OPINION

IVF with preimplantation genetic screening, a promising new treatment with unexpectedly negative health outcomes: the Hippocratic role of Data Monitoring Committees

Willem M. Ankum1,4, Johannes B. Reitsma2 and Martin Offringa3

1Department of Obstetrics and Gynaecology, Academic Medical Centre, University of Amsterdam, the Netherlands; 2Department of Clinical Epidemiology and Biostatistics, Academic Medical Centre, University of Amsterdam, the Netherlands; 3Department of Paediatric Clinical Epidemiology, Academic Medical Centre, University of Amsterdam, the Netherlands

4Correspondence address. E-mail: w.m.ankum@amc.uva.nl

A recently published randomized controlled trial showed preimplantation genetic screening (PGS) as part of an IVF programme to reduce ongoing pregnancy rates by 1/3 in comparison to the control group without PGS: rate ratio (RR) 0.69 (0.51–0.93), P = 0.01. A masked interim analysis already showed significant differences between treatment arms: RR 0.58 (0.35–0.94), P = 0.02. Despite this finding, the trial’s Data Monitoring Committee decided not to stop, but to continue the trial. This paper argues why this decision was sound, since it was based on (i) explicit statistical criteria and (ii) the trade-off between risks and benefits for current and future IVF patients. The trial’s findings confront the medical community once again with the general problem of new technologies being implemented without randomized evidence of effectiveness.

Keywords: IVF; preimplantation genetic screening; masked interim analysis; Data Monitoring Committee; randomized controlled trial

Introduction

The role of a Data Monitoring Committee (DMC) as part of a randomized controlled trial (RCT) is to protect participating patients from unnecessary harm and to decide whether or not a trial should be discontinued before its planned completion (Armitage, 1991). In making this decision, DMCs must combine interim results on efficacy with data on side effects.

A recently published RCT showed preimplantation genetic screening (PGS) as part of an IVF programme to be harmful (Mastenbroek et al., 2007). Rather than improving pregnancy rates as suggested by earlier uncontrolled studies, PGS was shown to significantly reduce pregnancy rates by 1/3 in comparison to the control group receiving IVF without PGS.

Halfway through the trial, a blinded interim analysis was carried out. Despite a significant difference in ongoing pregnancies between treatment arms, the DMC decided not to stop, but to continue the trial. Was this a sound decision?

Materials and Methods

Mastenbroek et al. (2007) conducted a multicentre, randomized, double-blind, controlled trial comparing three cycles of IVF with and without PGS in women 35 through 41 years of age. The primary outcome measure was an ongoing pregnancy at 12 weeks of gestation.

The live birth rate was a secondary outcome. According to the power calculation, 372 women would be needed to detect an increase in ongoing-pregnancies from 40 to 55% after three cycles of IVF with PGS, with a power exceeding 80% at a significance level of 0.05.

The trial’s DMC, which was established at the launch of the study, had planned a masked interim analysis when about half of patients were included, thereby using the O’Brien–Fleming procedure for stopping a trial prematurely because of superiority of one treatment above the other, based on the primary outcome parameter (O’Brien and Fleming, 1979). Critical P-values for stopping were determined at P = 0.0052 for the first interim analysis and P = 0.048 for the final analysis. The required dataset was programmed by the investigators, but carried out by an outsider who printed the masked data, merely indicating treatment ‘A’ and ‘B’ with respective ongoing pregnancy rates.

Results

The masked interim analysis was carried out when 187 patients had been included with sufficient follow-up to reach the primary endpoint. It showed treatment A to be superior to B (P = 0.02, Table I).

Based on the pre-defined O’Brien and Fleming statistical stopping boundary, the DMC decided not to discontinue the trial and judged it unnecessary to un-blend the data.
The trial’s investigators were advised to continue the inclusion of patients. In order not to introduce any bias in their doing so, the DMC provided no information about the interim results. Eventually 408 women were included in the trial: 206 assigned to PGS and 202 assigned to the control group. They underwent 836 cycles of IVF: 434 cycles with, and 402 cycles without PGS. The ongoing-pregnancy rate was significantly lower in women assigned PGS, i.e. 52/206 (25%) versus 74/202 (37%) in those not assigned PGS (Table I).

