

Towards better meta-analyses in assisted reproductive technology: Fixed, random or multivariate models?

Philippe Leheret

Philippe Leheret, Faculty of Medicine, the University of Melbourne, Southbank 3006, Victoria, Australia

Philippe Leheret, Faculty of Economics, UCL Louvain University, B-7000 Mons, Belgium

Author contributions: This author is the exclusive author of this whole research.

Conflict-of-interest statement: The author declares no competing interests.

Data sharing statement: None.

Open-Access: This article is an open-access article which was selected by an in-house editor and fully peer-reviewed by external reviewers. It is distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>

Correspondence to: Dr. Philippe Leheret, PhD, Professor of Statistics, Faculty of Medicine, the University of Melbourne, 801/250 St Kilda Rd, Southbank 3006, Victoria, Australia. philippe.leheret@gmail.com
Telephone: +61-3-96999411
Fax: +61-3-96999411

Received: May 15, 2015

Peer-review started: May 20, 2015

First decision: July 26, 2015

Revised: September 27, 2015

Accepted: October 16, 2015

Article in press: October 19, 2015

Published online: December 26, 2015

Abstract

AIM: To study the validity of the fixed, random, and multivariate meta-analytical models applied in meta-

analyses in artificial reproduction technique.

METHODS: Based on common characteristics of *in vitro* fertilization (IVF) meta-analyses, we simulated a large number of data to compare results issued from the fixed model (FM) with the random model (RM). For multiple endpoints meta-analysis (MA), we compared the univariate RM with the multivariate model (MM). Finally, we illustrate our findings in re-analyzing a recent MA.

RESULTS: In our review, although a homogeneous effect was excluded in 89% of the MAs (11%), FM was utilized in 41 studies (82%). From simulations, a concordance of $59\% \pm 6\%$ was found between the two tests, with up to 65% of falsely significant results with FM. The *Q*-test on studies characterized by substantial heterogeneity falsely accepted homogeneity in 46% of studies. Comparing separate univariate RM and MM on multiple endpoints studies, MM reduces the between endpoint discrepancy (BED) of 68%, and increases the power of $57\% \pm 8\%$. In the example dealing with the controversial effect of luteneizing hormone supplementation to follicle stimulating hormone during ovarian stimulation in IVF cycles, MM reduced BED by 66%, and consistent effects were found for all the endpoints, irrespective of partial reporting.

CONCLUSION: The FM generally may produce falsely significant differences. The RM should always be used. For multiple endpoints, the MM constitutes the best option.

Key words: Meta-analysis; Random model; Fixed model; Assisted reproductive techniques; *In vitro* fertilization

© The Author(s) 2015. Published by Baishideng Publishing Group Inc. All rights reserved.

Core tip: The numerous meta-analyses (MA) published in assisted reproduction technology (ART) are often

characterized by conflicting results. This paper provides evidence that the choice of the meta-analytical model constitutes a major concern. We first identified a general profile of characteristics of the ART studies, compare different models by simulation and resolve a practical case. MA based on the fixed model produce severe biases and falsely significant differences. Better results derive from the random model. For partially reported multiple endpoints, the multivariate model takes advantage of the between-endpoint inter-correlation and provides consistent estimates, better precision, and higher power.

Lehert P. Towards better meta-analyses in assisted reproductive technology: Fixed, random or multivariate models? *World J Meta-Anal* 2015; 3(6): 225-231 Available from: URL: <http://www.wjgnet.com/2308-3840/full/v3/i6/225.htm> DOI: <http://dx.doi.org/10.13105/wjma.v3.i6.225>

INTRODUCTION

Highly controversial in the 1990s, meta-analysis (MA) has become a widely recognized technique to synthesize evidence from clinical trials. Although considered by many clinicians as mixing apples and oranges^[1], evidence based medicine groups have contributed greatly to the acceptability of the approach. As a result, MA considerably impacts drug prescription and clinical practice maybe even more than isolated trials. Thus, like clinical trials, meta-analyses should be conducted with the highest quality and methodological standards.

