Statistical models for data from the alternating sequence design.

NOTES, 2025.07.15

This is a very 'niche' topic, but I liked it for its many teaching points, and also for the way, in the simpler case, Claire Weinberg was able to trick GLIM (or any glm software today) into fitting a (very natural and clever) random effects model.

I remember Sholom Walcholder, who was a classmate of hers at U Washington, telling me that Claire was very proud of showing how the beta-geometric model behaved as the successive cycles unfolded, and how frustrated she was at Sholom for not having looked at or taken a keener interest in her paper!)

But I suspect that niche / esoteric nature also made it harder to publish in mainstream statistical journals. Indeed, as I now go back over the various versions of this manuscript, I find that eventually we went to a speciality medical journal.

I don't seem to have kept copies of the reviews we got, but it seems we eventually gave up.

One other thing that hampered us was the availability of real data. The area of assisted reproduction is quite competitive, and quite 'commercial' and so authors are not inclined to share data (or for that matter, to commit to a protocol that keeps them from experimenting as they go).

I was the one who came upon this topic, and then brought in as collaborators experts in random effects modelling (RP), MCMC methods (ND), and gynaecology/obstetrics (MHM).

If people wish to 'rejuvenate' this project, I would be delighted to hear from them, or just to see them take it further on their own.

Sincerely,

James Hanley
webpage: https://jhanley.biostat.mcgill.ca | email: `james.hanley@mcgill.ca`

SUBMITTED TO STATISTICS IN MEDICINE ***JULY, 2003***

Data-analysis options for comparisons of assisted reproductive technologies

JAMES A. HANLEY*, NANDINI DENDUKURI,
ROBERT PLATT, AND
MARIE-HéLèNE MAYRAND

McGill University, Department of Epidemiology and Biostatistics,
1020 Pine Avenue West, Montreal, Quebec, Canada
James.Hanley@McGill.CA

SUMMARY

In the 'alternating sequence design' used to compare success rates with assisted reproductive technologies, women or couples are randomized to receive either the standard or experimental treatment in the first cycle, and -- if they do not become pregnant-- crossed between standard and experimental treatments after each successive cycle. Norman and Daya (*Fertility and Sterility*, 2000) have shown that, in the presence of heterogeneity of fertility, and an effective treatment, the overall efficacy of the experimental treatment is *overestimated* by this design. They advised that in order to achieve an accurate estimate of efficacy, the trial should be run for at least three cycles and all data from even-numbered cycles be excluded from the analysis, which should then be restricted only to odd-numbered cycles. In this paper, we describe approaches that make use of the data from *all* cycles to produce estimates that are both less biased and more precise. The methods are generalizations of those applicable to the 'constant sequence' design, where naive methods that do not take account of the heterogeneity produce *underestimates* of treatment efficacy.

*Corresponding author. telephone +1 (514) 398 6270; fax: +1 (514) 398 4503.

Short title : : Trials of assisted reproductive technologies

Keywords: fertility; experimental design; bias; precision; heterogeneity; generalized linear models.

---

* to whom correspondence should be addressed.
Tel No. +1 (514) 398-6270 Fax No. +1 (514) 398-4503

1. INTRODUCTION

Two different experimental designs have been used to evaluate the efficacy of assisted reproductive technologies (Daya et al. 1993). One is the parallel-design or 'constant-sequence' randomized trial, in which the experimental treatment is administered for one or -- if unsuccessful -- more cycles to a fraction (usually one half, randomly chosen) of the eligible patients, and the control treatment for the same number of cycles to the remaining fraction. The other is the 'alternating-sequence' design, in which some of the women or couples are randomized to receive the standard, and the others to receive the experimental treatment in the first cycle. Those who do not become pregnant are crossed to the opposite treatment after each successive cycle.

The relative merits of these two designs have been keenly debated (Daya 1993; Khan *et al.* 1996; te Velde 1998; Cohlen *et al.* 1998; Daya 1999, Norman and Daya 2000). Some of the arguments focus on efficiency and sample sizes: if the experimental therapy is effective, the alternating design results in more pregnancies than the constant sequence design, and is more attractive to couples. Others have to do with possible biases in the resulting estimates of efficacy. The first suggestion of bias came from comparisons of results of actual trials that used the two different designs to evaluate the same procedure (Khan *et al*., 1996). The authors noticed that, relative to those seen in parallel trials, treatment effects of the more effective treatment were higher in -- i.e. overestimated by -- crossover trials. Subsequent Monte Carlo evaluations (Cohlen *et al*., 1998), simulating patients from a heterogeneous subfertile population, indicated that while results from parallel trials appeared to slightly *under*estimate efficacy, the alternating sequence design did indeed seem to slightly -- but in their opinion not materially -- *over*estimate it. Thus, they advised that "because of its practical advantages and because more pregnancies are achieved, a crossover design should be the first choice in infertility research"

The clearest understanding of the exact origin, nature and extent of the biases of estimates from these two designs is found in the calculations of Norman and Daya (2000). As shown in the first row of Table 1, they assumed a heterogeneous population, where fecundability i.e., the per-cycle

probability of getting pregnant, varied from couple to couple. To simplify matters, they assumed that with the less effective (Control) treatment, 80% of couples had, at each cycle, a 10% probability of becoming pregnant and the remaining 20% of couples had a 40% probability, i.e.,

$$\text{fecundability} = \begin{matrix} 0.1 \text{ for } 80\% \text{ of couples} \\ 0.4 \text{ for } 20\% \text{ of couples} \end{matrix}$$

Thus the overall average fecundability is 16%, and the standard deviation is 12%. They further assumed that the more effective experimental therapy had a constant 'relative risk' or Efficacy ("*E*") of 2, i.e., that at each cycle, a couple's probability of becoming pregnant was doubled. Using *expected*, rather than *random* numbers of pregnancies at each cycle, they showed that the estimates of the efficacy of the more effective treatment from *both* designs are biased -- the parallel design *underestimates* (apparent efficacy: $E = 1.83$, calculations not shown here but discussed later) and the alternating sequence design *overestimates* (apparent efficacy $E = 2.10$, middle column Table 1). However, they noted that the bias in the alternating design is limited to the data from even-numbered cycles.

