[redacted] **MARCH 2019 MEETING**

AD HOC EXPERT GROUP ON SYSTEMATIC REVIEW AND META-ANALYSIS OF STUDIES ON ORAL HORMONE PREGNANCY TESTS

NOT FOR PUBLICATION

COMMISSION ON HUMAN MEDICINES

Title of paper: Evaluation of systematic review and meta-analysis of studies on oral hormone pregnancy tests, including Primodos **Type of paper: For advice**

Products(s): Oral Hormone Pregnancy Tests, including Primodos	Assessors: [redacted] Dr Katherine Donegan Dr Jane Woolley
Active constituent(s): various	Previous Assessments: N/A
MAH(s): Alinter Group Bayer plc GlaxoSmithKline UK Marshall's Pharmaceuticals Ltd Merck, Sharpe and Dohme Ltd Pfizer Piramal Healthcare Ltd Sanofi	Legal status: No longer licensed
Therapeutic classification: Hormone pregnancy tests	

Table of Contents

1. Issue

A study by Heneghan et al. published in the online journal F1000 Research concluded that use of oral hormone pregnancy tests (HPTs) in pregnancy is associated with increased risk of congenital anomalies. This was a systematic review and meta-analysis of observational case-control and cohort studies that included data from pregnant women that were exposed to oral HPTs within the estimated first three months of pregnancy and compared with a relevant control group.

The HPTs have not been available in the UK since 1978. However, the progestogenic and estrogenic components of HPTs are currently found in a range of widely-used authorised gynaecological medicines across the EU including oral contraceptives (OCs), hormone replacement therapies, treatments for endometriosis, disorders of menstruation, period delay and some cancers.

The Commission on Human Medicines (CHM) therefore advised that an ad hoc meeting of experts should be convened to carefully evaluate the new analysis and its findings.

2. Terms of reference for the review

The ad hoc Expert Group is asked to advise the CHM on the systematic review and meta-analysis of *Heneghan et al*, 2018 and in particular:

- the suitability and robustness of the methodology, including the selection and application of the data quality score
- any clinical implications.

3. Summary

The background to the issue which is the subject of the paper under consideration is detailed in Section 4. Section 5 summarises the publication as presented by the study authors and highlights points for consideration. Section 6 discusses contextual issues relevant to the interpretation of data on this issue, touches on the guidelines available to conduct meta-analysis, and more general issues related to meta-analyses of epidemiological data. Section 7 considers the findings of the meta-analysis and the potential implications.

4. Background

4.1 Hormone Pregnancy Tests

A range of HPTs were widely used within Europe to diagnose pregnancy from the late 1950s until 1978. HPTs contained natural or synthetic sex steroid hormones, usually a progestogen in combination with an estrogen. They were also licensed for the treatment of secondary amenorrhoea. Whilst the roles of progesterone and

estrogens in supporting implantation and early placental development have been defined, their role in fetal organogenesis and development remains less clear.

During the 1950s and 1960s, access to family planning advice and effective contraception was limited and abortion (other than in extreme medical circumstances) was illegal in the UK until 1967. Whilst pregnancy testing became more available from the 1920s onwards, it did not become mainstream or universal until much later. Most women did not usually attend antenatal clinics or consult a doctor about their pregnancy before the second or third trimester.

There is some evidence that when HPTs first became available in the UK in the 1950s, testing for pregnancy was intended for women who were considered more at risk of having a complicated pregnancy (Gal 1972, Michaelis 1983). The alternative to HPTs was a physical examination by the doctor or a relatively slow and expensive laboratory test. At the time of their introduction, and despite questions about their reliability in diagnosing pregnancy, HPTs were therefore considered to have several advantages over the alternatives and recognised as offering a more accessible, quicker and cheaper method of diagnosing pregnancy than the alternatives.

Exact usage data are not known but one source (Gal 1978) has estimated that almost 8 million women in the UK were prescribed an HPT, of which about a million prescriptions were for diagnosing pregnancy.

Against a background of heightened awareness of the possible teratogenic effect of medicines taken in pregnancy (through recent experience with thalidomide) a great many studies, letters and reviews have been written on the use of HPTs since they were first introduced to the market. In October 1967, the first observational study to suggest a link between use of HPTs in pregnancy and congenital anomalies in the child exposed in utero was published in a letter to the journal Nature (Gal et al, 1967). This study stimulated major research interest in the issue and many further epidemiological studies investigating a possible association between HPTs and a range of congenital anomalies were published thereafter.

4.2 Primodos

The most frequently used oral HPT in the UK, Primodos, contained two hormones – norethisterone acetate (NETA;10mg per tablet) and ethinylestradiol (EE; 0.02mg per tablet). NETA, a prodrug of norethisterone (NET), is a progestogen derived from nortestosterone that also has weak oestrogenic and androgenic properties. The action that NET exhibits is therefore complex and will depend on its dose, route of administration, duration of use, the presence or absence of other hormones and the presence or absence of different hormone receptors.

EE is a semi-synthetic estrogen with actions similar to those of natural estradiol. EE exerts potent estrogenic effects through its action at the estrogen receptors and has similar or slightly stronger estrogen agonist activity than the naturally occurring estrogens.

One Primodos tablet was taken on two consecutive days by women suspected to be pregnant. In women who were not pregnant, a withdrawal bleed would occur a few days later. A conservative estimate of the likely window for use of Primodos was from the week of the woman's first missed period to the end of the first trimester; that is, 4

to 12 weeks of pregnancy (2 to 10 developmental weeks). Other oral HPTs marketed in the UK similarly contained high doses of a progestogen and an estrogen.

4.3 Currently available products containing NET(A) and EE

HPTs including Primodos have not been marketed in the UK for 40 years; however, NETA or NET are currently the progestogenic component of a number of oral contraceptives and hormone replacement therapies, and a common treatment for menstruation disorders (**Table 1**). Typical daily doses range from 0.35mg per day in Noriday contraceptive to 60mg per day in Utovlan, a treatment for disseminated breast cancer. Similarly, EE is frequently used as the oestrogenic component of combined oral contraceptive preparations; a typical daily dose is 20 to 40 µg.

Though no currently licensed medicines in the UK contain the same combination and dosage of progestogens and estrogens as were present in HPTs such as Primodos, varying combinations and dosages of similar progestogens and estrogens are therefore used daily by many millions of women.

Table 1 Examples of oral medicines containing NET(A) in combination with EE or estradiol

Product name	Progestogen/estrogen (per tablet)	Posology	Indication
Primodos	NETA 10 mg EE 20 µg	1 tablet on each of 2 consecutive days	hormone pregnancy test - discontinued
Loestrin 30	NETA 1.5 mg EE 30 µg	1 tablet daily (21 days per cycle)	combined oral contraceptive
Norimin	NET 1mg micrograms EE 35 µg	1 tablet daily (21 days per cycle)	combined oral contraceptive
Noriday 350	NET 350 µg	1 tablet daily	progestogen-only oral contraceptive
Primolut N	NET 5 mg	10-15 mg daily (for 4-6 months)	endometriosis
		15 mg daily (for 10 days)	dysfunction uterine bleeding menorrhagia
		15 mg daily (for 3 days)	postponement of menstruation
		15 mg daily (for 20 days)	dysmenorrhoea
Utovlan	NET 5 mg	40 - 60 mg daily	disseminated carcinoma of the breast
Elleste Duet Conti	NET 1mg estradiol 2mg	1 tablet daily	hormone replacement therapy (HRT)

4.4 Congenital anomalies

Congenital anomalies are defined by the WHO as “structural or functional anomalies (for example metabolic disorders) that occur during intrauterine life and can be identified prenatally, at birth or sometimes may be detected later in infancy, such as hearing defects” (WHO congenital anomalies factsheets).

Information from Europe (Moore et al, 2008) and the USA (FDA 2005) suggests a total prevalence of major congenital anomalies of between 24 and 40 per 1 000 births (2.4% to 4%). In both territories, congenital heart defects appear to be the most common post-natal anomaly, followed by limb defects, anomalies of the urinary system and nervous system defects.

The cause of at least half (60%) of all post-natal congenital anomalies remains unknown with the other half having genetic or environmental causes or both, as shown in Table 2. Many genetic conditions occur in individuals with no prior family history.

Table 2. Causes of post-natal human congenital anomalies[1]

Cause of congenital anomalies	Proportion of all congenital anomalies
Unknown aetiology	60%
Multifactorial (genetic and environmental)	20%
Environmental agents of which: Recognised teratogen Maternal illness Infection at birth	7–10% 2% 3% 2%
Genetic mutations	8%
Chromosomal abnormalities	6% (prenatally 30%)

[1]adapted from Emery’s Elements of Medical Genetics, 10th Edition (Mueller and Young, 1998)

Single gene defects and most chromosomal defects occur prior to conception and, in many cases of congenital anomaly, one or both of these possible causes should be ruled out before alternative aetiologies are considered. As genetic research continues to progress, it is likely that more congenital anomalies will be identified as having a genetic cause.