Trials may provide conflicting results, due to various reasons such as patient selection, trial conduct, duration, and sample size. Controversial results between MAs conducted on the same subject are more worrisome, as a MA constitutes, in essence, a synthesis from existing evidence. And yet, such differences are often observed among MAs, generating doubts on the validity of the results and the way they were identified.

The most known reasons of controversy are differences in study selection or elimination depending on whether or not they were published, blind, randomized, with a sufficient methodological quality risk of bias between studies.

Much less discussed, the choice of the meta-analytical method may involve potentially strong differences on results. In restraining to assisted reproduction technology (ART) context, at least two important concerns may be mentioned:

(1) A majority of MAs used the traditional and simplest fixed model (FM), in which the studied treatment effect is assumed constant across any study. An essential specificity of ART is the considerable difference of practice, procedures, medication use and knowhow among studies, countries or centres causing very heterogeneous performances^[2,3]. To which extent this assumption of constant effect remains bearable,

although in most MAs this assumption was not tested? Should other models such the random model (RM) admitting an heterogeneous treatment effect be more adapted, while being more conservative?

And (2) In most of the studies, several endpoints are evaluated, such as the number of retrieved oocytes, embryos, implantation rate, and pregnancy ratios. Their separate analyses involve difficulties in the discussion and become non-comparable for endpoints reported by different the number of selected trials (NST). This is called partial reporting and is very common in ART MAs. Separate MAs for each endpoint where high correlation and partial reporting co-exist produce a frequent paradox characterized by conflicting results on correlated endpoints, simply due to non-comparable power. The trickiest case is live birth, the ultimate endpoint in ART, much less reported than other endpoints, compared with earlier markers like clinical pregnancy necessitating much less follow up. Unlike univariate MA, the multivariate MA is a recent proposal briefly introduced here below, taking advantage of between endpoints correlations. To which extent using simple univariate models remains acceptable in such conditions, or is it worthwhile to turn to multivariate approach?

This research is based on the hypothesis that ART studies are characterized by a homogeneous profile of characteristics enabling the adapted choice of a meta-analytical model. In a first stage, we determined this profile based on a sample of MAs selected from a literature review, at a second stage we conducted simulation studies based on this profile to compare the models, and we apply these principles on a study case.

MATERIALS AND METHODS

We first attempted to identify a general specific profile of MAs in ART, including the NST, level of heterogeneity, number and kind of endpoints, the used model and its options. As lots of recalculations were necessary on each study, a random sample of 50 MAs was extracted from a list found from literature research (MEDLINE, EmBase, Google) irrespective of publication (paper/abstract), date or language, by using the key words list [(MA or systematic review) and (IUI or IVF or ICSI)].

We assessed whether a specific profile of MA studies has incidence on the adequacy of MA models, by using the specific profile collected from our MA sample, and conducting simulations in replicating these specific conditions on a multitude of generated samples. For all these samples we compared the FM with the RM and for multiple endpoints, the simple univariate calculation with the multivariate MA, on various criteria: (1) the magnitude of the difference between the estimates, and the direction of bias; (2) the difference in precision of estimates, and consequences on statistical power; and (3) for multiple endpoints, consistency between endpoints with respect to correlations between endpoints.

Finally, we illustrate our findings on a real study

case in re-analyzing a recently published MA in which we compare the results found according to the studied models.

Statistical analysis

Our simulation program was carried out with the statistical package R (release 3.01)^[4], univariate and multivariate models (MM) calculated with metafor^[5] and mvmeta^[6] packages, respectively.

We compared models on the following characteristics: the relative deviation between the estimated effects (EE) derived from two models A and B was calculated as $RD_{AB} = 100 \cdot |EE_A - EE_B| / EE_A$. The concordance between two tests was defined as the mean proportion of concordant decisions for the same data ($P < 0.05$ cutoff) over all the simulated tests, the relative precision of an EE as the ratio $RP = EE / CIL$ ($CIL = EE$ 95% half confidence interval length) over all the simulated data. For multiple endpoints study we defined two indexes: (1) provided that the highest the correlation between any two endpoints X and Y, the smallest the difference $EE_X - EE_Y$ should be, we define the between endpoint discrepancy (BED) index BDI as the mean of deviations $|EE_X - EE_Y|$ over all pairs of endpoints (X, Y) and all the studies weighted by the coefficient of determination R^2 (X, Y); and (2) As partial report of endpoints affects the comparability between endpoints, we determined the sensitivity of power to NST by the correlation coefficient R (RP, NST).

