Evidence-based reproductive health: testing times for treatments

Robbie Foy, MRCGP, MFPHM, *Senior Lecturer in Primary Care, Centre for Health Services Research, University of Newcastle-upon-Tyne, Newcastle-upon-Tyne;* **Mike Crilly,** MRCGP, MFPHM, *Senior Lecturer in Clinical Epidemiology and Public Health Medicine, Department of Public Health, Aberdeen University Medical School, Aberdeen;* **Susan Brechin,** MRCOG, MFFP, *Clinical Senior Lecturer in Sexual and Reproductive Health, Aberdeen University, Department of Obstetrics and Gynaecology, Aberdeen Maternity Hospital, Aberdeen, UK*

Correspondence: *Dr R Foy, Centre for Health Services Research, University of Newcastle-upon-Tyne, 21 Claremont Place, Newcastle-upon-Tyne NE2 4AA, UK. Tel: +44 (0) 191 222 7214. Fax: +44 (0) 191 222 6043. E-mail: r.c.foy@ncl.ac.uk*

(Accepted 16th May 2003)

Journal of Family Planning and Reproductive Health Care 2003; **29**(3): 165–168

Clinical scenario

A 30-year-old woman consults you. She has been successfully using a progesterone-only pill (POP) containing levonorgestrel (LNG) (Microval®) but is anxious that her POP has a higher failure rate than her previous combined oral contraceptive pill (COC). She would prefer to continue some form of POP, having disliked depot medroxyprogesterone acetate (DMPA) in the past. She is unable to use combined oral contraception (COC) due to migraine with focal aura and has recently heard that a better POP is now available.

You recall attending a recent update meeting on contraception where a couple of colleagues were enthusiastically advocating the new POP (Cerazette®), which contains 75 µg (microgrammes) desogestrel (DSG) and apparently has the efficacy of a COC. During lunch you had visited pharmaceutical company stalls and picked up the usual useful freebies, including some literature on Cerazette, a pocket calculator and a pen that writes a message backwards in Mandarin:

梦幻可成真; 求愿须谨慎。

Based upon your colleagues' suggestions, you prescribe the DSG-only pill, Cerazette. But you are unsure if it is really any better than other POPs. One of the adverts for Cerazette identifies a promising study title: 'A double-blind study comparing the contraceptive efficacy, acceptability and safety of two progestogen-only pills containing DSG 75 µg/day or LNG 30 µg/day'.[1]

You request a copy via the postgraduate library. When it arrives you are not quite certain how to make sense of it all. Fortunately, a colleague who is an evidence-based medicine enthusiast recommends downloading a guide to appraising randomised controlled trials (RCTs) from this website: www.phru.org.uk/~casp/appraisa.htm.[2,3] The key questions from this guide are summarised in Figure 1. So, that evening you sit down at home with a glass of wine, a family-value-size bar of chocolate and the calculator to make sense of the article.

Screening questions
- Did the study ask a clearly focused question?
- Was this a randomised controlled trial (RCT) and was it appropriately so?

If the answer is 'Yes' to both, it is worth proceeding with the remaining questions.

Validity (trustworthiness)
- Were participants appropriately allocated to intervention and control groups?
- Were participants, staff, and study personnel 'blind' to participants' study group?
- Were all of the participants who entered the trial accounted for at its conclusion?
- Were the participants in all groups followed up and data collected in the same way?
- Did the study have enough participants to minimise the play of chance?

The results
- How are the results presented and what is the main result?
- How precise are these results?

Applying the study findings
- Were all the important outcomes considered so the results can be applied?

Figure 1 *Questions to help make sense of randomised trials (adapted from www.phru.org.uk/~casp/appraisa.htm and Guyatt et al.[2,3])*

Did the study ask a clearly focused question?

The first article in this series highlighted the usefulness of the Population, Intervention, Comparison and Outcome ('PICO') approach to structuring clinical questions.[4] Your focused clinical question is: In women eligible for the POP ('Population'), does the use of DSG-only pill ('Intervention') compared to a conventional POP ('Comparison' with LNG) reduce the risk of pregnancy ('Outcome')? Although not discussed with your patient, she may also be interested to know if switching to a DSG-only pill is likely to affect her bleeding pattern.

