Title: Identifying signals of potentially harmful medications in pregnancy: use of the double false discovery rate method to adjust for multiple testing

Authors: Alana Cavadino[1-2], David Prieto-Merino[3-4], Joan K. Morris[1]

[1] Wolfson Institute of Preventive Medicine, Queen Mary University of London, UK

[2] Section of Epidemiology and Biostatistics, School of Population Health, University of Auckland, New Zealand

[3] Department of Medical Statistics, London School of Hygiene and Tropical Medicine, London, United Kingdom

[4] Applied Statistics in Medical Research Group, Catholic University of Murcia (UCAM), Spain

ORCID: Alana Cavadino, 0000-0002-5709-367X

Corresponding author: Professor Joan K Morris. Centre for Environmental and Preventive Medicine, Wolfson Institute of Preventive Medicine, Barts and the London School of Medicine and Dentistry, Queen Mary University of London, Charterhouse Square, London EC1M 6BQ.
Telephone: +442078826274, Fax: +442078826270, Email: j.k.morris@qmul.ac.uk

Author Contributions: AC was responsible for study design, statistical analysis, interpreting the findings, and writing the manuscript. JM was responsible for study design, supervising the study, advising on interpretation of the results and contributing to all revisions of the manuscript. DP advised on statistical methods and the interpretation of the results, and contributed to revisions of the manuscript.

Running title: Identifying harmful medications in pregnancy.
Number of tables: 1
Number of figures: 3
Number of references: 33
Abstract word count: 250
Total word count: 4719

Keywords: Congenital anomalies, medication safety, methodology, pharmacoepidemiology, pharmacovigilance, pregnancy

Abstract

Aims

Surveillance of medication use in pregnancy is essential in order to identify associations between first trimester medications and congenital anomalies (CAs). Medications in the same Anatomical Chemical Therapeutic classes may have similar effects. We aimed to use this information to improve the detection of potential teratogens in CA surveillance data.

Methods

Data on 15,058 malformed foetuses with first trimester medication exposures from 1995-2011 were available from EUROmediCAT, a network of European CA registries. For each medication-CA combination, the proportion of the CA in foetuses with the medication was compared to the proportion of the CA in all other foetuses in the dataset. The Australian classification system was used to identify "high risk" medications in order to compare two methods of controlling the false discovery rate (FDR): a "single" FDR applied across all combinations, and a "double" FDR incorporating groupings of medications.

Results

There were 28,765 potential combinations (523 medications x 55 CAs) for analysis. An FDR cut-off of 50% resulted in a reasonable effective workload, for which single FDR gave rise to 8 medication signals (3 "high risk" medications) and double FDR 50% identified 16 signals (6 "high risk"). Over a range of FDR cut-offs, double FDR identified more "high risk" medications as signals, for comparable effective workloads.

Conclusions

The double FDR method appears to improve the detection of potential teratogens in comparison to the single FDR, while maintaining a low risk of false positives. Use of double FDR is recommended in routine signal detection analyses of CA data.

What is already known on this subject

- Continued post-marketing surveillance of medication safety in pregnancy is important due to a lack of safety information for medications taken during pregnancy.
- EUROmediCAT performs systematic signal detection using data from population-based congenital anomaly registries.
- The current EUROmediCAT method examines each medication separately, without taking similarities amongst medications into account.

What this study adds

- Information about which Anatomical Chemical Therapeutic class a medication belongs to can be used to improve the detection of potential teratogens in the analysis of first trimester exposure in foetuses with congenital anomalies.
- Identified associations between medications and congenital anomalies require validation from independent data sources.

Introduction

The first trimester of pregnancy is an essential stage of development for the foetus, when organs undergo critical steps in their development and most non-inherited congenital anomalies (CAs) occur. Certain medications can cause CAs when taken during early pregnancy, yet exposure to prescription and over-the-counter medication during pregnancy is common [1, 2]. Despite this widespread use, information on the safety of medicines in human pregnancy is lacking, particularly for new products; a key reason is that pregnant women are usually exempt from pre-marketing medication safety studies. When a medication is licensed for marketing, little or nothing is generally known about the safety for its use during pregnancy. Whilst information about reproductive toxicity is sometimes available from animal studies, these are limited in their ability to predict risks of CA in humans [3]. As such, post-marketing studies and the continued surveillance of medication use in pregnancy are essential in order to detect new teratogens.

EUROmediCAT is a network of European CA registries, which has developed a systematic signal detection method to identify medications potentially associated with an increased risk of specific CAs [4, 5]. This method searches thousands of medication-CA combinations to identify a small set of medication-CA combinations that are more likely than the other combinations to contain some true associations between the medications and the CAs. These combinations are called signals, and once identified they need to be investigated in detail [6]. In order to only identify a small set of signals a false discovery rate (FDR) procedure is used. The FDR is the proportion of false positive results among the set of all positive results, e.g. an FDR of 10% means that up to 10% of identified associations are expected to be false positives. The Anatomical Therapeutic Chemical (ATC) system is used to code medications [6], with those in the same ATC classes often working in similar ways. However, this information is not incorporated in the current EUROmediCAT signal detection method, which examines each medication and each CA separately [5]. In this study, we aimed to refine signal detection methods for CA data by using a “double” FDR procedure, which takes groupings of ATC medications or groupings of CA into account. If similar medications are likely to have similar teratogenic effects or if similar anomalies are likely to be caused by the same medications, then the double FDR will use this information, and may be a more efficient method to identify signals compared to the single FDR method. The double FDR was first proposed by Mehrotra and Heyse in 2004 [7]; here we use an improved version of the double FDR as presented by Mehrotra and Adewale in 2012 [8].

Methods

Study population

Built on the European Surveillance of Congenital Anomalies (EUROCAT) network of population-based CA registries, EUROmediCAT includes registries with information on first trimester medication use. Consistency of inclusion criteria, data collection and recording across EUROCAT registries is monitored using data quality indicators [9]. EUROmediCAT data on malformed foetuses with first trimester medication exposures from 1995-2011 were available, for a population coverage of around 7 million births from 11 European countries, across 13 registries that agreed to participate in this study (Table 1). Ethical and data access approvals were obtained for each database from the relevant governance infrastructures. A large proportion of this EUROmediCAT dataset has been previously analysed for signal detection purposes [5].