RESULTS

Description of the MA sample

Our sample consisted of MAs published between 1997 until 2014, 12 were congress abstracts out of which 9 were available in poster proceedings. Thirty-seven limited selection to randomized controlled trials. Eight treatments were compared. Forty-eight are based on literature findings, and 2 on individual patient data. The median NST was 13 (IQ = 7-17). Forty-three studies analyzed multiple endpoints. These endpoints were either continuous (drug dosage and duration, estradiol, hormonal values), counts (number of oocytes, metaphase II oocytes, embryos, transferred embryos), ratios (embryo quality, implantation rate, etc.) or binary endpoints: Biochemical pregnancy [positive pregnancy test (β -hCG) 15-20 d post-hCG administration], clinical pregnancy (ultrasound scan with at least one sac with heartbeat 35-42 d post-hCG administration), ongoing pregnancy (viable pregnancy 10-12 wk after embryo transfer), live birth, multiple birth, occurrence of ovarian hyperstimulation syndrome, ectopic pregnancy or miscarriage. The median effect size converted to Risk Ratio was $RR = 1.35$ (IQ = 1.11, 1.53). The most referenced endpoints were the number of oocytes, biochemical or clinical pregnancy (76%), the least referenced was live birth (21%). The median number of referenced endpoints was 7 (IQ = 4, 12).

Comparing RM and FM

In our study sample, FM was used as the main model in 41 studies (82%). This choice was not justified for 22 MAs (54%). Q-test was mentioned in 34 studies but discussed only in 8 studies, and FM was used in spite of a detected significant heterogeneity in 13 studies. The forest plot was available for almost all ($n = 48$) MAs and allowed recalculation of the non-reported Q-test and I^2 statistics. The homogeneity assumption was rejected by the Q-test in 35 studies (70%), although in the 15 other MAs, NST was less than 7. The overall mean I^2 statistic was 58% (SD = 12) and 52% (SD = 12) for the 15 studies not rejected by the Q-test. Out of the 41 studies based on the FM, 29 (71%) were characterized by $I^2 > 40$. Based on 10000 simulated samples based on the identified profile, our results are summarized as follows:

Difference of EEs: The mean relative deviation between FM and RM was $RD = 4.3\% + 2.1\%$, 46% exceeding a deviation of 5%, higher differences observed for larger heterogeneity ($RD = 8.3\% + 3.2\%$ when $I^2 > 60\%$).

Power and precision: The mean concordance between RM and FM was $59\% \pm 6\%$, the ratio of the relative precision to $RP_{RM/FM} = 1.39 \pm 0.12$, and the ratio of power to $P_{RM/FM} = 1.33 + 0.09$, thus expected 33% more significant results found with FM. For $NST \leq 7$, this difference was larger ($P_{RM/FM} = 1.65 \pm 0.13$).

Q-test fallacy: For MAs characterized by $I^2 = 30\%$, 50% and 75% considered as a moderate, substantial and strong heterogeneity, the mean power (at 0.05 level) to reject homogeneity was 0.32, 0.54 and 0.89 when $NST = 7$, and 0.53, 0.66 and 0.97, when $NST = 17$. Thus using Q-test on studies characterized by at least substantial heterogeneity falsely accepted homogeneity in 46% and 34% of studies for $NST = 7$ and 17 respectively, corresponding to the quartiles of the NST distribution of our MA sample.

Partial reporting of multiple endpoints

For these studies (86% of our sample), separate univariate MAs were conducted for each endpoint. The median ratio between the NST available for the most and less reported endpoint in each study was 4.2 (IQ = 2.1, 8.4). The within-study correlations between endpoints were not available from our MA sample. We approximated these values in estimating correlations between endpoints in available retrospective studies.

EE difference: The mean relative differences between RM and MM were $RD = 12.3\%$ (IQ = 1.1, 35.3), RD strongly increasing with the magnitude of the correlations and the partiality of reporting.

