Health of children born to subfertile couples
Seggers, Jorien

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2015

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):
CHAPTER 7

IS OVARIAN HYPERSTIMULATION ASSOCIATED WITH HIGHER BLOOD PRESSURE IN 4-YEAR-OLD IVF OFFSPRING? PART 2: AN EXPLORATIVE CAUSAL INFERENCE APPROACH

S. La Bastide-Van Gemert
J. Seggers
M.L. Haadsma
M.J. Heineman
K.J. Middelburg
T.J. Roseboom
P. Schendelaar
E.R. Van den Heuvel
M. Hadders-Algra

Human Reproduction 2014; 29: 510-517
Chapter 7

Part 3 — The Groningen ART Cohort Study

**Abstract**

**Study question:** What causal relationships underlie the associations between ovarian hyperstimulation, the *in vitro* procedure, parental-, fertility- and child characteristics, blood pressure (BP) and anthropometrics of 4-year-old IVF children?

**Summary answer:** Causal models compatible with the data suggest the presence of positive direct effects of controlled ovarian hyperstimulation as applied in *in vitro* fertilization (COH-IVF) on systolic blood pressure (SBP) percentiles and subscapular skinfold thickness.

**What is known already:** Increasing evidence suggests that IVF is associated with higher blood pressure and altered body fat distribution in offspring, but underlying mechanisms describing the causal relationships between the variables are largely unknown.

**Study design, size, duration:** An assessor-blinded follow-up study in which 194 children were assessed. The attrition rate until the 4-year-old assessment was 10%.

**Participants/materials, setting, methods:** We measured blood pressure and anthropometrics of 4-year-old singletons born following COH-IVF (n=63), or born following modified natural cycle IVF (MNC-IVF, n=52), or born to subfertile couples who conceived naturally (Sub-NC, n=79). Primary outcome measures were the SBP and diastolic blood pressure (DBP) percentiles. Anthropometrics included triceps and subscapular skinfold thickness. Causal inference search algorithms and structural equation modeling were applied.

**Main results and the role of chance:** Explorative analyses suggest a direct effect of COH on SBP percentiles and on subscapular skinfold thickness. This hypothesis needs confirmation with additional, preferably larger, studies.

**Limitations, reasons for caution:** Search algorithms were used as explorative tools to generate hypotheses on the causal mechanisms underlying fertility treatment, blood pressure anthropometrics and other variables. More studies using larger groups are needed to draw firm conclusions.

**Wider implications of the findings:** Our findings are in line with other studies describing adverse effects of IVF on cardiometabolic outcome, but this is the first study suggesting a causal mechanism underlying this association. Perhaps ovarian hyperstimulation negatively influences cardiometabolic outcome via changes in the early environment of the oocyte and/or embryo, possibly resulting in epigenetic modifications of key metabolic systems that are involved in BP regulation. Future research needs to confirm the role of ovarian hyperstimulation in poorer cardiometabolic outcome and should investigate the underlying mechanisms. Our proposed causal models provide research hypotheses to be tested with new data from preferably larger studies.
**INTRODUCTION**

Recent reports have evoked concerns that the use of *in vitro* fertilization (IVF), with or without intracytoplasmic sperm injection (ICSI), is associated with birth defects and worse cardiometabolic outcome in offspring, such as higher blood pressure (BP), altered body fat distribution and vascular dysfunctions (for more details and references, see ‘Is ovarian hyperstimulation associated with higher blood pressure in 4-year-old IVF offspring? Part 1, multivariable regression analysis’) (Seggers et al. 2014). These associations could not be completely explained by known risk factors for hypertension and adiposity, such as preterm birth and low birthweight. Various other mechanisms—such as ovarian hyperstimulation affecting the early intra-uterine environment—may play a role, as well as the underlying subfertility. However, the exact causal mechanism underlying these factors remains largely unknown.

The aim of this prospective assessor-blinded study was to examine the effect of ovarian hyperstimulation, the *in vitro* procedure and a combination of these two on BP and anthropometrics of 4-year-olds. Our primary outcome measures were systolic BP (SBP) and diastolic BP (DBP) percentiles. The traditional statistical approach addressing this research question is the use of a series of multivariable linear regression analyses, in which the effect of fertility treatment (conventional IVF with controlled ovarian hyperstimulation (COH-IVF) or modified natural cycle IVF (MNC-IVF)) on outcome (BP or anthropometrics) is calculated while correcting for a number of subsets of (presumably confounding) background variables (Seggers et al. 2014). The results from these analyses showed a consistent adverse effect of COH-IVF on SBP percentiles compared to MNC-IVF, but the results shed no light on the causal mechanism underlying the variables, as in general, multivariable linear regression analysis is unable to distinguish confounding from intermediate effects and has limited ability to unravel causal mechanisms. Moreover, in the case where underlying causal mechanisms are unknown, multivariable regression analysis can easily lead to incorrect causal interpretations (Glymour et al. 2005).

