Contemporary Clinical Trials Communications 16 (2019) 100426

Contents lists available at ScienceDirect

Contemporary Clinical Trials Communications

journal homepage: www.elsevier.com/locate/conctc

Review article

The crossover design for studies of infertility employing in-vitro fertilization: A methodological survey

Dalton R. Budhram[a,1,*], Daniel Shi[a,1], Sarah D. McDonald[a,b,c], Stephen D. Walter[a]

[a] Department of Health Research Methods, Evidence, and Impact, McMaster University Faculty of Health Sciences, Hamilton, ON, L8S 4K1, Canada
[b] Department of Obstetrics and Gynecology, McMaster University Faculty of Health Sciences, Hamilton, ON, L8S 4K1, Canada
[c] Department of Radiology, McMaster University Faculty of Health Sciences, Hamilton, ON, L8S 4K1, Canada

ARTICLE INFO

Keywords:
Methodology survey
Crossover design
In-vitro fertilization
Infertility
Missing data
Clinical trials

ABSTRACT

Background: Infertility has become increasingly common worldwide. There is a need for the infertility literature to evaluate new interventions with IVF. The crossover design presents many methodological advantages for IVF trials. In addition to providing a within-person comparison of outcomes, it offers participants the opportunity to potentially benefit from more than one available treatment. However, infertility studies present a unique challenge in terms of bias: successful participants do not cross over to the second treatment group.

Objectives: The main objective of our study was to survey the methodological features of crossover trials for infertility with in-vitro fertilization (IVF) based interventions. A secondary focus was reporting key results.

Study design & setting: We conducted a methodological survey by systematically searching Medline and Embase databases. The capture-recapture technique was used to estimate the number of relevant studies that were not retrieved by our search strategy. We employed the Cochrane risk of bias tool to assess methodological rigour. Crossover-specific methods features were summarized. Treatment effects for pregnancy outcomes across studies are also presented.

Results: 15 studies met inclusion criteria. Most studies were deemed to have high or unclear risks of bias, usually because of incomplete reporting of outcome data and assessment procedures. 13 studies did not employ crossover-specific methods to analyze outcome data by period, which may bias treatment effect estimates. Four studies reported pregnancy outcome data with sample sizes from both treatment periods. Of these four studies, three reported that the control intervention was favoured.

Conclusions: The main limitation of our survey was the small sample size of studies. Future reviews should be larger and seek to encompass a broader range of the infertility literature. Despite the issues identified in the included trials, consideration should still be given to using the crossover design in future infertility research. Employing crossover-specific analysis methods, such as accounting for participant non-completion, along with strict adherence to CONSORT reporting guidelines, may significantly reduce the risk of bias in individual studies.

1. Background

Infertility is common, especially with increasing numbers of women delaying their childbearing until later ages [1]. In a 2010 survey, 16% of couples reported not achieving pregnancy despite not using contraception for 12 months [1]. In-vitro fertilization (IVF) is the most widely used intervention for infertility, even when compared to intra-uterine insemination and ovulation induction [2]. By the end of 2013, five million IVF babies were born worldwide [3]. Accordingly, there is a need for infertility trials to evaluate new interventions, and the use of a crossover design presents many advantages. In standard crossover trials, participants are randomized to receive two or more alternative treatments in successive time periods [4]. The crossover design is attractive because it offers participants the opportunity to potentially benefit from more than one available treatment [5]. Additionally, the crossover approach provides a within-person comparison of outcomes, reducing the impact of between-person variation. This typically leads to a more precise estimate of treatment benefit [6]. However, while the crossover design presents these methodological advantages [7], it is more susceptible to problems than parallel group designs when missing data is present [8], especially if inappropriate analysis methods are utilized [9].

* Corresponding author. Department of Integrated Science, McMaster University, 1280 Main St W, L8S 4L8, Canada.
E-mail addresses: dbudhram@qmed.ca (D.R. Budhram), dshi@qmed.ca (D. Shi), mcdonals@mcmaster.ca (S.D. McDonald), walter@mcmaster.ca (S.D. Walter).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.conctc.2019.100426
Received 20 March 2019; Received in revised form 21 July 2019; Accepted 4 August 2019
Available online 21 August 2019

In infertility studies, participants leave the study once a successful outcome occurs; women who become pregnant after the first treatment do not cross over to the second treatment. The interpretation is then problematic, because the differential selection of participants who continue to the second and later time periods can bias the estimated treatment effect. Note that *period* here refers to duration of treatment in the crossover design, as opposed to the menstrual cycle [10]. The extent to which proposed analytic approaches limit this bias and improve precision of the estimated treatment effect are largely unknown [9,10].

Previous literature has suggested that the treatment effect may be overestimated in a crossover design if a naïve data analysis is employed. In particular, Khan et al. concluded that crossover designs for infertility interventions are inappropriate when pregnancy is the main outcome measure [11]. However, more recent literature has proposed that these crossover trials should be regarded as parallel group trials with additional information, as opposed to crossover trials with missing data. While the first treatment period is analogous to a parallel group trial, the second treatment period provides additional information that permits within-person comparisons for some of the participants. This approach provides a novel way to accommodate crossover trials in infertility, reducing the risk of overestimation of treatment effects [12]. Some other approaches to avoid overestimation due to participant non-completion in crossover trials in general have also been proposed: (1) re-randomization designs, (2) the logistic mixture model, (3) the beta-binomial mixture model, and the Mantel–Haenszel analysis method [7]. Takada et al. compared five study designs: (1) Two-period, two-treatment comparison; (2) crossover; (3) 1:1 re-randomization; (4) 2:1 re-randomization; and (5) 1:2 re-randomization and conducted simulations to identify the most appropriate design and analysis methods, and concluded that crossover designs have highest power and the smallest bias [7]. However, it remains unclear whether a crossover design is appropriate for infertility studies in particular. While the potential for use of the crossover design is high, no study to date has surveyed its application specifically in the evaluation of IVF interventions. It is useful to evaluate the methodological features and the types of data analysis currently being used,to determine if the crossover trial is being used effectively in infertility research. The present study's aim was to conduct a methodological survery to assess the rigour of the current literature, thereby informing the conduct and direction of future infertility trials. A secondary focus was to describe the key results of the current literature.

2. Methods

The present study surveyed the methodological features of crossover trials in infertility studies employing in-vitro fertilization (IVF) based interventions. The types of outcomes employed in these infertility trials was also of interest. A secondary focus was reporting the key results of included studies (the estimated effect sizes of the main outcomes).

