BMJ Health & Care Informatics

Ontology to identify pregnant women in electronic health records: primary care sentinel network database study

Harshana Liyanage, John Williams, Rachel Byford, Simon de Lusignan

To cite: Liyanage H, Williams J, Byford R, *et al*. Ontology to identify pregnant women in electronic health records: primary care sentinel network database study. *BMJ Health Care Inform* 2019;**26**:e100013. doi:10.1136/bmjhci-2019-100013

► Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136bmjhci-2019-100013).

HL and JW contributed equally.

Accepted 28 May 2019

Department of Clinical and Experimental Medicine, Faculty of Health and Medical Sciences, University of Surrey, Guildford, Surrey, UK

Correspondence to
Dr Harshana Liyanage, Department of Clinical and Experimental Medicine, University of Surrey, Guildford GU2 7XP, UK;
h.s.liyanage@surrey.ac.uk

ABSTRACT

Objective To develop an ontology to identify pregnant women from computerised medical record systems with dissimilar coding systems in a primary care sentinel network.

Materials and methods We used a three-step approach to develop our pregnancy ontology in two different coding schemata, one hierarchical and the other polyhierarchical. We developed a coding system–independent pregnancy case identification algorithm using the Royal College of General Practitioners Research and Surveillance Centre sentinel network database which held 1.8 million patients' data drawn from 150 primary care providers. We tested the algorithm by examining individual patient records in a 10% random sample of all women aged 29 in each year from 2004 to 2016. We did an external comparison with national pregnancy data. We used χ^2 test to compare results obtained for the two different coding schemata.

Results 243 005 women (median age 29 years at start of pregnancy) had 405 591 pregnancies from 2004 to 2016 of which 333 689 went to term. We found no significant difference between results obtained for two populations using different coding schemata. Pregnancy mean ages did not differ significantly from national data.

Discussion This ontologically driven algorithm enables consistent analysis across data drawn from populations using different coding schemata. It could be applied to other hierarchical coding systems (eg, International Classification of Disease) or polyhierarchical systems (eg, SNOMED CT to which our health system is currently migrating).

Conclusion This ontological approach will improve our surveillance in particular of influenza vaccine exposure in pregnancy.

Summary box

What is already known?

- Data recorded about pregnancies are of low quality. Poor information sharing between the multiple agencies who provide care during and after pregnancy contributes to this.
- Ontological approaches can greatly enhance case finding in situations where the attributes used to identify cases are poorly recorded.
- Algorithmic approaches to pregnancy case finding have been limited to date.

What does this paper add?

- This is the first time that the combination of an algorithmic search based on a systematic ontological approach has been used to identify pregnancies based on patient-level coded information.
- This the first study to describe the application to patient-level data of a reliable search algorithm which determines the start and end dates of pregnancies for individual women, either with reference to a single point in time or to a defined period of time.
- The pregnancy toolkit described is enabling the first near-real-time monitoring dashboard for influenza vaccine uptake in primary care.

How might it impact on clinical practice in the foreseeable future?

- The pregnancy toolkit provides a robust solution to the problem of reliable identification of discrete pregnancies in patient-level data and should be of particular value to sentinel networks.
- Examples of potential future use include the surveillance and monitoring of infectious diseases and adverse events in pregnancy, adherence to guidelines relating to drug use in pregnancy, associations between drug misuse during pregnancy and birth abnormalities, quality of postnatal care for women with gestational diabetes and pre-gestational type 2 diabetes.

BACKGROUND AND SIGNIFICANCE

WHO recommends influenza vaccination in pregnancy,[1] but monitoring is challenging because pregnancy is often poorly recorded in medical record systems. The reliable identification of pregnant women is frequently required for surveillance and monitoring of infectious diseases and adverse events.[2 3] Pregnant women who are not immunised against influenza remain at risk and sentinel networks need to be able to monitor this.[4–7] In England, the Royal College of General Practitioners (RCGP) Research and Surveillance Centre (RSC) monitors influenza and other infection and contributes to the study of influenza vaccine effectiveness[8]; but is not currently reporting vaccine uptake or effectiveness in pregnancy.

Pregnancy case finding is challenging because multiple agencies provide care

during and after pregnancy, each with their own recording systems in different locations using varying formats exacerbating poor information sharing. Over time, English general practitioners (GPs) have become progressively less involved in obstetric care, but the GP is still the first point of contact in 66% of all pregnancies with 90% subsequently undergoing a postnatal check.[9] It has been shown in the Netherlands that pregnancies are often not recorded in coded GP data.[10] In many pregnancies, data are stored non-electronically in a patient's handheld notes to try to overcome some of the difficulties with data sharing.[11]

Algorithmic approaches to pregnancy case finding have been limited.[12–16] Studies have explored the safety of medication use during pregnancy.[3 17–19] A review of automated methods used to ascertain the beginning and duration of pregnancy in health databases[20] revealed that most studies used a wide range of markers to identify pregnancies[3 19] while others were limited to specific markers (eg, gestational age).[17] One study, using a rule-based algorithm, experienced a high false-positive rate attributed to incomplete recording of key clinical events.[21] Others used standardised methods to quantify maternal mortality and 'near misses'[22–24] based on the use of coded death certificates rather than on identifying the timing and duration of discrete pregnancies in patient-level data. A pregnancy inferencing algorithm was developed using a common data model (CDM) to detect pregnancies across databases.[16] There is no obvious gold standard against which to measure the results of these algorithmic approaches.

Ontological approaches can greatly enhance case finding when the attributes used to identify cases are poorly recorded. They have been used successfully to identify and classify patients with chronic diseases such as diabetes in GP databases.[25 26]

To address the lack of reporting of influenza and vaccine uptake in pregnancy in the RCGP RSC sentinel network, we developed an ontologically driven algorithm to identify pregnant women.

MATERIALS AND METHODS

Overview

Our conceptual approach was to look for data that signified the start and end of pregnancy, and whether or not the pregnancy had run to term (full length). We also applied minimum and maximum time intervals between the possible start and end dates for each pregnancy. We accommodated two differently structured clinical coding systems in use by practices in the RCGP RSC.