5. Research for consideration

Oral hormone pregnancy tests and the risks of congenital malformations: a systematic review and meta-analysis. (Heneghan et al, 2018)

The evidence to be considered is the systematic review and meta-analysis by Heneghan et al, 2018. This is published in F1000Research, which is an Open research publishing platform which offers immediate publication of articles, whilst being peer reviewed, hence allowing amendments after publication.

The following assessment refers to version 2 of the paper, which was published on 29th January 2019 (first version on 31st October 2018). The changes from version 1

were mostly minor amendments and included a revision of labelling of effect estimates in the forest plots and an accompanying updated excel data sheet. Version 2 of the paper can be found in Annex 1.

Heneghan et al. conducted a systematic review of cohort and case-control studies to study the association between HPTs and congenital malformations and used meta-analysis to obtain summary estimates of the likelihood of an association. Potential biases in these estimates were assessed. The methodology, findings and conclusions of the authors are summarised below, and points that the Expert Group may wish to consider have been highlighted.

Methods:

Data Sources

The authors searched Medline, Embase and Web of Science as well as regulatory documents online, including the UK Government's 'Report of the Commission on Human Medicines' Expert Working Group on Hormone Pregnancy Tests' to retrieve relevant articles. Date limits or language restrictions were not applied, and search terms included (Primodos OR Duogynon OR "hormone pregnancy test" OR "sex hormones" OR "hormone administration" OR "norethisterone" OR "ethinylestradiol") AND pregnancy AND (congenital OR malformations OR anomalies). In addition, the authors performed additional searches for comparable high-dose HPTs available at the same time as Primodos.

Study selection

The authors applied the following criteria:

Inclusion criteria:

- Studies of women who were or became pregnant during the study and were exposed to oral HPTs within an estimated first 3 months of pregnancy
- Studies with a relevant control group
- Publication in any language.

Exclusion criteria:

- Studies in which intervention was an oral hormone taken for other reasons (e.g. oral contraception)
- Studies from which it was not possible to extract data on HPTs

Outcomes

'All major congenital malformations' was considered as the primary outcome of interest and the outcomes were also categorised into congenital cardiac, gastrointestinal, musculoskeletal, nervous system, urogenital and the VACTERL syndrome (Vertebral defects, Anal atresia, Cardiovascular anomalies, Tracheoesophageal fistula, Esophageal atresia, Renal abnormalities and Limb defects.

Risk of bias assessment

Two reviewers extracted the data based on the inclusion and quality assessment criteria and any discrepancies were resolved through discussion with other authors. Data was extracted about study type, number of exposed and unexposed pregnancies (to HPTs) and types of outcomes. Furthermore, if available, data was extracted on ascertainment of cases, age, parity, setting, exposure to other medications and confounding variables. In case-control studies, if data were reported on more than one control group, data was extracted for non-disease/non-abnormality controls, and control groups were combined if necessary.

The authors assessed the quality of the included studies using the Newcastle-Ottawa Scale (NOS) for non-randomised studies included in systematic reviews. The scale assesses three main aspects of a study with 8 criteria: 1) selection of study groups (cases and controls and/or exposed and non-exposed) in terms of case definition, representativeness of the population, and choice of comparator, 2) comparability of study groups and control of the most important confounder and 3) ascertainment of the outcome/exposure including potential issues of follow up and recall bias. Each positive criterion scores 1 point, except comparability of study groups, which can score up to 2 points (one for 'study controls for the most important factor' and one for 'study controls for any additional factors'). The maximum NOS score is 9 and, in the Heneghan study, a score of 1 to 3 points was considered to indicate a high risk of bias based on a previous research article (Lunny et al, 2013). For the 'comparability of study groups' criterion i.e., 'study controlled for the most important factor', the items that were reported in the original paper were selected and any disagreement between reviewers was resolved through consensus using a third author. The authors examined whether there was a linear relation between methodological quality and study results, by plotting the odds ratios against the NOS scores, and assessed correlations of NOS scores with several confounding variables that were collected.

Statistical methods

The authors calculated study-specific odds ratios and associated confidence intervals for all outcomes. A random-effects model was used for the meta-analysis and heterogeneity was assessed using the I^2 statistic and publication bias using funnel plots. Several sensitivity analyses were conducted: 1) excluding single studies to judge the stability of the effect and explore effect on heterogeneity, 2) excluding studies of low quality from the analysis and 3) excluding studies with zero events from the analysis.

Meta-regression was also performed to assess whether the observed heterogeneity could be explained by differences in NOS score in which the NOS score was used as a covariate against the log OR as weights. Cochran Q test was used to look at the

timing of HPT administration in relation to pregnancy and organogenesis and study design as part of a subgroup analysis.

The authors followed the reporting guidelines of the Meta-Analysis of Observational Studies in Epidemiology (MOOSE).

Patient involvement

Members of the campaign group 'Association for Children Damaged by HPTs' were acknowledged as being involved in the original discussions of the review and provided input to outcome choices, the search, the location of study articles and translations.

Results

Initially 409 items were retrieved for screening out of which 354 were excluded since they were not relevant to the aim of the review. The authors assessed the full texts of 37 articles and identified 24 articles for inclusion. These included 26 studies of which 16 were case-control and 10 were prospective cohort studies, and two were unpublished. These studies were published between 1972 and 2014 and included a total of 71,330 women. The case-control studies included 28,761 mothers, 594 of whom were exposed to HPTs while the cohort studies included 42,569 mothers, 3,615 of whom were exposed to HPTs.

The comparator groups for cohort studies tended to be women recruited from antenatal clinics or birth centres. The choices of controls in the case-control studies ranged from healthy infants born on a date close to the case infants to infants with malformations other than those under investigation. Further information on the characteristics of the included studies are listed in Table 2 of the Heneghan et al. paper (Annex 1).

Quality assessment

Three studies (Laurence 1971, Fleming 1978 and Haller 1974, the latter two were unpublished) were assessed as being at high risk of bias (NOS score of 3 or below). Twelve studies were judged to be at low risk of bias (NOS 7 to 9). The median NOS score was 5 (mean 6.1) and ranged from 2 to 9. A breakdown of the scores according to the specific NOS criteria is shown in table 3 of the Heneghan et al. paper (Annex 1).

For the NOS item assessing the comparability of cases and controls based on design or analysis (item 5), 12 case-control studies and six cohort studies were judged to have controlled for the most important factor (item 5a) and nine case-control studies and four cohort studies were judged to have controlled for important additional factors. Seven studies (Laurence 1971, Levy 1973, Tummler 2014, Fleming 1978, Haller 1974, Meire 1978 and Roussel 1968) did not report the confounding variables collected. There was a high positive correlation between NOS scores and the number of confounding variables collected (r=0.83).

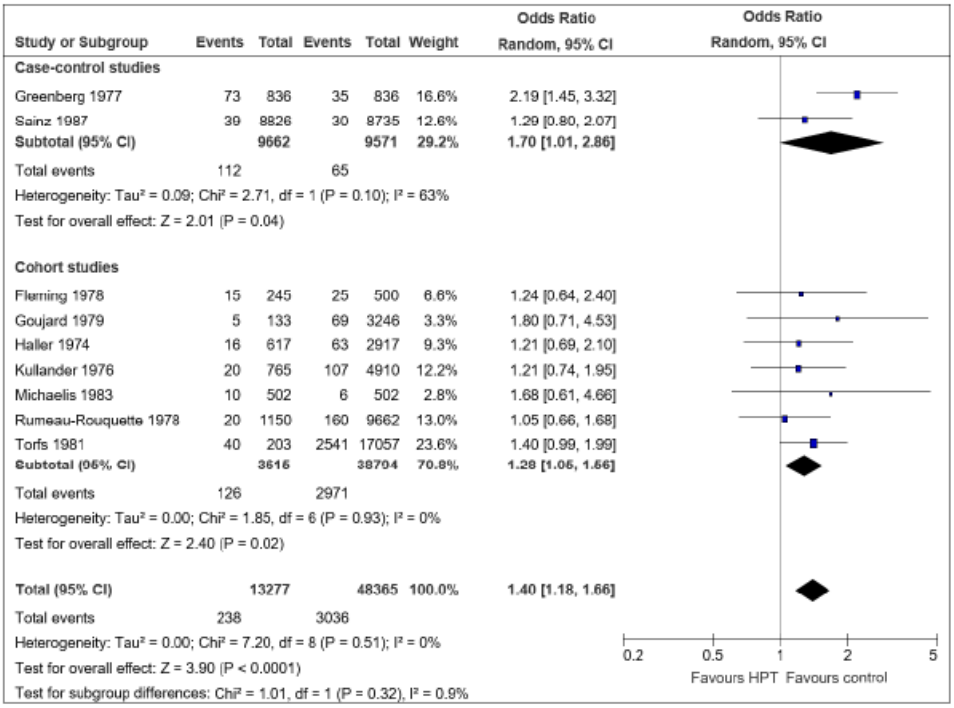
Funnel plots were presented to look at publication bias for the outcomes 'all congenital malformations' and 'congenital heart disease' but due to the insufficient number of studies more advanced statistical methods were not used.