Search strategies were developed with the aid of a research librarian to retrieve eligible studies on current infertility interventions that had employed a crossover design [see Additional file 1]. Search strategies did not limit results by outcome variable (i.e. pregnancy, live birth, etc.), as we aimed to maintain the generalizability of our review and it was unclear what the expected outcomes in the literature would be. Searches were conducted in Medline (1946–2017) and Embase (1974–2017) databases on April 4, 2017. Duplicate studies with multiple reports were only included once in the analysis. We used the capture-recapture technique, a method designed to inform researchers about when it would be appropriate to stop searching for more literature, by estimating the amount of relevant literature not retrieved. In particular, we used the numbers of papers independently identified by Medline and Embase to estimate how many relevant studies had not been found in either database [13].

A second search strategy [see Additional file 1] was developed *post hoc* to provide insight into the sensitivity of the first strategy. We searched Medline and Embase using the same infertility content and intervention terms as the first strategy, but the methods filters were instead designed to retrieve parallel group trials. Hence, the second strategy was useful in assessing the article retrival of the first strategy.

Studies retrieved in Medline and Embase databases were assessed on the basis of title and abstract for relevance independently by two reviewers (DB & DS). The full texts of the articles deemed to be relevant by title and abstract were then independently reviewed. In the event that title and abstract review was not conclusive, the full text of these articles was reviewed. Any discrepancies between the two reviewers were discussed and the relevant articles were then re-evaluated to determine if a consensus could be reached. In the event of persistent discrepancies, the conflict was resolved by a third reviewer (SDW or SDM). Our inclusion criteria for qualitative synthesis (risk of bias assessment and extraction of methodological features) were: (1) randomized crossover trial for the evaluation of infertility treatments involving IVF published 1994–2017, and (2) live births, positive pregnancy test, or other pregnancy related surrogates for positive outcome of the intervention including temporary pituitary suppression, serum concentrations of luteinizing hormone (LH), follicle stimulating hormone (FSH), testosterone, and progesterone (P), fertilization rate, embryo quality, implantation rates, ovarian responsiveness, and endothelin-1, uterine artery pulsatility index, endometrial thickness, morphological changes to sperm, endometrial histology, and expression of estrogen and progesterone receptors must be the primary outcomes in the study. Valid surrogate outcomes were included *post hoc* due to a wide range of primary outcomes reported in the literature sample. For simplicity, studies were limited to treatments involving IVF, to avoid complications associated with having multiple interventions.

The published manuscripts of selected studies were critically appraised and the following methods and features were extracted: (1) planned and actual sample sizes, (2) sample size calculation, (3) power calculation, (4) reporting of missing data pattern, (5) statistical analysis method, (6) effect estimates, (7) primary outcomes, (8) washout period, and (9) carryover effect. The Cochrane Collaboration's tool to assess risk of bias (RoB) and an extraction table (see Table 3) were used by two independent reviewers to identify and evaluate the design aspects and the methodological rigour of the included studies. As the Cochrane Collaboration's RoB tool addresses the main sources of bias in randomized trials that use the standard parallel group design, the present study has also included other important crossover design features (washout periods and carryover effects) in the other potential sources of bias section.

We also recorded treatment effects across the studies which reported pregnancy outcomes and sample sizes for treatment groups.

3. Results

Our search strategies in Medline and Embase yielded 37 and 87 studies respectively (Fig. 1). Of these 124 studies, 29 were duplicates (23.4%). After duplicate papers were removed, 95 were found to be eligible for title and abstract review and 19 of those were determined to be relevant (20.0%). Most of the excluded studies were removed on the basis of having irrelevant outcomes and methods, and because a wide variety of other (non-crossover) study designs was retrieved. Of the 19 studies, four (21.1%) were excluded during the full-text review because they reported outcomes that were unrelated to pregnancy or used interventions not involving IVF. Thus, 15 studies were finally included [see Additional file 2]. The capture-recapture technique was performed using the following values: duplicates in Medline and Embase (n = 4), found in Embase and not Medline (n = 8), found in Medline and not Embase (n = 3). The numbers of studies found in each database alone were multiplied and subsequently divided by the number of duplicates. This calculation revealed that an estimated six potentially eligible studies had not been retrieved by the search strategy in either database, assuming that errors of omission were independent between the two

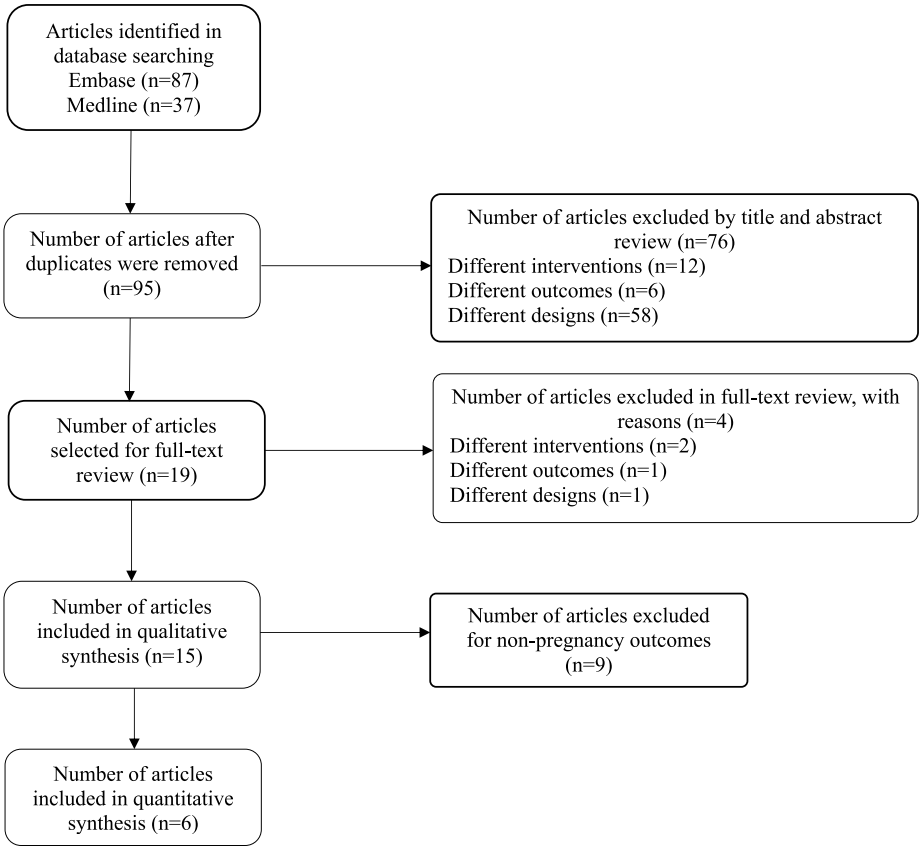

Fig. 1. Study flow diagram of included and excluded crossover trials with in-vitro fertilization.

databases.