Study cohort

We applied our approach to data from the RCGP RSC. At December 2016, the database consisted of coded primary care data from 178 primary care providers, in the UK described as general practices, with a total population of around 1.7 million patients.[27 28] We extracted data which had been collected from 2004 to 2016. The main clinical coding systems used in the database were Read version 2 (Read v2), a hierarchical system with disease headers, symptom codes not unlike a clinical modification of the International Classification of Disease (ICD); and Clinical Terms version 3 (CTV3), a polyhierarchical system, sharing many of the characteristics of the Systematised Nomenclature of Medicine Clinical Terms (SNOMED CT).[29]

Three-step ontological approach to identify pregnancy cases

We used a three-step ontological process[1]: ontology construction—identification of key concepts that define a case and their inter-relationship[2 30 31]; coding layer derived by application of the ontology to the coding schemata[3]; a logical data extraction process, which we implemented as an algorithm.[12 16 30 31]

Step 1: ontological layer

This process involved a literature review and identifying key concepts and their relationships. We searched the evidence base to identify similar ontologically based work associated with patient-level data and found none. Therefore, we consulted a group of experienced GPs to assist in generating the initial concept definition for the pregnancy ontology. We considered levels of certainty:

- *Definite*: case ascertainment with a high degree of certainty (eg, using concepts declaring a diagnosis).
- *Probable*: case ascertainment with a moderate degree of certainty (eg, using concepts related to a pattern of symptoms and signs suggestive of a case).
- *Possible*: case ascertainment with a low degree of certainty (eg, using concepts related to laboratory tests without clear indication of result).

Because our prime objective was to maximise the certainty around pregnancy detection, we limited the ontology to definite concepts. We used the Protégé ontology development environment and Web Ontology Language specification. The pregnancy ontology is published online.[32]

Step 2: coding layer

We mapped from ontological concepts to clinical codes classifying individual codes according to the degree of certainty with which they mapped from ontological concepts[33 34]:

- Direct mapping: ontological concept maps directly to specific coding scheme term(s).
- Partial mapping: ontological concept can only be mapped to term(s) in the coding system which is/are incompletely or partially representative of the ontological meaning
- No clear mapping: ontological concept cannot be mapped to any term(s) in the coding scheme with any certainty.

Each ontological concept had none, few or many representations in a given coding scheme.

The ontology was applied to Read v2 and CTV3 coding schemes which have different hierarchical structures.[35]

Read v2 is hierarchical like a branching tree; codes are meaningful and determine the hierarchical positioning of concepts whereas in CTV3 the code is meaningless and does not determine hierarchical position. The application of the ontology to the two coding schemes therefore differed significantly. We identified codes that directly mapped from all of the ontology main concepts. Codes sets were originally generated by searching the relevant parts of the Read v2 hierarchy. Subsequently, for each overarching ontological concept, the resulting data were categorised into subconcepts. The existence of these more granular subconcepts facilitated identification and addition of missing codes, re-allocation of codes to more suitable main concepts and complete removal of those codes that did not truly fit the definitions of any of the main concepts. The process was then repeated for the CTV3 clinical terminology.

Step 3: logical data extract layer

We developed a pregnancy case identification algorithm to handle the extraction and processing of the codes resulting from the preceding steps taking into consideration the range of possible ways that pregnancies could be recorded in routine data. We optimised both the algorithm and the ontology by conducting iterative developmental testing, searching for anomalies, identifying causes and implementing changes to improve sensitivity and specificity.

Internal and external validation

In the absence of any reliable external gold standard, we limited ourselves to testing the outputs of the algorithm by inspecting the individual records in a random 10% sample of all women in the RCGP RSC database aged 29, the median age for start of pregnancy, for each year from 2004 to 2016. We checked that all pregnancies identified by the algorithm were clinically valid. We also checked those cases with at least one pregnancy ontology code entry where the algorithm had not identified pregnancy. Review of resulting descriptive statistics provided a degree of internal validation. For external validation, we compared average age at time of pregnancy with ONS over the period 2004 to 2016. We adjusted RCGP RSC ages by adding 40 weeks because ONS takes age at time of delivery while RCGP RSC takes age at start of pregnancy. We used χ^2 test where appropriate to compare data from the two different coding schemata.

Ethical approval and consent to participate

Better identification of pregnant women and whether they have been exposed to influenza vaccine is a key public function of the RCGP RSC. This development was approved by the RCGP study approval committee. Use of the online Health Research Authority Decision tool (1 August 2018) indicated that further approval was not required for this research project which is exclusively based in England. This complies with the requirements for ethical review set out in section 2.3 of the Governance Arrangements for Research Ethics Committees, published by the UK Health Departments in May 2011.

RESULTS

Ontological layer

The resulting ontology had three overarching main concepts:

1. Definite pregnancy: Any concept indicating a definite state of pregnancy.
2. Term pregnancy (delivered): Any concept indicating that a potentially viable baby had been delivered.
3. Non-term pregnancy: Any concept indicating the early end of a pregnancy (most commonly by miscarriage/termination).

Each of these main concepts was further subdivided into a number of subcategories (online supplementary file 1). We noted that a relatively small number of subcategories accounted for more than 99% of the representation of each main concept. This was true for each individual coding scheme (figure 1) although the relative contributions of the subcategories differed between the coding schemes. Identifying the most important subcategories