Between endpoint discrepancy: By using RR to quantify EE, the mean BDI was 0.08 ± 0.04 and 0.25

Table 1 Luteneizing hormone supplementation effect: Comparison between fixed, random and multivariate models

		BPR				CPR				OPR				LBR			
Overall	FM	1.17	1.01	1.28	0.03	1.08	0.99	1.17	0.03	1.09	0.93	1.19	0.25	1.28	0.98	1.56	0.07
	RM	1.14	0.95	1.38	0.16	1.07	1	1.19	0.05	1.05	0.93	1.19	0.45	1.23	0.88	1.72	0.23
	MM	1.23	0.94	1.61	0.12	1.09	1	1.2	0.06	1.13	1.03	1.24	0.01	1.13	1.01	1.28	0.04
POR	FM	1.22	0.97	1.53	0.1	1.30	1.03	1.64	0.03	1.29	0.82	2.02	0.28	1.81	0.99	3.31	0.06
	RM	1.19	0.87	1.64	0.27	1.30	0.99	1.64	0.04	1.29	0.82	2.02	0.28	1.70	0.78	3.69	0.27
	MM	1.37	1.00	2.21	0.04	1.27	1.01	1.61	0.04	1.38	1.06	1.81	0.02	1.53	1.09	2.15	0.01

Luteneizing hormone supplementation effect for overall population and POR models. Comparison between fixed (FM), random (RM), and multivariate (MM) models. Values are risk ratio 95%CI, and *P*-value. BPR: Biochemical pregnancy rate; CPR: Clinical pregnancy rate; OPR: On going pregnancy rate; LBR: Live birth rate; POR: Poor ovarian responder.

Table 2 Between endpoint correlation

	CP	OP	LB	NST	<i>I</i> ²
BP	0.95 (0.91, 0.97)	0.91 (0.84, 0.95)	0.87 (0.79, 0.93)	22	41.2
CP	-	0.96 (0.92, 0.97)	0.92 (0.85, 0.95)	39	31.8
OP	-	-	0.96 (0.92, 0.98)	13	46.36
LB	-	-	-	8	40.1

Within-study correlations (95%CI). NST: Number of studies reporting each endpoint; *I*²: Heterogeneity index; BP: Biochemical pregnancy; CP: Clinical pregnancy; OP: On going pregnancy; LB: Live birth.

± 0.06 for MM and RM, MM reducing the discrepancy index of 68% compared with RM.

Sensitivity to NST: The mean correlation *R* (RP, NST) across any two pairs of endpoints and for every study was 0.75 ± 0.07 and 0.27 ± 0.5 for RM and MM, respectively, thus MM having a beneficial effect on reduction of the sensitivity to NST of 64%.

Power and precision of estimates: The mean concordance of decision between RM and MM was $57\% \pm 8\%$. The ratio of the 95%CI length (CIL) of FM on RM was $CIL_{RM/MM} = 0.65 \pm 0.13$ and the mean power ratio $P_{MM/RM} = 1.57 \pm 0.38$, thus in average 57% more decisions of significant differences in using MM.

Study case

To illustrate these general principles, we re-analyzed the data of a recent MA^[7]. Adding recombinant human luteinizing hormone (r-hLH) to recombinant human follicle-stimulating hormone (r-hFSH) during ovarian stimulation has motivated numerous studies and conflicting MAs, suggesting a particular benefit on poor responders (POR) compared with normal responders (NOR). The selection was comprised of RCTs on women (18-45 years) undergoing IVF/ICSI treated with r-hFSH plus r-hLH (FL group) or r-hFSH alone (F group). From 40 RCTs (6443 patients), we identified 45 separate reports on 31 NOR and 15 POR studies. Clinical pregnancy rate (CP) was significantly higher in FL group both for the overall sample and POR subgroup. However, the analysis failed to find significant results for the other pregnancy markers, due to partial reporting, biochemical BP, clinical CP, ongoing OP pregnancies and live birth LB, reported

in the selected studies with 22, 39, 13 and 8 studies, respectively. We applied three methods on these data (Table 1): The univariate analyses based on FM and RM, and a multivariate meta-analysis (MM).

