

The Identification of Pregnancies Within The Full Feature-General Practice Research Database

Scott Devine, PhD¹; Suzanne West, PhD¹; Elizabeth Andrews, PhD²; Patricia Tennis, PhD²; Susan Eaton, MSPH³; John Thorp, MD¹; Andrew Olshan, PhD¹

¹University of North Carolina School of Public Health, Chapel Hill, North Carolina, United States; ²RTI Health Solutions, Research Triangle Park, NC, United States;

³The UK Medicines and Healthcare Products Regulatory Agency, London, United Kingdom

ABSTRACT

Background: The Full Feature-General Practice Research Database (FF-GPRD) is a large, anonymized, longitudinal electronic patient medical records database providing clinical information from general practitioners. The GPRD data contain approximately 46 million patient-years of follow-up, representing 10.11 million unique patients. Over 460 general practices in the UK currently submit data to the FF-GPRD on 3.23 million patients. An algorithm for the identification of pregnancies in the GPRD would allow researchers to use this data for pharmacoepidemiology research of pregnancy and birth outcomes.

Objectives: Our objective was the development of an algorithm for identifying and classifying pregnancy outcomes using the FF-GPRD.

Methods: We developed an algorithm that identifies pregnancy outcomes within a woman's medical record while handling conflicts and eliminating redundancy of medical coding. The algorithm identified pregnancies in 15-to-45-year-old women between January 1, 1987 and September 14, 2004. We identified live births, stillbirths, spontaneous abortions, and terminations within a woman's clinical record and established the appropriate ultimate pregnancy outcome based upon an event hierarchy established by the authors. The algorithm was evaluated for accuracy through a series of alternate analyses and reviews of electronic records.

Results: We analyzed 98,922,326 records from 980,474 individuals and identified 255,400 women who had 374,878 pregnancies. There were 271,613 live births, 2,106 pre- or postterm births, 1,191 multi-fetus deliveries, 55,614 spontaneous abortions or miscarriages, 43,264 elective terminations, 7 stillbirths in combination with a live birth, and 1,083 stillbirths or fetal deaths. At least one marker of pregnancy care was identifiable for 330,153 pregnancies. Eighty-four percent of these 330,153 pregnancies had data available going back at least 180 days prior to the first marker of pregnancy care.

Conclusions: We believe our pregnancy identification algorithm will be a useful tool for researchers. Its hierarchical approach to identifying the pregnancy outcome builds upon other methods while implementing additional steps to minimize potential misclassification of pregnancy outcomes.

CONFLICT OF INTEREST

This research was funded by the Agency for Healthcare Research and Quality as part of the UNC Center for Education and Research on Therapeutics (Alan Stiles, CERTs PI; award number 2 U18 HS10397) at the University of North Carolina at Chapel Hill. Susan Eaton, MSPH, is an employee of The UK Medicines and Healthcare Products Regulatory Agency.

BACKGROUND

Large automated electronic medical records databases are extremely valuable for the study of medication use during pregnancy; several recent studies have highlighted their usefulness.^{1,2} For these databases to be useful, they must provide comprehensive health care information about women before and during pregnancy and at delivery. Often a challenge exists for researchers because the time period in which a woman is pregnant is not easily identifiable in the database, requiring researchers to develop approaches to identify these records. We present a detailed report on a procedure for identifying pregnancy time periods in the Full Feature-General Practice Research Database (FF-GPRD). This procedure was developed and used in a study of neural tube defect identification in the GPRD.³

METHODS

Data Source

The GPRD is a large anonymized, longitudinal database of patient's electronic medical records. It provides clinical information based on general practitioner (GP) records. The GPRD data contain approximately 46 million patient-years of follow-up, representing 10.1 million unique patients.⁴ Over 460 general practices in the UK are currently submitting data to the GPRD on 3.2 million patients or approximately 5% of the UK population.^{4,5}

Identification of Pregnancies

Among those registered in GPRD, we searched for records with one of 5,266 medical codes indicative of a pregnancy after January 1, 1987 in women aged 15 to 45 years. An electronic-record-based, three-step pregnancy identification procedure was designed to handle the complexities of identifying pregnancies in the GPRD data.

