnancy episodes, outcomes, and mother–infant pairs in the Vaccine Safety Datalink. Vaccine. 2013; 31 (27):2898–903. https://doi.org/10.1016/j.vaccine.2013.03.069 PMID: 23639917