Through a bootstrapped replication ($n = 1000$), we assessed the superiority of the r-hLH supplementation effect in testing three models: (1) effect on overall population (intercept only); (2) restricted effect on POR only; and (3) effect on overall population with an additional effect on POR. By selecting our model based on a significant decrease of the Akaike Information Criteria value, model (2) was the more likely, but not significantly better than model (1), whereas model (3) was rejected, thus we restrained our analysis in testing the two first models (1) and (2).

Within-study correlations were unknown and handled following various techniques: Among possible handling techniques^[8], we narrowed the range of possible values based on existing data on 10 centers at our disposal in conducting sensitivity analyses by imputing values within the 95%CI. Strong correlations were found between the endpoints (Table 2).

The proportion of concordant decisions ($P < 0.05$) was 5/8, 1/8, and 1/8 between FM-RM, RM-MM and FM-MM, respectively. The relative precision of estimates were $RP = 0.26, 0.34$ and 0.19 for FM, RM and MM, respectively. Among the three significant differences detected by RM, only one was confirmed by RM, highlighting FM anticonservatism, in this heterogeneous example ($I^2 > 40\%$). RM detected only one significant difference for CPR, but this is also by far the most reported endpoint. MM detected consistent estimates with respect with high correlations and similar significant values unrelated to NST. Overall, the BED index of each model was $BDI = 0.61, 0.43$ and 0.16 for FM, and the sensitivity to NST was $0.85, 0.54$ and 0.18 . Thus MM reduced BDI and sensitivity to NST to 63% and 66% respectively. MM reducing sensitivity of 66% compared with RM. For LB in particular, characterized by the smallest NST, univariate models found a strong although non-significant values of 1.81 and 1.7 , inconsistent with other endpoints, in spite of high correlations. Finally, we repeated this analysis in adding the number of oocytes, the number of embryos, and implantation rate (not reported). Very few differences were found

on these endpoints, RR on LB values were 1.51 (1.1, 2.03), $P = 0.006$, and 1.11 (0.97, 1.23), $P = 0.06$. For the POR and overall model, respectively. Our analysis provides evidence of a clinically relevant effect of live birth for POR supplemented with r-hLH, compared with r-hFSH alone, this effect remaining consistent for all the pregnancy endpoints.

DISCUSSION

Random or FM

In our MA sample, the FM was overwhelmingly preferred (82%), often unjustified, in spite of evidence of heterogeneity (only 7.2% of studies such that $I^2 < 40\%$). The heterogeneity of effect among studies may be explained by the strong multifactorial variability observed in pregnancy predicting models^[2,3] in which center variability was found as the major predictor, followed by the patient mix (age, ovarian reserve, etc.). The reasons behind center heterogeneity are many: Differences between used medication and dosage, staff expertise, differences in protocols and standard therapy, and patient mix differences. Although a fixed effect remains possible, the accumulation of causes potentially generating a variable treatment effect designates RM as the most likely model.

To which extent their results derived from FM can be considered as a reasonable approximation? Our simulation suggests a deviation higher than 5% for 46% of the studies and higher deviations expected for high heterogeneity ($I^2 > 60\%$). The difference is more worrying concerning precision and power: The between-study τ^2 variance assumed in the RM increases the standard error and the confidence interval length, thus produces generally more conservative tests in particular for small NST. Previous papers warn that although the power of FM is generally better, the gain of power becomes uncertain for RM in particular when NST is small and I^2 increases^[9]. Our simulation provides more accurate conclusions for ART: FM and RM provide inconsistent results with 33% and 65% more significant results in using FM for all the studies and small studies (NST < 7) respectively. The problem of discordance between the two tests is more concerning than the difference between estimates: FM is expected to falsely find significant differences in situations where heterogeneity is present as it is generally the case in ART studies.

Another crucial conclusion is the *Q*-test fallacy: In spite of its popularity, previous researches^[9,10] warned on the high sensitivity of the power of this test with NST. Our simulation clearly highlights that, although it is the commonest test to select between RM and FM, the *Q*-test is not reliable in falsely accepted homogeneity in 46% and 34% of studies for NST = 7 and 17 respectively.