Although rapidly gaining ground in epidemiology, methods for causal inference are still fairly uncommon in the epidemiology of developmental disabilities in children (Day and Reynolds 2013). This is unfortunate, as recent developments in statistical perspectives on causality provide possibilities for analysis of observational data from which this research field could benefit (La Bastide-Van Gemert and van den Heuvel 2013). In general, knowledge of the underlying causal mechanism is necessary to determine the correct set of confounding variables for which correction should take place in order to obtain the desired unconfounded effect estimates from conventional analyses (Greenland et al. 1999). Where still too little is known about the causal mechanism to postulate a specific causal hypothesis to be tested, a more explorative approach could be applied as a first step.
In this paper we explore possible causal mechanisms underlying fertility treatment, parental-, fertility- and child characteristics, blood pressure (BP) and anthropometrics of 4-year-old IVF children, using causal graphs and causal inference search algorithms. Search algorithms try to answer the question: which (classes of) causal graphs are consistent with the given data? Given the observational data, they produce a set of causal models that, under various model assumptions (see ‘Methods’), could underlie the observations and at the same time may exclude a group of causal models that are not likely to have generated the same data. Thus, these algorithms can be a valuable explorative tool for the development of causal hypotheses when little of the underlying mechanisms is known and where one has to rely on observational data alone (La Bastide-van Gemert et al. 2014).

**Methods**

**Recruitment and examination**

We measured BP and anthropometrics of 4-year-old singletons born following COH-IVF (n = 63), born following MNC-IVF (n = 52), or born to subfertile couples who conceived naturally (Sub-NC, n = 79). Primary outcome measures were the SBP and DBP percentiles. Anthropometrics included triceps and subscapular skinfold thickness. More elaborate details on recruitment and examination can be found in Part 1 of this study (Seggers et al. 2014).

**Causal inference search algorithms**

Causal inference search algorithms were performed to simultaneously investigate the effects of fertility treatment groups on BP and anthropometrics and to explore causal relationships with parental, fertility, and child variables (Pearl 2009a, Shipley 2000, Spirtes et al. 2000). They are explorative tools, based on the theory of causal graphs. This theory mathematically connects causal graphs and statistical models or distributions as described by Pearl and Spirtes et al. (Pearl 2009a, Spirtes et al. 2000). Causal graphs describe the causal relations between the variables in a causal model. Such a graph consists of vertices denoting variables, connected with edges which can be oriented. An oriented edge, denoted by an arrow, graphically represents a direct causal relationship (i.e. not mediated by any of the other variables in the model). A short introduction into the terminology and theory of causal diagrams is given in the Supplementary data. Easily accessible overviews on the subject can also be found elsewhere (Pearl 2009b, Spirtes 2010).
Various causal inference search algorithms exist. They vary in search approach, in complexity of the calculations, and in data assumptions made. Their result consists not of just one graph, but of a minimal sized class of potential graphs, showing what can be conclusively causally deduced regarding the causal model, based on the underlying data under the model assumptions. The graphs belonging to such a class are observationally equivalent, i.e. they cannot be statistically distinguished on the basis of the observational data alone, as they each generate the same pattern of correlation. An example illustrating how these search algorithms work is given in the Supplementary data.

As this is an explorative analysis, we applied several search algorithms and compared the results. More details of the particular algorithms (CPC, CFCI, GES) used are provided again in the Supplementary data. Adding specific background information and assumptions (from theory, previously established results or due to time constraints, i.e. due to the logical time ordering of variables) narrows down the various possible causal graphs found by the algorithms. We only added background knowledge to the algorithms based on the logical time-ordering of variables. This keeps the algorithms from exploring nonsensical graphs in which for example child’s height at age 4 has an effect on his or her birthweight.

The parental and fertility variables used in the causal models were: high level of education, smoking during pregnancy, maternal pre-pregnancy BMI, age at time of conception, time to pregnancy, and study groups (COH-IVF, MNC-IVF, Sub-NC). The child variables were: gender, gestational age, birthweight, age at examination, height, weight, BMI, waist circumference, subscapular skinfold thickness, triceps skinfold thickness, SBP percentile and DBP percentile. To disentangle the effect of ovarian hyperstimulation from that of the in vitro procedure, we defined two dummy variables as follows: 1) the variable ‘IVF’ denoting whether or not IVF was applied, 2) the variable ‘COH-IVF’ denoting whether or not controlled ovarian hyperstimulation was applied (in addition to IVF). Using these variables distinguishes the three groups COH-IVF (COH-IVF = IVF = 1), MNC-IVF (COH-IVF = 0, IVF = 1), Sub-NC (COH-IVF = IVF = 0), as described in Part 1 (Seggers et al. 2014)