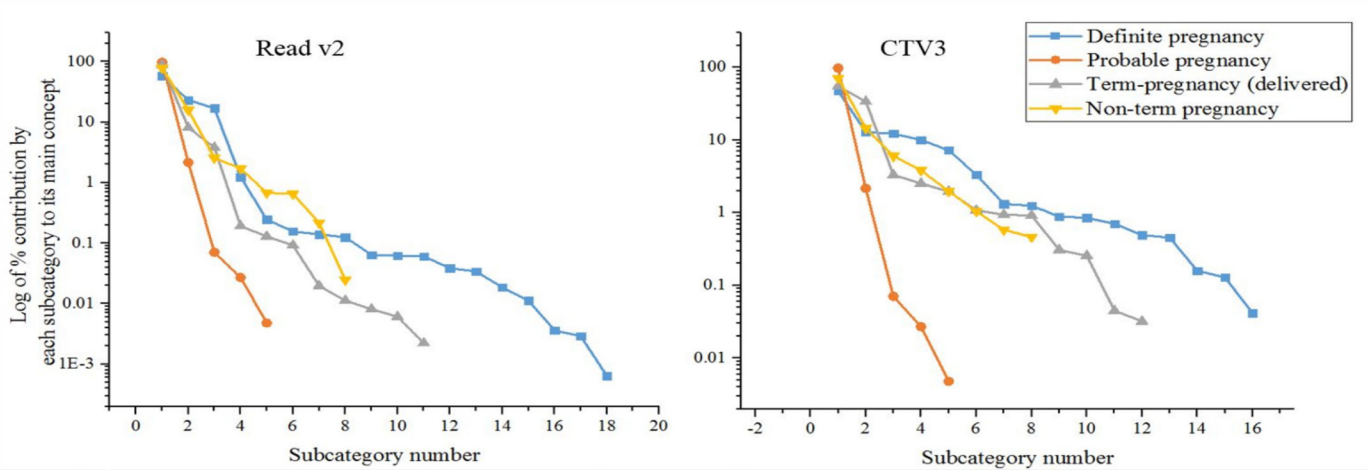


Figure 1 Subcategories for two coding schemes.

Table 1 Comparison of findings by coding scheme for 2015

	Read v2	CTV3	All systems
Demographics			
No of women registered	719 451	48 347	767 798
Summary of age (minimum/mean/maximum)	0/40.67/110	0/41.86/105	0/40.75/110
Pregnancy outcome type			
All pregnancies	47 278	2760	50 038
Term pregnancies (delivered)	41 735	2676	44 411
Non-term	5543	84	5627
χ^2=53.834, p<0.001			
Pregnancy cases resulting in delivery			
No of women pregnant	40 398	2574	42 972
Women with one pregnancy	39 065	2472	41 537
Women with two pregnancies	1329	102	1431
Women with three pregnancies	4	0	4
χ^2=0.46906, p=0.4934			

assisted in applying the ontology to a new coding scheme. The number of subcategories accounting for 99% of the representation of each overarching concept was 8 for 'Definite pregnancy', 5 for 'Term pregnancy (delivered)' and 7 for 'Non-term pregnancy'. Details of the numbered subcategories in order of importance are in online supplementary file 1.

Coding layer

Table 1 provides a comparison of findings by coding scheme (Read v2 or CTV3). The number of women registered under the CTV3 scheme was much smaller than for Read v2. With the exception of non-term pregnancies, the numbers were within very similar proportions. A χ^2 test performed with Yates continuity correction revealed no significant difference between Read v2 and CTV3 systems for term and non-term pregnancies per 1000 women (p=0.17).

The graphs in figure 2 compare the total pregnancy counts per woman (expressed as a percentage of the total number of pregnant women) for the pregnancy outcomes term pregnancy (delivered) and non-term. It can be seen that, as expected, for both outcomes, the number of women rapidly drops away for increasing total pregnancy counts. The rates of drop-off for the two different coding schemes fit very closely especially for term-pregnancy (delivery) outcome.

Logical data extract layer

We designed the pregnancy case identification algorithm to move forward in time through individual medical records looking for any codes defined by the pregnancy

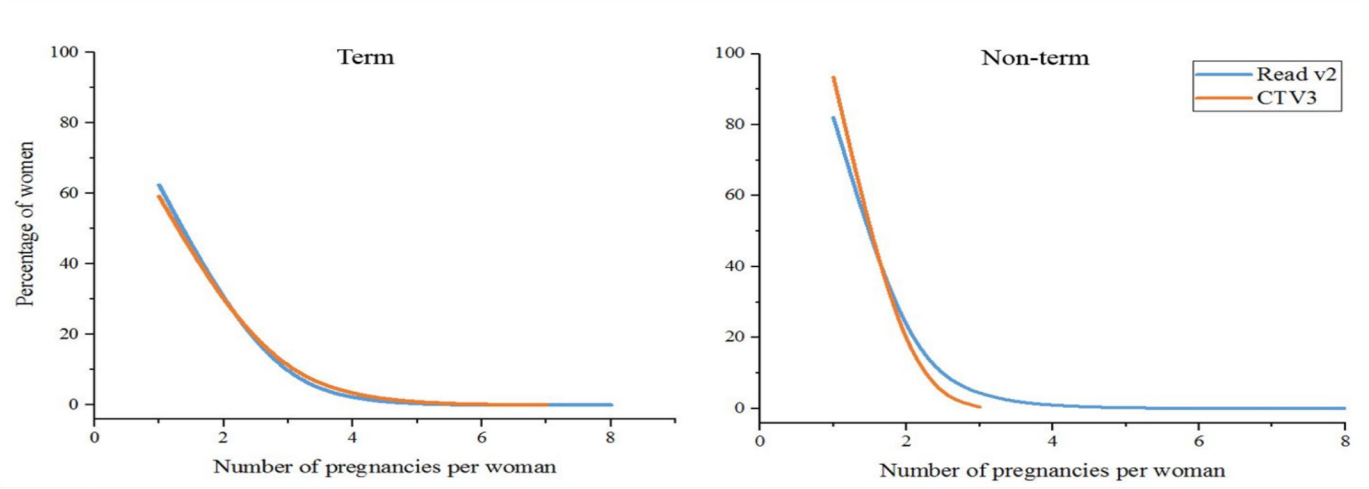


Figure 2 Graph of pregnancy counts per woman for term pregnancy (delivered) and non-term pregnancies (for Read v2 and CTV3 systems). Term pregnancies (deliveries) (for number of pregnancies from 1 to 4): χ^2=6.4592, p<0.1; non-term outcome (for number of pregnancies 1–3): X^2=57.834, p<0.001.