-- Table 1 about here --

Despite the greater bias in the parallel design, Norman and Daya limited discussion of their concerns to -- and aimed their cautions at proponents of -- the alternating sequence design. They suggested a compromise between patient preference for this design and the statistical bias: "The objective of obtaining an accurate estimate of the effect of treatment, but also allowing all subjects to have the opportunity to receive the experimental treatment in at least one cycle, can now be achieved with the alternating-sequence design trial. The proviso is that the trial should run for at least three cycles and all data from the even-numbered cycles would have to be excluded from the analysis, which would be restricted only to the odd-numbered cycles." They concluded by advising that "When multiple cycles of treatment are undertaken to evaluate the efficacy of infertility therapy, the alternating-sequence design with restriction of the analysis to only the odd-numbered treatment cycles provides an unbiased estimation of the treatment effect".

This bias-avoiding strategy is unlikely to be an acceptable option for most investigators, patients and ethics review committees, and prompts the obvious questions: must we discard 'biased' cycles and compensate for the decreased precision by increasing the numbers of couples enrolled? if we know the form of the bias, why can't we remove it statistically?

The purpose of this paper is to do just that. We describe two approaches; both use only the 'aggregated by cycle' data. They draw on, and adapt when necessary, a dispersed statistical literature that -- unfortunately -- does nor seem to have 'crossed over' into fertility research. For illustration, we will first use the same simulated data used by Norman and Daya (Table 1).

2. PREAMBLE: HOMOGENEOUS FECUNDABILITY

Let p_C denote a selected woman's fecundability i.e., her per-cycle probability of getting pregnant, with the standard treatment (t=0). For now, assume that there is no variation in p_C across women, i.e., that Var[p_C] = 0. Let p_E denote this woman's fecundability with the experimental treatment (t=1). The comparison between p_E and p_C can be expressed in different ways by selecting different forms for the function g in the regression equation $g[\,p_E\,] = g[\,p_C\,] + \beta \times t$. For example, β is the absolute difference in fecundability if we select $g[\,]$ to be the *identity* function; $\exp[\beta]$ is the fecundability ratio if g is the *ln* function, or the fecundability odds ratio if g is the *logit* function. For this paper, we, like Norman and Daya use the *fecundability ratio* scale, whereas Cohlen *et al* used the odds ratio scale.

Suppose that one such woman, alternating from the experimental treatment in cycle 1, became pregnant on this treatment in the 5th cycle.

Cycle:	1	2	3	4	5
Treatment:	Experimental	Control	Experimental	Control	Experimental
Outcome:	—	—	—	—	+
Probability(+) :	p_E	p_C	p_E	p_C	p_E
Probability(Outcome):	$1 - p_E$	$1 - p_C$	$1 - p_E$	$1 - p_C$	p_E

+ or — denotes whether woman did or did not become pregnant in that cycle

The observed data can be modeled as a sequence of independent Bernoulli trials with alternating probabilities of success. The likelihood is the product of the probabilities of the 5 individual outcomes; it can also be re-arranged and written as a product of two binomial-like likelihoods, corresponding to $s_C = 0$ successful cycles, preceded by $u_C = 2$ unsuccessful ones, when the success probability is p_C ; and $s_E = 1$ successful cycle, preceded by $u_E = 2$ unsuccessful ones, when the success probability is p_E , i.e.,

$$L \propto (1 - p_C)^2 \times (1 - p_E)^2 p_E$$

If p_C and p_E are constant from woman to woman, so that all woman-cycles within the same treatment condition are exchangeable, then the likelihood based on the data from *several* such women can again be written as the product of two binomial-like likelihoods

$$L \propto (1 - p_C)^{U_C} p_C^{S_C} \times (1 - p_E)^{U_E} p_E^{S_E}$$

where $U_C = \Sigma u_C$ and U_C and S_C and S_E are the TOTAL numbers of unsuccessful and successful cycles on C and E respectively, i.e., summed over all women and all cycles. They lead to simple closed-form MLE point estimators of p_C and p_E , namely the total numbers of pregnancies divided by the total numbers of cycles, and likelihood-based interval estimates for any of the comparative parameters defined by g[ ] above.

3.1 HETEROGENEOUS FECUNDABILITY

In reality, fecundability with the standard treatment *does* vary across the source population of women, i.e., var[p_C] > 0. We denote this variation by the probability distribution function f[p_C]. In their example, Norman *et al.* took p_C to have a 2-point distribution.

IF the latent p_C's and p_E's of the n women studied could be known, the likelihood would simply be

$$L \propto \Pi \{ (1 - p_C)^{u_C} p_C^{s_C} \times (1 - p_E)^{u_E} p_E^{s_E} \}$$

with the product, Π, taken over the *n* women studied. However, since these 'parameters' are not known, the likelihood also involves the parameters of f[ ]. Although it is easy to write the likelihood, it involves an n-dimensional integral over the latent fecundability values for each woman, and so ML estimators of the model parameters, and of the relevant *comparative* parameter, no longer have a closed form.