Women assigned to PGS also had a significantly lower live birth rate: 49/206 (24%) versus 71/202 (35%, P = 0.01). The study concluded PGS not to increase, but to significantly reduce ongoing pregnancy and live birth rates in women undergoing IVF.

**Discussion**

The masked interim analysis showed treatment B to be superior to treatment A in terms of ongoing pregnancies, the only evaluable outcome at that time, at a significance level of P = 0.02. This raises the question as to whether the trial should have been stopped at that point.

In a meta-analysis, Montori et al. showed that the majority of trials that had been discontinued earlier than originally planned because of apparent benefit of one of the compared treatments—mostly the experimental new one—often failed to report adequate information about the decision to stop early (Montori et al., 2005). The majority of these trials showed implausibly large treatment effects, particularly when the number of events was small. Apparently, the benefit of treatment resulted from catching a ‘random high’, a chance finding by sheer (bad?) luck. The meta-analysts advised to view the results of prematurely discontinued trials with scepticism.

To contain this threat of an apparent benefit by chance, we feel that stopping rules should always be very conservative. This is why we chose a P-value of 0.005 as our criterion. Seeing P = 0.02 halfway through the trial, at 187 patients, it would have been both unethical and scientifically inappropriate to stop. If the blinded DMC had called a halt to what turned out to be a positive trial, this move would have been severely questioned. If stopped for a negative effect, the trial would have been dismissed by the protagonists of PGS as unacceptably flawed.

Some authors have argued that the task faced by DMCs in deciding whether to stop or continue a trial, apart from statistical stopping boundaries, still largely depends on good clinical judgement as well (DeMets et al., 1999; Pocock, 2005, 2006).

When the present trial started, PGS had already been adopted by various units providing IVF, and for some of those was—and continues to be—a main asset from a commercial perspective. What would have happened, in the case at hand, if the trial had been stopped prematurely?

If PGS would have appeared to be the superior treatment in the interim analysis, given the final result, patients not receiving this adjunct to standard IVF would have been worse off in the trial. For those patients being treated outside the trial, however, harm could only be amended by introducing PGS at a large scale, implying an increased workload for IVF technicians, a need for more skilled personal and rising costs of IVF. Obviously, this would have been attainable only after a prolonged period of time and at considerable investments.

If, however, the superior treatment arm at interim analysis would have appeared to be the control group without PGS, 50% of the future IVF patients included in the study, i.e. 110 women, would be harmed by the addition of PGS. In that case, those patients being treated outside the trial by standard IVF without PGS would have been better off, both now and in the future.

With these considerations in mind, the DMC decided not to truncate the trial and to adhere to the strict stopping boundary, i.e. discontinuation only when the difference between treatment strategies reached a P-value <0.005, which it did not.

By allowing continued patient inclusion, the investigators were enabled to reach sufficient statistical power, thereby providing more convincing data.

To the astonishment of many, instead of improving outcomes as earlier suggested, the addition of PGS turned out to result in far worse pregnancy rates than conventional IVF treatment. Not surprisingly, this finding lead to disbelief, denial and criticism from those who, despite the lack of randomized data, had already adopted PGS as part of their IVF work-up.

The result of the PGS trial alludes to the ancient dictum ascribed to Hippocrates: *Primum non nocere* (First, do no harm). Once again, this trial has confronted the medical community with the question as to whether any new technology should be implemented in daily clinical practice before being properly compared with the standard treatment. In view of the present data, it seems even questionable if any future research in this field can be done without meeting ethical and medico-legal objections.

The case at hand illustrates two important dilemmas that occur frequently in modern medicine, i.e. when to stop an ongoing trial, and when to introduce a new technology. The obvious answer applies to many other situations: Not too soon, and not without good clinical and scientific judgement.

**References**