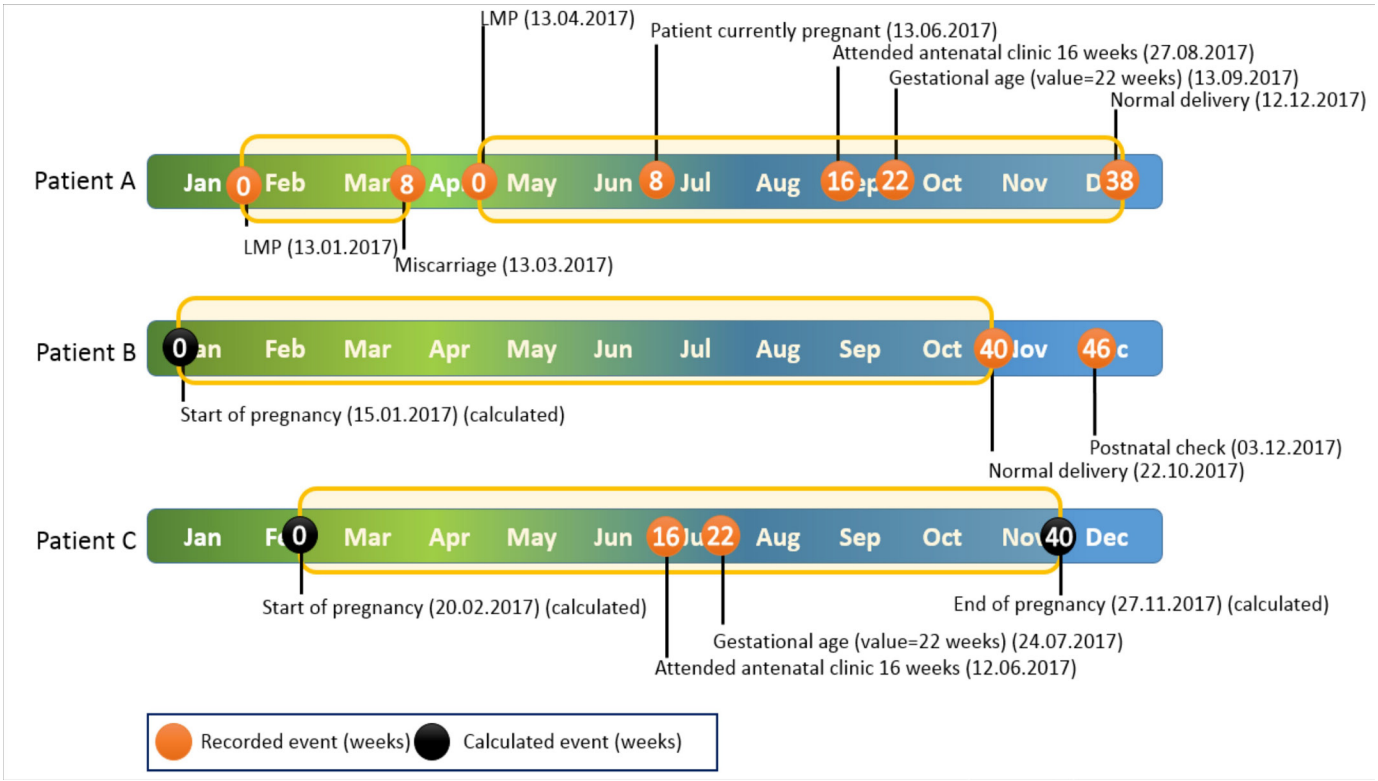


Figure 3 Timeline of three example patients and their pregnancies as identified by pregnancy ontology during 2017. Patient A had two pregnancies with clearly recorded start and end indicator events recorded. Patient B has one pregnancy with only the end of pregnancy indicator events recorded. Patient C has one pregnancy with both start and end dates calculated. LMP, last menstrual period.

ontology. Codes relating to the ontological definite pregnancy main concept were primarily used to calculate the start date of a pregnancy whereas codes relating to the other two ontological main concepts were used to determine the end date and also the pregnancy outcome (eg, delivery of baby or non-term, typically as a result of a miscarriage).

Description of resulting algorithm

The algorithm was designed to search through time until it found a first code indicating pregnancy and, based on the characteristics of the concept represented by that code, to assign pregnancy start and end dates and also a search period end date set beyond the expected pregnancy end date (figure 3). The search was extended in this way to ensure that later codes relating to a given pregnancy were appropriately processed and not misinterpreted as representing the start of a subsequent pregnancy. Where successive codes were found in a pregnancy search period, the algorithm recalculated the start and/or end dates, depending on the characteristics of the code. Where that calculation resulted in an earlier start and/or end date, then generally these were to be adopted. The precise processing behaviour of the algorithm on encountering codes during a search period depended on the ontological categories of the codes found.

Details of the rules governing algorithm processing, along with configurable table-driven parameters, can be found on the ontology website.[36]