Not counting the inefficient method produced by Norman and Daya, two broad data-analysis approaches can be used. The *first* of these, the subject of this paper, uses the row-by row data summaries in Table 1, i.e., it uses the data, aggregated over women, for each treatment in each cycle, to estimate an efficacy ratio. This approach does not incorporate woman-specific and woman-cycle-specific covariates. In section 3, we describe two such methods to estimate a fecundability *ratio,* assumed to be common over women, from *aggregated* data. One method makes no assumptions about the form of f[ ]. The other is based on a specific parametric model for f[ ] -- the beta-geometric. Using only standard software for generalized linear models, Weinberg and Gladen were able, using a helpful re-parametrization, to obtain the ML estimates of its parameters from data with the same structure as those generated by the constant-sequence design. Here, we propose an ad-hoc modification to their method that allows us to accommodate data from the alternative-sequence design.

The *second* broad approach uses as the unit of analysis the (Bernoulli) *data from each separate woman-cycle*. This much more general approach will be described in a separate paper.

3.2 METHODS

Unspecified-form for f[p]: Let E denote the ratio of p_E to p_C for all values of p_C. Table 2 gives, for each of the first three cycles, for *any* distribution f[p_C], and any legitimate *E*, the expected 'numerators' and 'denominators' defining the proportion of those entering the cycle who become pregnant in that cycle. The extension to cycles 4 and beyond is obvious, although the algebra becomes tedious. Of note is the fact that the two success proportions in cycle '*k*' involve the first *k*

moments of the distribution of p_C. Other authors (e.g., Lau, 1996) have also noted this in the simpler, constant-condition or constant-sequence, situation. Thus, each cycle adds two new data points and one new parameter, so that the total of 2*K* datapoints from *K* cycles can be modeled by K + 1 parameters. Thus, if *K* is 3 or more, the remaining *K*-1 degrees of freedom can be used to assess the fit of the model.

-- Table 2 about here --

Since it is difficult to fit the *K*+1 parameters (*E*, and the moments p_1 to p_K) using standard statistical software, we obtained MLE's of the parameters by numerically finding the roots of the derivatives of the log likelihood. We used *Mathematica* (program available from http://www.epi.mcgill.ca/hanley/ software/alt_seq.html) to do so. From the information matrix, we obtained the standard error for the estimate of *E*. The results of this procedure, applied to the data in Table 1, are shown in Table 3. Like the procedure of Norman and Daya, this method correctly 'recovers' *E*. But -- because it uses data from *all* cycles -- it produces smaller standard errors. Thus, this increased precision can be achieved without having to forego accuracy, even if higher order moments (of order 3 or more in our example) -- of decreasing magnitudes, since p is bounded by 0 and 1 -- are omitted (i.e., set to zero in the likelihood).

-- Table 3 about here --

There are considerable practical technical difficulties in fitting such a high-order nonlinear model; moreover, the number of parameters (moments) relative to the numbers of observations is large, and the software is inaccessible to most end-users. In order to provide a method that could be implemented in mainstream statistical packages, we modified the generalized linear model suggested by Weinberg and Gladen(1986), which is based on a *specific*, but natural, *distributional form* for f[*p*].

Adaptation of Weinberg and Gladen's 'Beta-Geometric' Generalized Linear Model: Before describing our adaptation, we revisit the simpler parallel design considered by Weinberg and

Gladen(1986). They were concerned with fecundability, measured over as many as 12 cycles, in smokers relative to non-smokers, i.e., with data analogous in structure to that for the 'constant-sequence' experimental design. They took f[p], the fecundability distribution in non-smokers, to be a Beta distribution, with the location and shape governed by the two traditional parameters a and b. These parameters correspond to $\mu_p = a/(a+b)$ and $\sigma_p^2 = ab/((a+b)^2(a+b+1))$. They showed that in those (previously unsuccessful) couples who enter cycle k, the -- *now conditional* -- distribution of p is shifted towards zero but *remains a Beta distribution*, now with parameters a and $b + (k\text{-}1)$. By re-expressing the parameters a and b, they further showed that the mean probability p of success in this cycle among those who enter this cycle, which we denote as $p_{[\text{cycle}]}$, is related to the number of previously unsuccessful cycles (cycle -1) via the simple reciprocal link:

$$1 / p_{[\text{cycle}]} = c + d\,(\text{cycle} - 1).$$

The parameter c is the expected number of cycles to become pregnant for couples with a per-cycle probability of $p_{[1]}$ i.e., $c = 1 / p_{[1]}$. The parameter d measures the spread of the *initial* distribution of p (if there is no heterogeneity, the number of cycles to pregnancy reduces to the same geometric random variable for each couple). Thus, the parameters of this specialized model can be fit to the datapoints (successes in cycle/number who undergo cycle) in any software which allows binomial regression with an inverse (i.e., power^{-1}) link. Using indicator variables and product terms, Weinberg and Gladen extended the model to fit two *separate* Beta distributions for the probabilities among smokers ($I_{\text{smoking}}=1$) and non-smokers($I_{\text{smoking}}=0$), via a single equation i.e.,

$$1 / p_{[\text{cycle, group}]} = c_1 + d_1\,(\text{cycle} - 1) + (c_2 - c_1)I_{\text{smoking}} + (d_1 - d_2)(\text{cycle} - 1)I_{\text{smoking}}\ .$$

Consider now the *alternating-sequence* design, where the *initial* distribution of p_C in the source population is again Beta[a,b]. Conditionally on entering the different cycles, the expected proportions who become pregnant in these cycles do *not* follow the simple inverse linear patterns above. But they do have a *predictable*, if slightly imperfect mathematical, pattern: As an example, Figure 1 shows the inverses of the expected proportions under an initial Beta[3,15] distribution of p_C, and a constant fecundability-ratio E=2. The parameter values a=3 b=15 used to generate the

sequences of proportions yield a mean[p_C] = 1/6 = 0.17 and a SD[p_C] = 0.085, somewhat 'tighter' than the SD of 0.12 in the 2-point distribution used by Norman and Daya. These values correspond to the values $c=(a+b)/a = 6$ and $d = 1/a = 1/3$ in Weinberg and Gladen's representation.