Association of exposure to HPT with the risks of malformations

All malformations

Nine studies (two case-control and seven cohort studies), examined the association of pregnancy with all congenital malformations, the primary outcome of interest. These studies included a total of 61,642 mothers of infants of whom 3,274 were exposed to HPTs. A statistically significantly pooled odds ratio was obtained (1.40, 95%CI 1.18, 1.66) with no evidence of important statistical heterogeneity (I^2=0%), reported as implying a 40% increased risk of all congenital anomalies with exposure to oral HPTs (fig 2 from paper). For the two case-control studies the pooled OR was 1.70 (95%CI 1.01, 2.86) with substantial statistical heterogeneity (I^2=63%). For the seven cohort studies the pooled OR was 1.28 with no evidence of statistical heterogeneity (95%CI 1.05, 1.56, I^2 = 0%).

In a post-hoc sensitivity analysis, removing the studies that did not collect any confounding variables (Haller 1974, Fleming 1978, both assessed as low quality using the NOS criteria) gave a similar pooled effect estimate compared to the main analysis (OR 1.44, 95%CI 1.18, 1.75; I^2=11%). In the meta-regression, no association was observed between total NOS score and increased risk (p=0.51) and the test for subgroup differences was not significant (p=0.32).



Hormone Pregnancy Tests (HPT); Confidence Interval (CI)

Figure 2. Association of exposure to oral HPTs in pregnancy with all malformations in the offspring.

Congenital heart malformations (Figure 3 in Annex 1)

Five case-control studies and two cohort studies looked at congenital heart malformations with a total of 19,267 mothers and 218 exposed to oral HPTs. The pooled OR was significantly increased at 1.89 (95%CI 1.32, 2.72, I^2=0%) and similar results were obtained when removing one study (Levy 1973) that collected no confounding variables (OR=1.88, 95%CI 1.25, 2.85, I^2=12%). The pooled OR for

case-control studies was statistically significantly increased (1.87, 95%CI 1.23, 2.85, I^2=9%); the pooled OR for the cohort studies was increased but not significantly so (OR=1.95, 95%CI 0.44, 8.69, I^2=32%). The meta-regression was not significant (p=0.94).

Nervous system malformations (Figure 4 in Annex 1)

A total of 12,486 mothers, with 127 exposed to HPTs, were included in three case-control studies and two cohort studies looking at the association between HPT exposure and nervous system defects. The pooled OR was significantly increased at 2.98 (95%CI 1.32, 6.76) with a high statistical heterogeneity ($I^{2=}$78%). In a post-hoc sensitivity analysis, the statistical heterogeneity was removed after excluding two studies (Laurence 1971, Roussel 1968) that did not collect any data on confounding variables (OR=6.04, 95%CI 3.33, 10.78, I^2=0%).

Gastrointestinal malformations (Figure 5 in Annex 1)

Three studies (one case-control and two cohort) reported on the association between exposure to oral HPTs and gastrointestinal malformations, with a total of 2,722 mothers of which 79 were exposed to HPTs. The increased pooled OR obtained was not statistically significant (OR=4.50, 95% CI 0.63, 32.20) with high statistical heterogeneity (I^2=54%).

Urogenital malformations (Figure 6 in Annex 1)

A non-significant pooled OR of 2.63 (95%CI 0.84, 8.28, I^2=0) was obtained from one case-control and one cohort study looking at the association between HPT exposure and urogenital malformations.

Musculoskeletal malformations (Figure 7 in Annex 1)

Three case-control studies and one cohort study reported on the association of HPT exposure and musculoskeletal malformations, with 79 exposed women out of a total of 2,464 mothers. The cohort study (Torfs et al 1981) had zero events and removal of this study did not affect the effect estimate (OR=2.24, 95%CI 1.23, 4.08, I^2=0).

VACTERL (Figure 8 in Annex 1)

Two case-control studies (Nora et al 1975, Nora et al 1978) reported a pooled OR of 7.57 (95% CI 2.92, 19.07, I^2=0) for the association between HPT exposure and VACTERL, which was based on 135 women and 27 exposed to HPTs.

Points for consideration:

The authors have used recognised systematic review and meta-analysis approaches, including established statistical models. In the context of the particular issue being examined here and the underlying epidemiological data used, certain aspects on the implementation of these methods are presented below.

Implementation of the study selection criteria:

The presence of a control group was stated as one of the main inclusion criteria. The letter by Meire et al (1978), reported one case of oesophageal atresia out of 20 exposed to HPTs (out of a total of 500 women). Though not specifically stated by Meire to be a cohort study, the meta-analysis refers to the comparator group as

being 0 out of 480 women unexposed to HPTs (for which no cases of malformations were mentioned). A similar study by Oakley et al (1973), which interviewed 436 women who gave birth to children with anomalies and compared the prevalence of HPT exposure in the first trimester across different types of malformations (NTDs, cleft lip/palate, oesophageal atresia, limb reduction deformities etc) was not included in the meta-analysis. Oakley et al. (1973) found no significant difference in the rate of exposure in any one malformation group compared to the other and concluded that the study provided no evidence of an association given that the likelihood that HPTs could cause an increase in all malformation was unlikely on biological grounds.

Inclusion criteria comprised studies of women exposed to oral HPTs in the first trimester of pregnancy and studies which also had a control group. For two studies data on oral HPTs specifically (rather than injectable HPTs which contained progesterone and estradiol instead of synthetic hormones), was not clear (Goujard et al 1979, Michaelis et al 1983). The two Nora et al. case-control studies (1975, 1978) included data on exposure to unspecified estrogens+progestogens, with data on number exposed to oral HPTs explicitly provided only for the cases. The authors state, in their meta-analysis extraction sheet, that the 15 cases who were exposed to OCs were excluded in Nora et al (1978) case-control studies 2 and 3. This implies that in addition to women with exposure to HPTs, women were also included if they had exposure to hormones given to prevent threatened abortion and to clomiphene+progestogen for infertility.

Some studies (e.g. Goujard et al. 1979 for cardiac defects, Roussel et al. 1968 for all congenital anomalies) also reported observations on different congenital anomalies that were being reviewed in the meta-analysis, but these observations do not appear to have been presented within the relevant anomaly category analyses.

Of note, the identified study by Tummler et al (2014) has not been included in any analyses. This is appropriate since the study was based on retrospective spontaneous case reports and was designed to explore reporting bias.

Interpretation of the Newcastle Ottawa Scale (NOS) scoring criteria:

The NOS criteria are designed to be relevant for general case-control and cohort studies. The studies are not particularly detailed but need careful interpretation to ensure they are robustly applied. Two criteria (three NOS points) where interpretation is particularly important, given the purpose of the underlying studies, are discussed below.

a) Controlling for confounder vs collection of data on characteristics: To reflect the importance of accounting for potential differences between exposed and non-exposed populations or cases and controls, two items in the NOS data quality system relate to studies controlling for the 'most important factor' and the 'most important additional factor' or `other additional factors'. The Heneghan et al. publication does not state what factors were considered by the authors to be the most important confounders and whether these were prespecified or consistent across studies.

   For the association between HPT exposure and congenital anomalies it may be considered that maternal age, reproductive history including parity and previous miscarriages, family history of anomalies, history of

threatened abortion, concurrent infections, folic acid supplementation, smoking, geographical location, educational level, and socioeconomic status, amongst others, are important potential factors that may act as confounders and therefore need to be controlled for. The NOS scoring system specifically awards points to studies controlling for the confounding factors.

In this meta-analysis, 19 of the 26 studies were given 1 point for controlling for the 'most important factor' and 13 of the studies were given 1 point for 'important additional factors'. Of those awarded points for these criteria, it is important to note that while some state that they collected data on at least a subset of the potential confounders listed above, few matched or adjusted for more than basic characteristics related to the mother (e.g. age). Some stated that there were no differences between groups for some characteristics although this was not always statistically explored (e.g. Greenberg et al. 1977, Hadijigeorgiou et al. 1982, Hellstrom et al. 1976). The NOS manual suggests that in the review of studies, 'statements of no differences between groups or that differences were not statistically significant are not sufficient for establishing comparability'. Others demonstrated some differences between groups (e.g. Janerich et al. 1977, Michaelis et al. 1983) but did not adjust for these variables.

b) Outcome of interest not present at start of study (cohort studies): All the cohort studies in the meta-analysis score 1 point for this criterion. As this is a study in pregnancy the interpretation of this criteria is particularly important. As HPTs would have been used to confirm pregnancy the presence or absence or a malformation at the point of exposure would not have been known. If the start of the study is considered to be the point of conception, then a malformation can clearly not be already present. It is also not clear if studies generally excluded genetic cases since those with a genetic link would be present at conception and hence study start. This criterion is considered to be more relevant for studies on other drug safety issues and reflects a lack of robustness and generalisability in the NOS criteria.