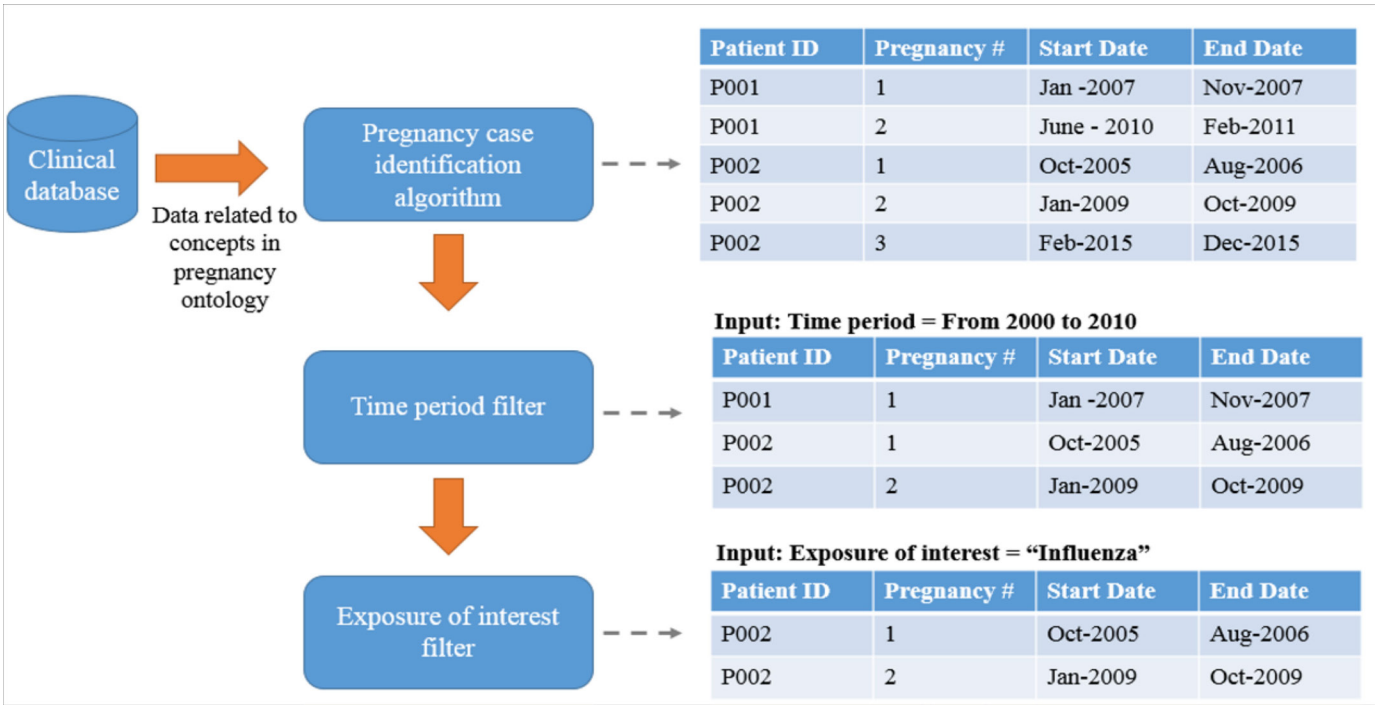
Inputs and outputs

The minimum required inputs relating to every woman in the source database were a unique identifier, a date of birth formatted to support calculation of age at the time of any event, and the combination of code, associated term and clinically relevant date for every instance of every pregnancy ontology dataset code.

The output from the algorithm was directed to two tables. The 'Pregnancy Episode' table held data about every pregnancy episode found including patient ID, pregnancy number, patient age at the start of the pregnancy, pregnancy start and end dates, and the pregnancy outcome. The 'Pregnancy Episode Codes' table held information about every pregnancy ontology dataset code encountered during any single pregnancy episode including the event date and a composite key made up of patient and pregnancy number.

Pregnancy case identification workflow

The pregnancy case identification workflow started with the creation of the input table as described above. The input table was then parsed to identify pregnancies, resolve any ambiguous cases and finally write the results into the two output tables. Once the pregnancy case



Patient ID	Pregnancy #	Start Date	End Date
P001	1	Jan -2007	Nov-2007
P001	2	June - 2010	Feb-2011
P002	1	Oct-2005	Aug-2006
P002	2	Jan-2009	Oct-2009
P002	3	Feb-2015	Dec-2015

Patient ID	Pregnancy #	Start Date	End Date
P001	1	Jan -2007	Nov-2007
P002	1	Oct-2005	Aug-2006
P002	2	Jan-2009	Oct-2009

Patient ID	Pregnancy #	Start Date	End Date
P002	1	Oct-2005	Aug-2006
P002	2	Jan-2009	Oct-2009

Figure 4 Pregnancy ontology function pipeline.

identification algorithm had been executed, the output tables could be repeatedly reused until the next source database update (eg, influenza in pregnancy) (figure 4).

Algorithm output check

Each pregnancy identified by the pregnancy case identification algorithm in the 10% random samples of women aged 29 was carefully checked and found to be clinically credible. The records of women with at least one pregnancy ontology code but no pregnancy were also checked. There was no evidence that the algorithm had missed any clinically identifiable pregnancy in this group.

Descriptive statistics

The application of the toolkit to the whole RCGP RSC database detected 405 591 pregnancies with start dates ranging from 28 February 2004 up to 11 November 2016. Table 2 shows details of the entire content of the resulting output tables.

The majority of pregnancies ended with a delivery outcome. In total, 12 018 pregnancies with unclassified outcome had not yet resolved either into a term pregnancy (delivered) or non-term pregnancy at the time that the data were extracted. The median duration of term (delivered) pregnancies was 280 days (range, 176–308). Pregnancies with a non-term outcome had a shorter median duration of 98 days (range, 42–279) as would be expected. The median number of ontologically relevant coded entries detected and processed per pregnancy was 3 (range, 1–49). We also examined the toolkit output table data for single calendar years along with data from the RCGP RSC database relating to all women actively registered in each of those years. Table 2 shows the results for the calendar year 2015. Overall, 767 798 women accounted for 50 038 pregnancies. A total of 42 972 women had 44 411 term pregnancies (deliveries). Moreover, 41 357 women had one pregnancy with delivery outcome. There were 1431 women who had two pregnancies with delivery outcome. Four women had three pregnancies with delivery outcome. Though unlikely, it is logically possible for a woman to have had three pregnancies with delivery outcome reported within 12 months. In such cases, the tool would have picked up the delivery at the end of the earliest pregnancy, the start and end of the second pregnancy, and the beginning of the third pregnancy (ie, with most of the pregnancy plus delivery occurring during the subsequent year). We checked all full years of data from 2005 to 2015 and obtained similar results. In no case was any woman reported in any one year to have had more than three pregnancies resulting in deliveries. We also checked the number of pregnancies reported for each woman over the period 2004 to 2016. The number of women rapidly fell away with increasing pregnancy count regardless of the outcome (table 3). Mean ages at time of pregnancy over successive years from 2004 to 2016 showed a consistent upward trend in both Office of National Statistics (ONS) and RCGP RSC data (figure 5). After adjustment, there was no significant difference between ages derived from RCGP RSC and ONS (p=0.218).

DISCUSSION

To our knowledge, this is the first time that a systematic ontological approach has been used to identify pregnancies based on patient-level coded information. We believe that our system is capable of producing reliable results

Table 2 Overall results from database (2004 to 2016)

No of pregnancy cases				
Women with at least one pregnancy	243 005			
Total no of pregnancies found	405 591			
Term pregnancies (delivered)	333 689			
Non-term pregnancies	59 884			
Unclassified pregnancies	12 018			
GP first aware within 12 weeks (%)				
All pregnancies	75.9			
Term pregnancies (delivered)	79.7			
Non-term pregnancies	50.2			
Duration of pregnancies (days)	Minimum	Median	Mean	Maximum
All pregnancies	42	280	243.4	308
Term -pregnancies (delivered)	176	280	269.2	308
Non-term pregnancies	42	98	92.31	279
Summary of ages for pregnancy cases				
Age at start of any pregnancy (2004 to 2016)	10	29	29.16	69
Age at start of first pregnancy (2004 to 2016)	10	29	29.29	69
Age at start of any pregnancy—2005	11	31	30.61	69
Age at start of any pregnancy—2015	10	29	28.97	69
No of clinical codes processed per pregnancy	1	3	3.99	49
Pregnancies with two or more codes	282 197			
Pregnancies with only one code	123 394			

GP, general practitioner.

across different coding systems used within the same healthcare setting. Full details of the pregnancy ontology toolkit and its functionality can be found on our ontology webpage and online supplementary file 2.[36] Our literature review identified examples of the multiple problems encountered in attempts to identify discrete pregnancies in patient-level data. There were many references to the poor quality of the data recorded about pregnancies, often worsened by a general lack of communication between the multiple agencies involved in the delivery of patient care.[3 10 13–17 20 21 23] Apart from the Matcho study,[16] we did not find any other ontologically based solution developed to overcome these problems.