Our primary search strategy yielded only a small number of crossover trials. In particular, no studies published between 2005 and 2010 met our inclusion criteria. Our second search strategy was used to determine if the reason for the sparseness in the data was due to methodological restrictions (specifically, limiting the search to crossover trials), or to a general lack of clinical trials (with any design) for infertility during that period. Samples of the relatively large numbers of studies found by the second strategy (Medline, n = 1258; Embase, n = 2601) were randomly drawn (Medline, n = 30; Embase, n = 30). These samples were assessed in accordance with our inclusion criteria. Only one crossover trial that had been identified by the second strategy, but not the first strategy, was found. This study, which was published in 1991 [14], does not suggest any published crossover literature employing IVF between 2005 and 2010. These findings suggest that only a small body of crossover literature was not retrieved by our primary search strategy, which may be a result of indexing studies with different terms. 12 of the 15 included studies were retrieved by the second search strategy in both databases. The overlap of literature (80%) is encouraging but again suggests that the terms used to index crossover trials may be variable.

The results of the risk of bias assessment that was applied to crossover trials are presented in Figs. 2 and 3, which are based on the reviewers' (DB and DS) judgment for each risk of bias item across all included studies.

3.1. Allocation

Six studies were deemed to have a low risk of bias for the method of assigning the sequence of treatment allocations, by having used computer-generated randomization sequences or a table of random numbers. One study was at a high risk of bias because treatment allocation was performed according to birth date [15].

All the studies deemed to have a low risk of bias for sequence generation were also deemed to have a low risk of bias for allocation concealment. The most common method used to conceal allocation in these studies was sealed, opaque envelopes.

3.2. Blinding

Most of the studies (9/15) did not report that participants or outcome assessors had been blinded to treatment, and were thus deemed to have an unclear risk of bias. Studies which reported an open-label trial design were judged to be at high risk. Studies that reported blinding of both participants and outcome assessors were judged to be at low risk. However, it is important to note that in many cases, it may not have been possible to blind the physicians or participants, depending on the nature of the intervention.

3.3. Incomplete outcome data

We judged 10 studies to have high risk of bias due to having incomplete outcome data. These studies have missing data for reasons other than successful outcomes in the first period (pregnancies), and these reasons were not reported in sufficient detail. Two studies [16,17] included only those patients who had completed both treatment periods in their analysis. More common reasons for missing data included (but were not limited to) participants' refusal to continue, and non-compliance with study protocol. Three studies were deemed to have a low risk of bias. In particular, one study used a partial crossover design, in which only the controls in the first period were crossed over in the second treatment period [18]; thus, this study did not have missing data. Another study reported no pregnancies in the first period and therefore had no missing outcome data [19]. In the final study, the only reason for missing data was when a participant achieved pregnancy as a successful outcome [20].

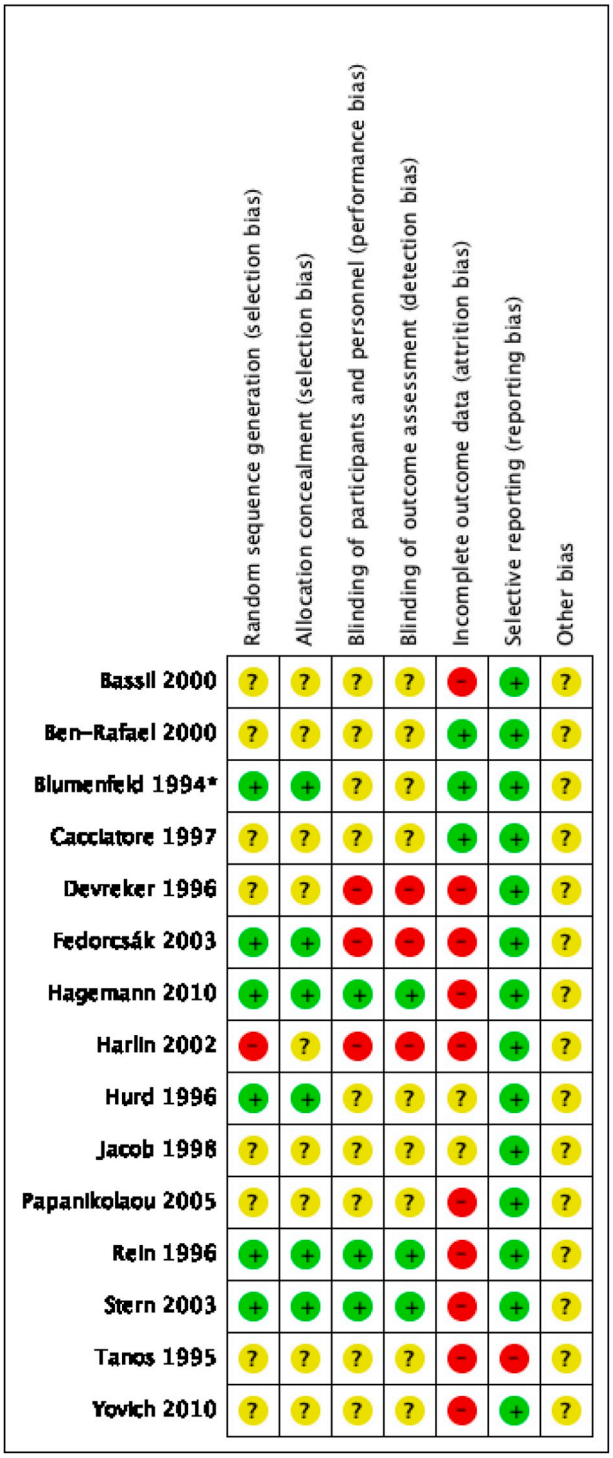

Fig. 2. Risk of bias summary: Review of authors' judgments for each included study.

3.4. Selective reporting

The reviewers judged all studies, with one exception, to be free of selective reporting, because all primary outcome data was reported. The one exception did not report data for patients who did not complete both treatment periods [17].

3.5. Other potential sources of bias

All 15 studies were judged to have an unclear other risk of bias. None of the studies reported their washout periods or attempted any assessment of a potential carryover effect in their analysis. Only one study [21] reported their funding source; however, it did not describe the role of their sponsor or report any possible conflicts of interest.

3.6. Other key methodological features

We surveyed the design features of crossover trials and identified additional elements that were not captured by the Cochrane Collaboration's assessment for risk of bias (see Tables 1–4).