Each search algorithm uses a particular threshold for announcing a certain effect between variables in the causal model ‘significant’ (i.e. alpha values for CPC and CFCI and penalty discounts for GES, see the Supplementary data). Varying these thresholds provides insight in the certainty of the found edges in the graphs and therefore in the robustness of the resulting models (Shipley 2000). For the GES algorithm penalty discounts ranging from 0.5 to 2 and for the CPC and CFCI algorithm alphas ranging from 0.05 to 0.25 were applied using the freeware program TETRAD, version 4.3.10-6 (The Tetrad Project).
The resulting graphs represent each an observationally equivalent class of causal models, which means that other graphs from that class are equally likely to have generated the data (see the Supplementary data). By choosing one directed acyclic graph representing such a class, model fit could be calculated and compared between the found classes of models. In acknowledgment of the ongoing debate on model fit indices for structural equations models in general and their relative worth when dealing with small, real life samples, we used the χ² test statistic, χ²/df as well as the Bayesian Information Criterion (BIC) to compare models (Bollen 1989, Bollen and Long 1993, Shipley 2000). The null hypothesis of the χ² test states that the population covariance matrix over the measured variables is equal to the covariance matrix of the model, with χ²/df < 2 indicating a good model fit. We additionally applied structural equation modeling to a graph representing one of the better fitted classes of causal models to estimate the size of the causal effects (Bollen 1989, Fox 2006). We used the sem library in R, version 2.15.0 (R Development Core Team).

Results

An overview of the description of background characteristics, outcome variables and the results of the conventional multivariable linear regression analysis approach can be found in Part 1 (Seggers et al. 2014).

By comparing the various graphs resulting from the performed causal inference search algorithms (GES and CPC) the most prominent and consistent direct effects could be distinguished. In general, an effect which is consistently not found by the various searches also constitutes a result from the analyses, as it can be considered an indication of no direct causal effect. Both types of results (presence or absence of specific effects) from our explorative analyses can then be interpreted as potentially new causal hypotheses worth further investigation.

In a large majority of the causal models found in our analyses, a positive, direct effect of COH-IVF on SBP was shown. Some of the models also showed a direct, non-mediated effect of COH-IVF treatment on subscapular skinfold thickness. All other effects of COH-IVF treatment on anthropometrics outcome variables in our models were intermediated by either gestational age or birthweight, and no other direct effect of COH-IVF treatment on outcome variables was found. We found the effect of COH-IVF on SBP percentiles in most graphs to be confounded by time to pregnancy (due to the existence of the path COH-IVF ← IVF ← Time to Pregnancy → SBP (Pearl 2009a). Application of the FCI-algorithm yielded no clear indication of the presence of latent variables (data not shown), although this cannot be ruled out.

Apart from these findings, the models also supported other already known mechanisms, such as the effect of parental education on smoking behaviour during
pregnancy, the effect of birthweight on child’s height at age 4, as well as various direct effects between closely related outcome variables such as a direct effect of subscapular skinfold thickness on triceps skinfold thickness, and a direct effect of SBP percentiles on DBP percentiles.

A graph with a good model fit representing most of these consistent explorative findings described above is provided in Figure 1. It represents the class of (observational equivalent) graphs found by the GES search with penalty discount 0.7. The specific class of models to which Figure 1 belongs to is not very large: in fact, it only contains two distinctive graphs, varying only in the direction of the arrow between maternal education and paternal education. In the graph, the numbers accompanying each arrow (reflecting the size of the causal effect) are the regression weights and their standard errors as estimated using structural equation modeling. The direct effect of COH-IVF is estimated at an increase of 12.69 points on the SBP percentile score. The combined effect of both IVF-groups is estimated at a decrease of 2.66 points on the SBP percentile score. This small beneficial effect of IVF on SBP is the result of the MNC-IVF group. This model also showed a direct, unmediated positive effect of ovarian hyperstimulation on subscapular skinfold thickness. Ovarian hyperstimulation had an indirect effect (via SBP percentiles) on DBP percentiles. The expected causal association between gestational age and birthweight on BP levels is represented in Figure 1 by the arrows pointing from COH-IVF to those variables, illustrating how COH-IVF is – as a classic confounder – closely related to both shorter gestational age and higher SBP percentiles. Other effects can be found in Figure 1.

Figure 1 represents just one of the classes of graphs found by the GES-algorithm. The dotted arrow-lines (i.e. the ones from smoking to subscapular skinfold thickness, from paternal age at conception to waist and from COH-IVF to subscapular skinfold thickness) indicate these direct effects to be the first to disappear when increasing the strictness with which edges in the graph are removed within the algorithm. Model fit values of the graph without these three arrows are slightly worse, but still reasonably good compared to the ones of the graph including these three edges. The disappearance of these three effects indicates less certainty regarding the existence of these three particular effects compared to the other effects depicted in the graph. When further increasing the penalty discount, both effects of COH-IVF and IVF on SBP disappear from the resulting graph, but simultaneously model fit decreases when compared to the ones described previously.