The low median and mean numbers of codes processed by the toolkit for each detected pregnancy reflect the sparsity of pregnancy-related data available in the RCGP RSC

Table 3 Table of pregnancy count, all term pregnancies (delivered) and non-term pregnancies

Number of pregnancies	All pregnancies	Delivered	Non-term	Ongoing
1	136 488	138 833	40 044	12 018
2	136 784	127 268	13 670	0
3	77 157	48 405	4176	0
4	34 040	14 396	1360	0
5	13 675	3585	415	0
6	4920	900	132	0
7	1673	238	21	0
8	576	64	48	0
9	207	0	18	0
10	60	0	0	0
11	11	0	0	0
Totals	405 591	333 689	59 884	12 018

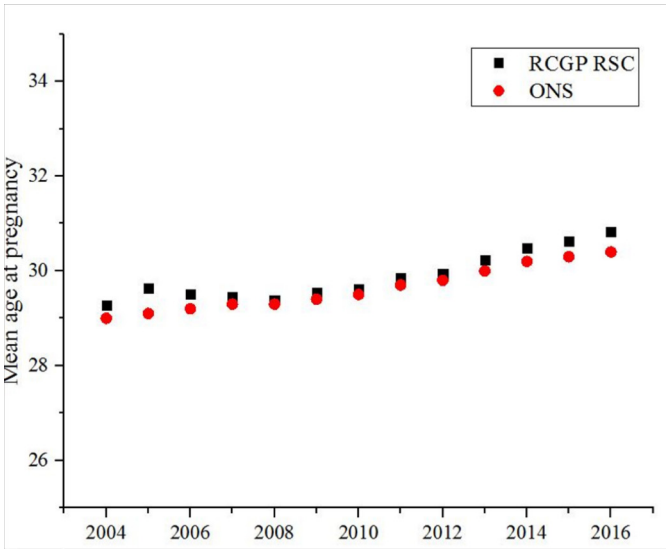


Figure 5 Comparison of mean ages at time of pregnancy between Office of National Statistics (ONS) and Royal College of General Practitioners Research and Surveillance Centre (RCGP RSC) for 2004–2016.

database. The algorithm was functioning in most cases on no more than two or three codes and in many cases on just one code per detected pregnancy (figure 6) while having the flexibility to handle much larger numbers of codes when available. In many cases, there was no event coded in the data to represent either the pregnancy start date or the pregnancy end date so that one or both of these had to be calculated. This will clearly have had an adverse effect on the precision of these dates and thus also of the duration of the pregnancy. The paucity of the data also made it necessary to fine tune the parameter table to obtain plausible durations and also to ensure that search periods were optimised to avoid either misinterpreting late-coded entries as new pregnancies or alternatively incorrectly including them into a previous pregnancy when in fact they signalled a true new pregnancy.

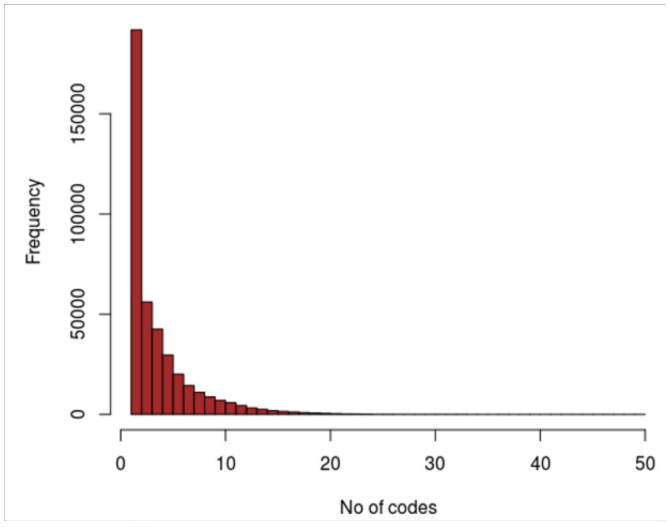


Figure 6 Number of coded entries processed per pregnancy in the database.

We started with a hierarchical (Read v2) terminology and our ontological approach enabled us to extend this into a polyhierarchical terminology (CTV3). In the coming year we will use this same approach to incorporate SNOMED CT[37 38] concepts when UK GP systems have migrated to that coding scheme. Others could use our approach to work with ICD or one of its clinical modifications.

Age at start of any pregnancy ranged from 10 to 69 (table 2), but of the overall 405 591 pregnancies, more than 99% had ages in the range 15–44. The small numbers at the extremes were checked (RCGP RSC database review) and found to be genuine. There were differences relating to coding schemes and systems in use even allowing for the much smaller number of pregnancies relating to CTV3 use. Most notably, there were fewer non-term pregnancies in the CTV3 group than in the Read v2 group (figure 2). This cannot be explained by differences in the coding schemes because CTV3 provides the same or a greater range of concepts as Read v2. However, the system supporting CTV3 is significantly different from those supporting Read v2, and the difference may be due to a more stringent application of Information Governance preventing the extraction of sensitive codes relating to termination and abortion. In contrast, for pregnancies resulting in delivery per woman, we demonstrated that the algorithm performed consistently across the two coding schemes Read v2 and CTV3 (figure 2).