The pattern in Figure 1 matches that seen earlier in Table 1: the ratios of the inverse proportions (and thus the proportions themselves) are constant at E=2 in odd-numbered cycles, and greater than 2 in even-numbered cycles -- but to a *decreasing* extent with *increasing* cycles. This pattern suggests, as a *pragmatic* approximating model, the following generalized linear model,

$$1 / p_{[\text{cycle, condn}]} = c_1 + d_1 (\text{cycle} - 1) + (c_2 - c_1)\text{I} + (d_1 - d_2)(\text{cycle} - 1)\text{I} + g_1 z + g_2 \text{I} z .$$

As before, the indicator I indicates the experimental condition. The regressor

$$z = \begin{cases} 1/(\text{cycle}-1) & \text{in even-numbered cycles} \\ 0 & \text{otherwise.} \end{cases}$$

and the corresponding regression coefficients g_1 and g_2 allow decreasing 'corrections' for the even-numbered cycles (to conserve degrees of freedom, one may wish to use opposite signs, and a single g for z). The model can again be fit using any Bernoulli regression software that allows the reciprocal link. The estimate of E is obtained from the ratio c_1 / c_2 since $c_1 = 1 / p_{[1]}$ and $c_2 = 1 / (E \times p_{[1]})$; the parameter estimates c_1 and c_2 can (in the absence of a subgroup where p_C=0) be interpreted as the average number of cycles to become pregnant for a (random) couple with 'average' fertility, if they remained on treatment 1(C) or 2(E) throughout. In the example, the best fit of the above equation to the 5 sets of proportions in Table 1 yields c_1 = 6.01 cycles and c_2 = 3.01 cycles, to give, to 2 decimal places, $E = (1/c_2) / (1/c_1)$ = 1.99 (see Table 4).

4. EFFICACY ESTIMATES FROM A CLINICAL TRIAL

We compared the various estimates in the context of the sample sizes, fertility rates, and logistical constraints encountered in practice. We did so by applying them to data from a study that evaluated whether a second generation protocol improved the success rate with donor insemination that used frozen semen (Brown *et al*., 1988). This is an important contemporary issue, since the time-window

needed to screen semen for HIV infection now precludes the use of fresh semen. Earlier, in 1984, in what seems to have been the first documented use of the alternating sequence design, the same team had achieved a fecundability rate of only 5.0 pregnancies per 100 cycles with a first-generation protocol, versus 18.9 per 100 using fresh semen (Richter *et al*., 1984]).

The alternating sequence design was again used in the second-generation protocol. Again, for each woman, the semen was from her matched donor for the first six cycles, and if pregnancy was not achieved by then, the donor was changed. The data for the first six cycles are shown in Table 6. However, as is obvious from the tabulated denominators for fresh and frozen semen cycles, the same practical difficulties were encountered as those mentioned in the first study: "Cryopreserved semen was frequently substituted in a cycle scheduled to be fresh because the donor was not available". Detailed information on which sequence was actually followed by each woman is no longer available (S. Shapiro, personal communication, 2002). .

The estimates of the relative efficacy of frozen semen are given at the bottom of Table 6. Those produced by the proposed methods are closest to the null, suggesting that these methods removed some of the bias induced when one ignores the heterogeneity of fecundability. Possibly by chance, given its poor precision, the estimator advocated by Norman and Daya was furthest form the null, further than both the crude and the Mantel-Haenszel estimates. The estimate from the Generalized Linear Model applied to the 'aggregated-by-cycle' data was the closest to the null, but had a higher SE, possibly because of the large number of parameters (6) fitted to the 12 datapoints.

5. DISCUSSION

The alternative sequence design allows investigators to recruit greater sample sizes to compare the performance of assisted reproductive technologies. The two statistical options we have presented allow clinical trials to benefit from the full efficiency of these larger sample sizes, but without sacrificing statistical accuracy. Norman and Daya's method of avoiding bias squanders the statistical advantage of this design. Given that most contemporary trials in this area are already

based on fewer than 25 pregnancies over *all* cycles, one cannot afford to decrease precision any further by omitting data from even-numbered cycles.

The degree of bias in efficacy estimates from a naive analysis is a function of two factors, the degree of heterogeneity in p and the difference in treatment efficacy; both must be substantial in order to produce a serious bias. Norman and Daya's concerns, and resulting advice, were based on an extreme 2-point distribution of p, and a treatment efficacy that *doubled* the p=40% in the high fertility subgroup to a 'biologically nearly impossible' p = 80%. Even then, the bias from using the aggregated-by-cycle data was less than the imprecision induced by the sample sizes used in practice. And, with the same E=2, the bias was much smaller when Norman and Daya took a more conservative 2-point distribution where f[0.025] = 0.8 and f[0.1] = 0.2, so that p is closer to zero, with mean[p] = 0.04, SD[p]=0.03. Moreover, if, rather than *increasing* p, an experimental treatment -- such as frozen semen -- *reduces* it, the degree of bias in the naive estimate of E is also less, because of the smaller impact of the differential removal of the most fertile at the end of each odd-numbered cycle. These 'low-bias' conditions would also apply if this design were used for comparisons of *contra*ceptive methods, where in contrast, the probability of an *un*wanted pregnancy is already low with most methods.

Some investigators place a high premium on avoiding even small biases. Others may wish to accommodate the actual sequence of treatments received, intended or otherwise. For example, in the study by Brown et al. there were unavoidable deviations from the planned alternating-sequence protocol. Yet others may use variations on the alternating design: for example, in Ecochard *et al.* (2000), half the patients received one treatment for the first two cycles and the competing treatment for the next two cycles, whereas the other half followed the opposite sequence. For data from these more complex designs, investigators will probably wish to use random effects models for binary (Bernoulli) data that allow one to model the cycle-by-cycle sequence of outcomes for each woman using a full regression approach that makes use of *all of the individual level data* for each woman

(any baseline covariates, the cycle-by-cycle treatment indicators and any other available cycle-dependent covariates).