Congenital anomaly data

EUROCAT registries use multiple sources of information to ascertain all CA cases, including live birth, foetal death and termination of pregnancy for foetal anomaly. Data is obtained through active case finding and voluntary reporting [10]. All EUROCAT registries use a standardised methodology, i.e. standard variables, definitions and coding instructions [11]. CAs are coded according to EUROCAT coding guide version 1.4, which uses the International Classification of Diseases (ICD) version 10 with the British Paediatric Association (BPA) extension. ICD10-BPA codes are used to group cases of CA into standard subgroups, described in detail in the EUROCAT coding guide 1.4 (http://www.eurocat-network.eu/content/Section%203.3-%2027_Oct2016.pdf) [11]. Cases with genetic conditions (chromosomal anomalies, skeletal dysplasia, genetic syndromes or microdeletions) were excluded from this study. Foetuses with isolated congenital dislocation of the hip as their only major CA were also excluded, since the aetiology of this CA is mechanical. Foetuses exposed only to vitamins, minerals and/or folic acid were excluded. Foetuses with a congenital heart defect (CHD) but no information regarding the specific type of CHD were combined into a separate subgroup (unspecified CHD, n=542) for analysis purposes. We performed signal detection for 55 EUROCAT CA subgroups that were analysed in the previous EUROmediCAT signal detection analysis, but here including two fewer CAs due to a more recent version of EUROCAT coding being used for the extraction of data.

Exposure data

Maternal medication exposure data in EUROmediCAT are obtained from prospectively recorded maternity records (medical files); additional data sources include general practitioner records, maternal interviews and infant medical records [12, 13]. Data collection and processing methods vary, with completeness of information and the type of medications recorded (e.g. over-the-counter, prescribed use and/or actual use) differing between registries [12]. Inclusion in the EUROmediCAT database requires first trimester exposures, defined as the first day of a woman's last menstrual period to the end of her twelfth gestational week [11]. The EUROCAT coding guide instructs registries to calculate length of gestation by using "best estimate based on last menstrual period and/or ultrasound determination". An unlimited number of exposures are recorded using the hierarchical ATC system, with up to 7 digits and free text information where required. Medications were analysed using 7-digit ATC5 codes; 5-digit ATC4 level codes were used only where a more detailed ATC5 code was unavailable. We were not able to differentiate between exposures to the same medication at different dosages since EUROmediCAT registries do not routinely include this information. ATC codes subject to alterations over time were identified from the WHO website [14], with older codes updated where available. Cases exposed exclusively to codes with less than 5 digits (not coded to at least ATC4), topical medications (including eye drops and other non-oral/non-IV medications) and medications taken in the 2nd/3rd trimester or with unknown timing were also excluded. Medications recorded in less than 3 foetuses were not investigated, but foetuses exposed to these medications were included as controls in analyses.

Statistical analysis

For each medication-CA combination, the proportion of CA cases in foetuses with the medication exposure was compared to the proportion of that CA in all other foetuses in the dataset (i.e. those without that specific medication, but exposed to at least one other medication). An unadjusted, one-sided Fisher's exact test was used. For each medication-CA combination, registries with no exposures and registries with no cases were excluded from that particular analysis. Outcomes were reported using proportional reporting ratios (PRRs) and P-values. All analyses were performed using Stata version 12 [15].

A “single” Simes adjustment [16] across all medication-CA combinations was used to control the FDR, as per existing EUROmediCAT methodology [5]. We compared this to a “double” FDR procedure [8], incorporating an additional step to consider groupings of medications.

The **single FDR** procedure is specified for m tests and an FDR cut-off α as follows

- Order P-values P_i according to their magnitude $P_1 < \cdots < P_m$
- Let $\tilde{P}_i$ denote the FDR adjusted value for P_i, where

  $\tilde{P}_m = P_m$

  $\tilde{P}_i = min\left(\frac{m}{i}P_i\,, \tilde{P}_{i+1}\right) \quad for\ i\ \leq m-1$
- Null hypotheses with $\tilde{P}_i \leq\ \alpha$ are rejected; these combinations are the potential signals

The **double FDR** is specified for $i = 1, \ldots, n$ medication groups each with g_i $(j = 1, \ldots, g_i)$ medications as follows

Step 1: Perform a single FDR adjustment (as above) within each group i to get $\tilde{P}_{ij}$, then let P_i^* denote the smallest FDR-adjusted P-value in each group

$$P_i^* = \min(\tilde{P}_{ij}; 1 \leq j \leq g_i)$$

Apply an FDR adjustment to the P_i^* to get the set of representative FDR-adjusted P-values $\tilde{P}_i^*$, where $P_1^* < \cdots < P_n^*$ and

$$\tilde{P}_n^* = P_n^*$$

$$\tilde{P}_i^* = min\left(\frac{n}{i}P_i^*\,, \tilde{P}_{i+1}^*\right) \quad for\ i\ \leq n-1$$

All groups i where $\tilde{P}_i^*\ \leq \alpha$ are taken to the second step.

Step 2: Let $F \equiv \{P_{ij} \mid \tilde{P}_i^*\ \leq \alpha\}$ be the family of P-values from groups flagged by $\tilde{P}_i^*$ in step 1. Then apply a single FDR procedure across all P-values in F such that $\tilde{P}_{ij}^{(F)}$ is the FDR-adjusted P-value for all $P_{ij} \in F$. Null hypotheses are rejected for all tests where $\tilde{P}_{ij}^{(F)} \leq \alpha$; these combinations are the potential signals.

The FDR cut-off is defined for a pre-specified proportion α between 0 and 1; e.g. $\alpha = 0.1$ corresponds to an FDR of 10%, meaning that up to 10% of combinations identified as signals are expected to be false positives. The overall FDR level for double FDR is at most α at the group level no matter how many groups contain at least one true signal, hence it may be possible that the overall FDR exceeds α in some scenarios [8]. A lack of low powered associations violates the underlying assumptions of a multiple testing procedure by markedly shifting the distribution of P-values towards zero; associations for combinations with less than 3 exposed cases were therefore retained in the multiple testing procedures, but were not flagged as signals for further consideration.