The article appears to answer these questions. The study population comprised healthy woman aged 18–45 years (average age 30 years) with normal menstrual cycles recruited from 44 centres in six Northern European countries. The largest single group of women were those switching from COCs and other POPs (37% 'switchers'). The study also included breastfeeding women and those starting oral contraception ('starters'). Women had to be willing to complete a daily diary documenting their bleeding pattern and adherence to POP. Thus the women represented a selected group of highly motivated women (as trial participants often are). The intervention comprised DSG 75 µg daily whilst the comparison was LNG 30 µg daily. The outcomes measured included self-reported bleeding patterns, pregnancy and other adverse experiences assessed after 3, 7 and 13 months.

Was this a RCT?

Although a range of study designs can be used to assess the effects of treatments, a RCT provides the most trustworthy (valid) evidence of effectiveness of treatments compared with other study designs.

In a RCT one group receives the treatment being tested, whilst a comparison group receives an alternative treatment. The key is that the likelihood of each participant receiving the trial treatment is determined by chance (random allocation). A broad range of health care interventions can be tested in RCTs, such as specific

therapies (e.g. drug therapy, counselling) or broader clinical policies (e.g. screening). The alternative intervention may include the standard treatment, a placebo or no treatment. Comparing the outcomes in the different groups allows an assessment of the results of the trial.

Unlike the dramatic therapeutic benefits observed with the introduction of penicillin or insulin, most effective treatments tend to have modest but important benefits.[5] Rigorous research methods that employ a carefully designed RCT are required to reliably detect such effects. For a RCT to be ethical and feasible to conduct, genuine uncertainty that a treatment works is essential. Treatments are sometimes advocated on the grounds of prevailing theory, or even 'common sense.' For example, the reported beneficial effects of hormone replacement therapy (HRT) on lipids – a proxy measure for coronary heart disease (CHD) risk – do not necessarily translate into a reduction in CHD events.[6]

Sometimes RCTs are prohibitively difficult or inappropriate to conduct. This is especially so when we hope to detect small differences or improvements for relatively rare outcomes, such as venous thromboembolism. In such cases, evidence about the benefits and risks of treatments might be required from large cohort or case-control studies. Such non-randomised studies, however, tend to show larger (or even false-positive) treatment effects compared with RCTs.[7] When these once-promising effect sizes are put to the sword of randomised evaluations they tend to diminish, sometimes altogether. For example, the reduction in the risk of CHD suggested by observational studies of HRT has not been confirmed by RCTs.[8–10]

Judging validity

Questions about validity are concerned with how trustworthy study findings are, based on the details of how the trial was conducted reported in a paper's methods section. Just as a designer label does not guarantee quality, the use of 'randomisation' in a study does not guarantee bias-free results. Attention needs to be paid to the adequacy of the randomisation process and the levels of care and attention subsequently received by study participants. Poorer quality randomised trials produce results that tend to overestimate the benefits of treatments compared with well-conducted trials.[7]

Were participants appropriately allocated to intervention and control groups?

In the DSG study, women were randomly allocated in a ratio of 3:1 (resulting in a DSG group of 989 and an LNG group of 331 women), presumably to provide more reliable follow-up data for women on DSG.

True randomisation usually entails allocating participants on the basis of computer-generated lists of random numbers. Quasi-randomisation involves allocating participants to study and control groups in a predictable manner (e.g. according to date of birth). The problem with predictable allocation is that prior knowledge of the allocation sequence might influence when and whether a clinician recruits participants. Clinicians naturally might wish to ensure that their patients receive what they judge to be the best treatment. This can lead to those patients with the most (or sometimes least) favourable prognoses receiving the study treatment.

Safeguards also need to be built into the process of trial entry. The use of sealed envelopes containing randomly generated treatment allocations is prone to tampering; clinicians may continue opening envelopes until the preferred treatment for a particular patient is found. Remote allocation, whereby the clinician contacts a centre 'hotline' by telephone or via an Internet website to ascertain the patient's allocation status following consent, can assure the independence of randomisation.

How women were randomly allocated in the DSG trial is not adequately described. No details are provided concerning how the random allocation schedule was created, concealed or administered. However, the baseline characteristics of the groups (age, menstrual patterns and previous pregnancies) following randomisation were similar, suggesting that the trial's randomisation process successfully produced two comparable groups.

In a carefully conducted RCT, the treatment allocation is made after trial entry has been decided upon. This sequence ensures that foreknowledge of what the next treatment is going to be cannot affect the decision to enter the participant into the trial. If it did, those allocated one treatment might differ systematically from those allocated another. For example, if a trial selectively recruited more highly motivated women to one group rather than another, subsequent pregnancy rates may differ because of this. Random allocation is intended to distribute such potential confounding factors evenly across the two groups. Any differences in outcome are therefore more likely to be attributable to the treatment alone rather than the characteristics of women allocated to each group. Some imbalance in confounding factors between the treatment and control groups can still occur by chance, although this becomes less likely with larger clinical trials.