Contrary to misconceptions, the alternating sequence design is not a full crossover study. Nor does it carry the full statistical efficiency usually associated with self-matched comparisons. Thus, the sample size and power calculations/projections are best done by analogy with unmatched designs. Another misconception is that the outcomes of an individual woman's 'multiple' cycles in the same woman are not statistically independent. Conditional on the (unmeasurable) p (and thus on $E \times p$) which is specific to that woman, the cycles *do* constitute independent Bernoulli trials with alternating probabilities.

The approaches we have described are also applicable to data analyses for the parallel- or 'constant-sequence' randomized trial design. Since this competing design has the same data structure as Gladen and Weinberg's example of pregnancy rates in smokers and non-smokers, their 'Beta-Geometric' Generalized Linear Model is immediately applicable without modification. The method based on moments is also applicable. Applied to Norman and Daya's 'constant-sequence' example (and the 'data' in their Table 1), both of our methods recover estimates closer to the true E=2, whereas a naive analysis produces an attenuated E=1.83.

In their main example, Norman and Daya assume that treatment increases p_C a constant-fold, i.e., by the same multiple, E , for all values of p_C . Just as they do, we too find it difficult to imagine that a treatment which improves p by 2-fold, from 0.1 to 0.2, with also raise the per-cycle success probability from 0.4 by the same factor to 0.8. Although, following clinical trial practice, we have used 'relative risks' (i.e., ratios of proportions) as measures of efficacy, it is preferable to allow an unrestricted range of p, and of the efficacy measure E, , by using the also more biologically plausible 'constant OR' model.

Norman and Daya claim that "the assumptions of a constant drug efficacy is not necessary". In an appendix, they purport to show algebraically that, for any distribution of fertility $f[p_C]$, and for any

value of drug efficacy, $E(p_C)$, where t*he efficacy is a function of fertility*, "the outcome rates in the odd cycles in an alternating sequence are unbiased", i.e., that "the results will hold true regardless of the relationship between efficacy and fertility." In fact, the ratio from alternate cycles will *not* continue to be unbiased in the event that the treatment effect is variable in different fecundability groups. This is illustrated in Table 6 by slightly perturbing Norman and Daya's simulated example so that the risk ratio in the low fecundability group is 2.5, while the risk ratio in the high fecundability group remains at 2. This is closer to what happens in reality where there is likely to be a greater relative shift in the low fecundability groups, compared to the high fecundability groups. The true average risk ratio across the population is thus $2.5 \times 0.8 + 2 \times 0.2 = 2.4$. However, from Table 6 we can see that this estimate is not obtained even in the first cycle. This is because the aggregate ratios at each cycle are the ratios of the expected probabilities of successes rather than the expectation of the ratios of the success probabilities, i.e.

$$\frac{\Sigma\, p\ .\ \mathrm{E}[p]\ .\ \mathrm{f}[p]}{\Sigma\ \mathrm{p}.\ \mathrm{f}[p]} \neq \Sigma\ \mathrm{f}[p].\ \mathrm{E}[\mathrm{p}]$$

Further, it appears that the odd cycles will tend to underestimate the true risk ratio while the even cycles tend to overestimate it. If the study continues to a point when only women in the low fecundability group remain, then the ratio approach the true ratio of 2.5 in both odd and even cycles

This contrary finding is an additional impetus to consider a general regression model that allows not just between-individual heterogeneity, and covariates at the woman-cycle level, but also more flexibility in the specification of the comparative parameter. We will report on our experiences with random effects models for binary data in a subsequent paper.

ACKNOWLEDGMENT

This work was supported by individual operating grants from the Natural Sciences and Engineering Research Council of Canada and a team grant from Le Fonds Québécois de la recherche sur la nature at les technologies.

REFERENCES

BROWN, C.A., BOONE, W.R., AND SHAPIRO, S.S. (1988). Improved cryopreserved semen fecundability in an alternating fresh-frozen insemination program. *Fertility and Sterility* **50**, 825-827.

COHLEN, B.J., TE VELDE, E.R., LOOMAN. C,W,N,, EIJCHEMANS, R., AND HABBEMA, J.D.F. (1998). Crossover or parallel design in infertility trials? The discussion continues. *Fertility and Sterility* **70**, 40-45.

DAYA, S. (1993). Is there a place for the crossover design in infertility trials? *Fertility and Sterility* ***59*****,** 6–7.

DAYA, S. (1999). Differences Between Crossover and Parallel Study Designs—Debate? (letter) *Fertility and Sterility* **71**, 771-772.

ECOCHARD, R. AND CLAYTON, D.G. (2000). Multivariate parametric random effect regression models for fecundability studies. *Biometrics* 56, 1023-1029.

ECOCHARD, R, MATHIEU, C., ROYERE, D., BLACHE, G., RABILLOUD, M., AND CZYBA, J.C. (2000). A randomized prospective study comparing pregnancy rates after clomiphene citrate and human menopausal gonadotropin before intrauterine insemination. *Fertility and Sterility* **73**, 90-93.

KHAN, K.S, DAYA S, COLLINS J.A., AND WALTER S.D. (1996). Empirical evidence of bias in infertility research: overestimation of treatment effect in crossover trials using pregnancy as the outcome measure. *Fertility and Sterility* **65**, 939–45.

LAU, T.S.(1996). On the heterogeneity of fecundability. *Lifetime Data Analysis* **2**, 403-415.

NORMAN, G.R. AND DAYA, S. (2000). The alternating-sequence design (or multiple-period crossover) trial for evaluating treatment efficacy in infertility. *Fertility and Sterility* **74**, 319-324.