Grouping of medication-anomaly combinations

Medications were grouped according to pharmacological subgroups using 4-digit ATC3 codes (n=116 groups). ATC4 (chemical subgroup) codes were considered too detailed for grouping purposes, resulting in a large number of groups (n=244). ATC2 (therapeutic main group) levels would provide a smaller number of groups (n=61), but these would include a range of medications and were considered too heterogeneous for grouping purposes. Both ATC2 and ATC4 were considered in sensitivity analysis. Certain CAs are thought to be more sensitive (compared to other types of CAs) to medications in general, and it may be likely that a number of CAs have no signals at all. We therefore also considered grouping medication-CA combinations by CA.

Evaluation and comparison of signal detection methods

The Australian government department of health provides a database of recorded pregnancy related risks associated with medicines [17]. All medications are divided into five main lettered categories: category A medications are considered safe for use during pregnancy; category B have not shown evidence of harmful effects to human foetuses; category C may carry harmful effects to human foetuses, without evidence of causing CAs; categories D and X have evidence of moderate to high teratogenic risk. Specific CAs are not identified. This database was used to independently identify medications for which there has been evidence of high teratogenic risk by grouping medications as “high risk” (categories D and X) or “low risk” (categories A, B or C). Medications that could not be mapped to a risk category (i.e. that were not present) in this database were grouped as “unclassified risk”. Medication names were matched to EUROmediCAT data using substance names. The total number of “high risk” medications identified by single and double FDR were compared, and the proportion of all the “high risk” medications in the data that were identified as signals was calculated. This was defined as

$$\textbf{Identification rate} = \frac{\text{Number of "high risk" medications identified as signals}}{\text{Total number of "high risk" medications in the data}}$$

The total number of medications in the set of signals (the “effective workload”) identified by each method was also considered, since each medication signal requires further investigation regardless of the number of different CAs it may be associated with.

Results

Description of signal detection dataset

A total of 31,197 foetuses with at least one first trimester medication exposure and a major CA born from 1995-2011 were extracted from the EUROmediCAT central database for 13 registries. Of these, 905 foetuses with isolated congenital dislocation of the hip, 1,219 with no ATC4 or ATC5 exposures recorded and 452 with only topical medication exposures were excluded, leaving 28,621 foetuses with valid medication exposures (Table 1). Foetuses with exposures only occurring outside the first trimester of pregnancy (n=1,490) or with unknown timing (n=12,073) were further excluded, leaving 15,058 foetuses for analysis. Some registries had considerable data loss where it was not possible to verify when the reported medications had been taken (Table 1), and this has been discussed previously [5]. However, the distribution of types of CA were similar for those pregnancies included and excluded due to unknown timing, suggesting that cases remaining in the dataset for these registries aren't prone to selection biases in this respect.

On average, there were 1.6 recorded ATC non-topical medication exposures per pregnancy, ranging from one (in 65% of cases) up to 16 (one case). The total number of exposures to medications appearing at least 3 times in the dataset was 22,624; this included 523 ATC medications, of which 39 (7.5%) were coded only to ATC4. With 55 CAs for analysis, this gave 28,765 potential medication-CA combinations. Using ATC3 grouping resulted in 116 distinct groups with an average of 9 (range 1-20) unique medications and 487 (range 53-1,086) medication-CA combinations per group.

Of the 523 medications in the data, 297 (57%) were "low risk", 44 (8.4%) were "high risk" and 182 (35%) had "unclassified risk" according to the Australian risk categorisation database. Three medications mapped to a code in both the "low risk" and "high risk" group depending on their dosage. As there is no information on dosage in EUROmediCAT data, these medications weren't assigned to either risk category but were instead coded as "unclassified risk". Of the 116 ATC3 groups, 94 (81%) contained no "high risk" medications, 13 (11%) contained only one, and 9 (8%) groups contained two or more "high risk" medications.

Comparison of single and double FDR procedures to adjust for multiple testing

There were no cases for 369 specific medication-CA combinations, hence Fisher's exact test was performed for 28,396 combinations. Figure 1 is a smileplot [18], plotting the PRR against the P-value for each combination. Without adjustment for multiple testing, 670 combinations were significant at the conventional 5% level (points above the lower dashed line in Figure 1). For a cut-off of 50%, single FDR resulted in 10 medication-CA combinations being identified as signals (diamond symbols in Figure 1), involving 8 medications (1 medication had signals for 3 CAs). Double FDR 50% resulted in 25 medication-CA combinations being identified as signals (+ symbols in Figure 1), involving 16 medications (4 medications had signals for 2 CAs, 1 medication for 3 CAs and 1 medication for 4 CAs) in 5 ATC3 groups.

Double FDR 50% signals included 5 antiepileptic medications (4 “high risk”, 1 “low risk”), 4 insulin medications (2 “low risk”, 2 “unclassified risk”), 3 sex hormones (1 “high risk”, 2 “unclassified risk”), 2 antiasthmatic medications (both “low risk”) one “unclassified risk” gynaecological medication and one “high risk” medication for acid related disorders. Single FDR 50% identified only 7 of these medication signals over 10 combinations, including 2 of the antiepileptics (both “high risk”), 2 antiasthmatics (both “low risk”), one “unclassified risk” insulin-related medication and one “unclassified risk” sex hormone. Single FDR 50% also identified one further signal for a “low risk” anxiolytic medication, which was not a signal using double FDR 50% and was in an ATC3 group including 8 other medications that were not signals using either method.

Figure 2 presents the number of signals in each risk category using ATC3 groupings for single and double FDR across a range of cut-offs from 5-50%. Higher cut-offs resulted in a greater number of signals being identified for both methods. A maximum of 3 “high risk” signals were identified by single FDR for any cut-off from 15-50%. Only one additional “low risk” medication was identified as a signal for an FDR cut-off of 35% and above. Six “high risk” medications were identified by double FDR 30%, above which level only one further (“low risk”) medication signal was identified. The proportion of signals that were “high risk” was highest for double FDR at a cut-off of 10% or less; above this threshold the single and double FDR had similar proportions of “high risk” signals.