Univariate compared with multivariate approach

We limited the comparison between RM and MM. The median ratio of 4.2 between the available NST between

the most and less reported endpoints implies expected consequences on precision, power and estimates consistency:

(1) Between endpoint discrepancies: MM model has been demonstrated to provide a unique solution offering optimal consistency between estimates^[11]. In our results, considerable relative deviations between RM and MM were observed (RD = 18.3% + 7.3%). Moreover, taking into account the correlation between endpoints, the mean discrepancy index BDI was 0.08 ± 0.04 and 0.25 ± 0.06 for MM and RM, respectively. Thus MM increases the consistency between endpoint of 68% compared with RM.

(2) MM provides optimal estimates in reducing bias and sensitivity to partial reporting of multiple endpoints^[11]. We confirm this result from our simulations: The mean correlation R (RP, NST) were 0.75 ± 0.07 and 0.27 ± 0.5 for RM and MM, respectively, thus MM reducing the undesired sensitivity to NST = 64%.

(3) The standard error of the estimates are always better in MM compared with univariate models MAs^[11]. In our simulation, the power of MM is 57% more than RM. As both tests are applicable, univariate RM is more conservative than MM.

And (4) Feasibility and assumptions: MM requires more assumptions than univariate approach in particular the multivariate normal assumption, another important constraint is the knowledge of within-study correlations, which is generally difficult to obtain, as ideally these values necessitate individual patient data. Alternative techniques are possible to substitute approximations^[11], and sensitivity analyses around these estimates are needed to assess the stability of the results to these approximations.

Although MM is characterized by multiple advantages, it was rarely used in practice in ART. The reasons include tradition, defiance against apparently more complex model, underlying hypotheses sometimes difficult to assess, more data necessary often non available (such as within study correlations for which simplifying their values needs further research), and the lack of easy software to implement this technique.

In conclusion, the concept of MA is now widely accepted, but many methodological aspects remain controversial and the choice of an unjustified model may result into strongly biased results. This paper highlights this particular aspect in ART where the frequent use of the fixed traditional FM is source of important biases, due to the strong observed heterogeneity and partial reporting of multiple endpoints. Based on our results, we suggest the following implications in practice:

(1) RM must be regarded as the appropriate model for MAs in ART research. FM should be considered only upon robust a priori justification concerning the homogeneity of the studied question and confirmed by the observed I^2 . The *Q*-test should definitely be disregarded. More conservative than FM, RM has the same power in case of homogeneity of effect, in which case the two results coincide. Thus selecting RM does

not involve loss of power.

(2) The level of heterogeneity needs to be reported and discussed. The I^2 statistics and Tau value (standard deviation of the between study size) constitute good descriptive measurements of heterogeneity. Failing to provide heterogeneity level may induce misinterpretations. An effect size of $RR = 1.50$ seems conclusive, however, a large value of I^2 means an important dispersion of this value. In that sense, the Tau value allows calculation of the proportion of studies for which RR becomes non-clinically meaningful.

(3) MAs may be characterized by very large between-study heterogeneity ($I^2 > 75\%$); this happens particularly when important differences of selection criteria are observed between studies. These studies, not uncommon in ART (19% of our sample), are subject to controversy, some arguing that the summary of results is based on non comparable studies. The RM remains fully applicable in these cases, with an obvious loss of power, price to pay to demonstrate the generalizability of the efficacy of a treatment across heterogeneous situations.

And (4) MAs involving partially reported multiple endpoints suffer from chaotic difference between the effects on the studied endpoints, the precision of the endpoints badly affected by the available NST. MM takes advantage of the between-endpoint inter-correlation by borrowing strength from all the other endpoints, to provide a consistent, unique and comparable estimate for all the endpoints, thereby compensating for the effect due to unequal sample sizes. The MM is much more consistent, not affected by partial reporting, and with more accurate estimates on all the endpoints.

ACKNOWLEDGMENTS

The author would like to thank Dr. Dan Jackson (MRC Biostatistics Unit, Cambridge, United Kingdom) for fruitful discussion on multivariate MA questions.

COMMENTS

Background

Meta-analysis (MA) is a widely accepted technique employed in the synthesis and evaluation of evidence from past clinical trials. The growing impact of MA on clinical practice requires that they should be conducted according to high quality standards. Numerous MAs have been published in assisted reproductive technology (ART) and conflicting results are not uncommon. A rarely discussed reason is the choice of the meta-analytical model: Irrespective of expected heterogeneity of the studied effect and partially reported multiple endpoints, the fixed model (FM) is almost always used. The objective of the present study is to assess the extent to which this approach or other models are more appropriate.