- Step 1 – Identify and remove duplicate end-of-pregnancy codes for each woman.** For each woman, duplicate records were addressed within each of three pregnancy categories—(1) stillbirths; (2) spontaneous and elective terminations; and (3) live births—by designating her earliest end-of-pregnancy code within a category as the index event, then disregarding all subsequent end-of-pregnancy codes in that same category within a predetermined time frame.
- Step 2 – Apply a hierarchical coding scheme to select the final pregnancy outcome for each pregnancy.** A hierarchy was devised through a review of 10,000 electronic records of women with a suspected pregnancy. We ordered each woman's event codes chronologically and then ranked them based on this hierarchy: stillbirths (Category 1), spontaneous and elective terminations (Category 2), and live births and deliveries (Category 3). Each end-of-pregnancy record date was compared to every other end-of-pregnancy record in the woman's profile and the highest category event within a prespecified number of days from another lower category event was selected.
- Step 3 – Use final pregnancy outcome to identify the first pregnancy-care marker for each pregnancy.** For each end-of-pregnancy event identified after applying Steps 1 and 2, the earliest marker of pregnancy care within 280 days of end of pregnancy, or if appropriate, between two end-of-pregnancy dates, was selected. End-of-pregnancy medical codes were then matched to a pregnancy-care marker creating a complete pregnancy profile.

Figure 1. Implementation of Hierarchical Coding Scheme to Select the Final Pregnancy Outcome in the General Practice Research Database



RESULTS

Figure 2. Succession of Electronic Medical Records Through Pregnancy Identification Procedure

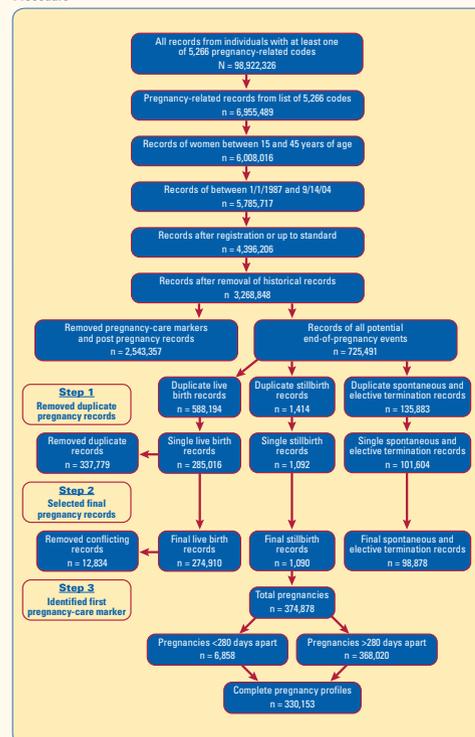


Table 1. End-of-Pregnancy Events Found by the Identification Procedure in Women Aged 15 to 45 Years in the General Practice Research Database, 1987-2004

Event Type	Total N (% w/ PCM)
Stillbirths	1,090 (90.0)
Elective terminations	43,264 (74.1)
Spontaneous terminations	55,614 (76.5)
Multibirth deliveries	1,191 (91.1)
Pre-/Postterm deliveries	2,106 (92.2)
Live birth/deliveries	271,613 (92.6)
Total	374,878 (88.1)

* PCM = pregnancy-care marker.

Table 2. Distribution of Pregnancy in Women Aged 15 to 45 Years in the General Practice Research Database, 1987-2004

Number of Pregnancies	Women with Pregnancies n (%)	Pregnancies by These Women n (%)
1	169,869 (66.5)	169,869 (45.3)
2	60,930 (24.0)	121,860 (32.5)
3	17,879 (7.0)	53,637 (14.3)
4	4,839 (2.0)	19,356 (5.1)
5	1,353 (0.5)	6,765 (1.8)
6	383 (0.1)	2,298 (0.6)
7	106 (0.0)	742 (0.2)
8	26 (0.0)	208 (0.0)
9	7 (0.0)	63 (0.0)
10	8 (0.0)	80 (0.0)
Total	255,400 (100.0)	374,878 (100.0)

- The mean number of pregnancies per woman was 1.5, with a median of 1 and a maximum of 10.
- Of the women with more than four pregnancies, the mean, median, and range in years between the first and last pregnancy was 7.7, 7.5, and 1.2 years to 14.5 years, respectively.

Figure 3. The Distribution of Days Between First Pregnancy-Care Marker and its Associated End-of-Pregnancy Event in the General Practice Research Database, 1987-2004

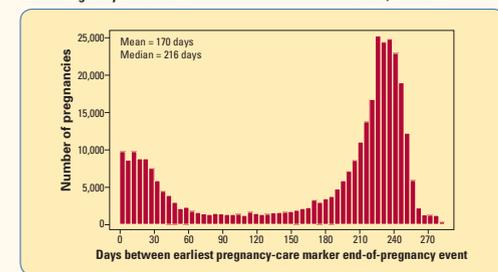


Table 3. First Pregnancy-Care Marker Details Stratified by End-of-Pregnancy Event Types in the General Practice Research Database, 1987-2004