Figure 3 shows the identification rate for single and double FDR with cut-offs from 5-50%. Double FDR identified more of the “high risk” medications than single FDR, for a range of FDR cut-offs and comparable effective workloads (the number of medications identified as signals).

Sensitivity analyses

FDR procedures were repeated using ATC2 and ATC4 codes and CAs to group medication-CA combinations. Grouping by ATC2 provided 61 groups with an average of 20 (range 1-54) medications per group. ATC4 gave 244 groups, with an average of 3 (range 1-8) medications per group. Both these groupings resulted in similar or lower effective workloads across the different levels of FDR cut-off; however, the number and proportion of “high risk” medications detected was generally lower compared to ATC3 groupings. For FDR 25%, for example, ATC3 groupings identified a 11 medication signals of which 5 were “high risk”, compared to 12 (5 “high risk”) and 6 (2 “high risk”) for ATC2 and ATC4 groupings, respectively. Grouping of medication-CA combinations by the 55 CAs an average of 518 (range 325-523) unique medication-CA combinations per group; this resulted in a greater number of signals across all levels of FDR cut-off, with no increase in the number of “high risk” medications identified. This might be influenced by the use of a risk classification system to judge the methods, since these are not specific to the type of CA. Overall, however, grouping by ATC2, ATC4 or by CA did not show an improvement over ATC3 grouping.

Of 523 medications in the analysis, 35% had “unclassified risk” (not present in the Australian classification system database). Sensitivity analyses were conducted to assess the effect of a number of hypothetical situations. Firstly, medications without a risk category were assumed to either be all “high risk” or all “low risk”, representing the two most extreme situations. We also considered the effect of assuming that a medication with “unclassified risk” was “high risk” if there was at least one other “high risk” medication in that group, and “low risk” if all medications in the same group had “low” or “unclassified risk”. For these three situations, comparisons between single and double FDR were the same as described previously, and conclusions remained unchanged (data not shown).

Discussion

We compared two FDR procedures for use in EUROmediCAT’s signal detection process to assess whether the FDR multiple testing adjustment could be improved by incorporating medication groupings. The double FDR identified more “high risk” medications as signals than the currently used single FDR method, for a range of FDR cut-offs and comparable resulting workloads. Using double FDR a cut-off level of 50% was judged appropriate, as it resulted in a reasonable effective workload (16 medication signals over 25 combinations) requiring follow up. The only medication that was a signal using single but not double FDR was in the “low risk” category. All but one of the 16 additional signals from double FDR were in ATC3 groups including at least one other signal, highlighting how the double FDR picks up more signals in groups where there are other (stronger) signals.

Results were also compared to those from the previous EUROmediCAT signal detection analysis (using single FDR with a 50% cut-off [5]), for which the dataset differed by omission of two registries and inclusion of an additional year of data for another two registries. Of the 26 combinations identified as signals in the present analysis, 20 were signals in the previous EUROmediCAT analysis. Combinations that were signals here but not in the previous analysis (where their FDR-adjusted P-values did not reach statistical significance using single FDR) included antiepileptic medications, which are well-established as being a teratogenic group of medications [19-21], and progesterone sex hormones, which have been previously linked to an increased risk of hypospadias [22] and CHDs [23]. All signals identified in this study have been reported previously in the literature; however, as the EUROmediCAT database increases in size annually, future analyses are expected to identify new associations, particularly for newer medications. The identification of signals using double FDR is more powerful than single FDR methodology, and will provide additional independent information to confirm the results of hypothesis driven studies for specific types of medications, as well as evidence to perform new hypothesis driven studies.

Statistical considerations

The choice of FDR cut-off level should balance the proportion of false negative and false positive associations. If this is set too high, the resulting workload for follow up of associations may be impractical, and there is potential for unwarranted anxiety to pregnant women if false positive associations are reported. A low FDR cut-off, conversely, could miss important signals and result in delay of detecting such teratogens until more data (i.e. more exposures to harmful medications) are available. The FDR 50% used here might be considered a higher than acceptable rate of likely false positives, but when considering this in contrast to a lack of any adjustment for multiple testing (where the FDR approaches 100%), a cut-off of 50% is an improvement. The choice of FDR cut-off should be re-evaluated frequently, and in practice should depend on resources available for follow-up of signals. One-sided tests are used for CA signal detection as only foetuses with a CA are included; if a medication reduces the risk of a CA there will, by definition, be fewer cases in the study population and therefore a low power to detect preventive medications. Two combinations indicated preventive associations using both FDR procedures; these associations are likely due to chance or biases occurring from the study design (as controls all have at least one CA and one medication exposure). Furthermore, the comparison is to other CAs and medications in the database, hence preventive associations would not imply that the overall risk of a CA is likely to be lower. Adjustment for registry was not done as numbers were already small for the majority of combinations.

Evaluation of signal detection methods

Evaluation of signal detection methods in CA data is difficult due to the lack of a "gold standard" to classify risks for medications in pregnancy according to CAs. A definitive measure of how many teratogens are missed by a signal detection method, and possible reasons for such lack of detection, cannot therefore be obtained. To evaluate EUROmediCAT's signal detection process we used the Australian classification system to assign risk categories to medications. Two important limitations of this risk classification are pertinent here; firstly, there was no available classification of pregnancy-related risk for around a third of medications in the EUROmediCAT database. It might have been possible to garner further information on some of these unclassified medications by also using the Swedish and US risk classification systems, as done in a previous web-based study of medication safety of in Europe [24]. However, after including information from all three classification systems in their study, 23.2% of medications still remained unclassified [24], indicating that only a relatively small improvement would be made by including additional information from these other databases. We therefore considered the use of one database sufficient for our main purpose of directly comparing the FDR methods. Secondly, an important limitation of these risk classification systems in general is that specific CAs are not identified for "high risk" medications. However, the teratogenic risk of a medication is almost always specific to one (or more) particular CA(s), rather than an increased risk of malformations in general [25]. The identification rate used in this study does not differentiate between a medication associated with only one CA and one associated with a number of CAs.