Research frontiers

Although developed since more than 10 years, multivariate MA constitutes a promising approach for multiple endpoints MAs. In summary, multivariate model (MM) supposes a prior knowledge of each within-study correlation between endpoints, and assumes the existence of an additional unknown between-study random effect [as the univariate random model (RM)]. By fixing this random effect to zero, the MM generalizes the univariate fixed model (FM). The endpoints are assumed to be distributed according a multivariate normal

distribution, the individual effect estimates are calculated as weighted mean over all the studies and in addition to univariate approach taking into account the correlation between endpoints. The multivariate MM model palliates the apparent deficiencies of separate univariate analysis for multiple endpoints MAs. However, these preliminary theoretical considerations show that the advantages of these models depend on study parameters, the number of studies (NST), considered endpoints, the level of partial reporting, the between study heterogeneity and the between endpoint correlations between effects.

Innovations and breakthroughs

The validity of models were discussed in statistical articles, rarely in applications, and this research provides evidence of the importance of this choice. The most known reasons of controversy in results are differences in study selection or elimination depending on whether or not they were published, blind, randomized, with a sufficient methodological quality risk of bias between studies. Much less discussed, the choice of the meta-analytical method may involve potentially strong differences on results. In restraining to ART context, at least two important concerns may be mentioned: (1) A majority of MAs used the traditional and simplest FM, in which the studied treatment effect is assumed constant across any study. An essential specificity of ART is the considerable difference of practice, procedures, medication use and know-how among studies, countries or centres causing very heterogeneous performances. To which extent this assumption of constant effect remains bearable, although in most MAs this assumption was not tested? Should other models such the RM admitting an heterogeneous treatment effect be more adapted, while being more conservative? (2) In most of the studies, several endpoints are evaluated, such as the number of retrieved oocytes, embryos, implantation rate, and pregnancy ratios. Their separate analyses involve difficulties in the discussion and become non-comparable for endpoints reported by different NST. This is called partial reporting and is very common in ART MAs. Separate MAs for each endpoint where high correlation and partial reporting co-exist produce a frequent paradox characterized by conflicting results on correlated endpoints, simply due to non-comparable power. The trickiest case is live birth, the ultimate endpoint in ART, much less reported than other endpoints, compared with earlier markers like clinical pregnancy necessitating much less follow up. Unlike univariate MA, the Multivariate MA is a recent proposal taking advantage of between endpoints correlations. To which extent using simple univariate models remains acceptable in such conditions, or is it worthwhile to turn to multivariate approach? This research is based on the hypothesis that ART studies are characterized by a homogeneous profile of characteristics enabling the adapted choice of a meta-analytical model. In a first stage, the author determined this profile based on a sample of MAs selected from a literature review, at a second stage the author conducted simulation studies based on this profile to compare the models, and the authors apply these principles on a study case.

Applications

These results may have a very important practical implication for ART/*in vitro* fertilization (IVF) researchers: The conclusions are very simple and strictly specific for MAs in this pathology: The FM has always been classically used and may provide important bias. The MM is the only model allowing a clear solution for multiple endpoints MA. For a better clinical understanding, I provide a practical example based on IVF data easily interpretable for clinicians.

Terminology

For reader less familiar with meta-analytical models, the author summarizes the principles of the FM, RM and MM in appendix. In summary FM assumes a fixed effect across all the studies. The overall estimated effect of the studied treatment compared with control is estimated by the mean of all the study estimates weighted by the reciprocal of their variance. An alternative to FM is the RM generalizing FM by assuming that the effect varies across studies, according to a normal distribution with unknown mean and variance to be estimated. The Q-test can be used before and compares the adequacy of FM and RM in testing the significance of the dispersion of the EE, however this test is known oversensitive with the NST. Another heterogeneity measurement, I^2 statistic, evaluates the percentage of variation attributable to the between study heterogeneity. Various rules were based on I^2 , in particular a value exceeding 40% evidencing a substantial heterogeneity should motivate the choice of RM. An introductory and seminal approach to MM can be found in. In summary,

MM supposes an prior knowledge of each within-study correlation between endpoints, and assumes the existence of an additional unknown between-study random effect (as the univariate RM). The multivariate MM model palliates the apparent deficiencies of separate univariate analysis for multiple endpoints MAs. However, these preliminary theoretical considerations show that the advantages of these models depend on study parameters, the NST, considered endpoints, the level of partial reporting, the between study heterogeneity and the between endpoint correlations between effects.