	Number of Events w/ PCM	Missing PCM	Mean Number of Days	Median Number of Days	Inter-quartile Range
All EOP Events	330153 (88.1)	44725 (11.9)	170	216	149
Stillbirths	981 (90.0)	109 (10.0)	163	170	92
Elective terminations	32056 (74.1)	11208 (25.9)	45	21	24
Spontaneous abortions	42547 (76.5)	13067 (23.5)	46	26	33
Multiple live births	1085 (91.1)	106 (8.9)	188	203	44
Pre-/Postterm births	1942 (92.2)	164 (7.8)	185	197	67
Live births	251542 (92.6)	20071 (7.4)	207	226	39

EOP = end of pregnancy; PCM = pregnancy-care marker.

Table 4. Duration of Available General Practice Research Database Data Prior to First Pregnancy-Care Markers in the General Practice Research Database, 1987-2004

Days of Data Before First PCM	Stillbirths and Fetal Deaths n (%)	Spontaneous and Elective Abortions n (%)	Live Births/Deliveries n (%)
>30 days	866 (88.3)	69,692 (93.4)	222,590 (87.4)
>60 days	830 (84.6)	68,043 (91.2)	214,593 (84.3)
>90 days	813 (82.9)	66,615 (89.3)	208,796 (82.0)
>180 days	771 (78.6)	62,665 (84.0)	194,512 (76.4)
360 days	679 (69.2)	56,064 (75.1)	171,417 (67.3)

PCM = pregnancy-care marker.

DISCUSSION

Strengths

- At least one pregnancy care record was available for 88% of all pregnancies, and 78% of the identified pregnancies had records accessible at least 300 days prior to the pregnancy outcome.
- Over 78% of the complete pregnancy profiles had medical history records going back at least 180 days before the first pregnancy-care marker.
- This algorithm reduces the chance of selecting an indeterminate pregnancy outcome as the final end-of-pregnancy code by detecting outcomes that were out of order in the woman's record.
- A strength of both the algorithm presented here and the GPRD is the ability to identify recorded spontaneous terminations.

Limitations

- There is a potential for the incomplete ascertainment of pregnancies in the GPRD, thus limiting its use for prevalence calculations.
- As for any electronic medical record or claims-based database, information on pregnancies that occurred prior to the patient's registration is incomplete.
- Although this algorithm attempts to minimize misclassification, the potential for misclassification of pregnancy outcomes with this algorithm still exists.

CONCLUSIONS

We were able to identify details on a large number of pregnancies in the GPRD. The details included information on potential exposures in a woman prior to and during all stages of pregnancy, and potential pregnancy outcomes not limited to live births. Details on other outcomes, including reported spontaneous abortions, are not readily available from other data sources. This resource should prove valuable for future research on medication exposures during pregnancy.

REFERENCES

- Andrade SE, Raebel MA, Morse AN, Davis RL, Chan KA, Finkelstein JA et al. Use of prescription medications with a potential for fetal harm among pregnant women. *Pharmacoepidemiol Drug Saf*. 2006;15(6):546-54.
- Cooper WO, Hernandez-Diaz S, Arbogast PG, Dyer S, Gideon PS, et al. Major congenital malformations after first-trimester exposure to ACE inhibitors. *N Engl J Med*. 2006 Jun 8;354(23):2443-51.
- Devine ST, West S, Andrews E, Tennis P, Eaton S, Thorp J, Olshan A. Validation of Neural Tube Defects in the General Practice Research Database. *Pharmacoepidemiology and Drug Safety*. 2007 International Conference on Pharmacoepidemiology & Therapeutic Risk Management; Quebec City, Canada; August 20, 2007.
- Medicines and Healthcare products Regulatory Agency (MHRA). Full-featured General Practice Research Database. The premier source of healthcare data from primary care. What is GPRD? <http://www.gprd.com/html/index.asp?main=whatsGPRD.htm>. 2004. Accessed May 20, 2004.
- MHRA. Full-featured General Practice Research Database. The premier source of healthcare data from primary care. Facts and Figures. <http://www.gprd.com/whygprd/factsandfigures.asp>. Accessed January 23, 2006.

CONTACT INFORMATION

Scott Devine, RPH, MPH, PhD

2101B McGavran-Greenberg Hall
School of Public Health, CB#7435
University of North Carolina
Chapel Hill, NC 27599

Phone: 314-997-4456
Email: sdevine@unc.edu

Presented at: 23rd International Conference on Pharmacoepidemiology and Therapeutic Risk Management August 19-22, 2007
Quebec City, Quebec, Canada