Two of 5 “low risk” signals identified using double FDR with a 50% cut-off were insulin medications. It is not surprising to find the insulin medications in the low risk category according to the Australian classification system because although the association between insulin medications and CHDs is well documented, this is thought to be a result of the fact that the mother has diabetes rather than the insulin medications themselves [26-28]. This example of potential confounding by indication highlights that signal detection cannot distinguish between cause and effect, and emphasises the importance of further evaluating signals from a medical and biological perspective. The number of combinations identified as “high risk” may be overestimated, since categorisations are specific only to the medications and not the type of CA. Conversely, medications classified as B or C in the Australian database may be given their classification due to a lack of power with information available at the time of categorisation, and as such we cannot definitively say that these medications are known to be “low risk”. Due to these issues, the Australian categorisation system is not as precise as would be ideal for use as a “gold standard” in classifying teratogenic risk. However, the aim of this study was to directly compare the single and double FDR, and each procedure should have the same lack of data and power for known teratogens. We cannot provide an absolute measure of how good either method is for two main reasons. Firstly, there is a lack of a “gold standard”. Secondly, the number of women exposed to well-established teratogens (such as thalidomide or isotretinoin) during pregnancy is now very low due to their known risks, hence the power to detect signals for such medications in more recent data is low. It is therefore important to adopt more efficient methods of signal detection than are currently being used, rather than delay adoption due to a lack of evidence based on relatively low overall detection rates across the methods.

Strengths and weaknesses

A major strength of EUROmediCAT data is the detailed and standardised coding of CAs across registries [13], combined with detailed information regarding first trimester medication exposures. Good agreement between the medication actually used and that recorded in one EUROmediCAT registry has also been previously demonstrated [29]. Whilst there are variations across the registries in the way data regarding maternal medication use is collected, and in the quality and completeness of these data, there is no reason to believe that either the single FDR or the double FDR method will be more or less robust to such variations between registries. A potential weakness of EUROmediCAT data is known under ascertainment of some medications [12, 30], which may reduce the sensitivity of signal detection analyses. Another important limitation is the lack of information regarding the dosage of medication exposures, which is particularly relevant as medications in different dosages can have different teratogenic effects. Timing of exposures could not be confirmed in a number of cases (a high proportion for the Polish registries), resulting in a loss of power and a source of potential bias. After data cleaning, all included cases were supposedly confirmed first trimester exposures, however it is not possible to know that the mothers took the medication during the critical period for development of each specific CA, particularly since these can differ according to the type of CA [31]. In these analyses, the large number of statistical tests involving very small numbers of exposures prohibits accurate adjustment for potential confounders, therefore multiple comparisons procedures are performed on unadjusted results. Nonetheless, heterogeneity between registries is considered in the next stage of the signal detection process; the first EUROmediCAT signal detection analysis was published alongside an accompanying paper that evaluated identified signals in detail, considering, amongst other factors: heterogeneity between registries, registry-specific effects, and potential confounding by indication or by co-exposure [32]. An advantage of case-malformed control studies is that some biases associated with the use of a healthy control population (e.g. recall and recording biases) are likely to be minimised. On the other hand, this type of study cannot produce medication exposure risk estimates for CAs compared to healthy (i.e. non-exposed and non-malformed) controls [33]; evidence of an increased risk of any CA in this study is relative only to the risks associated with other CAs.

Conclusion

The double FDR identified a greater number and proportion of “high risk” medications as signals than the currently used single FDR, for a range of FDR cut-offs and comparable resulting workloads. We therefore recommend that double FDR be used in routine signal detection analyses of CA data. This may help identify potentially teratogenic medications that could be missed by the existing single FDR procedure, whilst retaining an acceptable level of false positive associations. Signal detection analyses of CA data must continue to be accompanied by detailed follow up of any new potential signals identified and timely dissemination of results; together, this will help patients make appropriate decisions to balance any risks and benefits of medication use during their pregnancy, based on the most up-to-date information available.

Acknowledgements

We sincerely thank all of the people throughout Europe who have been involved in providing and processing information for both EUROCAT and EUROmediCAT, including affected families, clinicians, health professionals, medical record clerks and registry staff. EUROCAT registries are funded as fully described in Paper 6 of Report 9 “EUROCAT Member Registries: Organization and Activities” available at http://onlinelibrary.wiley.com/doi/10.1002/bdra.20775/pdf. This work was supported by the Medical Research Council [grant number 1504916].

Conflicts of Interest

There are no conflicts of interest to declare.