The study settings were all high-income countries including USA (n = 3), Finland, Norway, Belgium (n = 2), Brussels, Ireland, Luxembourg, Sweden, Israel (n = 2), New Zealand, and Australia, and generally the IVF treatments were performed for female infertility of varying duration and cause (Table 1). Seven studies did not report sample sizes by intervention group (Table 2). This incomplete reporting proved to be problematic in determining effect sizes. Thus, these studies could not be included in Fig. 4. Two additional studies were not included in Fig. 4 due to pregnancy outcomes not being reported in the full-texts. Another finding was that nearly all studies (n = 13) had relatively small sample sizes (< 200 participants) in the first period (Table 2). There was a considerable diversity of interventions between studies, but all pertain to IVF treatment (Table 3). Seven studies reported a significant difference ($p < 0.05$) for their primary outcomes, while five studies reported a non-significant difference. In these five studies, only one reported a p-value. A majority of studies that analyzed pregnancy outcomes and reported sample sizes by group for both periods (3/4) found that the control intervention was favoured (Fig. 4).

Table 5 summarizes key analysis features. Nine studies employ either precision estimates (n = 7) including standard deviation of the mean (SD) and standard error of the mean (SEM), or statistical estimation methods (n = 2) including confidence intervals (CI) and Bayesian analysis for their primary outcome. Reports from only four studies included a power calculation, but only three of these studies also included a sample size calculation. Investigators for seven studies did not report sample sizes by intervention group for each period. In some cases, data were also not reported by period. Many authors of study reports neglected the possibility of period and carryover effects by only reporting results aggregated over all periods. In reports from 13 studies, data were pooled across periods (results from treatment groups aggregated across periods to generate a total number of pregnancies). However, in reports from three of these studies, data were not reported separately for each period. Only two studies mentioned the possibility of a carryover effect [16,20], and one study reported their washout period [20]. These shortcomings, compounded with small to moderate sample sizes in 13 studies, the absence of sample size and power calculations in 11 studies, and the overall high risk of bias found in 10 studies, make it difficult to obtain a valid estimate of the treatment effect in these studies.

4. Discussion

We found that several studies had unclear risks of bias in many domains. Incomplete outcome data and incomplete reporting of key methodological features were particularly prevalent. Studies (n = 10) were judged to have a high risk of bias due to incomplete outcome data, which increases the possibility of a biased effect estimate. The insufficient reporting made it difficult to accurately assess the risk of bias of these studies and prevents readers from doing a fully valid data analysis. Particularly, sample sizes by group were reported infrequently. Most trials (n = 13) had small sample sizes in their first treatment periods (< 200 participants) and all trials had small sample sizes in their second treatment periods. Furthermore, most trials (n = 11) did not report any power calculations. The results presented in these studies may therefore be limited in their ability to declare statistical signficance and they are insufficient to fully inform clinical

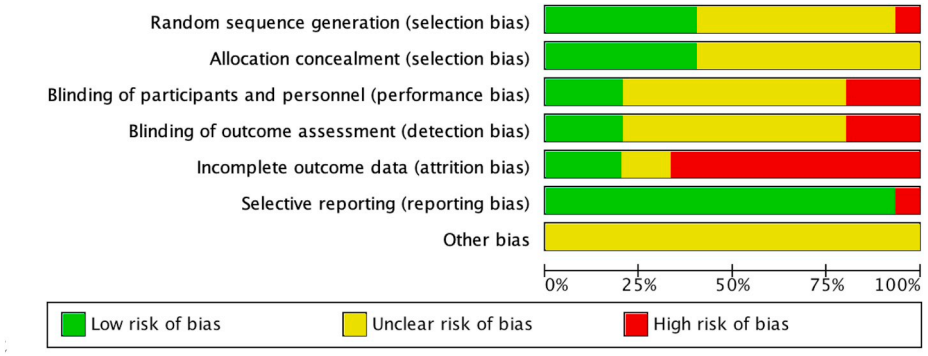

Fig. 3. Risk of bias graph: Review of authors' judgments across all included studies.

Table 1
Summary of patient population, setting, and duration.

Paper	Population	Setting	Study Duration
Blumenfeld, 1994	Women (mean age 32.5 years for clonidine negative patients, mean age clonidine positive patients not reported) with varied diagnoses of long-standing infertility of 2–16 years (underwent between three and 50 previous cycles of ovulation induction with HMG/HCG)	Medical Centre, Israel	NR
Tanos, 1995	Women with infertility (mean age 32.1 years), bilateral obstructed tubes and normal ovarian function; did not receive infertility treatment for 3 months prior to study	IVF clinic, Ein-Kerem, Jerusalem, Israel	October 1, 1993 to March 30, 1994
Devreker, 1996	Women with infertility, on their first IVF attempt (mean age 32.8 years; mean duration of infertility 4.8 years).	IVF unit at an academic hospital, Brussels, Belgium	NR
Rein, 1996	Women (mean age 34 years), varied diagnoses of infertility.	Tertiary care centre – Brigham and Women's hospital, Boston, USA	1991 to 1993
Hurd, 1996	Women, various diagnoses of infertility – intervention 1 group (no support): 35 years, mean age – intervention 2 group (luteal support): 33 years, mean age	The University of Michigan Medical Center, USA	October 1992 to September 1994
Cacciatore, 1997	Women with infertility (mean age 31.4 years), 16 had primary infertility, 2 had secondary infertility	Academic research centre in Helsinki, Finland	NR
Jacob, 1998	Women (mean age of intervention 1 - FSHr: 35, mean age of intervention 2 - HMG: 34.48), various diagnoses of infertility	Human Assisted Reproduction Unit, Dublin, Ireland	September 1996 and mid-February 1997
Bassil, 2000	Women with their first IVF attempt, mean age 37.6 years	Hospital Centre, Luxembourg	NR
Ben-Rafael, 2000	Normogonadotropic, normogonadal men with oligoteratoasthenozoospermia and at least one previous IVF attempt in which fertilization failed or the fertilization rate was $< 30\%$, primary or secondary infertility for at least one year, age range 18–55 years	IVF Unit, Golda Campus, Rabin Medical Center, Petah Tikva, Israel	Study initiated before 1993
Harlin, 2002	Women (mean age of intervention 1 is 34 years, mean age of intervention 2 is 33.2 years), various diagnoses of infertility	Clinic, Stockholm, Sweden	February 1997 to September 1999
Stern, 2003	Women with IVF implantation failure (mean age 35.2 years), various diagnoses of infertility	A hospital infertility clinic and associated IVF service, Australia and New Zealand	January 1998 to June 2001
Fedorcsák, 2003	Insulin-resistant women with infertility, polycystic ovary syndrome (mean age 30–31 years)	IVF unit in Oslo, Norway	April 2000 to April 2001
Papanikolaou, 2005	Patients with male or tubal (or a combination of both) infertility and primary or secondary infertility (mean age 30.7 years)	Centre for Reproductive Medicine at of the Dutch-speaking Brussels Free University, Belgium	April 2003 to March 2004
Hagemann, 2010	Women with zona pellucida thickness $\geq 13\,\mu m$ for any embryos (< 38 years)	Washington University Infertility Center, USA	April 2004 to February 2007
Yovich, 2010	Poor prognosis for pregnancy in women (defined by past failure to conceive and poor quality embryos), mean age 37.5 years	PIVET Medical Centre, Australia	January 2002 to December 2006

IVF: In-Vitro Fertilization.
NR: Not Reported in full-text paper.
HMG/HCG: Human Menopausal Gonadotrophin/Human Chorionic Gonadotrophin.
FSHr: Follitrophin beta.

decision making.