Peer-review

The author uses a mathematical approach to demonstrate existing concerns with meta-analyses conducted with studies on ART. In general, the author applies the findings from the analyses to support existing recommendations for "best practices" when performing a MA.

REFERENCES

- 1 **Eysenck HJ**. Meta-analysis and its problems. *BMJ* 1994; **309**: 789-792 [PMID: 7950571 DOI: 10.1136/bmj.309.6957.789]
- 2 **Arvis P**, Lehert P, Guivarc'h-Levêque A. Simple adaptations to the Templeton model for IVF outcome prediction make it current and clinically useful. *Hum Reprod* 2012; **27**: 2971-2978 [PMID: 22851717 DOI: 10.1093/humrep/des283]
- 3 **Paul SR**, Donner A. Small sample performance of tests of homogeneity of odds ratios in K 2 x 2 tables. *Stat Med* 1992; **11**: 159-165 [PMID: 1579755 DOI: 10.1002/sim.4780110203]
- 4 **R**, A Language and Environment for Statistical Computing, R Development Core Team, R Foundation for Statistical Computing. Vienna, Austria: 2010. Available from: URL: <http://www.r-project.org>
- 5 **Viechtbauer W**. Conducting meta-analyses in R with the metafor package. *J Stat Softw* 2010; **36**: 1-48 [DOI: 10.18637/jss.v036.i03]
- 6 **Gasparrini A**, Armstrong B, Kenward MG. Multivariate meta-analysis for non-linear and other multi-parameter associations. *Stat Med* 2012; **31**: 3821-3839 [PMID: 22807043 DOI: 10.1002/sim.5471]
- 7 **Lehert P**, Kolibianakis EM, Venetis CA, Schertz J, Saunders H, Arriagada P, Copt S, Tarlatzis B. Recombinant human follicle-stimulating hormone (r-hFSH) plus recombinant luteinizing hormone versus r-hFSH alone for ovarian stimulation during assisted reproductive technology: systematic review and meta-analysis. *Reprod Biol Endocrinol* 2014; **12**: 17 [PMID: 24555766 DOI: 10.1186/1477-7827-12-17]
- 8 **Nam IS**, Mengersen K, Garthwaite P. Multivariate meta-analysis. *Stat Med* 2003; **22**: 2309-2333 [PMID: 12854095 DOI: 10.1002/sim.1410]
- 9 **Hardy RJ**, Thompson SG. Detecting and describing heterogeneity in meta-analysis. *Stat Med* 1998; **17**: 841-856 [PMID: 9595615 DOI: 10.1002/(SICI)1097-0258(19980430)17:8<841::AID-SIM781>3.0.CO;2-D]
- 10 **Borenstein M**, Hedges LV, Higgins JP, Rothstein HR. A basic introduction to fixed-effect and random-effects models for meta-analysis. *Res Synth Methods* 2010; **1**: 97-111 [PMID: 26061376 DOI: 10.1002/jrsm.12]
- 11 **Jackson D**, Riley R, White IR. Multivariate meta-analysis: potential and promise. *Stat Med* 2011; **30**: 2481-2498 [PMID: 21268052 DOI: 10.1002/sim.4172]

P- Reviewer: Wang R

S- Editor: Gong XM L- Editor: A E- Editor: Jiao XK





Published by **Baishideng Publishing Group Inc**

8226 Regency Drive, Pleasanton, CA 94588, USA

Telephone: +1-925-223-8242

Fax: +1-925-223-8243

E-mail: bpgoffice@wjgnet.com

Help Desk: <http://www.wjgnet.com/esps/helpdesk.aspx>

<http://www.wjgnet.com>