References

1. Lupattelli A, Spigset O, Twigg MJ, Zagorodnikova K, Mardby AC, Moretti ME, et al., *Medication use in pregnancy: a cross-sectional, multinational web-based study.* BMJ Open, 2014. **4**(2): p. e004365.
2. Daw JR, Hanley GE, Greyson DL, and Morgan SG, *Prescription drug use during pregnancy in developed countries: a systematic review.* Pharmacoepidemiol Drug Saf, 2011. **20**(9): p. 895-902.
3. Wilson JG, *The evolution of teratological testing.* Teratology, 1979. **20**(2): p. 205-11.
4. de Jong-van den Berg L, Bakker M, and Dolk H, *EUROmediCAT: European surveillance of safety of medication use in pregnancy.* Pharmacoepidemiol Drug Saf, 2011. **20 (S1)**: p. 46–7.
5. Luteijn JM, Morris JK, Garne E, Given J, de Jong-van den Berg L, Addor MC, et al., *EUROmediCAT signal detection: a systematic method for identifying potential teratogenic medication.* Br J Clin Pharmacol, 2016.
6. WHO Collaborating Centre for Drug Statistics Methodology. *ATC Structure and principles*. 2011 2011-03-25 [cited 2016 21 April]; Available from: http://www.whocc.no/atc/structure_and_principles/.
7. Mehrotra DV and Heyse JF, *Use of the false discovery rate for evaluating clinical safety data.* Stat Methods Med Res, 2004. **13**(3): p. 227-38.
8. Mehrotra DV and Adewale AJ, *Flagging clinical adverse experiences: reducing false discoveries without materially compromising power for detecting true signals.* Stat Med, 2012. **31**(18): p. 1918-30.
9. Loane M, Dolk H, Garne E, and Greenlees R, *Paper 3: EUROCAT data quality indicators for population-based registries of congenital anomalies.* Birth Defects Res A Clin Mol Teratol, 2011. **91 Suppl 1**: p. S23-30.
10. Greenlees R, Neville A, Addor MC, Amar E, Arriola L, Bakker M, et al., *Paper 6: EUROCAT member registries: organization and activities.* Birth Defects Res A Clin Mol Teratol, 2011. **91 Suppl 1**: p. S51-s100.
11. EUROCAT Central Registry, *EUROCAT Guide 1.4: Instructions for the registration and surveillance of congenital anomalies*. 2013, EUROCAT Central Registry, University of Ulster.
12. Bakker M and Jonge Ld. *EUROCAT Special Report: Sources of Information on Medication Use in Pregnancy*. 2014; Available from: http://www.eurocat-network.eu/content/Special-Report-Medication-Use-In-Pregnancy.pdf.
13. Boyd PA, Haeusler M, Barisic I, Loane M, Garne E, and Dolk H, *Paper 1: The EUROCAT network--organization and processes.* Birth Defects Res A Clin Mol Teratol, 2011. **91 Suppl 1**: p. S2-15.
14. WHO Collaborating Centre for Drug Statistics Methodology. *ATC alterations from 1982-2016*. 2015 18 December 2015 [cited 2016 11 August]; Available from: http://www.whocc.no/atc_ddd_alterations__cumulative/atc_alterations/.
15. StataCorp, *Stata Statistical Software: Release 12.* 2011, College Station, TX: StataCorp LP.
16. Benjamini Y and Hochberg Y, *Controlling the False Discovery Rate - a practical and powerful approach to multiple testing.* Journal of the Royal Statistical Society Series B-Methodological, 1995. **57**(1): p. 289-300.
17. Australian Government Department of Health. *Prescribing medicines in pregnancy database.* 2016 [cited 2016 19th August]; Available from: https://www.tga.gov.au/prescribing-medicines-pregnancy-database.
18. Newson R, *Multiple-test procedures and smile plots.* Stata Journal, 2003. **3**(2): p. 109-132.
19. Bruni J and Willmore LJ, *Epilepsy and pregnancy.* Can J Neurol Sci, 1979. **6**(3): p. 345-9.
20. Petersen I, Collings SL, McCrea RL, Nazareth I, Osborn DP, Cowen PJ, et al., *Antiepileptic drugs prescribed in pregnancy and prevalence of major congenital malformations: comparative prevalence studies.* Clin Epidemiol, 2017. **9**: p. 95-103.

21. Veroniki AA, Cogo E, Rios P, Straus SE, Finkelstein Y, Kealey R, et al., *Comparative safety of anti-epileptic drugs during pregnancy: a systematic review and network meta-analysis of congenital malformations and prenatal outcomes.* BMC Med, 2017. **15**(1): p. 95.
22. Carmichael SL, Shaw GM, Laurent C, Croughan MS, Olney RS, and Lammer EJ, *Maternal progestin intake and risk of hypospadias.* Archives of Pediatrics & Adolescent Medicine, 2005. **159**(10): p. 957-962.
23. Zaqout M, Aslem E, Abuqamar M, Abughazza O, Panzer J, and De Wolf D, *The Impact of Oral Intake of Dydrogesterone on Fetal Heart Development During Early Pregnancy.* Pediatr Cardiol, 2015. **36**(7): p. 1483-8.
24. Tronnes JN, Lupattelli A, and Nordeng H, *Safety profile of medication used during pregnancy: results of a multinational European study.* Pharmacoepidemiol Drug Saf, 2017. **26**(7): p. 802-811.
25. Mitchell AA, *Research challenges for drug-induced birth defects.* Clin Pharmacol Ther, 2016. **100**(1): p. 26-8.
26. Allen VM, Armson BA, Wilson RD, Allen VM, Blight C, Gagnon A, et al., *Teratogenicity associated with pre-existing and gestational diabetes.* J Obstet Gynaecol Can, 2007. **29**(11): p. 927-44.
27. de Jong J, Garne E, Wender-Ozegowska E, Morgan M, de Jong-van den Berg LT, and Wang H, *Insulin analogues in pregnancy and specific congenital anomalies: a literature review.* Diabetes Metab Res Rev, 2016. **32**(4): p. 366-75.
28. Zabihi S and Loeken MR, *Understanding diabetic teratogenesis: where are we now and where are we going?* Birth Defects Res A Clin Mol Teratol, 2010. **88**(10): p. 779-90.
29. de Jonge L, de Walle HE, de Jong-van den Berg LT, van Langen IM, and Bakker MK, *Actual Use of Medications Prescribed During Pregnancy: A Cross-Sectional Study Using Data from a Population-Based Congenital Anomaly Registry.* Drug Saf, 2015. **38**(8): p. 737-47.
30. de Jonge L, Garne E, Gini R, Jordan SE, Klungsoyr K, Loane M, et al., *Improving Information on Maternal Medication Use by Linking Prescription Data to Congenital Anomaly Registers: A EUROmediCAT Study.* Drug Saf, 2015. **38**(11): p. 1083-93.
31. Czeizel AE, *Specified critical period of different congenital abnormalities: a new approach for human teratological studies.* Congenit Anom (Kyoto), 2008. **48**(3): p. 103-9.
32. Given JE, Loane M, Luteijn JM, Morris JK, de Jong van den Berg LT, Garne E, et al., *EUROmediCAT signal detection: an evaluation of selected congenital anomaly-medication associations.* Br J Clin Pharmacol, 2016.
33. Prieto L and Martinez-Frias ML, *Case-control studies using only malformed infants who were prenatally exposed to drugs. What do the results mean?* Teratology, 2000. **62**(1): p. 5-9.