RICHTER, M.A., HANING, R.V., AND SHAPIRO, S.S. (1984). Artificial donor insemination: fresh versus frozen semen: the patient as her own control. *Fertility and Sterility* **41**, 277-280.

SCHEIKE, T.H. AND JENSEN, T.K.(1997). A discrete survival model with random effects: an application to time to pregnancy. *Biometrics* 53, 318-329.

TE VELDE, E.R., COHLEN B.J., LOOMAN C.W., AND HABBEMA, J,D. (1998). Crossover designs versus parallel studies in infertility research. [letter] *Fertility and Sterility* **69**, 357-358.

WEINBERG, C.R. AND GLADEN B.C. (1986). The beta-geometric distribution applied to comparative fecundability studies. *Biometrics* **42**, 547-560.

ZHOU, H., WEINBERG, C.R., WILCOX, A.J., AND BAIRD, D.D. (1996a). A random effects model for cycle viability in fertility studies. *Journal of the American Statistical Association* 91, 1413-1422.

ZHOU, H. AND WEINBERG, C.R.(1996b). Modeling conception as an aggregated Bernoulli outcome with latent variables via the EM algorithm. *Biometrics* 52, 945-954.

ZHOU, H. AND WEINBERG, C.R.(1996c). Potential for bias in estimating human fecundability parameters: a comparison of statistical models. *Statistics in Medicine* 18, 411-422.

Table 1: Cycle-specific ratios of expected pregnancy proportions if the alternating sequence design is applied to a population of 2000 which is heterogeneous with respect to spontaneous fecundity [20% with higher, 80% with lower fecundity]. The course of the 1000 randomly allocated to undergo the 'control' treatment in the first cycle is tracked in bold. The entries at each cycle are the expected numbers of couples from the higher- and lower fecundity subpopulations who attempt to (and, in parentheses, the numbers who do) become pregnant. Table adapted from Figure 2 and Table 2 of Norman and Daya*.

	Treatment Received in the Indicated Cycle						
	Control				Experimental		
	Sub-population (fecundity)				Sub-population (fecundity)		
Cycle	High	Low	Combined	Ratio	Combined	High	Low
1	**200** (***80***)	**800** (***80***)	**1000** (***160***)		1000 (*320*)	200 (*160*)	800 (*160*)
			16%	2.00	32%		
2	40 (*16*)	640 (*64*)	680 (*80*)		**840** (***240***)	**120** (***96***)	**720** (***144***)
			11.8%	2.43	**28.6%**		
3	**24** (***9.6***)	**576** (***57.6***)	**600** (***67.2***)		600 (*134.4*)	24 (*19.2*)	576 (*115.2*)
			11.2%	2.00	22.4%		
4	4.8 (*1.9*)	460.8 (*46.1*)	465.6 (*48.0*)		**532.8** (***114.9***)	**14.4** (***11.5***)	**518.4** (***103.7***)
			10.3%	2.10	**21.6%**		
5	**2.9** (***1.2***)	**414.7** (***41.5***)	**417.6** (***42.6***)		417.6 (*85.2*)	2.9 (*2.3*)	414.7 (*82.9*)
			10.2%	2.00	20.4%		
1-5	271.7 (*108.7*)	2891.5 (*289.2*)	3163.2 (*397.8*)		3390.4* (*894.8*)	361.3 (*289.0*)	3029.1 (*605.8*)
			12.6%	2.10*	26.4%*		

Calculations were carried out on a spreadsheet and thus differ slightly from those of Norman and Daya. (* N & D also made an error in calculating an overall denominator of 3444.4 for the 'experimental' cycles, and thus an overall rate of 26.0% and an overall ratio of 2.06.

Table 2: The expected numerators, denominators, and success probabilities for each of the first three cycles, as a function of the efficacy, E, and the (absolute) moments of the distribution of p, the fecundability under the standard ["control" (C)] treatment. For simplicity, the subscript "C" is omitted . The results hold for *any* distribution of p, The course of those randomly allocated to undergo the 'control' treatment in the first cycle is tracked in bold.

		Intervention
Cycle	Control	Experimental
1	$\frac{\Sigma\, \boldsymbol{p \,.\, \mathrm{f}[p]}}{\mathbf{1}} = \boldsymbol{p_1}$	$\frac{\Sigma\, \mathrm{E} \,.\, p \,.\, \mathrm{f}[p]}{1} = E \,.\, p_1$
2	$\frac{\Sigma\, p \,.\, (1 - E \,.\, p).\mathrm{f}[p]}{\Sigma\ \ (1 - E \,.\, p).\mathrm{f}[p]}$ $= \frac{p_1 - E \,.\, p_2}{1 - E \,.\, p_1}$	$\frac{\Sigma\, \boldsymbol{E \,.\, p \,.\, (1 - p) \,.\, \mathrm{f}[p]}}{\Sigma\ \ \boldsymbol{(1 - p) \,.\, \mathrm{f}[p]}}$ $= \frac{\boldsymbol{E \,.\, p_1 - E \,.\, p_2}}{\boldsymbol{1 - p_1}}$
3	$\frac{\Sigma\, \boldsymbol{p \,.\, (1 - p) \,.\, (1 - E \,.\, p) \,.\, \mathrm{f}[p]}}{\Sigma\ \ \boldsymbol{(1 - p) \,.\, (1 - E \,.\, p) \,.\, \mathrm{f}[p]}}$ $= \frac{\boldsymbol{p_1 - p_2 - E \,.\, p_2 + E \,.\, p_3}}{\boldsymbol{1 - p_1 - E \,.\, p_1 + E \,.\, p_2}}$	$\frac{\Sigma\, E \,.\, p \,.\, (1 - p) \,.\, (1 - E \,.\, p) \,.\, \mathrm{f}[p]}{\Sigma\ \ (1 - p) \,.\, (1 - E \,.\, p) \,.\, \mathrm{f}[p]}$ $= \frac{E \,.\, p_1 - E \,.\, p_2 - E^2 \,.\, p_2 + E^2 \,.\, p_3}{1 - p_1 - E \,.\, p_1 + E \,.\, p_2}$

Cycle 1 starts with denominators of 1 (100%) in each group; it is assumed that there are no dropouts [i.e. women/couples who haven't yet gotten pregnant do not abandon the study] or that dropouts are 'at random' and unrelated to their values of p. Σ denotes summation or integration over the possible values of p. The symbols p_1 to p_3 are the first 3 moments of the distribution of p, the fecundability with standard treatment.