Application of the Newcastle Ottawa Scale (NOS) scoring criteria to individual studies:

Due to the lack of detail available in some of the underlying publications, classification of the quality of the HPT studies using the NOS can be challenging. A few points of uncertainty relating to the application of NOS criteria not already discussed in relation to their interpretation above are discussed below with examples.

a) Selection of comparator group: The selection of the comparator group, either controls or an unexposed cohort, is a crucial component in epidemiological studies and can potentially have a substantial impact on the results by introducing biases.

In the case-control studies, controls included: women who gave birth to healthy infants born on a date close to the case infants; women who gave birth to infants with malformations other than those under investigation; women who had the same postcode as those exposed; women who lived geographically far away; women who proactively sought a pregnancy test and were given a non-HPT test; and women who did not seek a pregnancy

test. In some studies selection of controls were not adequately defined (e.g. Gal et al 1972, Roussel et al 1968, Hellstrom et al 1977).

In the Gal et al (1972) study in which the selection of controls was allocated 1 NOS point, the cases were identified from two hospitals to which they were admitted for surgical treatment, but the publication does not specify where the control groups came from. A subsequent paper (Sainz et al., 1987) said of the Gal paper: "Some authors did not consider the selection of control groups to be adequate, as the cases came from an extensive area of England that included several hospitals and in contrast, the control groups just came from one single hospital and it is likely that the differences between the percentage of exposed children in both groups could be due to preferences for the use of one method or another for confirming pregnancy". If there is variation in the use of HPTs across different hospitals, then this could lead to issues with the comparability of the cases and the control groups.

For some studies included in the meta-analysis, control groups have been combined by the authors of the meta-analysis. The control group should be representative of the source population that produced the cases. If this holds true, then combining different control groups from the same study, as has been done in this meta-analysis, is unlikely to affect the results. However, if controls were not selected with care or, due to matching, the exposure distribution in the controls differed from the exposure in the population, then combining them could bias the effect estimates. In this meta-analysis the matched and random control groups in the study by Ferencz et al (1980) have been combined. The Torfs (1981) study, which was assigned a maximum score of 9 NOS points, also had two control groups, one in women who had pregnancy diagnosed through a serum test from 1959 to 1964, and the other through a urine test from 1965; these were compared with selection of the HPT group from 1962. Differences in the use or choice of test over time could also lead to a lack of comparability across these groups and indeed some differences in the cohorts are suggested in the paper.

b) Ascertainment of outcome/exposure: In the NOS scoring system ascertainment of exposure is given a score of one if the exposure is ascertained either from a secure record or structured interview (blinded for case/control status in case-control studies). In studies evaluating possible adverse effects on the developing fetus, reliable exposure ascertainment is crucial to minimise exposure misclassification and to assess timing of exposure to the critical period of organogenesis. In addition, bias can be introduced in studies in which giving birth to a child with an anomaly could influence recall of exposure.

   For several of the case-control studies, in which interviews were the main method of collecting data on exposure and which were given a score of one for this parameter, it is not clear from the original publications whether or not the interviewer was blinded to the case/control status (Ferencz et al (1980), Janerich et al (1974), Lammer et al (1986), Nora et al (1975), Nora et al (1978) and Rothman et al (1979)). The case series by Meire et al also provided no details about the interviews and was assigned a score of one for exposure ascertainment.

Statistical approach and methods used:

Heneghan et al. state that the Cochran Q test was used to look at the timing of organogenesis and study design as part of a subgroup analysis. However, the results for subgroup differences has only been mentioned for 'all congenital anomalies' which is assumed to be for study design.

Study Authors' discussion:

The study authors state that significant associations were observed for the primary outcome of all congenital malformations and separately for congenital heart malformations, nervous system malformations, musculoskeletal malformations and the VACTERL syndrome. Many of these pooled analyses had zero heterogeneity and the direction of effect favoured the controls in 30 out of 32 analyses undertaken. Sensitivity analyses also showed similar results and there was no relation between NOS score and increasing risk.

The authors go on to state that it has been suggested that there is no mechanistic argument for teratogenicity based on assumptions that a teratogenic effect of HPTs would be mediated by actions on estrogens and progestogen receptors and that concentrations of ethinylestradiol and norethisterone in the fetus would be too low to have a significant effect on those receptors. However, Gal had reported that that due to bleeding occurring in some pregnant woman soon after HPT exposure, the equilibrium of the uterus is affected. In terms of animal toxicity, they also highlight a study published in 2018 (Brown, et al., 2018) showed that components in Primodos are associated with dose-dependent and time-related damage in zebrafish embryos and affect nerve outgrowth and blood vessel patterning in zebrafish. Other animal studies have shown minimal effect on embryo development.

The authors include the following as strengths and weaknesses of their meta-analysis:

1) Causal associations in absence of randomisation is difficult, although for questions about harms the 'Oxford CEBM levels of evidence' puts systematic reviews of case-control studies on a par with systematic reviews of randomised trials.
2) The analysis was based on raw data from the publications and did not adjust for confounders, however most of the studies in the review used matched controls.
3) Susceptibility bias was another issue as women with threatened abortions might be more likely to present and take medication. Careful matching could have mitigated this and the previous issue, and 13 out of 16 studies controlled for the most important factor on the NOS scale.
4) The severity of malformations studied would have led to differing risk estimates across studies.
5) Bias could have been introduced by inappropriate methods of ascertainment of malformations and exposures.
6) Incomplete and uneven reporting as well as publication bias could have introduced bias and affected the effect estimates.
7) The NOS scoring system has been criticised and a major weakness is the possible low agreement between assessors, especially among those with limited experience in doing systematic reviews (Hartling, et al., 2013, Oremus, et al., 2012). But it is widely used in assessing quality of non-randomised

studies and a NOS score of 0 to 9 has been previously used as a potential moderator in meta-regression and been recommended by the Cochrane collaboration.

8) Sensitivity analyses i.e. changes in NOS score and subgroup differences, showed similar results to the main analysis.

9) The chance of publication bias was reduced by translation and assessment of unpublished data. Due to the large sample sizes in studies for all congenital malformations, congenital heart disease and nervous system malformations, small unpublished studies, if not highly heterogenous, would have little effect on the effect estimates. Due to the small sample sizes of studies looking at gastrointestinal, urogenital, musculoskeletal and VACTERL malformations the interpretation of the results should be treated with caution.

10) A significant strength of the study was the use of standard systematic review methods and a research question that solely focussed on exposure to HPTs.

11) While the effect of unmeasured confounder(s) cannot be ruled out, the lack of heterogeneity means such a confounder would have had to act in the same direction. Furthermore, confounding factors with variable effects on the pooled ORs would have led to a high degree of heterogeneity, which was not the case.

Authors conclusion

This systematic review shows an association of exposure to oral HPTs with congenital malformations.

Points for consideration:

The authors provide an extensive list of the strengths and limitations of their study although there is limited discussion on the potential extent of the impact of these on the conclusions drawn as is usual in a scientific publication where brevity is required. A wider consideration of the strengths and limitations of this meta-analysis is included in Section 6 of this paper. However, additional background details for one of the points highlighted by the authors is provided below in order to provide further clarification.

CEBM: As stated by the authors, the Oxford CEBM levels of evidence puts systematic reviews of case-control studies on a par with systematic reviews of clinical trials (level 1 of hierarchy). However, the case-control studies referred to are nested case-control studies where both cases and controls are drawn from the same population in a fully enumerated cohort. Of the case-control studies included in this meta-analysis only a couple (Fleming et al. 1968 and Kullander et al. 1976) could potentially be nested case-control studies but a lack of details means in those two publications makes this difficult to verify. CEBM places case-control studies at level 4 out of 5 of the hierarchy.

6. Assessor's discussion of the data

The systematic review and meta-analysis conducted by Heneghan et al. finds an increased risk of congenital malformations with use of oral HPTs in pregnancy.

In considering this finding, the Group may wish to take account of a number of broader aspects that are potentially relevant to the interpretation of the results as presented.

6.1 Scientific and epidemiological knowledge in 1960s – 1980s

The time period for the conduct of the studies, included in this meta-analysis, ranges from the 1960s to the 1980s. The quality and rigour of the study design and data collection and the statistical methods applied in epidemiological studies conducted in this time frame would, more likely than not, be inferior to those conducted today (characteristics of studies included in the meta-analysis can be found in Annex 2).