This 'methodological paranoia' is based upon strong empirical evidence as well as confidential confessions of trial clinicians.[11,12] Quasi-randomisation and less rigorous methods of allocating treatments are associated with inflated estimates of treatment effects.[7] The randomisation and allocation processes are not fully described in the DSG study. The lack of a clear description of the allocation processes in research papers is itself associated with inflated estimates of the effects of treatment.[12]

Were participants, staff and study personnel 'blind' to participants' study group?

The trial was 'double-blind' in that the women recruited and the clinicians providing care did not know who had been allocated to which treatment, as participants received 'identical-looking tablets'. Expectations of the effect of a treatment can influence outcomes. For example, a woman who knew she had been allocated to the latest POP might behave differently, possibly becoming less (or even more) reliable in taking her pill regularly. She might have greater expectations of the treatment and hence discount or not report adverse effects. Similarly, clinician knowledge of a participant's treatment status may influence the level of additional support provided ('performance bias') or the recognition and recording of harms and benefits ('observation bias'). Concealing which group the patient is allocated to (blinding or masking) reduces such biases and can be applied to patients, clinicians or investigators. Blinding is not always feasible for treatments with known side effects or if the intervention is obvious (e.g. surgery).

Investigators are also prone to observation bias. This is less of a problem with objective outcomes, such as mortality, but it can be problematic for subjective outcomes, such as pain or menstrual bleeding. The trial used standardised criteria for assessing bleeding patterns from women's self-completed diaries, which helps to reduce observational bias.

Were all the participants who entered the trial accounted for at its conclusion?

This is where the real fun usually begins. Adverse events or the ineffectiveness of treatment may cause participants to

Table 1 *Summary of a POP RCT comparing DSG with LNG in 1306 healthy women over a 13-month period (data extracted from Tables 1 and 2 of the original paper)*[1]

	DSG (n = 979)	LNG (n = 327)	Absolute difference	NNT (95% CI)
Discontinued POP ('any reason')	439 (44.8%)	129 (39.4%)	5.4%	19 (9 to ∞)
Discontinued POP due to 'irregular bleeding'	220 (22.5%)	59 (18.0%)	4.5%	23 (11 to ∞)
Pregnancy rate	3 per 728 WYs	4 per 258 WYs		
Pregnancy per WY	0.0041	0.0155	–0.0114	88 (29 to ∞)

95% CI, 95% confidence interval; DSG, desogestrel; LNG, levonorgestrel; NNT, number needed to treat; POP, progestogen-only pill; RCT, randomised controlled trial; WY, 'women-year' [e.g. a woman taking the POP for 6 months contributes 6/13 (or 0.46) women-years of follow-up].

drop out of a treatment arm. Hence all participants entering a trial should be accounted for at the end of the study. 'Attrition bias' occurs if participants dropping out of one study arm systematically differ from those in the other arm. Their exclusion from the final analysis will lead to the overestimation of any true treatment effect. If you are interested in the effectiveness of clinical actions (such as switching POPs), then study participants should be analysed in the group to which they were allocated, even if they do not receive the treatment. This is known as intention-to-treat analysis.

Inevitably, some people drop out of trials following randomisation, and it is sometimes neither feasible nor ethical to continue collecting outcome data. In this trial, 14 randomly allocated women never started DSG or LNG, leaving 979 and 327 in each group, respectively, at the outset. There were considerable losses to follow-up over the 13 months of the clinical trial (40% or more), and the group of women completing the trial are likely to systematically differ from those who entered the trial. There is no hard-and-fast rule on what level of follow-up is acceptable but at least 80% is preferable to reduce the likelihood of major attrition bias.[13]

Were the participants in all groups followed up and data collected in the same way?

There is therefore nothing to suggest any scope for performance bias, such as more intensive follow-up of those on the study treatment. Follow-up was standardised so that data were collected in the same manner for all women at 3, 7 and 13 months. Participants self-completed diary cards and were questioned on the occurrence of adverse events. How pregnancy was confirmed is not described and it is unclear if miscarriages were included.

Making sense of the results

Did the study have enough participants to minimise the play of chance?

Statistical power is the ability of a clinical trial to detect a difference between groups when a true difference really exists. This study was powered to detect a 7% difference in bleeding between women treated with DSG as opposed to LNG. But it was not powered to detect important differences in pregnancy outcomes. Thus at the outset of the trial any differences found in relation to pregnancy outcomes were likely to be attributable to chance.