Nine studies reported pregnancy outcomes. Of these nine, we summarized the pregnancy rate differences of four in Fig. 4, as they reported necessary data on pregnancy outcomes (pregnancy rates and sample sizes for both treatment groups across periods). Out of the four summarized studies, three reported that the control intervention was favoured. Various statistical methods were employed in most studies ($n = 9$) to estimate treatment effects of primary outcomes, with SEMs being the most common. The use of surrogate endpoints was relatively common, presumably because they are easier to measure than live births, and the implied shorter follow-up times reduce losses follow-up.

Based on these findings, we recommend that future crossover trials for infertility adhere to a standard for reporting guidelines for clinical trials (CONSORT) for randomized control trials, as applicable [22]. As suggested in a systematic review of chronic painful conditions, incomplete reporting of study features has been identified as a key problem in crossover trials outside of the infertility literature [23]. Furthermore, studies examining the crossover design substantiate that poor reporting of analysis features and results make it difficult to include these studies in meta-analyses [5,24]. Our review is consistent with these findings, and we suggest that the reporting of washout periods, carryover effects, and outcome data by group and period in particular should be improved [25]. Although there are no methods to reasonably deal with carryover effects, we believe that the authors should report the washout period or at least attempt to assess the impact of carryover effect on their results. A major issue with not accounting for the period

Table 2
Summary of sample sizes by period & intervention group.

Paper	Sample Size Int. 1 Period 1	Sample Size Int. 2 Period 1	Total Sample Size Period 1	Sample Size Int. 1 Period 2	Sample Size Int. 2 Period 2	Total Sample Size Period 2
Blumenfeld, 1994[c]	14	16	32[a]	15	9	24
Tanos, 1995	20	20	51[b]	20	20	40
Devreker, 1996	NR	NR	100	NR	NR	33
Rein, 1996[d]	9	9	18	8	8	16
Hurd, 1996[e]	NR	NR	93	NR	NR	24
Cacciatore, 1997[f]	NR	NR	18	NR	NR	18
Jacob, 1998	91	113	204	67	78	145
Bassil, 2000	NR	NR	27	NR	NR	20
Ben-Rafael, 2000	20	20	40	10	10	20
Harlin, 2002	266	170	436	NR	NR	40
Stern, 2003	74	69	143	45	38	83
Fedorcsák, 2003	9	8	17	5	4	9
Papanikolaou, 2005	NR	NR	12	NR	NR	11
Hagemann, 2010	49	54	103	10	8	18
Yovich, 2010	NR	NR	159	NR	NR	NR

NR: Not Reported in full-text paper.
Int. 1: Treatment Intervention.
Int. 2: Comparison Intervention.
[a] 32 patients were present in period 1. Some of these patients were reported to receive both interventions, and thus not included in either treatment group in this table.
[b] Participants were excluded if both cycles were not completed. Information regarding their treatment allocation was not reported.
[c] It was assumed that pregnancy was the only reason for dropout because other reasons were not reported.
[d] Studies which analyzed sample size only in terms of cycles, leading to sample size by cycle being reported in this table.
[e] Analysis was performed by cycle, but not enough information was reported to record sample size by cycle.
[f] The paper reported that women underwent 36 cycles in the study. It was assumed that cycles were evenly distributed across periods and that each woman underwent one cycle.

Table 3
Summary of interventions and main outcomes.

Paper	Interventions	Main Outcome Domain
Blumenfeld, 1994	IVF or *in vivo* fertilization with either GH co-treatment or HMG/HCG	Pregnancy
Tanos, 1995	IVF with either nafarelin or D-Trp6-LHRH[a]	Temporary Pituitary Suppression
Devreker, 1996	IVF with either long-acting or short-acting GnRHa	Serum concentrations of LH, E_2, and P, fertilization rate, embryo quality, implantation rates, pregnancy
Rein, 1996	IVF with either DEX or placebo	Ovarian responsiveness, implantation rates, and clinical pregnancy or live births
Hurd, 1996	IVF with either luteal support or no luteal support with both oral E_2 and vaginal P suppositories	Clinical Pregnancy
Cacciatore, 1997	IVF with either a spontaneous cycle or a gonadotropin stimulated cycle	Plasma levels of E_2, P, and endothelin-1; uterine artery pulsatility index; endometrial thickness
Jacob, 1998	IVF/intracytoplasmic sperm injections with either FSHr or hMG[a]	Fertilization
Bassil, 2000	IVF and GnRHa with either highly purified FSH or hMG during ovarian stimulation	Results of stimulation parameters and embryo quality
Ben-Rafael, 2000	75 IU of FSH or 150 IU of FSH before IVF treatment and IVF without treatment	LH, FSH, testosterone levels, morphologic changes in sperm, fertilization rates
Harlin, 2002	IVF with either Gonal-F or Puregon	Pregnancy and Delivery
Stern, 2003	IVF implantation failure patients with either subcutaneous unfractionated heparin and aspirin or placebo	Implantation
Fedorcsák, 2003	IVF either with metformin or without metformin	FSH dose and number of collected oocytes
Papanikolaou, 2005	IVF with either GnRH antagonist and FSHr ovarian stimulation or natural cycles	Endometrial histology, expression of estrogen and progesterone receptors
Hagemann, 2010	IVF with either assisted hatching or unhatched	Clinical Pregnancy
Yovich, 2010	IVF with either GH or no GH	Clinical Pregnancy

D-Trp6-LHRH: Decapeptyl.
DEX: Dexamethasone.
FSH: Follicle Simulating Hormone.
FSHr: Follitrophin beta.
GH: Growth Hormone.
GnRH: Gonadotrophin Releasing Hormone.
GnRHa: Gonadotrophin Releasing Hormone analogues.
HMG/HCG: Human Menopausal Gonadotrophin/Human Chorionic Gonadotrophin.
IVF: In-vitro fertilization.
LH: Luteinizing Hormone.
P: Progesterone.
[a] The study did not administer IVF, but rather the study population consisted of patients undergoing IVF.