More recently, there have been rapid advances in: the availability of digital databases to support high-quality routine and bespoke data collection, validation, and storage; the development of new epidemiological and statistical methods to adjust for biases and confounding; and increased understanding and knowledge of the applicability of these methods and the various potential biases that need to be accounted for which is reflected in guidance on how to conduct and report high quality epidemiological research (e.g. ENCePP methodological guide, ICPE Guidelines for Good Pharmacoepidemiology Practices, STROBE reporting guidelines). For example, statistical software was not routinely available to implement regression models and adjustment for confounding in particular would have been extremely difficult. Many of the studies included in the Heneghan et al. meta-analysis used basic methods to compare risk in different groups and/or presented proportions by confounders in simple tabulations. In nearly all studies presented in this meta-analysis, only simple non-parametric tests e.g. McNemar's test, Chi-squared test, and Fisher's exact test were used. Only the study by Ferencz et al (1980) additionally used multiple logistic regression to adjust for confounders. In many instances the statistical test was not mentioned.

A lack of clear reporting in the individual studies on many important aspects of study methodology, statistical analysis, confounders and selection of controls makes it challenging to appraise the studies accurately. Several studies were published as short letters to the editor for which details on the data collection methods, the analysis conducted, or any limitations are not available (e.g. Meire et al 1978, Levy et al 1973, Laurence et al 1971). This raises a question around how well the studies can be judged against the criteria of the Newcastle-Ottawa Score, which was designed in the context of modern reporting standards. Although sensitivity analyses excluded individual studies judged by the paper authors to be of low quality from the meta-analysis, the combined effect estimate would be expected to have been affected by the inherent flaws of most of the studies and any systematic biases compounded.

With time, there has also been a substantial increase in evidence and knowledge about the most important factors to consider in the study of potential associations between congenital anomalies and drug intake during pregnancy. The importance of folic acid intake during pregnancy to prevent neural tube defects in the baby is one important example; others include various other maternal nutritional factors (intake of iron, iodine) to ensure optimal maternal health during the pregnancy, and the detrimental effect of alcohol consumption, lifestyle drugs and smoking during pregnancy. In the majority of the studies these factors were not collected and controlled for in the analysis.

Although the same sorts of biases may also occur in epidemiology studies conducted today, there is greater awareness of this issue, methods for controlling the biases are more advanced, and studies are generally reported better. Accordingly, many of the later studies did attempt to address at least some of the concerns with the earlier studies.

6.2 Potential biases and confounding

Several methodological issues need to be considered for the studies identified in the review. While the potential for biases was largely unavoidable due to the time at which these studies were conducted the extent to which they could have an impact on the validity and strength of the statistical association reported in the meta-analysis still requires careful consideration. Some important issues are outlined below:

Recall bias

In nearly all the studies identified here, exposure status was obtained via an interview with the mother given that this data would not have been systematically recorded in a way in which it could be extracted for use in future research. When exposure is identified via the mother in this way, using normal babies as a control potentially increases the risk of recall bias as mothers of children with a malformation may be more likely to remember different exposures. A long delay between exposure and ascertainment may also result in exposure misclassification due to issues of recall. In retrospective studies, there may also be inadvertent increased pressure from unblinded investigators on mothers of babies with congenital anomalies to remember what medicines they took during pregnancy. As an example, Nora et al (1975) report that many of the initial answers that were negative in patients, were positive after "considerable probing".

Use of HPTs and characteristics of women (channelling bias)

At the time these studies were conducted, prior (or family) history of having a child with a congenital malformation or of single or multiple miscarriages would likely have increased the probability that a woman was offered, or sought, a pregnancy test as they were not routine, and prior history of a pregnancy with a malformation is a known risk factor for malformations in subsequent pregnancies. Many of the studies did not either match for reproductive history or adjust for it in the analysis (e.g. Janerich et al (1974), Janerich et al (1977), Lammer et al (1986), Laurence et al (1971), Levy et al (1973), Nora et al (1975), Nora et al (1978), Polednak (1983), Rothman et al (1979), Sainz et al (1987)).

Gal et al (1972a) reported a significant increased risk of spina bifida in babies of mothers exposed to HPTs. In a separate publication Gal states that in 18 (of the 19 mothers) exposed to HPTs, who had malformed babies and were included in that study, pregnancy was unwanted (Gal et al, 1972b). This raises the question of whether the women using HPTs had underlying complications that meant they were different in some way that may make them more predisposed to having infants with congenital defects and questions the robustness of the study finding for an increased risk of neural tube defects (NTDs) with HPTs. Similarly, in a retrospective study of cases of cleft lip and palate in Western Australia 18 of the 22 mothers who had received a pregnancy test were reported to have had unwanted pregnancy and had requested abortion several times (Brogan 1975, not included in Heneghan et al). In the study by Torfs et al (1981), the 3 groups receiving pregnancy tests (urine, serum and hormone) had much higher proportions of women who had reported a previous fetal loss, low birth weight infants or were more than 40 years old compared with women who had no

pregnancy test. Kullander et al (1976) highlighted a high incidence of Primodos usage in the group of induced abortions containing unwanted pregnancies. These examples suggest that clinical indication for the test is an important factor to be accounted for in the design and analysis of these studies.

Confounding

A number of risk factors could increase the probability of the occurrence of a congenital malformation including, as already highlighted, genetic factors, because a personal history of congenital malformations or history in a family is a strong predictor of future risk. Data on other relevant risk factors such as maternal nutritional status, smoking, alcohol use, folic acid use, concurrent infections, and exposure to pesticides and certain chemicals were not collected in majority of the studies. In the Heneghan paper unadjusted relative risks were reported with only limited adjustment considered through matching in the case-control studies.

In the studies included in the meta-analysis there is a general trend and consistency towards risk estimates greater than one. The authors argue the lack of heterogeneity adds weight to the findings as any residual confounding would have had to operate in the same direction. However, this is not implausible given the general lack of adjustment for many key potential confounders and the question whether unaccounted for confounding is sufficient to account for the observed effects which are small and of borderline significance remains, particularly given concerns of publication bias.

Publication bias

Within the publication, funnel plots designed to explore the risk of publication bias were presented for 'all congenital malformations' and cardiac malformations. Due to the low number of studies for the other malformations studied funnel plots were not carried out or presented and further statistical tests were not conducted due to the possibility of the tests being underpowered. The funnel plots presented in Heneghan et al. have not been interpreted by the authors and are difficult to assess due to the small numbers and the size of any potential risk based on the data that were published.

Publication bias cannot be excluded due to the age and timing of the studies which were carried out when the case of thalidomide was still relatively recent. Many of these studies were published after the products were removed from the market. It is therefore possible that questions regarding their safety were already known to authors. Further, the drive for transparency and publication of all data was not as it is now and observational studies showing non-statistically significant or marginal findings have historically been less frequently published than clinical trials or studies showing large effects meaning that we would have less confidence that all relevant data are available for inclusion in any review.

6.3 The Newcastle-Ottawa Scale (NOS)

One of the most important aspects of any systematic review and meta-analysis is assessing the risk of bias of the included studies. Several quality scales are available, including the NOS. This scale was developed as part of an ongoing collaboration between the Universities of Newcastle, Australia and Ottawa, Canada with a goal to provide an easy and convenient tool for quality assessment of non-randomised studies (case-control and cohort) to be used in a systematic review. The NOS was developed to "assess the quality of non-randomised studies with its design,

content and ease of use directed to the task of incorporating the quality assessments in the interpretation of meta-analytic results" (Wells et al). The scale is composed of three broad perspectives: 1) the selection of study groups 2) comparability of the study groups 3) ascertainment of either exposure or outcome of interest for case-control or cohort studies respectively. In total eight aspects of the study are considered. A points system is used, whereby a study can be judged to achieve a maximum of one point for each of seven items and a maximum of two stars for the item on comparability, to give a maximum NOS score of 9.

This scale has been criticised by some, on the low agreement between assessors doing the scoring (Hartling & Milne, 2013), but it has been suggested that this can possibly be ameliorated by training of authors. As with any scale, there is a subjective element to assigning scores, especially if there is lack of clarity either of what is being asked or of what is being assessed. The NOS manual is not very detailed and some aspects on interpretation and implementation of the criteria appear to lack clarity. The Cochrane Collaboration note its potential use as a tool and its simplicity but warn that it may need to be customised to the issue of interest.