In general, p would have a *continuous* distribution. However, without loss of generality, and in order to simplify the presentation, we have taken p to be a *discrete* random variable, and use *summation*, Σ, rather than *integration*, over the distribution of p.

Table 3: Maximum Likelihood Estimates(SE) of the efficacy parameter *E*, as a function of the number of data cycles used, and the number of moments used in the estimation; comparison with method of Norman and Daya. 'Data' from Table 1.

	Method based on moments					Norman and Daya	
	Numbers of moments estimated					(odd-numbered cycles)	
Cycles used	1	2	3	4	5	*Cycles used*	Estimate (SE)
1	2.00 (0.17)					*1*	2.00 (0.17)
1, 2	2.33 (0.18)	2.00 (0.13)					
1, 2, 3	2.18 (0.14)	2.06 (0.12)	2.00 (0.12)			*1, 3*	2.00 (0.15)
1, 2, 3, 4	2.28 (0.15)	2.01 (0.11)	2.02 (0.11)	2.00 (0.11)			
1, 2, 3, 4, 5	2.18 (0.13)	2.07 (0.11)	2.01 (0.10)	2.00 (0.10)	2.00 (0.11)	*1, 3, 5*	2.00 (0.14)

Figure 1: Motivation for adapting the Weinberg-Gladen model: shown are the reciprocals of the success probabilities at each cycle, under the alternating sequence design. Smaller dots: control condition; larger dots: experimental condition. Probabilities were generated under a (null) Beta[3,15] distribution of a first-cycle probabilities, and an 'efficacy' of E=2. Success probabilities themselves are shown at left. Cycle-specific ratios are 2, 2.18, 2, 2.13 and 2. Deviations of the reciprocals in even-numbered cycles from the virtually linear pattern of the reciprocals in odd-numbered cycles are a decreasing function of cycle.

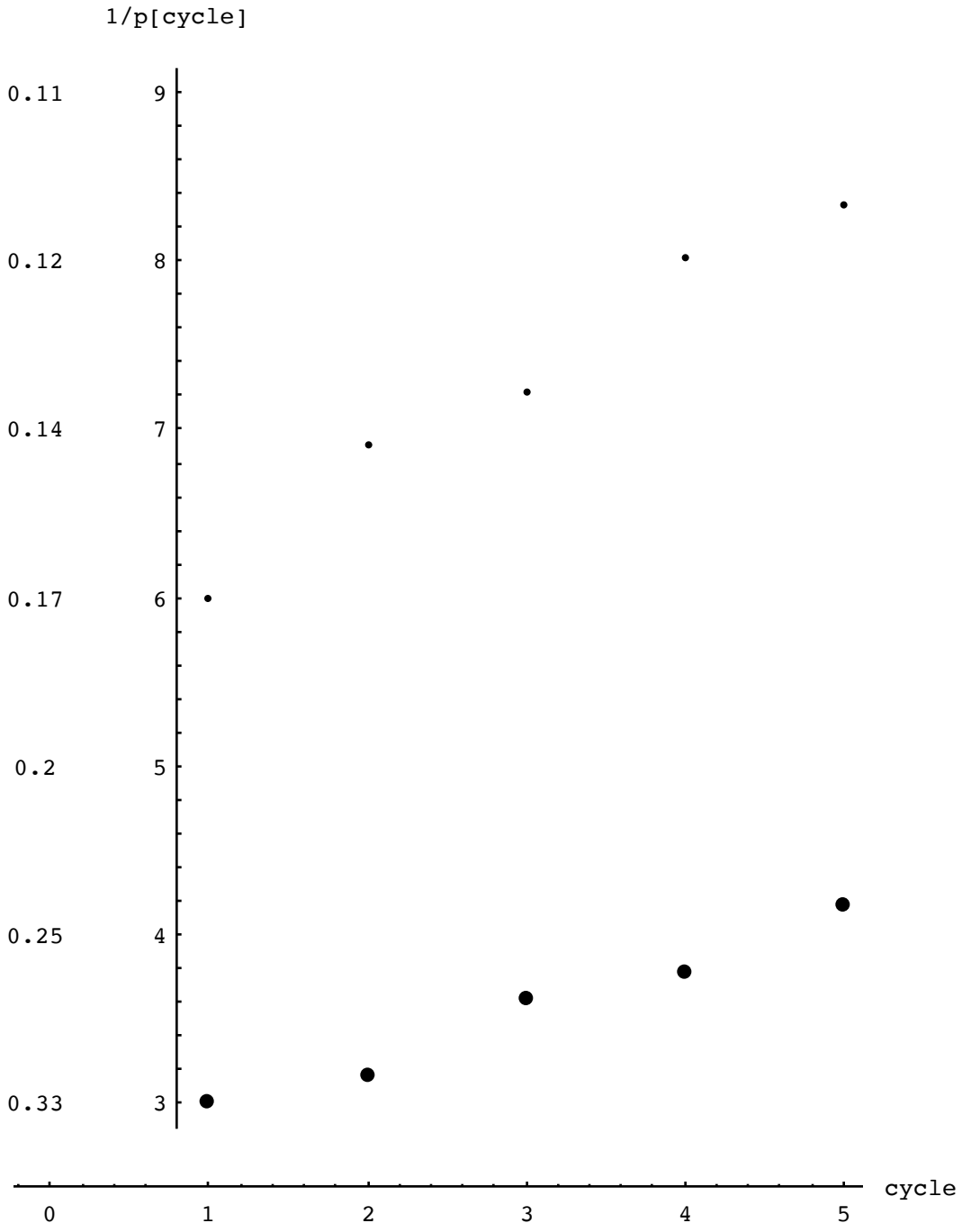


Table 4 Fits of adapted Weinberg-Gladen generalized linear model to Norman and Daya 'data' (a) Single Model for both treatments (b) Model fit separately for each treatment