Table 4
Summary of effect sizes.

Paper	Effect Sizes of Study Outcomes	P-Value
Blumenfeld, 1994	Pregnancy Rate for Clonidine Negative Patient Group:[c] GH co-treatment: 58.3% HMG/HCG: 0% Effect Size: 58.3%	NR
	Pregnancy Rate for Clonidine Positive Patient Group:[c] GH co-treatment: 0% HMG/HCG: 25% Effect Size: 25%	NR
Tanos, 1995	Temporary Pituitary Suppression (Main Study Outcome): Effect Size: 3.6 ampules per cycle - In favor of nafarelin (intervention 1)[b]	P = 0.0005
	Fertilization Rate (Clinically Relevant Outcome):[d] Effect Size: 16.2% - In favor of D-Trp6-LHRH (intervention 2)[b]	P = 0.001
Devreker, 1996	Pregnancy Rate:[c] Short-acting GnRH-a: 59.2% Long-acting GnRH-a: 39.6% Effect Size: 19.6%	P < 0.05
Rein, 1996	Clinical Pregnancy Rate:[d] Dexamethasone: 21% Placebo-controlled: 35% Effect Size: 14%	Not significant[a]
Hurd, 1996	Pregnancy Rate:[d] Control: 2% Luteal support: 16% Effect Size: 14%	P < 0.04
Cacciatore, 1997	E_2 Levels per cycle: Stimulated Cycles: 723.5 pg/ml Unstimulated Cycles: 101.0 pg/ml Effect Size: 622.5 pg/ml	P < 0.001
Jacob, 1998	Fertilization Rate (Main Study Outcome):[d] FSHr: 51.7% HMG: 53.4% Effect Size: 1.7%	Not significant[a]
	Clinical Pregnancy (Clinically Relevant Outcome):[d] FSHr: 14% HMG: 20% Effect Size: 6%	Not significant[a]
Bassil, 2000	Mean number of oocytes collected per cycle: FSH: 10.3 HMG: 7.3 Effect Size: 3	P = 0.02
	Clinical Pregnancy Rate (Clinically Relevant Outcome):[d] FSH: 33.3% HMG: 18% Effect Size: 15.3%	NR
Ben-Rafael, 2000	Fertilization Rates:[d] 75IU FSH: 19.7% 150IU FSH: 20.5% Control: 5.8%	
	Effect Size (75IU FSH - Control): 14.7%	P < 0.05
	Effect Size (150IU FSH - Control): 15.5%	P < 0.05
	Effect Size (75IU FSH − 150IU FSH): 0.8%	Not Significant[a]
Harlin, 2002	Pregnancy Rate:[d] Gonal-F: 26% Puregon: 28% Effect Size (P vs. G): 2%	Not Significant[a]
Stern, 2003	Implantation Rate:[f] Subcutaneous unfractionated heparin and aspirin: 6.8% Placebo: 8.5% Effect Size: 1.7%	Not Significant[a]
Fedorcsák, 2003	Number of collected oocytes per woman: Metformin: 8.6 Without Metformin: 4.6 Effect Size: 4	Probability of 0.61 that at least 10% more oocytes are collected using metformin.[g]
Papanikolaou, 2005	Endometrial thickness per cycle: GnRH agonist and FSHr ovarian stimulation: 8.9 mm Natural cycles: 8.2 mm Effect Size: 0.7 mm	Not Significant[a]
Hagemann, 2010	Clinical Pregnancy:[c] Assisted Hatching: 47% Unassisted Hatching: 50% Effect Size: 3%	Not Significant (P = 0.86)
Yovich, 2010	Clinical Pregnancy:[d] GH: 20% GHu: 32% No GH: 9%	
	Effect Size (GH - no GH): 11%	P < 0.05
	Effect Size (GH - GHu): 12%	P < 0.05
	Effect Size (GHu - no GH): 23%	P < 0.001

D-Trp6-LHRH: Decapeptyl.
FSH: Follicle Stimulating Hormone.
FSHr: Follitrophin beta.
GH: Growth Hormone.
GHu: Uncontrolled Growth Hormone.
GnRH: Gonadotrophin Releasing Hormone.
GnRH-a: Gonadotrophin Releasing Hormone analogues.
HMG/HCG: Human Menopausal Gonadotrophin/Human Chorionic Gonadotrophin.
PR: Pregnancy Rate.
[a] Studies which we report as not significant without p-values did not report p-values in text.
[b] Study did not report group specific outcome rates. However, difference between groups was reported.
[c] Rate was calculated by # of events/total number of women in treatment group.
[d] Rate was calculated by # of events/total number of cycles in treatment group.
[e] Rate was calculated by # of events/total number of transfers in treatment group.
[f] Rate was calculated by # of events/total number of embryos in treatment group.
[g] No p-value given as Bayesian statistics were used.

effect (where the success rates are different for the first versus the second treatment) is the possibility of a continuing additive difference in the effect of the first treatment into the second period, affecting women who have not achieved pregnancy with the first treatment and who therefore may, on average, have a lower overall chance of success.

It is also important to consider the analysis methods used to estimate treatment effect. Most studies (n = 13) aggregated outcome data across periods, but this approach fails to account for any potential period effects, and it potentially biases estimates of treatment effects, because of the differential selection of participants who have unsuccessful outcomes in the first period. Simpson's paradox indicates that the estimate from data aggregated in this manner may not even be in

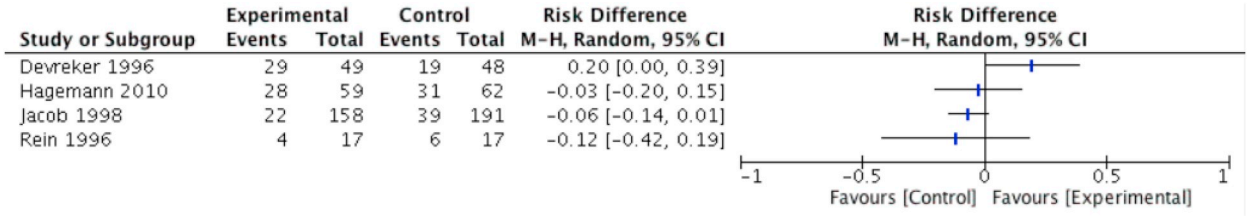

Fig. 4. Forest plot of included studies with pregnancy rate outcomes.
Legend: Studies are arranged by effect size (Control favoured to Experimental favoured).Events: Number of Pregnancies.Risk Difference: Difference in Pregnancy Rates Between Experimental and Control Group.Total: Sample size for corresponding group across treatment periods.Note – Bassil 2000, Hurd 1996, Harlin 2002, and Yovich 2010 reported pregnancy rates but not the sample sizes for both intervention groups across treatment periods and hence were not included in this analysis. Blumenfeld 1994 included participants from two different populations and hence were also not included in this figure.

the same direction as the period-specific results [26].