As mentioned before, the information provided by the resulting graphs is not only presented by the drawn arrows, but even more so by absent effects. Of course, such a finding cannot be interpreted as ‘hard evidence’, but should be seen as a possible suggestion of a direction for further research. With our 21 variables and taking into account the added time constraints, theoretically 158 edges could have been found by the algorithms, but none of our searches resulted in graphs with more than
Figure 1. A causal graph including the most consistent results of the causal effect search algorithms. The graph represents the equivalence class of causal models found as a result of the Greedy Equivalency Search (GES) search algorithm with penalty discount 0.7. The model fit of this graph was determined by Pearson’s chi-square statistics ($\chi^2 = 152.2$, df = 166, $p = 0.771$, $\chi^2$/df = 0.917), indicating a good model fit. Each arrow is accompanied by two numbers, the first represents the estimated regression weight and the second its standard error. These numbers reflect the size of the causal effect per unit change. A positive estimated regression weight in the graph means that higher values of the variable are associated with an increase in value of the variable to which the arrow is pointing. For example: on average, a higher maternal BMI is associated with higher weight of the child at age 4. A negative estimated regression weight in the graph means that higher values of the variable are associated with a decrease in value of the variable to which the arrow is pointing. For example: on average, a higher maternal age at conception is associated with a lower child’s BMI at age 4. Conditionally on the time to pregnancy and IVF, the effect of ovarian hyperstimulation (COH-IVF) on SBP is estimated at 12.694 (direct effect), indicating an increase in SBP percentiles. Conditionally on time to pregnancy, the total effect of IVF on SBP is estimated at $-2.657 (= -9.537 + 0.542 \times 12.694)$, indicating a decrease in SBP percentiles. Parental age and time to pregnancy calculated in years, age in months, gestational age in weeks, birthweight in grams, weight in kg, skinfold thicknesses and height in cm.
IVF and blood pressure — an explorative causal inference approach

40 edges. The absent 118 edges can be interpreted as consistencies between the different search results. We have drawn a selection of the most interesting of these absent edges in Figures 2a and 2b. Figures 2a and 2b show all direct effects from or towards Time to pregnancy and COH-IVF respectively, which were not found by any of the searches performed within our subfertile study population. Each of these edges in itself can be considered a plausible causal hypothesis to be tested with new data, such as: ‘there exist no direct causal effect of COH-IVF on Triceps skinfold thickness at age four’.

**Discussion**

Although the use of randomized experiments is the commonly most accepted method for inference on causality (Hill 1965), in many research settings this approach is often not feasible. Testing hypothesized causal relations using the observed data can provide a method to test the plausibility and consistency of causal models. In the last decade the use of causal graphs is more and more accepted and explored in the fields of both statistics and epidemiology (Greenland et al. 1999, Hernan et al. 2004, Van der Weele et al. 2008).

The results from the search algorithms indicate the presence of positive direct effects of COH-IVF on SBP percentiles and subscapular skinfold thickness. This is the first study using causal inference search algorithms combined with structural equation modeling in this context. This approach makes it possible to take into account and estimate the causal effects between all (background and outcome) variables in the model simultaneously. As such, this approach is able to detect possible confounding, to distinguish between the direct and indirect effects and to account for the influence of latent variables. Multivariable regression analysis cannot do that and its results are therefore prone to represent biased effect estimates and may possibly lead to incorrect conclusions regarding the observed associations (Greenland et al. 1999). Especially in this area, where the causal mechanism hiding behind COH-IVF treatment effects is not clear and outcome variables may be closely related in different ways, this approach may offer more insight and ways to disentangle the underlying causal relationships than the multivariable linear regression approach could provide.

Zooming in on the result from Figure 1, our data suggest that COH-IVF, i.e. IVF that includes ovarian hyperstimulation, is associated with higher SBP percentiles and thicker subscapular skinfolds in 4-year-old offspring. This is in line with other studies describing adverse effects of IVF on cardiometabolic outcome (Belva et al. 2007, Belva et al. 2012a, Ceelen et al. 2007, Ceelen et al. 2008b, Sakka et al. 2010, Scherrer et al. 2012). In addition, we found no adverse – and even slightly beneficial – effects of MNC-IVF, i.e. IVF without ovarian hyperstimulation, on SBP percen-
tiles. This suggests, for the first time, that ovarian hyperstimulation is involved in the poorer cardiometabolic outcome seen in IVF offspring. The suggested beneficial effect of the in vitro procedure on SBP percentiles was relatively small in comparison to the effect of COH-IVF (IVF: -2.7 percentiles; COH-IVF: +12.7 percentiles, see Figure 1). It is conceivable that the in vitro effect is spurious and may be related to the selection criteria for MNC-IVF. Moreover, an indirect adverse effect of COH-IVF on the triceps skinfold thickness (peripheral fat) as well as a direct adverse effect on subscapular skinfold thickness (truncal fat) was shown, underlining the findings from Part 1 that COH-IVF may be associated with the cardiometabolic syndrome (Seggers et al. 2014).