To improve accuracy, the Matcho study[16] deliberately excluded any pregnancy represented by less than two coded entries. If the same exclusion had been be applied to our UK RCGP RSC data, we would have lost about 30% of all of the pregnancies detected. The CDM approach used can be expected to lose detail as clinical terms are condensed into CDM concepts. In contrast, our ontological approach enables us to leverage the richness of the clinical data and enable more inferences to be made at a more granular level. This may be particularly valuable where the data are sparse. Our approach may be preferable for those studies that need more reliable detection of any pregnancies represented in the data while the Matcho study CDM approach may be more appropriate for studies where precise duration of identified pregnancies is more important than reliability of pregnancy detection.

We demonstrated that a high proportion of the representation of each ontological concept was accounted for by a small number of subcategories suggesting that the task of mapping the pregnancy ontology to other coding schemes could be simplified by excluding the lower order subcategories (figure 1). Furthermore, this optimisation should enable the algorithm to scale well with respect to big datasets.

Our internal validation was limited to checking the working of the algorithm and reviewing the findings on the RCGP RSC database. There was no external validation such as checking original medical records or comparison

with external data. Pregnancy data are available from secondary care. In England, these are called Hospital Episode Statistics,[39] but they are only available 3 to 6 months in arrears. Such data were not available for validation of term pregnancies identified in this study due to time taken to obtain approval and cost. In any case, they are not without problems.[40 41] However, such validation may be possible in future. The close agreement between ONS and adjusted RCGP RSC mean ages at time of pregnancy was reassuring.

The RCGP RSC network has a dashboard capability which enables us to feedback to practices and improve data quality. This pregnancy ontology will go live in our dashboard[42] across the 2018/2019 season and report vaccine exposure to pregnant women alongside older people, high-risk groups and children which are fed back already. Other uses planned or in train include adherence to guidelines relating to drug use in pregnancy, associations between drug misuse during pregnancy and birth abnormalities, quality of postnatal care for women with gestational diabetes and pre-gestational type 2 diabetes.

CONCLUSIONS

We have designed a reliable search algorithm for patient-level GP data which determines the start and end dates of pregnancies for individual women either with reference to a single point in time or to a defined period of time. We encourage researchers analysing pregnancy data to adopt our approach and for those in adjacent fields to extend the ontology further. Our pregnancy toolkit provides a robust ontologically based solution to the problem of reliable identification of discrete pregnancies in routine patient-level data and should be of particular value to sentinel networks.

Acknowledgements Patients for allowing their data to be used for surveillance and research. Practices who volunteer to be part of the RCGP RSC network and allow us to extract and used health data for surveillance, quality improvement and research. Filipa Ferreira (Project Manager), Ivelina Yonova (Senior Practice Liaison Officer) and other members of the Clinical Informatics and Health Outcomes Research Group at University of Surrey. Apollo Medical Systems for data extraction. Collaboration with EMIS, TPP, InPractice and Micro-test CMR supplier for facilitating data extraction. Colleagues at Public Health England. Bernadette Carpenter (BC), Julia Hine (JH) and Simon Jones (SJ) for preliminary research work. Mark Joy for statistical advice.

Contributors HL was involved with the study design and conception, framework/toolkit design, ontology development. JW was involved with the ontology development, code mapping development, data analysis and validation. RB was involved in algorithm development, validation and optimisation of the algorithm, data extraction. JW provided intellectual content and support. SdL conceived the original idea and developed this with BC, JH and SJ (see Acknowledgements), and provided intellectual contribution and support. All authors read and approved the final manuscript.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement All data generated or analysed during this study are included in this published article. The original dataset is available from SdL, but restrictions apply to the availability of the data because the RCGP RSC database is not publicly accessible. RCGP RSC data can be accessed through an established process: details of the application process can be found on the website (www.rcgp.org.uk/rsc) or by emailing Ivelina Yonova (i.yonova@surrey.ac.uk).

REFERENCES

1 Steinhoff MC, MacDonald N, Pfeifer D, *et al*. Influenza vaccine in pregnancy: policy and research strategies. *Lancet* 2014;383:1611–3.
2 Makelarski JA, Romitti PA, Caspers KM, *et al*. Use of active surveillance methodologies to examine over-reporting of stillbirths on fetal death certificates. *Birth Defects Res A Clin Mol Teratol* 2011;91:1004–10.
3 Naleway AL, Gold R, Kurosky S, *et al*. Identifying pregnancy episodes, outcomes, and mother–infant pairs in the Vaccine Safety Datalink. *Vaccine* 2013;31:2898–903.
4 Galvao TF, Silva MT, Zimmermann IR, *et al*. Influenza vaccination in pregnant women: a systematic review. *ISRN Prev Med* 2013;2013:1–8.
5 Ompad DC, Galea S, Vlahov D. Distribution of influenza vaccine to high-risk groups. *Epidemiol Rev* 2006;28:54–70.
6 Mak TK, Mangtani P, Leese J, *et al*. Influenza vaccination in pregnancy: current evidence and selected national policies. *Lancet Infect Dis* 2008;8:44–52.
7 Sakala IG, Honda-Okubo Y, Fung J, *et al*. Influenza immunization during pregnancy: benefits for mother and infant. *Hum Vaccin Immunother* 2016;12:3065–71.
8 Pebody RG, Sinnathamby MA, Warburton F, *et al*. Uptake and impact of vaccinating primary school-age children against influenza: experiences of a live attenuated influenza vaccine programme, England, 2015/16. *Euro Surveill* 2018;23.
9 Redshaw M, Henderson J. Safely delivered: a national survey of women's experience of maternity care. National Perinatal Epidemiology Unit. *Policy Research Unit - Maternal & Health Care* 2014. Contract No.: ISBN 978-0-9931267-2-7.
10 Feijen-de Jong EI, Baarveld F, Jansen DEMC, *et al*. Do pregnant women contact their general practitioner? A register-based comparison of healthcare utilisation of pregnant and non-pregnant women in general practice. *BMC Fam Pract* 2013;14.
11 Hawley G, Janamian T, Jackson C, *et al*. In a maternity shared-care environment, what do we know about the paper hand-held and electronic health record: a systematic literature review. *BMC Pregnancy Childbirth* 2014;14.
12 Hornbrook MC, Whitlock EP, Berg CJ, *et al*. Development of an algorithm to identify pregnancy episodes in an integrated health care delivery system. *Health Serv Res* 2007;42:908–27.
13 Devine S, West S, Andrews E, *et al*. The identification of pregnancies within the general practice research database. *Pharmacoepidemiol Drug Saf* 2010;19:45–50.
14 Hardy JR, Holford TR, Hall GC, *et al*. Strategies for identifying pregnancies in the automated medical records of the general practice research database. *Pharmacoepidemiol Drug Saf* 2004;13:749–59.
15 Margulis AV, Setoguchi S, Mittleman MA, *et al*. Algorithms to estimate the beginning of pregnancy in administrative databases. *Pharmacoepidemiol Drug Saf* 2013;22:16–24.
16 Matcho A, Ryan P, Fife D, *et al*. Inferring pregnancy episodes and outcomes within a network of observational databases. *Plos One* 2018;13:e0192033.
17 Li Q, Andrade SE, Cooper WO, *et al*. Validation of an algorithm to estimate gestational age in electronic health plan databases. *Pharmacoepidemiol Drug Saf* 2013;22:524–32.
18 Cea-Soriano L, García Rodríguez LA, Fernández Cantero O, *et al*. Challenges of using primary care electronic medical records in the UK to study medications in pregnancy. *Pharmacoepidemiol Drug Saf* 2013;22:977–85.
19 Mikolajczyk RT, Kraut AA, Garbe E. Evaluation of pregnancy outcome records in the German Pharmacoepidemiological Research Database (GePaRD). *Pharmacoepidemiol Drug Saf* 2013;22:873–80.
20 Margulis AV, Palmsten K, Andrade SE, *et al*. Beginning and duration of pregnancy in automated health care databases: review of