While we aimed to include studies with comparable outcomes, the infertility literature proved to be heterogenous. This is especially problematic for reviewers attempting a quantitative synthesis, which poses an issue for clinicians and other experts who wish to implement this research into practice. While some studies reported on clinical outcomes (particularly, pregnancies and live births), other studies reported on biochemical and histological outcomes (such as hormone serum concentrations, endometrial thickness, receptor expression, implantation rate, etc.). This heterogeneity could be due to various factors, including the limited consensus of reproductive endrocrinologist (REI) physicians on a singular standard outcome, the transfer of care of patients from REI physicians to obsetricians, and the willingness of journals to publish on a wide range of outcomes.

Previous studies have identified issues in using crossover methodology to assess infertility interventions and have proposed novel design and analytic methods. However, the current study is the first survey of infertility studies employing the crossover design, with a focus on reported results and a rigorous assessment of their methodological features. A limitation of the current study is the small sample of eligible crossover trials retrieved by our primary search strategy. In particular, crossover trials are less common than other clinical trial designs. Limiting the interventions to IVF-related treatments and selecting only pregnancy related surrogate outcomes further restricted our sample size. However, these restrictions were necessary to reduce heterogeneity, and thus bias. Our capture-recapture analysis revealed that a moderate proportion of relevant literature may not have been retrieved. The limited number of available databases searched by our primary strategy may be at least in part responsible. In general, methodological filters have poor sensitivity, and hence may have been a major contributor to missing these six potential studies [27]. Future systematic reviews should be conducted on the topic and search multiple databases.

Although the crossover approach is considered one of most rigorous designs because of its potential to make within-person comparisons [6], in the context of infertility trials, it may not yield high quality evidence because of the frequent occurrence of participant non-completion [28]. In the sample of trials that we assessed, none took into account the effects of missing data for reasons other than pregnancy. Methods that have been proposed to deal with missing data include sensitivity analysis, regression imputation, and multiple imputation [9,28]. While our sample of eligible literature was small, the use of the crossover design is nevertheless useful in infertility. Our second search strategy, which surveyed the frequency of the crossover design in infertility literature in relation to other clinical trial designs, identified 25 clinical trials with IVF in our random sample of 60 studies, and only one of these trials (4%) were of crossover design. Although the most recent paper that this review identified in the IVF literature was published in 2010, the crossover design has recently been used in the general infertility literature employing treatments other than IVF [29,30]. The issues that we have identified with binary outcomes and participant non-completion remain applicable to these other areas of infertility research.

Table 5
Summary of additional features important to crossover design.

Paper	Statistical Estimate of Treatment Effect (Y/N)[b]	Data Analysis by Period (Y/N)	Power Calc. Reported (Y/N)	Data for Each Period Reported (Y/N)	Mention of Washout Period (Y/N)	Mention of Carry-over Effect (Y/N)
Blumenfeld, 1994	N	N	N	N	Y	Y
Tanos, 1995	Y	N	N	N	N	N
Devreker, 1996	Y	N	N	N	N	N
Rein, 1996	Y	Y	Y	Y	N	N
Hurd, 1996	N	N	N	N	N	N
Cacciatore, 1997	Y	N	N	N	N	N
Jacob, 1998	N	N	Y	N	N	N
Bassil, 2000	Y	N	N	N	N	N
Ben-Rafael, 2000	Y	N	N	N	N	N
Harlin, 2002	N	N	N	Y	N	N
Stern, 2003	Y	N	Y	Y	N	N
Fedorcsák, 2003	N[a]	N	N	Y	N	Y
Papanikolaou, 2005	Y	N	N	N	N	N
Hagemann, 2010	Y	Y	Y	Y	N	N
Yovich, 2010	N	N	N	N	N	N

?: Given data from full-text of included studies, reviewers were unable to determine whether there was missing data for reasons other than pregnancy.
[a] Bayesian statistics were employed.
[b] Statistical estimation methods of treatment effect refers to primary outcomes and include: standard deviation of the mean, standard error of the mean, and confidence intervals.

5. Conclusion

We have completed the first methodological survey to date of infertility literature employing the crossover design. Despite the problems we have identified, serious consideration should still be given to using the crossover design in infertility research. Methods to account for missing data and more complete reporting of key methodological features may significantly reduce the risk of bias and improve the validity of these trials. Future reviews on crossover trials should continue to employ sensitive methodology filters in their search strategies, in order to encompass the entire body of relevant literature.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Authors' contributions

SDW and DB conceived and designed the study protocol. DB and DS carried out the protocol and analyzed and interpreted the data. SDW and SM contributed materials, analysis tools, and knowledge and expertise in their respective fields. SDW, SDM, DB, and DS wrote the paper.

Acknowledgements

Not applicable.

List of Abbreviations

CI	Confidence Interval
D-Trp6-LHRH	Decapeptyl
DEX	Dexamethasone
FSH	Follicle Simulating Hormone
FSHr	Follitrophin beta
GH	Growth Hormone
GHu	Uncontrolled Growth Hormone
GnRH	Gonadotrophin Releasing Hormone
GnRHa	Gonadotrophin Releasing Hormone analogues
HMG/HCG	Human Menopausal Gonadotrophin/Human Chorionic Gonadotrophin
PR	Pregnancy Rate
NR	Not Reported
Int	Intervention
IVF	In-vitro fertilization
LH	Luteinizing Hormone
P	Progesterone
SD	Standard Deviation of the Mean
SEM	Standard Error of Mean

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.conctc.2019.100426.