The use of search algorithms for causal inference has not been without criticism (Humphreys and Freedman 1996, Korb and Wallace 1997), mainly referring to the required data assumptions and their strictness (see the Supplementary data). Although a valid argument, the counter argument that similar data assumptions are being made for most tests and universally accepted methods in statistics could also be made. In allowing such limitations to prevent the application of statistical tests, very little statistics could be applied to real-world problems ever (Korb and Wallace 1997, Spirtes et al. 1997). As the method applied is an explorative one, results need to be interpreted with appropriate caution. Results should be interpreted as possible indications for new research hypothesis and do not necessarily render

Figure 2a. Pattern depicting all edges (direct effects) from or towards TTP which were not found by any of the searches performed.
IVF and blood pressure — an explorative causal inference approach

In the application of the here described methods, some issues need to be taken into account with regard to model assumptions and generalisability of the results. Within the field of machine learning, many simulation studies have been done regarding the application and performance of the methods here used (Ramsey 2010, Spirtes et al. 2000). Less is known about their performance with regard to the application on real-life data, although progress in this area is made in various research fields as more applications to real-life data have been performed (Acid et al. 2004, Chickering 2002, La Bastide-Van Gemert and van den Heuvel 2013, Mwebaze et al. 2010).

As with conventional regression analysis, all effects are assumed linear by the algorithms, an assumption which might not be strictly met by all variables. However, as Korb and Wallace pointed out, the fact that the true causal relationship is not linear does not necessarily mean that it cannot be detected using tests that assume linearity (Korb and Wallace 1997). Multivariate normality is another assumption which is not fully met by our variables, and extensions of the methods for arbitrary distributions are still developing (Hoyer et al. 2012). However, it can be shown that the algorithms tend to work well for uni-modal roughly symmetrical distributions.

Figure 2b. Pattern depicting all edges (direct effects) from or towards COH-IVF which were not found by any of the searches performed.
(Tetrad manual 2011), a criterion met by most of our (continuous) variables. For the limited number of included dichotomous variables, this was less of a problem as most of them could be considered exogenous and hence treated as continuous variables by the algorithm. This is similar to regression where binary explanatory variables can be treated as numerical variables in the analysis.

The issue of unmeasured confounding (i.e. the possible existence of latent variables disturbing the effects of interest) is also a concern not uncommon to more conventional statistical methods. In our case, unmeasured parental factors such as socioeconomic status could play such a role. CPC and GES are search algorithms which are not developed to specifically identify such latent variables, but in general, indications of their existence can be read off of their results (see the Supplementary data). We have run the CFCI algorithm as well, which is equipped to detect latent variables (Spirtes et al. 1995). However, no firm distinctive conclusions regarding latent variables could be drawn from the CFCI result. A likely explanation could be that the parental background variables included in our analysis (education levels, ages at conception) act as a proxy for other unmeasured socioeconomic variables, partly capturing their effects.

Although not all assumptions were met, the selected graph (Figure 1) has a good model fit. Due to our relatively small sample size and the somewhat larger amount of variables used, this could partly be an effect of over-fitting, a common worry when applying structural equation modeling techniques. However, our sample size was not too small to be able to detect smaller associations (Shipley 2000). This holds especially for the found effects (undirected edges between variables) in general. For the orientation of the found effects the error rate would be much larger for our sample size. However, due to our time constraints, the amount of possible mistakes made by the algorithms in orienting the effects remains small (Shipley 2000).

The fact that already known and expected mechanisms were detected by our models, further underlines the validity of the results. As explained, by varying the alpha values and penalty discounts in the algorithms, alternative possible graphs underlying the data can be calculated. The different models with the best model fit indices (see ‘Results’) all showed similar (and hence stable) mechanisms concerning the effect of COH-IVF on SBP, indicating a certain consistency of the found relations between these variables. Moreover, the resulting graphs did seem to consistently discriminate and eliminate a large number of theoretically plausible effects, again giving rise to confidence in the resulting causal hypothesis to be worth further investigating.

Despite the explorative character of the causal inference approach and the caution needed when interpreting the consequences, we do feel that the analyses described here can be a valuable tool in the development of causal hypotheses
in a field where little of the underlying mechanisms is yet known. Testing the validity of the here proposed causal model to eliminate possible spurious associations due to sampling error should be done using new data from multiple larger, carefully designed studies. Ideally, children of patients randomly assigned to COH-IVF or MNC-IVF, who underwent single embryo transfer (like in the INeS study), are followed (Bensdorp et al. 2009).