estimation methods and validation results. *Pharmacoepidemiol Drug Saf* 2015;24:335–42.
21 Strom BL, Schinnar R, Jones J, *et al*. Detecting pregnancy use of non-hormonal category X medications in electronic medical records. *J Am Med Inform Assoc* 2011;18(Supplement 1):i81–6.
22 Say L, Souza JP, Pattinson RC, *et al*. Maternal near miss—towards a standard tool for monitoring quality of maternal health care. *Best Pract Res Clin Obstet Gynaecol* 2009;23:287–96.
23 Deneux-Tharaux C, Berg C, Bouvier-Colle M-H, *et al*. Underreporting of pregnancy-related mortality in the United States and Europe. [Erratum appears in Obstet Gynecol. 2006 Jan;107(1):209]. *Obstet Gynecol* 2005;106:684–92.
24 Mitchell C, Lawton E, Morton C, *et al*. California pregnancy-associated mortality review: mixed methods approach for improved case identification, cause of death analyses and translation of findings. *Matern Child Health J* 2014;18:518–26.
25 Liaw S-T, Taggart J, Yu H, *et al*. Integrating electronic health record information to support integrated care: practical application of ontologies to improve the accuracy of diabetes disease registers. *J Biomed Inform* 2014;52:364–72.
26 Liyanage H, Krause P, De Lusignan S. Using ontologies to improve semantic interoperability in health data. *Jhi* 2015;22:309–15.
27 Correa A, Hinton W, McGovern A, *et al*. Royal College of General Practitioners Research and Surveillance Centre (RCGP RSC) Sentinel Network: a cohort profile. *BMJ Open* 2016;6:e011092.
28 de Lusignan S, Correa A, Smith GE, *et al*. RCGP Research and Surveillance Centre: 50 years' surveillance of influenza, infections, and respiratory conditions. *Br J Gen Pract* 2017;67:440–1.
29 de Lusignan S, Codes deLS. Codes, classifications, terminologies and nomenclatures: definition, development and application in practice. *Inform Prim Care* 2005;13:65–9.
30 Liyanage H, Liaw S-T, Kuziemsky C, *et al*. Ontologies to improve chronic disease management research and quality improvement studies - a conceptual framework. *Stud Health Technol Inform* 2013;192:180–4.
31 Liyanage H, Liaw S-T, Kuziemsky C, *et al*. The evidence-base for using ontologies and semantic integration methodologies to support integrated chronic disease management in primary and ambulatory care: realist review. *Yearb Med Inform* 2013;22:147–54.
32 Liyanage H. Pregnancy ontology, 2018. Available: https://bioportal.bioontology.org/ontologies/PREGONTO
33 Rollason W, Khunti K, De Lusignan S. Variation in the recording of diabetes diagnostic data in primary care computer systems: implications for the quality of care. *Jhi* 2009;17:113–9.
34 Liyanage H, de Lusignan S. Ontologies to capture adverse events following immunisation (AEFI) from real world health data. *Stud Health Technol Inform* 2014;197:15–19.
35 Department of Health (DH) RCoGPR, British Medical Association (BMA). The Good Practice Guidelines for GP electronic patient records Version 4 2011 [For Chapter 7 Clinical coding schemes]. Available: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/215680/dh_125350.pdf
36 Liyanage H. Pregnancy ontology toolkit, 2018. Available: https://clininf.eu/index.php/pregnancy-ontology/
37 De Lusignan S, Chan T, Jones S. Large complex terminologies: more coding choice, but harder to find data—reflections on introduction of SNOMED CT (Systematized Nomenclature of Medicine—Clinical Terms) as an NHS standard. *Jhi* 2011;19:3–5.
38 Digital N. SNOMED CT. Available: https://digital.nhs.uk/services/terminology-and-classifications/snomed-ct
39 Digital N. Hospital Episode Statistics (hES). Available: https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/hospital-episode-statistics
40 Dattani N, Macfarlane A. Linkage of maternity Hospital Episode Statistics data to birth registration and notification records for births in England 2005–2014: methods. A population-based birth cohort study. *BMJ Open* 2018;8:e017897.
41 Harper G. Linkage of maternity Hospital Episode Statistics data to birth registration and notification records for births in England 2005–2014: quality assurance of linkage of routine data for singleton and multiple births. *BMJ Open* 2018;8:e017898.
42 Pathirannehelage S, Kumarapeli P, Byford R, *et al*. Uptake of a Dashboard Designed to Give Realtime Feedback to a Sentinel Network About Key Data Required for Influenza Vaccine Effectiveness Studies. *Stud Health Technol Inform* 2018;247:161–5.