How are the results presented and what is the main result?

The presentation in the paper of data on vaginal bleeding patterns is complex and largely restricted to two subgroups (POP 'switchers' and 'starters'). Such subgroup analyses can be misleading and require cautious interpretation.[14] Such analyses need to be stated in advance, relate to the primary outcome measure and be limited in number. When a trial fails to detect a statistically important effect overall (as this trial did) subgroup analysis should not influence the conclusions drawn.[14]

For the assessment of the trial primary outcome of bleeding patterns, the discontinuation rate for both POPs over 13 months' follow-up was very high. Overall 44% stopped the POP to which they were originally randomised. Discontinuation due to 'irregular bleeding' was an absolute 4.5% higher in the DSG group than the LNG group (Table 1). This means that on average for every 23 patients allocated to DSG (instead of LNG), one additional women stopped her POP before 13 months due to bleeding problems. There was little difference between POPs with respect to all adverse events (around 40%), or serious adverse events (around 1.5%).

The secondary outcome of contraceptive efficacy was assessed by the occurrence of pregnancy during the 13 months of the trial 'in-treatment' period. The pregnancy rate from the trial was based on 'women-years' (Table 1). A woman taking a POP for 6 months would be counted as having contributed 6/13 (or 0.46) women-years over the 13-month follow-up period. The pregnancy rate was low for both groups: 0.41 pregnancies per 100 women taking DSG compared with 1.55 pregnancies per 100 women taking LNG (both averaged for 1 year of POP consumption) (Table 1).

The absolute difference between these two rates is 1.14 per 100 women (this is simply derived by subtracting one from the other). The number of women needed to treat (NNT) in order to avoid one pregnancy is thus 88 (i.e. 100 divided by 1.14, the 'reciprocal of the absolute difference'). Hence the trial indicates that some 88 women need to take DSG for 1 year (instead of LNG) in order to avoid one additional pregnancy (that would occur with LNG).

How precise are these results?

At the outset the trial was powered to detect a 7% difference between the two groups. It was not large enough to confirm that the actual difference found (4.5%) was due to the play of chance finding. The NNTs shown (Table 1) all include 'infinity (∞)', indicating that the results are compatible with there being no real difference between these two POPs. The lower limit of the NNTs puts a lower limit on the assessment of differences between these two POPs. For example, the NNT to prevent pregnancy is unlikely to be better than 29 women taking DSG for 1 year (instead of LNG) to prevent one additional pregnancy. Similarly, the NNT to cause irregular vaginal bleeding is unlikely to be worse than 11 (Table 1).

Were all the important outcomes considered so the results can be applied?

This question is largely concerned with whether you can apply the findings of the study locally. The trial participants represent a selective and more highly motivated group of woman than those seen in routine clinical practice and you therefore might not anticipate equally good Pearl indices in your population. Clearly you would not require any additional equipment or skills (e.g. counselling) in order to change your prescribing of POPs. However, cost inevitably becomes a consideration if you did wish to add it to your

repertoire. Based on the current *British National Formulary*, a year's supply of DSG would cost £115 compared with £9 for LNG.[15] A potentially lower failure rate for DSG would benefit women and result in health care savings which might outweigh the increased prescription costs.

One limitation of the trial was that it did not assess the influence of these POPs on the quality of life of the women in the study. The issue of bleeding patterns is only important insofar as it can affect women's quality of life. A further consideration is the safety of DSG, which has been associated with an increased risk of thromboembolism (pulmonary embolus and deep vein thrombosis) compared with LNG when used in the COC.[16] However, current evidence suggests that POPs are not associated with venous thromboembolism.[17]

Conclusions

The study has several strengths in that it was a double-blind RCT with a standardised assessment of participants who were managed in a similar manner with the exception of exposure to different POPs. Its weaknesses are a lack of description of the randomisation process, which is central to any RCT, and the very high attrition rates. For the primary outcome measure of vaginal bleeding DSG appears to cause more irregular bleeding that results in its discontinuation, whilst LNG appears to have a higher risk of pregnancy. But the trial was not large enough to exclude the play of chance as the explanation for both of these findings.

Strongly-held beliefs or conflicts of interest may influence the interpretation of study results. This can relate to both clinicians and investigators. Pharmaceutical industry-funded trials tend to report more favourable findings than those funded by other means and one of the trials authors is affiliated to the company that manufactures Cerazette.[18] After reading this paper, you are reminded of the old Chinese proverb:

谨慎求愿; 梦幻可成真。

"Be careful what you wish for; it may come true."