The NOS scale does not consider biological plausibility. When looking at risks of specific malformations the exact timing of any exposure is important to consider and should ideally be within one week of the critical period of organogenesis for the observed anomaly. Where detail on exact timing of exposures is not available this introduces some uncertainty about the plausibility of any observed effect. Some studies included in the meta-analysis did at least state timing of exposure to be first trimester, however there were others in which the timing of exposure was not clear or not mentioned (e.g. Hadijigeorgiou et al 1982, Laurence et al 1971, Meire et al 1978, Rothman et al 1979).

Comparability of cases and controls/cohorts

The comparability items of the NOS scale questions whether the study 'controls for the most important factor' and 'controls for additional factors'/`controls for the second most important factor'. The meaning of 'important' factor is not defined but should be specific to the research question. The most important confounders and the most important additional confounders should therefore be prespecified based on a careful consideration of the issue and prior to scoring the individual studies. It is not clear from the publication what the authors considered to be most important factor or factors. Usually, matching in case control studies is at least by age and sex, which is generally used to increase the efficiency of adjustment of confounding of these variables compared to unmatched case-control studies. Therefore, as age (and sex) are often potential confounders it has been argued that the comparability aspect of the NOS system has little importance. This is because the vast majority of case-control studies are assigned points for this aspect on the NOS scale, most likely based on matching by age, despite not sufficiently controlling for other potentially more important factors (Stang, 2010).

Judging which is the most important factor will not usually be straightforward. In the relationship between exposure to HPTs and congenital anomalies, a few factors may be considered to be important: maternal age, reproductive and family history, folic acid intake (for NTDs), concurrent infections, intake of other drugs or alcohol etc. In the publication by Heneghan et al. it is not clear what the important factors were or whether these were pre-specified. It is also not clear why the authors sometimes refer to confounding variables that were 'collected', rather than 'controlled' for in the original studies and how they interpreted this when implementing the relevant NOS criteria.

Thirteen out of the sixteen case-control studies and six out of the ten cohort studies included in the meta-analysis were assigned a score of 1 NOS point for the 'most important factor', although these studies were highly variable in the degree of adjustment/matching and the factors used. The latter category also comprised studies in which potentially important additional data on confounders was collected but not accounted for in the analysis.

Selection of controls/unexposed group

For case-control studies the NOS item on 'selection of controls' assesses whether the controls are derived from the same population as the cases and could have been cases had the outcome been present (online NOS manual, Wells et al). One NOS point is allocated if the controls were from the same 'community' as cases although `community' is not explicitly defined and as already discussed has in this case included control groups identified in different hospitals and using prospective vs retrospective methods. Control selection is an integral part of a case-control study and poor control selection can distort the effect estimates and provide contradicting results.

Other items of the NOS

Exposure ascertainment is another important aspect to consider, as misclassification of HPT exposure could lead to an underestimate or overestimate of the effect. Some studies were allocated one NOS point for this aspect, despite the lack of clarity in the original papers regarding how exposure was ascertained, or seemingly not meeting the definition of secure record or structured interview (blinded for case-control studies). Most case-control studies stated that the interviews conducted were not blinded, and in many studies it may be considered that there was not enough detail about the interviews conducted or robustness of methods to assign a score for ascertainment of exposure. Blinding might not always be possible, and it has been suggested that conducting highly standardised interviews performed by trained personnel, which are regularly monitored throughout the study, is a feasible approach for ascertainment of exposure (Stang, 2010). Also, as previously highlighted the overall validity of the implementation of the 'Outcome of interest not present at start of study' criteria for cohort studies is unclear.

Categorisation of overall risk of bias score

It is widely acknowledged that categorisation of the overall degree of bias in observational studies is subjective. Heneghan et al. referenced the study by Lunny et al (2013) to provide a classification of bias within a study as being low, medium or high. However, this ignores the relative importance of the different NOS criteria and the potential impact of deviation from them in the context of the issue and exact studies being considered. This means that a study that receives no points for either the selection of controls or comparability criteria but one point for the other (potentially less important) aspects the study can still have an overall score of 6 and be considered of low risk of bias i.e. high quality.

Lack of reporting of methodological details in published articles may potentially distort the assessment of risk of bias, as has been shown for RCTs (Deverreux, et al., 2004). A median NOS score has also been observed to be significantly higher among

reviewers compared to the authors of individual studies (Ka-Lok Lo, et al., 2014) implying that reviewers may not always have all information needed available from the published article in order to assess the risk of bias reliably. Many older studies, such as the ones included in this meta-analysis, suffered from lack of clear reporting of several characteristics of a study making it difficult to apply an accurate score.

A review of published literature on tools available to assess quality or susceptibility of bias in observational studies highlighted the lack of a single obvious candidate tool for assessing quality of observational studies (Sanderson et al, 2008). The authors suggested that a more transparent checklist approach that concentrates on the few, principal, potential sources of bias in a study's findings is preferable. They recommended that a tool should i) include a small number of key domains, ii) be as specific as possible (with consideration of the particular study design and topic area), iii) be a simple checklist rather than a scale and iv) show evidence of careful development.

To assess older studies specifically looking at HPTs, many of the published tools consider more complex criteria based on modern standards of epidemiological approach; ideally the criteria need to be developed, based on knowledge of the most important factors specific to the exposure, the population and the outcome.

More recently, the ROBINS-I tool designed for assessing the risk of bias in non-randomised studies of interventions has been published (Sterne et al, 2016). The tool reflects a shift in focus away from questions of methodological quality, as used in the Newcastle-Ottawa scale through a checklist and numerical scale, towards an assessment of the risk of bias across different domains. The tool is based on the underlying principle of a hypothetical pragmatic randomised trial, conducted on the same participant group and without features putting it at risk of bias, whose results would answer the question addressed by the observational study. It then focuses on different domains of potential bias and how they mean the study deviates from the target trial. The risk of bias should be interpreted as a risk of material bias i.e. that the impact of the bias on drawing valid conclusions is what should be considered. This means that identifying a serious risk of bias in one domain clearly highlights that the study as a whole is at risk of serious bias.

The authors of the ROBINS-I tool highlight that only exceptionally will a study be considered at low risk of bias due to confounding. As highlighted, confounding can have one of the most major impacts on the effect estimates of an epidemiological study, but this is not well reflected in the NOS which weights the potential for confounding in the final quality score using just two points, the same as for the follow-up of patients (when evaluating cohort studies) for example, which for studies for major congenital malformations is arguably considerably less relevant given that these will be established early on after birth.

If the ROBINS-I tool for non-randomised studies of interventions is considered in relation to the studies included in the analysis conducted and under review here, of the seven domains of potential bias, earlier discussion in this assessment suggests that the two where deviations from an ideal hypothetical pragmatic randomised controlled trial occur are `bias due to confounding' and `bias in classifications of interventions'. In particular, the studies will in the majority likely suffer from a serious risk of bias due to confounding given a lack of control for maternal history and/or the reason for receiving a test, or a critical risk of bias due to confounding given the use of unexposed comparator groups rather than controls who used an alternative type of pregnancy test. Further they would often be considered at a serious risk of bias in the classification of interventions, as exposure was predominantly identified

retrospectively via interview. They would also all be considered at serious risk of `bias in the selection of the reported results' but this reflects practice at the time they were conducted where there was considerably less emphasis placed upon transparency.

6.4 Usefulness of meta-analyses and the hierarchy of evidence

Meta-analysis of randomised controlled trials (RCTs) is considered to provide the strongest evidence on an intervention, however, due to ethical and methodological reasons, RCTs are not feasible in many situations. Observational studies constitute the majority of published clinical research and publication of meta-analyses solely of observational studies has increased over the years although they are still not commonly conducted. Systematic review and meta-analysis of multiple studies addressing the same clinical question can provide potentially strong evidence, however synthesis of evidence from observational studies differs from the approach used when examining evidence from RCTs. This is due to the observational nature of the study design which lacks the experimental element of random allocation whereby the study groups might differ with respect to many other factors, apart from the exposure.

Many factors need to be considered when planning to conduct a meta-analysis of observational data since there are important biases and differences in study designs. A recent study by Mueller et al. (2018) looked at the methodological recommendations on how to conduct systematic reviews and meta-analysis of observational data and found that there were some conflicting recommendations such as that relating to use of scales and summary scores to assess quality of studies, pooling results of different study designs, assessment of publication bias and use of statistical measures of heterogeneity.