In our study population, increased time to pregnancy – which may be used as a proxy for the severity of subfertility – was associated with (slightly) lower rather than higher blood pressure levels, suggesting no adverse effect of subfertility in our data. Previously, we found that a longer time to pregnancy was associated with less trait anxiety and better mental health of parents, one year after childbirth (Jongbloed-Pereboom et al. 2012). This might be an effect of self-selection: couples that are able to deal with a long period of subfertility and subsequent IVF-treatments, presumably cope well with stress. Via ‘nature’ and ‘nurture’ these parental characteristics may be associated with lower BP levels in offspring.

As our leading research question focussed specifically on unravelling the causal effect of ovarian hyperstimulation on outcome variables, we did not explicitly include the effect of intracytoplasmic sperm injection (ICSI) as a separate variable in the model. Including ICSI in the causal model would have further split up the here found effects of MNC-IVF and COH-IVF on outcome in a direct and an indirect (mediated by ICSI) effect, but it would not have essentially altered the interpretation and conclusion concerning the derived causal effects and would unnecessarily complicate the interpretation of the model.

In conclusion, the results of the present study suggest that COH-IVF is associated with higher SBP percentiles and increased truncal fat in 4-year-old offspring. Future research needs to confirm the here hypothesized causal role of ovarian hyperstimulation in poorer cardiometabolic outcome and its generalisability, and should further investigate the underlying mechanisms, using our result as a research hypothesis to be tested with new data using causal inference and structural equation modeling. Our findings emphasize the importance of cardiometabolic monitoring of the growing number of children conceived with IVF worldwide.

**Acknowledgements**

The study was supported by the University Medical Center Groningen, the Cornelia Foundation and the school for Behavioral- and Cognitive Neurosciences. We thank the parents and children who participated in the study.
Supplementary data

Causal graphs: terminology and theory

Causal models consist of a statistical model and a causal graph. Causal graphs describe the causal relations between the variables in a causal model. Such a graph consists of vertices denoting variables, connected with edges which can be oriented. An oriented edge, denoted by an arrow, represents a direct causal relationship. An example of such a graph is given in Figure 1, representing hypothetical causal relations between variables *Level of education, Smoking during pregnancy, Maternal age at conception, Subscapular skinfold thickness and Triceps skinfold thickness*.

In this example, the (hypothetical) causal relationship *Level of education* is a direct cause of *Smoking during pregnancy* (i.e. relative to the other variables in the model) is expressed as *Level of education* → *Smoking during pregnancy*. The variable *Level of education* is then called the *parent*; the variable *Smoking during pregnancy* is called the *child*.

A graph is called a Directed Acyclic Graph (DAG) when all edges are directed and the graph contains no feedback-loops. This means that starting at any vertex in the graph and walking through the graph respecting the direction of the arrowheads, it is not possible to visit any vertex twice (i.e. no variable can cause itself). A sequence of adjacent edges from one variable to another (not necessarily following the direction of the arrows) is called a *path*. A path from variable X to variable Y is called *directed* (or causal) when, following the path from X to Y, one only encounters arrows pointing towards Y and the path consists of directed edges only. If such a path between two vertices exists, then they are causally dependent. In Figure 1, the path *Level of education* → *Smoking during pregnancy* → *Subscapular skinfold thickness* is an example of such a directed path.
An ancestor of a vertex \( X \) is any vertex \( Y \) such that there is a directed path from \( Y \) to \( X \). Descendants are defined in a similar way. In our example, Level of education is a parent of Smoking during pregnancy and an ancestor of both Subscapular skinfold thickness and Triceps skinfold thickness. The variable Triceps skinfold thickness is a descendant of Level of education (and a child of Subscapular skinfold thickness).

A collider vertex on a path is a vertex with arrows pointing to it from two directions on that path. The variable Subscapular skinfold thickness on the path Smoking during pregnancy \( \rightarrow \) Subscapular skinfold thickness \( \leftarrow \) Maternal age at conception is an example of a collider. Note that Subscapular skinfold thickness is not a collider relative to the path Smoking during pregnancy \( \rightarrow \) Subscapular skinfold thickness \( \rightarrow \) Triceps skinfold thickness. An unshielded collider is a vertex \( Y \) on a path \( X \rightarrow Y \leftarrow Z \), such that \( Y \) is a collider on that path and there is no edge between \( X \) and \( Z \) in the graph. The variable Subscapular skinfold thickness on the path Smoking during pregnancy \( \rightarrow \) Subscapular skinfold thickness \( \leftarrow \) Maternal age at conception is such an unshielded collider.