References

[1] T. Bushnik, J.L. Cook, A.A. Yuzpe, S. Tough, J. Collins, Estimating the prevalence of infertility in Canada, Hum. Reprod. 27 (3) (2012) 738–746.
[2] E. Te Velde, D. Habbema, E. Nieschlag, T. Sobotka, A. Burdorf, Ever growing demand for in vitro fertilization despite stable biological fertility-A European paradox, Eur. J. Obstet. Gynecol. Reprod. Biol. 214 (2017) 204–208.
[3] E.I. Kamphuis, S. Bhattacharya, F. van der Veen, B.W. Mol, A. Templeton, I.V.F.G. Evidence Based, Are we overusing IVF? BMJ 348 (2014) g252.
[4] B. Jones, M.G. Kenward, Design and Analysis of Cross-Over Trials, third ed., CRC Press, 2014.
[5] T. Li, T. Yu, B.S. Hawkins, K. Dickersin, Design, analysis, and reporting of crossover trials for inclusion in a meta-analysis, PLoS One 10 (8) (2015) e0133023.
[6] S.S. Senn, Cross-over Trials in Clinical Research, Wiley, 2003.
[7] M. Takada, T. Sozu, T. Sato, Practical approaches for design and analysis of clinical trials of infertility treatments: crossover designs and the Mantel-Haenszel method are recommended, Pharm. Stat. 14 (3) (2015) 198–204.
[8] T.N. Bonten, B. Siegerink, J.G. van der Bom, [Cross-over studies], Ned. Tijdschr. Geneeskd. 157 (3) (2013) A5542.
[9] R.J.A. Little, D.B. Rubin, Statistical Analysis with Missing Data, second ed., Wiley, Hoboken, N.J., 2002.
[10] S. Daya, Pitfalls in the design and analysis of efficacy trials in subfertility, Hum. Reprod. 18 (5) (2003) 1005–1009.
[11] K.S. Khan, S. Daya, J.A. Collins, S.D. Walter, Empirical evidence of bias in infertility research: overestimation of treatment effect in crossover trials using pregnancy as the outcome measure, Fertil. Steril. 65 (5) (1996) 939–945.
[12] B. Makubate, S. Senn, Planning and analysis of cross-over trials in infertility, Stat. Med. 29 (30) (2010) 3203–3210.
[13] M. Kastner, S.E. Straus, K.A. McKibbon, C.H. Goldsmith, The capture-mark-recapture technique can be used as a stopping rule when searching in systematic reviews, J. Clin. Epidemiol. 62 (2) (2009) 149–157.
[14] W.C. Dodson, D.K. Walmer, C.L. Hughes Jr., S.E. Yancy, A.F. Haney, Adjunctive leuprolide therapy does not improve cycle fecundity in controlled ovarian hyperstimulation and intrauterine insemination of subfertile women, Obstet. Gynecol. 78 (2) (1991) 187–190.
[15] J. Harlin, A. Aanesen, G. Csemiczky, H. Wramsby, G. Fried, Delivery rates following IVF treatment, using two recombinant FSH preparations for ovarian stimulation, Hum. Reprod. 17 (2) (2002) 304–309.
[16] P. Fedorcsak, P.O. Dale, R. Storeng, T. Abyholm, T. Tanbo, The effect of metformin on ovarian stimulation and in vitro fertilization in insulin-resistant women with polycystic ovary syndrome: an open-label randomized cross-over trial, Gynecol. Endocrinol. 17 (3) (2003) 207–214.
[17] V. Tanos, S. Friedler, A. Shushan, N. Strauss, I. Hetsroni, A. Lewin, Comparison between nafarelin acetate and D-Trp6-LHRH for temporary pituitary suppression in in vitro fertilization (IVF) patients: a prospective crossover study, J. Assist. Reprod. Genet. 12 (10) (1995) 715–719.
[18] Z. Ben-Rafael, J. Farhi, D. Feldberg, B. Bartoov, M. Kovo, F. Eltes, J. Ashkenazi, Follicle-stimulating hormone treatment for men with idiopathic oligoteratoasthenozoospermia before in vitro fertilization: the impact on sperm microstructure and fertilization potential, Fertil. Steril. 73 (1) (2000) 24–30.
[19] B. Cacciatore, N. Simberg, A. Tiitinen, O. Ylikorkala, Evidence of interplay between plasma endothelin-1 and 17 beta-estradiol in regulation of uterine blood flow and endometrial growth in infertile women, Fertil. Steril. 67 (5) (1997) 883–888.
[20] Z. Blumenfeld, M. Dirnfeld, Y. Gonen, H. Abramovici, Growth hormone co-treatment for ovulation induction may enhance conception in the co-treatment and succeeding cycles, in clonidine negative but not clonidine positive patients, Hum. Reprod. 9 (2) (1994) 209–213.
[21] C. Stern, L. Chamley, H. Norris, L. Hale, H.W. Baker, A randomized, double-blind, placebo-controlled trial of heparin and aspirin for women with in vitro fertilization implantation failure and antiphospholipid or antinuclear antibodies, Fertil. Steril. 80 (2) (2003) 376–383.
[22] D. Moher, S. Hopewell, K.F. Schulz, V. Montori, P.C. Gotzsche, P.J. Devereaux, D. Elbourne, M. Egger, D.G. Altman, Consolidated standards of reporting trials G. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials, J. Clin. Epidemiol. 63 (8) (2010) e1–37.
[23] S. Straube, B. Werny, T. Friede, A systematic review identifies shortcomings in the reporting of crossover trials in chronic painful conditions, J. Clin. Epidemiol. 68 (12) (2015) 1496–1503.
[24] S.J. Nolan, I. Hambleton, K. Dwan, The use and reporting of the cross-over study design in clinical trials and systematic reviews: a systematic assessment, PLoS One 11 (7) (2016) e0159014.
[25] S. Wellek, M. Blettner, On the proper use of the crossover design in clinical trials: part 18 of a series on evaluation of scientific publications, Dtsch Arztebl Int 109 (15) (2012) 276–281.
[26] P. Armitage, G. Berry, J.N.S. Matthews, Statistical Methods in Medical Research,

Wiley, 2013.
[27] M.M. Leeflang, R.J. Scholten, A.W. Rutjes, J.B. Reitsma, P.M. Bossuyt, Use of methodological search filters to identify diagnostic accuracy studies can lead to the omission of relevant studies, J. Clin. Epidemiol. 59 (3) (2006) 234–240.
[28] R.J. Little, R. D'Agostino, M.L. Cohen, K. Dickersin, S.S. Emerson, J.T. Farrar, C. Frangakis, J.W. Hogan, G. Molenberghs, S.A. Murphy, et al., The prevention and treatment of missing data in clinical trials, N. Engl. J. Med. 367 (14) (2012) 1355–1360.
[29] S.A. Amer, J. Smith, A. Mahran, P. Fox, A. Fakis, Double-blind randomized controlled trial of letrozole versus clomiphene citrate in subfertile women with polycystic ovarian syndrome, Hum. Reprod. 32 (8) (2017) 1631–1638.
[30] R. Stanislavov, P. Rohdewald, Sperm quality in men is improved by supplementation with a combination of L-arginine, L-citrullin, roburins and Pycnogenol(R), Minerva Urol. Nefrol. 66 (4) (2014) 217–223.