Sometimes, authors and readers of papers find what they hope to find in the interpretation of study findings.

Resolution of the clinical scenario

On the evidence considered above – which appears to be the only trial of its kind in the world literature – you cannot tell if the DSG pill is superior or inferior to other POPs. From the evidence it appears that there may be a trade-off between higher protection against pregnancy with a higher risk of irregular bleeding sufficient to provoke discontinuation.

About randomised trials

This article covered only selected aspects of RCTs. For further accounts, we suggest the article by Greenhalgh on how to read papers about drug trials[19] or the articles by Guyatt and colleagues.[2,3]

Acknowledgements

The authors wish to thank Yuet Wan and colleagues from the London-South Thames Medicines Information Service for the use of their appraisal in the initial drafting of this article and Xui-Ping Li for translation.

Statements on funding and competing interests

Funding. None identified.

Competing interests. Mike Crilly has received payment from the 'Oxford Centre for EBM' for tutoring on their workshops.

References

1 Collaborative Study Group on the Desogestrel-containing Progestogen-only Pill. A double-blind study comparing the contraceptive efficacy, acceptability and safety of two progestogen-only pills containing desogestrel 75 µg/day or levonorgestrel 30 µg/day. *Eur J Contraception Reprod Health Care* 1998; **3**: 169–178.

2 Guyatt GH, Sackett DL, Cook DJ. Users' guide to the medical literature. II. How to use an article about therapy or prevention. A. Are the results of the study valid? *JAMA* 1993; **270**: 2598–2601.

3 Guyatt GH, Sackett DL, Cook DJ. Users' guide to the medical literature. II. How to use an article about therapy or prevention. B. What are the results and will they help me in caring for my patients? *JAMA* 1994; **271**: 59–63.

4 Crilly M, Foy R. Evidence-based family planning: finding answers to clinical questions. *J Fam Plann Reprod Health Care* 2003; **29**(2): 48–52.

5 Cochrane AL. *Effectiveness and efficiency. Random reflections on health services.* London: Nuffield Provincial Hospitals Trust, 1972.

6 Samaan SA, Crawford MH. Estrogen and cardiovascular function after menopause. *J Am Coll Cardiol* 1995; **26**: 1403–1410.

7 Kunz R, Oxman AD. The unpredictability paradox: review of empirical comparisons of randomised and non-randomised clinical trials. *BMJ* 1998; **317**: 1185–1190.

8 Barrett-Connor E. Hormone replacement therapy. *BMJ* 1998; **317**: 457–461.

9 Grady D, Herrington D, Bittner V, et al. Cardiovascular disease outcomes during 6.8 years of hormone therapy. Heart and Estrogen/Progestin Replacement Study Follow-up (HERS II). *JAMA* 2002; **288**: 49–57.

10 Writing Group for the Women's Health Initiative Investigators. Risks and benefits of estrogen plus progesterone in healthy post-menopausal women. *JAMA* 2002; **288**: 321–333.

11 Schulz KF. Subverting randomization in controlled trials. *JAMA* 1994; **274**: 1456–1458.

12 Schulz KF, Chalmers I, Hayes R, et al. Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA* 1995; **273**: 408–412.

13 Sackett DL, Richardson WS, Rosenberg W, et al. *Evidence-based medicine: how to practice and teach EBM*. London: Churchill Livingstone, 1997.

14 Assmann SF, Pocock SJ, Enos LE, et al. Subgroup analysis and other (mis)uses of baseline data in clinical trials. *Lancet* 2000; **355**: 1064–1069.

15 *British National Formulary*, No. 45. London: British Medical Association and the Royal Pharmaceutical Society of Great Britain, March 2003.

16 Jick H, Jick SS, Gwewich K. Risk of idiopathic cardiovascular death and non fatal venous thromboembolism in women using oral contraceptives with differing progestogen components. *Lancet* 1995; **346**: 1589–1593.

17 World Health Organization. Cardiovascular disease and use of oral and injectable progestogen only contraceptives and combined injectable contraceptives. Results of an international, multicentre, case control study. *Contraception* 1998; **57**: 315–324.

18 Friedberg M, Saffran B, Stinson TJ, et al. Evaluation of conflict of interest in economic analyses of new drugs used in oncology. *JAMA* 1999; **282**: 1453–1457.

19 Greenhalgh T. How to read a paper: papers that report drug trials. *BMJ* 2003; **315**: 480–483.