A path between two variables which contains a collider is called a blocked path. Otherwise, it is called unblocked. In Figure 1, the path Level of education \( \rightarrow \) Smoking during pregnancy \( \rightarrow \) Subscapular skinfold thickness \( \leftarrow \) Maternal age at conception is a blocked path, since Subscapular skinfold thickness is a collider on that path. We can block an unblocked path by conditioning on a variable on that path. Colliders play an essential role within the search algorithms.

**Principles of causal inference search algorithms**

Let us again consider our hypothetical example Figure 1 and assume this DAG to obey the Causal Markov Condition. This implies that conditioned on its parent(s), each variable in the graph is independent of its non-descendant(s), i.e. conditioned on (or corrected for) the variable Subscapular skinfold thickness (its parent), Triceps skinfold thickness does not depend on any of the variables Level of education, Smoking during pregnancy and Maternal age at conception.

Under the Causal Markov Condition, a given directed graph dictates conditional independence relationships in the joint probability distribution of observed variables. In deriving these consequences a property called directed separation (\( d \)-separation) is extremely useful. It provides a way to distinguish conditional independencies between two vertices in a causal model. We can determine whether two vertices are \( d \)-separated given a conditioning set of other variables by looking for unblocked paths between them. When every path between them is blocked given that certain set of variables we condition on, the vertices are said to be \( d \)-separated and the variables are conditionally independent given that set of variables. In Figure 1, Triceps skinfold thickness and Maternal age at conception are \( d \)-separated by the variable Sub-
**scapular skinfold thickness**, i.e. according to the causal relationships as depicted in the graph, when conditioned on *Subscapular skinfold thickness, Triceps skinfold thickness* and *Maternal age at conception* are no longer statistically associated.

Search algorithms try to answer the question: which (classes of) causal graphs are consistent with the given data? To illustrate how (constraint-based) search algorithms work, we again consider the hypothetical graph depicted in Figure 1. Let’s assume that this graph – which is unknown to us observers – represents the true causal mechanism underlying our data. All we observe are the conditional (in)dependencies in the data. Each d-separation relation which can be derived from a causal graph must also imply a corresponding statistical independency in the observed data, and vice versa (Pearl 2009a, Shipley 2000). This fact is used by the constraint-based search algorithm to derive the possible underlying causal graphs from the observed conditional (in)dependencies in the following way. The algorithm starts with what is called the complete, undirected graph (Figure 2).

Theory tells us that two variables in a graph remain connected when the association between them cannot be removed by conditioning on any subset of the other variables (Pearl 2009a). This can be verified using a conditional independence test appropriate for the type of data used. In our example this means that due to the true underlying mechanism as depicted in Figure 1, we will find *Triceps skinfold thickness* to be independent of *Level of education, Smoking during pregnancy* and *Maternal age at conception* when conditioned on *Subscapular skinfold thickness*. Hence, the algorithm concludes that the three corresponding edges in Figure 2 can be removed, resulting in the undirected graph in Figure 3a.

**Figure 2.** Complete, undirected graph with variables *Level of education, Smoking during pregnancy, Maternal age at conception, Subscapular skinfold thickness* and *Triceps skinfold thickness*. 
As Subscapular skinfold thickness is conditionally independent from Level of education when simultaneously conditioning on both variables Smoking during pregnancy and Maternal age at conception according to the true underlying mechanism from Figure 1, the algorithm also knows to remove the edge between Subscapular skinfold thickness and Level of education. A similar argument leads to the removal of the edge between Smoking during pregnancy and Maternal age at conception. All other edges in the graph remain, as those dependencies between the variables cannot be eliminated by conditioning on any subset of the other variables. These steps result in the so-called undirected dependency graph (Figure 3b), a graph already similar to the true underlying graph from Figure 1, but without any of the arrows.

The algorithm then proceeds with the orienting phase, in which the remaining edges are directed as far as possible, again based on the conditional dependencies observed in our data. Various mechanisms and algorithms are being used in this step. We illustrate only part of the orienting phase by describing the part which makes use of unshielded patterns, such as the one represented by the triplet Level of education - Maternal age at conception - Subscapular skinfold thickness. In this step, the algorithm examines whether the middle variable in the unshielded pattern, here Maternal age at conception, is a collider. D-separation implies that if Maternal age at conception were a collider in the underlying true causal graph, conditioning on it would introduce an association between its parents Level of education and Subscapular skinfold thickness. Moreover (and typical for the effect of conditioning on a collider), this association would be impossible to eliminate when we would condition on any set of variables including Maternal age at conception. However in our example, the association between Level of education and Subscapular skinfold thickness
disappears when we condition on *Smoking during pregnancy* (and of course on *Maternal age at conception*). Therefore, we conclude *Maternal age at conception* to be anything but a collider, which is called a definite non-collider.