Table 1. Description of data from 13 EUROmediCAT registries for the comparison of false discovery rate methods in the analysis of safety of medication use during first trimester of pregnancy

EUROCAT Registry	Birth years included	Foetuses with CAs and at least one valid exposure	Foetuses with CAs following data cleaning by timing of exposure [a]	Data loss by data cleaning (%) [a]	Total eligible ATC coded exposures	Total ATC codes excluding those with <3 exposures	Average ATC coded medication exposures per pregnancy
Belgium, Antwerp	1997-2011	349	347	1	508	478	1.46
Croatia, Zagreb	1995-2011	198	190	2	243	218	1.28
Denmark, Odense	1995-2011	240	240	0	367	346	1.53
France, Paris	2001-2011	658	658	0	970	897	1.47
Italy, Emilia Romagna [b, c]	1995-2011	2,350	2,349	0	3,860	3,736	1.64
Italy, Tuscany	1995-2011	1,083	1,033	4	1,417	1,352	1.37
Malta	1996-2011	306	305	0	461	453	1.51
Netherlands, North Netherlands	1995-2011	2,451	1,848	25	3,133	2,933	1.70
Norway	2005-2010	3,051	3,051	0	5,535	5,481	1.81
Poland, Wielkopolska	1999-2011	3,180	469	85	640	632	1.36
Poland (excluding Wielkopolska)	1999-2010	12,389	2,214	82	2,788	2,741	1.26
Switzerland, Vaud	1997-2011	309	297	1	458	433	1.54
UK, Wales	1998-2011	2,057	2,057	0	3,030	2,924	1.47
Total	1995-2011	28,621	15,058	47	23,410	22,624	1.55

[a] *After exclusion of CA registrations with only medication exposures of unknown timing*

[b] *During the period 1995 to 2004 Emilia Romagna database had space for only 5 medications to be recorded*

[c] *Terminations of pregnancy for foetal anomaly were excluded from the Emilia Romagna registry as information on medications is only available for live and still births*

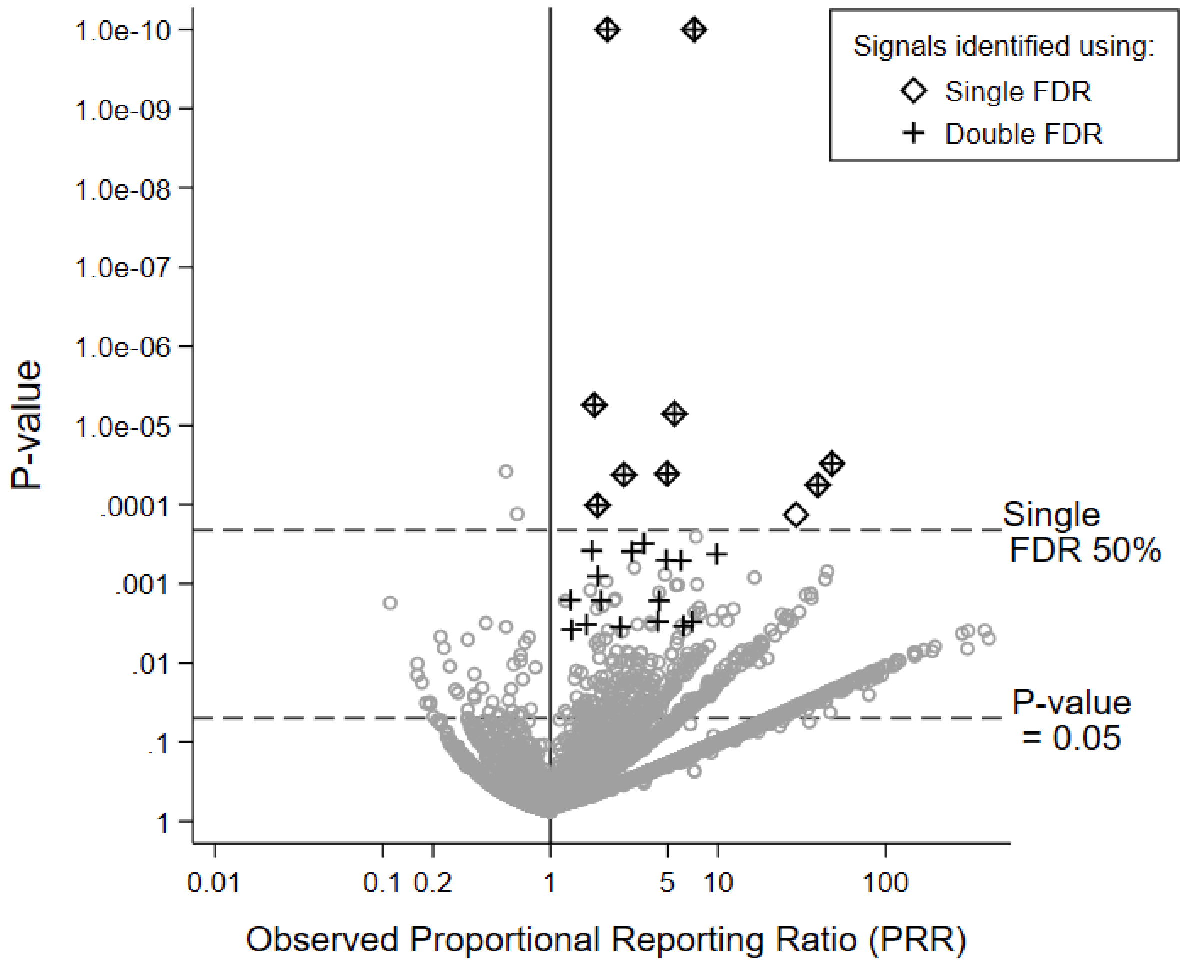

Figure 1. Smileplot highlighting signals identified from 28,396 medication-CA combinations by single and double FDR procedures, using a cut-off of 50% and ATC3 grouping (P-values truncated at 1.0E-10)