When assessing the conduct of a meta-analysis, regulators may refer to some of the following guidelines although it is acknowledged that meta-analyses and systematic review methodologies are less well developed for observational studies than they are for randomised controlled trials. The European Network of Centres for Pharmacoepidemiology and Pharmacovigilance (ENCePP), which is a network coordinated by the European Medicines Agency (EMA), publishes a guide on Methodological Standards in Pharmacoepidemiology which includes guidance on conducting systematic reviews and meta-analyses of completed comparative pharmacoepidemiological (observational) studies of safety outcomes and is supported by the EMA Guideline on Good Pharmacovigilance Practice of non-interventional post-authorisation safety studies. The Cochrane collaboration publishes a handbook for Systematic reviews of interventions which includes guidance on assessing risk of bias in non-randomised studies and synthesis of data from non-randomised studies (Higgins & Green, 2011). The Council for International Organisations of Medical Sciences (CIOMS) has published a guide on Evidence synthesis and meta-analysis for drug safety (CIOMS, 2016), which has a section of specific issues in meta-analysis of observational studies. All of these cover similar aspects to consider with respect to the conduct of meta-analyses of observational studies. Of the specific major points mentioned in the guidelines, Heneghan et al included a clear definition of the research question, inclusion and exclusion criteria, an adequate search strategy and adequate extraction of data (studies were reviewed by two researchers) however the exact implementation of the stated definitions and approaches is a little unclear in places as already discussed. The authors did not publish their protocol (as required by many journals although still not completely standard practice) and/or the statistical analysis plan which would have been useful

in helping to understand the exact methodologies used including what was prespecified in terms of the definitions for different malformations and the variables defined as the most important risk factors for example. In terms of study assessment tools, even though the NOS scale is recommended by the Cochrane collaboration it has been advised that items still may need to be customised to the review question of interest. The ENCePP guideline advises to steer away from 'summarising overall quality of a study through a single score which obscures the assessment of the individual study components.' Therefore, the applicability of NOS score to the issue under review, needs careful consideration. ENCePP guidance also highlights the need to consider the limitations of the individual studies and to consider pre-specifying sensitivity analyses to consider potential sources of bias.

Assessing heterogeneity

It is important to consider both statistical heterogeneity, in terms of the magnitude of the effect estimate, and clinical and methodological heterogeneity. As defined by Cochrane, variability in the participants, interventions and outcomes studied may be described as clinical diversity (sometimes called clinical heterogeneity), and variability in study design and risk of bias may be described as methodological diversity (sometimes called methodological heterogeneity). Clinical and methodological heterogeneity can lead to statistical heterogeneity. Appraisal of the similarity of studies is important in order to ascertain whether to include (or exclude) a certain study and the robustness of combining results into a meta-analysis. This is to determine whether there are genuine differences underlying the results of the studies, or whether the variation in findings is compatible with chance alone. Meta-analysis should only be considered when a group of studies is sufficiently homogeneous in terms of participants, interventions and outcomes to provide a meaningful summary. In this meta-analysis, the lack of detail provided in the individual publications may limit the ability to be confident that the studies are suitably homogeneous in these respects.

The Heneghan et al. publication, puts emphasis on the lack of statistical heterogeneity for many of the results. Statistical heterogeneity, which is measured using the I^2 statistic, assesses the variation in the study effect estimates but not in other important study characteristics. A low I^2 does not provide any information about whether the studies were conducted in similar populations or using similar methodology. Further, as with any metric, I^2 has some uncertainty but confidence intervals for the I^2 have not been provided here to assess this. Ioanndis et al (2007) showed that confidence intervals for the I^2 statistic were often wide and emphasised that putting too much trust in homogenity of effects may give a false sense of reassurance and that confidence intervals should be presented.

Results of the meta-analysis

The meta-analysis found a 40% increased risk of 'all congenital malformations' associated with exposure to oral HPTs in the first trimester of pregnancy, and a significant increase in risk for congenital heart malformations, nervous system malformations, muculoskeletal malformations and the VACTERL syndrome.The question is whether the issues highlighted above: the lack of clear reporting in many of the studies; uncertainty over the adequacy of controlling for confouding; questions over comparability of controls and unexposed cases; the dated statistical analysis techniques; and high within-study variance in many cases, make interpretation of these findings more challenging.

Interpreting the results of all malformations together (i.e. combining different congential anomalies) is not straightforward due to the different weeks of organogenesis of various organ development. The majority of studies did not stratify the data according to the possible week of fetal development that exposure took place, possibly because this was unknown.

The authors conducted post-hoc analysis for some anomalies in which the “studies that collected no confounding variables” were excluded, and they found that the risk estimate did not change substantially. However, the perceived limitations of the NOS for this particular issue have been highlighted.

Evidence Hierarchy

The Oxford CEBM levels of evidence of hierarchy puts systematic reviews of nested case-control studies on a par with systematic reviews of clinical trials. While this may be true in some instances and is likely relevant to recently conducted studies subjected to much higher standards of conduct and reporting, it may not necessarily apply when the quality of the individual case control studies is questionable or unclear as seen with most of the HPT studies, particularly as the CEBM hierarchy is referring to only nested case control studies, for which exposure and outcome ascertainment is likely to be good and sufficient data on confounding is both available and accounted for. Further, guidance documents on the hierarchy state that it is designed as a short cut for time-constrained researchers to find the likely best evidence. While such tools will often be as accurate as more complicated decision processes, they will not be definitive.

The classic hierarchy of evidence has been criticised for always placing randomised controlled trials higher than observational studies as there is a growing recognition that observational studies, particularly those showing a large effect or those showing a smaller effect size but with high standards of design and analysis which sufficiently account for potential bases, may provide stronger evidence of a causal association than poorly conducted non-significant randomised trials. Of note, the magnitude of the effect suggested by this meta-analysis is relatively small and the methods used in the individual studies are less advanced than studies conducted to modern standards.

The Grading of Recommendations Assessment, Development, and Evaluation (GRADE) working group have developed an alternative approach to grading the quality of evidence coming from systematic reviews. While the apriori ranking for observational studies is low this can be upgraded if there is a large consistent effect, evidence of a dose response, or sufficient adjustment for confounders that reduces but does not remove an effect.

7. Conclusion

The paper by Heneghan et al. Version 2 published on 29 January 2019 states that “the evidence of an association between exposure to HPTs and congenital malformations has previously been deemed weak” but “this systematic review shows an association of oral HPTs with congenital malformations” and their results “show the benefit of undertaking systematic reviews”.

The use of systematic reviews and advancing meta-analysis techniques can clearly be beneficial in supporting robust evidence synthesis and they are increasingly used

to support regulatory decision making, though could potentially be used more routinely.

The statistical meta-analysis regression methodology used by Heneghan et al. (2018) is robust and appropriately applied including subgroup analysis by study design and excluded studies judged to be at high risk of bias.

Nevertheless, the key consideration is whether a meta-analysis of studies conducted and reported according to the standards of the time of these, with important identified methodological considerations, is the most appropriate way of combining and evaluating the available evidence to assess if receiving a HPT increased a woman's risk of having a child with a major congential malformation, or whether it would not necessarily overcome each study's limitations.

8. Advice sought

On the basis of the evidence considered, the ad hoc Expert Group is asked to advise on:

- the suitability and robustness of the methodology, including the selection and application of the data quality score
- any clinical implications

References

Brogan, W., F. (1975). Correspondence: Cleft lip and palate and pregnancy tests. Medical Journal of Australia: 44

CIOMS (2016). Evidence synthesis and meta-analysis: Report of CIOMS Working Group X. Available at: https://cioms.ch/shop/product/evidence-synthesis-and-meta-analysis-report-of-cioms-working-group-x/ (Accessed on 25/02/2019)

Devereux, P.,j., Choi, P.,T.,L., El-Dika, S. et al (2004). An observational study found that authors of randomised controlled trials frequently use concealment of randomisation and blinding, despite the failure to report these methods. Journal of Clinical Epidemiology 57: 1232-1236. Available at: https://www.jclinepi.com/article/S0895-4356(04)00176-3/pdf (Accessed on 18/01/2019)

ENCePP Guide on Methodological Standards in Pharmacoepidemiology. Available at: http://www.encepp.eu/standards_and_guidances/methodologicalGuide.shtml (Accessed on: 07/02/2019)

FDA, Reviewer Guidance: Evaluating the risks of drug exposure in human pregnancies, C.f.D.E.a. Research, Editor. 2005: Rockville, USA.

Ferencz, C., G. M. Matanoski, P. D. Wilson, J. D. Rubin, C. A. Neill and R. Gutberlet (1980). Maternal hormone therapy and congenital heart disease. Teratology 21(2): 225-239.

Fleming, D.M. (1978). Abnormal Outcome of pregnancy after exposure to sex hormones. Personal communication (Provided by Bayer).

Gal, I., Kirman, B. and Stern, J. (1967). Hormonal pregnancy tests and congenital malformation. Nature (Lond.), 216.83.

Gal I (1972a) Risks and benefits of the use of hormonal pregnancy test tablets. Nature 240: 241-242.