A similar reasoning holds for the variables *Level of education* and *Smoking during pregnancy*, hence we would conclude both of them to be definite non-colliders. The variable *Subscapular skinfold thickness* in the unshielded pattern *Maternal age at conception — Subscapular skinfold thickness — Smoking during pregnancy* is detected to be collider: the association found between *Maternal age at conception* and *Smoking during pregnancy* when conditioning on *Subscapular skinfold thickness* remains, even when we condition on *Level of education, Triceps skinfold thickness* or both (in addition to *Subscapular skinfold thickness*).

Therefore, we add arrows to the edges pointing towards *Subscapular skinfold thickness*. This would conclude the search algorithm. What is conclusively found by the algorithm is depicted by the partially directed graph of Figure 4.

---

**Figure 4.** Partially directed graph with variables Level of education, Smoking during pregnancy, Maternal age at conception, Subscapular skinfold thickness and Triceps skinfold thickness, depicting the result of the causal inference algorithm.

The open circles at the end of the edges denote the fact that there could either be an arrow or an empty space. Definite non-colliders are denoted by the drawn bars over (or under) a variable and two open-circle ends. The definite non-collider *Level of education* as represented by the pattern:

*Smoking during pregnancy o—o Level of education o—o Maternal age at conception* from Figure 4 represents one of the three following situations:

*Smoking during pregnancy o→ Level of education —o Maternal age at conception*

*Smoking during pregnancy o–Level of education ←o Maternal age at conception*

*Smoking during pregnancy o—Level of education —o Maternal age at conception*

Note that the latter is indeed the one from the true underlying mechanism we were trying to detect. In a similar way the other open circle ends of the defi-
nite non-colliders could be replaced, as long as no new colliders are introduced. A double-headed arrow $\leftrightarrow$ denotes the possible existence of one of more unmeasured common ancestors (latent causes) between two variables. Thus the sequence Level of education $\leftrightarrow$ Maternal age at conception could imply the existence of an unmeasured confounding variable $U$ such that the following holds:

$$\text{Level of education} \leftarrow U \rightarrow \text{Maternal age at conception}$$

Note that in this way the graph in Figure 4 represents a class of graphs, as all causal graphs obeying the found associations as depicted would be equally likely to be the true graph (including the ‘true’ one from Figure 1). These graphs are observationally equivalent, i.e. they cannot be statistically distinguished on the basis of observational data, as they each generate the same observed pattern of correlation. This is why in general the result of causal inference algorithms is represented by a class of graphs, rather than one particular graph.

**The causal inference search algorithms used in this study**

We applied the Conservative Peter-Clark (CPC) algorithm, an algorithm proven to be more conservative and reliable in the orientation phase compared to the PC algorithm in the case where model assumptions are not completely met (Ramsey et al. 2006). This algorithm searches for acyclic graphs only and the variables are assumed to be continuous with a joint normal probability distribution. All direct causal influences are assumed linear. Moreover, it is assumed that probabilistic independencies among the variables are only those which can be derived from the causal graph under the Causal Markov Condition. Theory regarding the testability and possible weakening of these assumptions are still being developed (Zhang and Spirtes 2008).

In addition, we applied the Greedy Equivalency Search (GES) algorithm, which uses optimisation of a modified version of the Bayesian Information Criterion (BIC) of the causal models as its scoring metric, in which the penalty discount plays a role. The penalty discount represents the measure of strictness with which edges in the graph are removed within the algorithm’s procedure. The higher the values of the discount, the more certain we are about the edges that remain in the graph. To be more precise, the modified version of the BIC used is defined by $L-c^*K^*\ln(N)$, with $L$ denoting the maximized value of the likelihood function of the model, $K$ the number of parameters to be estimated, $N$ the number of variables in the model and $c$ the penalty discount. The GES algorithm runs under the same data assumptions as the CPC algorithm.

In both algorithms, latent variables between the observed variables are assumed non-existing: all common causes of the variables in the graph are themselves in the graph (causal sufficiency). The occurrence of saturated patterns or ‘spaghetti patterns’ in the resulting graphs (in which all vertices between three variables are
connected) could be an indication of the presence of latent variables which are a common cause of all three variables. To explore this possibility, we applied the Conservative Fast Causal Inference (CFCI) algorithm. Unlike the CPC algorithm, the FCI algorithm does allow for the existence of latent common causes (variables not observed in the data). Its result is a class of partially oriented graphs. It also assumes the variables to be continuous with a joint normal probability distribution and all direct causal influences to be linear.

Under the various model assumptions, the algorithms have been proven to be point wise consistent: they converge almost surely to the correct answer (Robins et al. 2003, The Tetrad Project). Varying the thresholds of announcing a particular effect between the variables in the causal model 'significant' provides insight in the certainty of the found edges in the graphs and therefore in the robustness of the resulting models (Shipley 2000). This is either done by varying the alpha (in CPC and CFCI) or the penalty discount (in GES).
Part 4

The PGS Follow-Up Study