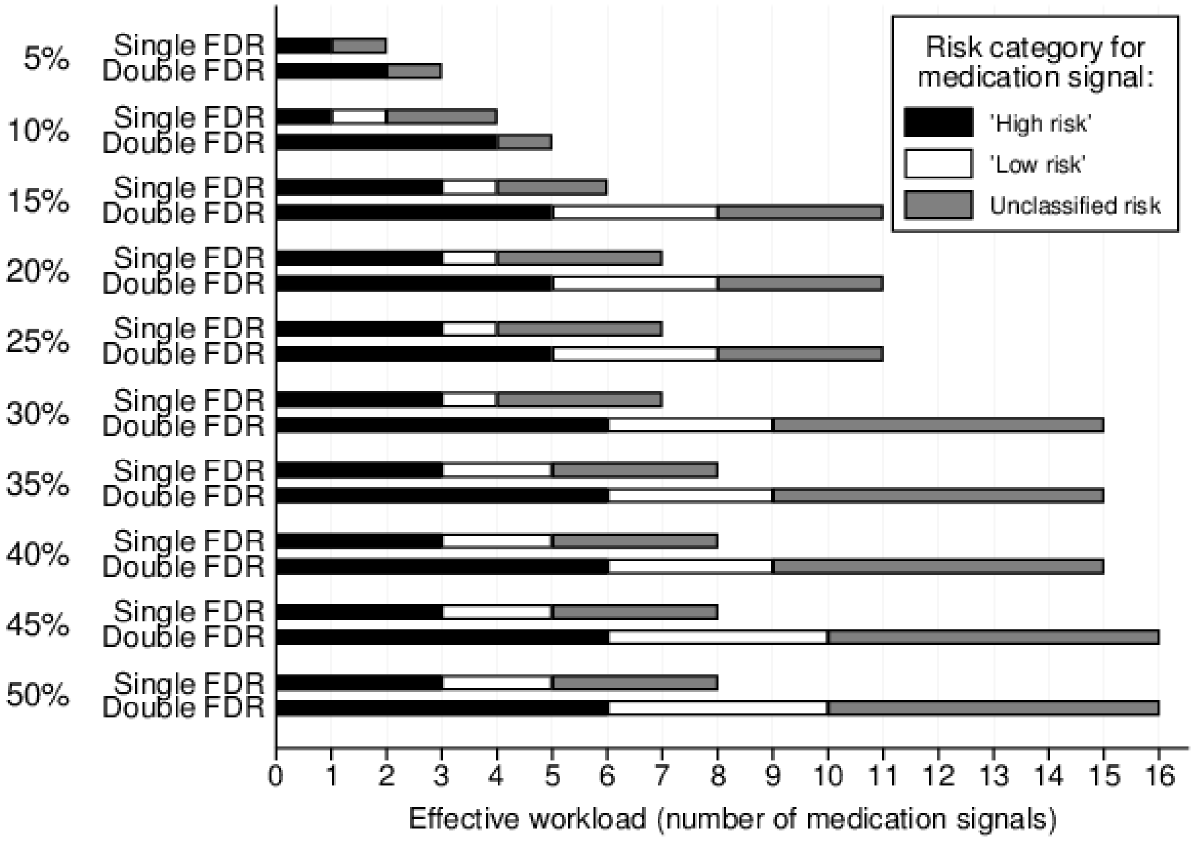

Figure 2. Effective workload and the number of medication signals in each risk category as identified by the Australian risk categorisation system database for single and double FDR, at FDR cut-offs ranging from 5% to 50%

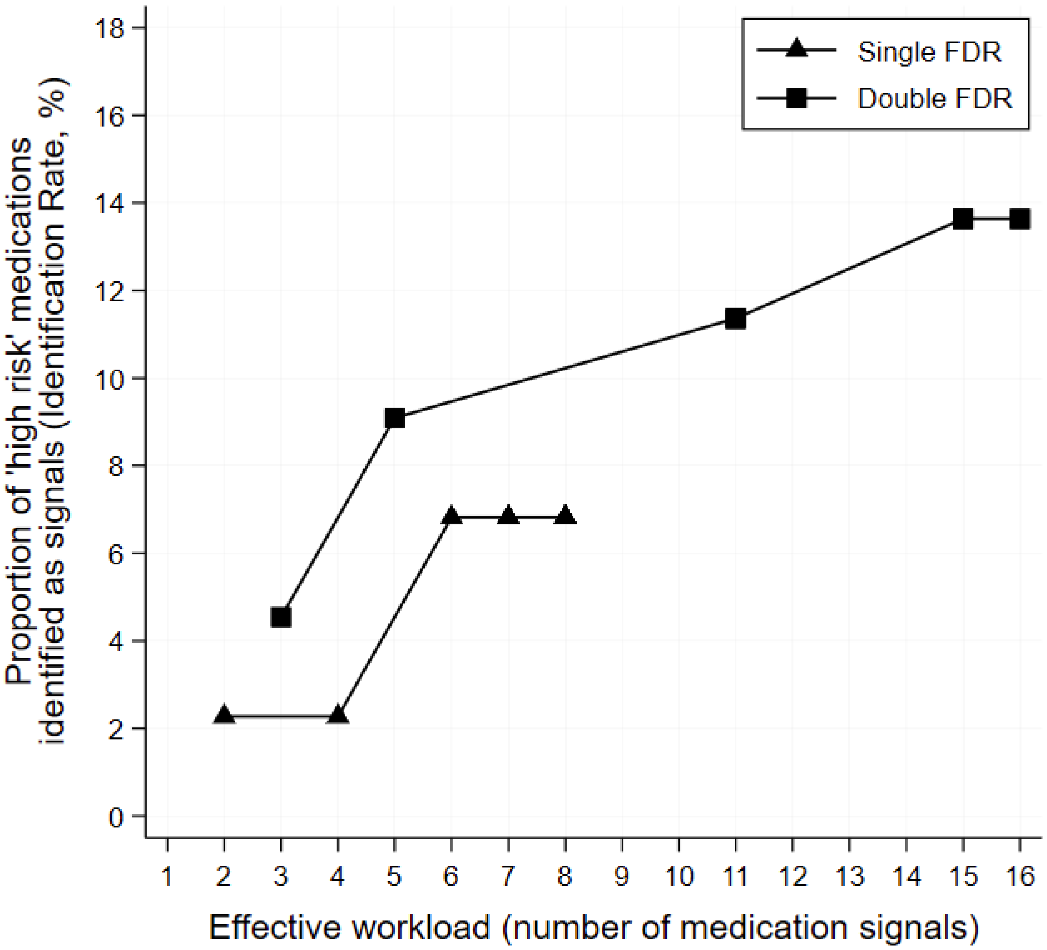

Figure 3. Proportion of “high risk” medications in the EUROmediCAT signal detection dataset that were identified as signals (identification rate) against the total number of medication signals (effective workload) for single and double FDR, with each point corresponding to a level of FDR cut-off ranging from 5% to 50%