Gal I (1972b) Hormonal imbalance in human reproduction. Advance in Teratology (volume five): 161-173

Gal I (1978) Teratological adverse drug effects: Review of evidence implicating hormonal pregnancy tests. Available at: https://mhra.filecamp.com/public/files/2ou7-p1dlcbo2#/public/file/2oux-r491isbm

Goujard, J. Rumeau-Rouquette, C., Saurel-Cubizolles, M.J. (1979). Hormonal tests of pregnancy and congenital malformations. Journal de gynecologie obstetrique et biologie de la reproduction 8: 489-196 (French)

Greenberg, G., Inman, W.H.W., Weatherall, J.A.C, Adelstein, A.M., Haskey, J.C. (1977). Maternal drug histories and congenital anomalies. BMJ 2: 853-856.

Guidelines for Good Pharmacoepidemiology Practices, Available at: https://www.pharmacoepi.org/resources/policies/guidelines-08027/ (Accessed on 07/02/2019)

Hadjigeorgiou, E., A. Malamitsi-Puchner, D. Lolis and P. Lazarides (1982). Cardiovascular birth defects and antenatal exposure to female sex hormones. Developmental pharmacology and therapeutics 5: 61-67.

Haller, J. (1974). Hormontherapie wahrend der graviditat. Deutsches Arzteblatt 14: 1013-1015. (German)

Hartling, L., Milne, A., Hamm, M., P. et al (2013). Testing the Newcastle Ottawa Scale showed low reliability between individual reviewers. J Clin Epidemiol 66(9): 982-993. Available at: https://www.sciencedirect.com/science/article/pii/S0895435613000899 (Accessed on 26/02/2019)

Hellstrom, B., Lindsten, J., Nilsson, K., Reid, I. (1976). Prenatal sex-hormone exposure and congenital limb reduction defects. Lancet 2: 372-373.

Heneghan, C., Aronson, J.K., Spencer, E., Holman, B., Mahtani, K.R., Perera, R., Onakpoya, I. (2018). Oral hormone pregnancy tests and the risks of congenital malformations: a systematic review and meta-analysis. F1000Research 7: 1725. Version 2, Available at: https://doi.org/10.12688/f1000research.16758.2 (Accessed on 06/02/2019)

Higgins JPT, Green S (editors). Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 [updated March 2011]. The Cochrane Collaboration, 2011. Available at: http://handbook.cochrane.org (Accessed on 25/02/2019)

Ioannidis, J., P., A., Patsopoulos, N., A., Evangelou, E. (2007). Uncertainty in heterogeneity estimates in meta-analyses. BMJ 335(7626): 914-916. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2048840/ (Accessed on 25/02/2019).

Janerich, D. T., J. M. Piper and D. M. Glebatis (1974). Oral contraceptives and congenital limb-reduction defects. N Engl J Med 291(14): 697-700.

Janerich, D. T., J. M. Dugan, S. J. Standfast and L. Strite (1977). Congenital heart disease and prenatal exposure to exogenous sex hormones. Br Med J 1(6068): 1058-1060.

Ka-Lok Lo, C., Mertz, D., Loeb, M (2014). Newcastle-Ottawa Scale: comparing reviewers' to authors' assessments. BMC Medical Research Methodology 14: 45.

Available at: https://bmcmedresmethodol.biomedcentral.com/track/pdf/10.1186/1471-2288-14-45 (Accessed on 18/01/2019).

Kullander, S. and B. Kallen (1976). A prospective study of drugs and pregnancy. Acta Onstet Gynecol Scand 55: 221-224.

Lammer, E. J., J. F. Cordero and M. J. Khoury (1986). Exogenous sex hormone exposure and the risk for VACTERL association. Teratology 34(2): 165-169.

Laurence M., Miller M., Evans K., Carter C. (1971). Hormonal pregnancy tests and neural tube malformations. Nature 233: 495-496.

Levy, E. P., A. Cohen and F. C. Fraser (1973). Hormone treatment during pregnancy and congenital heart defects. Lancet 1(7803): 611.

Lunny, C., Knopp-Sihota, J.A., Fraser, S.N. (2013). Surgery and risk for multiple sclerosis: a systematic review and meta-analysis of case-control studies. BMC Neurol 13: 41

Meire, F. and K. Vuylsteek (1978). Continued use of hormonal pregnancy tests. British Medical Journal 1:856.

Melsen, W.G., Bootsma, M.C.J., Rovers, M.M., Bonten, M.J.M (2014). The effect of clinical and statistical heterogeneity on the predictive values of results from meta-analysis. Clinical Microbiology and Infection 20(2): 123-129.

Michaelis, J., H. Michaelis, E. Gluck and S. Koller (1983). Prospective study of suspected associations between certain drugs administered during early pregnancy and congenital malformations. Teratology 27: 57-64.

Moore K.L., P.T.V., *Before we are born: essentials of embryology and birth defects*. Seventh ed. 2008: Elsevier Inc

Mueller, M., D'Addario, M., Egger, M. et al (2018). Methods to systematically review and meta-analyse observational studies: a systematic scoping review of recommendations. BMC Medical Research Methodology 18: 44. Available at: https://doi.org/10.1186/s12874-018-0495-9

Nora, A. H. and J. J. Nora (1975). A syndrome of multiple congenital anomalies associated with teratogenic exposure. Arch Environ Health 30(1): 17-21.

Nora, A.H. and J.J. Nora (1978). Maternal exposure to exogenous progestogen/estrogen as a potential cause of birth defects. Advances in planned parenthood 12: 156-159.

Oakley, G.P., Flynt, J.W. (1973). Hormonal pregnancy tests and congenital malformations. Lancet 2: 256-257.

OCEBM Levels of evidence. CEBM. 2016. Available at https://www.cebm.net/2016/05/ocebm-levels-of-evidence/ (Accessed on 07/02/2019)

Polednak, A. P. and D. T. Janerich (1983). Maternal characteristics and hypospadias: a case-control study. Teratology 28(1): 67-73.

Report of the Commission on Human Medicines (CHM) Expert Working Group on Hormone Pregnancy Tests. Available: https://www.gov.uk/government/publications/report-of-the-commission-on-human-medicines-expert-working-group-on-hormone-pregnancy-tests

Rothman, K. J., D. C. Fyler, A. Goldblatt and M. B. Kreidberg (1979). Exogenous hormones and other drug exposures of children with congenital heart disease. Am J Epidemiol 109(4): 433-439.

Roussel (1968). GP Survey – An investigation into the effects of oral pregnancy tests on the incidence of CNS malformations. Unpublished – obtained from Bayer.

Rumeau-Rouquette et al (1978). Malformations congenitales risques perinatales. L'institut National de la Sante et de la recherce medicale (INSERM) (French)

Sainz, M.P., Rodriguez Pinilla, E., Martinez-Frias, M., L. (1987). Progestogens and estrogens in high doses (hormone pregnancy tests): the risk of appearance of spina bifida with anencephaly. Medicina clinica 89: 272-274. (Spanish)

Sanderson, S., Tatt, I.D., Higgins, J., P.,T (2007). Tools for assessing quality and susceptibility to bias in observational studies in epidemiology: a systematic review and annotated bibliography. International Journal of Epidemiology 36(3): 666-676. Available at: https://academic.oup.com/ije/article/36/3/666/653571 (Accessed on: 15/01/2019).

Stang, A. (2010). Commentary: Critical evaluation of the Newcastle-Ottawa scale for the assessment of the quality of nonrandomized studies in meta-analyses. Eur J Epidemiol 25: 603-605.

Sterne, A.C., Hernan, M.A., Reeves B.C. et al (2016). ROBINS-I: tool for assessing risk of bias in non-randomised studies of interventions. BMJ 355: i4919. Available at: https://www.bmj.com/content/355/bmj.i4919 (Accessed on 07/02/2019)

Strengthening the reporting of observational studies in epidemiology (STROBE) Statement. Available at https://strobe-statement.org/index.php?id=strobe-home (Accessed on 07/02/2019)

Torfs, C.P., L. Milkovich and B.J. Van den Berg (1981). The relationship between hormonal pregnancy tests and congenital anomalies: A prospective study. American Journal of Epidemiology 113(5): 563-574.

Tummler, G., RiBmann, A., Meister, R., Schaefer, C. (2014). Congenital bladder exstrophy associated with Duogynon hormonal pregnancy tests – Signal for teratogenicity or consumer report bias? Reproductive toxicology 45: 14-19.

Wells, G.A., O'Connell, D., Peterson, J., Welch, V., Losos, M., Tugwell, P. The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomized studies in meta-analyses. Available at http://www.ohri.ca/programs/clinical_epidemiology/oxford.asp (Accessed on 30/01/2019)

WHO. Congenital anomalies. https://www.who.int/en/news-room/fact-sheets/detail/congenital-anomalies (Accessed on 04/02/2019)

Annexes

Annex 1: Heneghan et al (2018). Oral hormone pregnancy tests and the risks of congenital malformations: a systematic review and meta-analysis (version 2). F1000Research 7: 1725.

Annex 2: Characteristics of studies included in the Heneghan et al (2018) meta-analysis.