(a) Single Model

Parameter	Estimate (SE)
INTERCEPT	6.32 (0.45)
Cycle -1	1.02 (0.30)
(Expt'l) Treatment	-3.16 (0.48)
Treatment*(Cycle -1)	-0.50 (0.32)
Z (see footnote 1)	1.15 (0.98)
Treatment*Z	-1.33 (1.01)

(b) Separate models[2]

	Control	Expertimental	Ratio(SE of *ln* Ratio)
INTERCEPT	6.33 (0.45)	3.15 (0.14)	$2.00(0.08^2)$
Cycle -1	1.02 (0.30)	0.52 (0.10)	
Z	1.15 (0.98)	-0.18 (0.23)	

---

1. Z = 1/(cycle-1) if even-numbered cycle, 0 otherwise

2. The ratio estimate, from the single model, is 6.32/(6.32-3.16); since the covariance between the 6.32 and (6.32-3.16) is virtually zero, the variance of the ratio can be simplified by fitting separate models for control and experimental cycles. The variance of *ln* ratio can therefore be computed as

$$[\ (0.45/6.33)^2 + (0.14/3.15)^2\ ]^{1/2} = \ 0.08$$

Table 5: Pregnancies and Fecundability in Cycles of Insemination with Fresh and Frozen Semen, Together with Estimates of Efficacy. The course of the patients who underwent insemination with fresh semen in the first cycle is tracked in bold. Data from Table 1 of Brown et al (1986).

	Fresh semen			Frozen semen		
	Number of			Number of		
Cycle	Patients	Pregnancies	Fecundability	Patients	Pregnancies	Fecundability
1	**163**	**57**	**0.350**	125	18	0.144
2	69	18	0.261	**130**	**12**	**0.092**
3	**73**	**20**	**0.274**	87	8	0.092
4	59	12	0.203	**69**	**9**	**0.130**
5	**51**	**12**	**0.235**	50	1	0.020
6	51	12	0.235	**28**	**2**	**0.071**
1-6	466	131	0.281	489	50	0.102

Efficacy (SE), estimated from ...	**Frozen / Fresh**
"Crude": 50/489 = 0.102 ÷ 131/466 = 0.281	0.36(0.14)
Odd-numbered cycles [27/262 ÷ 89/287]	0.33(0.18)
Mantel-Haenszel Risk Ratio [strata: cycles]	0.37(0.15)
Method I [4 moments]	0.39(0.15)
Adaptation of beta-binomial [2.88/7.00*]	0.41(0.24)[1]

(SE) is the SE of the *ln* of the ratio (in some cases, back-calculated from SE of ratio itself)

[1] See footnote 2 to Table 4..

Deviance / df = 0.95; Chi-square goodness of fit statistic = 5.3 (6 df).

Table 6: Simulated example with varying risk ratio in each fecundability group; otherwise, same setup as in Table 1.

	Treatment Received in the Indicated Cycle						
	Control				Experimental		
	Sub-population (fecundity)				Sub-population (fecundity)		
Cycle	High	Low	Combined	Ratio	Combined	High	Low
1	**200 (*80*)**	**800 (*80*)**	**1000 (*160*)**		1000 (*360*)	200 (*160*)	800 (*200*)
			16%	2.25	36%		
2	40 (*16*)	600 (*60*)	640 (*76*)		**840 (*276*)**	**120 (*96*)**	**720 (*180*)**
			11.8%	2.79	**32.9%**		
3	**24 (*9.6*)**	**540 (*54*)**	**564 (*63.6*)**		564 (*154.2*)	24 (*19.2*)	540 (*135*)
			11.2%	2.43	27.3%		
4	4.8 (*1.9*)	405 (*40.5*)	409.8 (*42.4*)		**500.4 (*133*)**	**14.4 (*11.5*)**	**486 (*121.5*)**
			10.3%	2.57	**26.6%**		
5	**2.9 (*1.16*)**	**364.5 (*36.45*)**	**367.4 (*37.61*)**		367.4 (*93.429*)	2.9 (*2.32*)	364.5 (*91.125*)
			10.2%	2.48	25.4%		
1-5	271.7 (*108.7*)	2709.5 (*270.6*)	2981.2 (*379.3*)		3271.8 (*1016.625*)	361.3 (*289.0*)	2910.5 (727.625)
			12.7%	2.45	31.1%		

Design and data analysis options for comparisons of assisted reproductive technologies

James Hanley, Nandini Dendukuri, Robert Platt, Marie-Hélène Mayrand

Dept of Epidemiology and Biostatistics, McGill University

Background

Statistical Methods

Results

Applications

Lessons

BACKGROUND

Two experimental designs to evaluate efficacy of assisted reproductive technologies [Daya et al. *Fertility and Sterility*, 1993;59:6–7].

Parallel-design or 'constant-sequence' randomized trial

Experimental treatment administered for one or -- if unsuccessful -- more cycles to a fraction (usually 1/2, randomly chosen) of the eligible women/couples

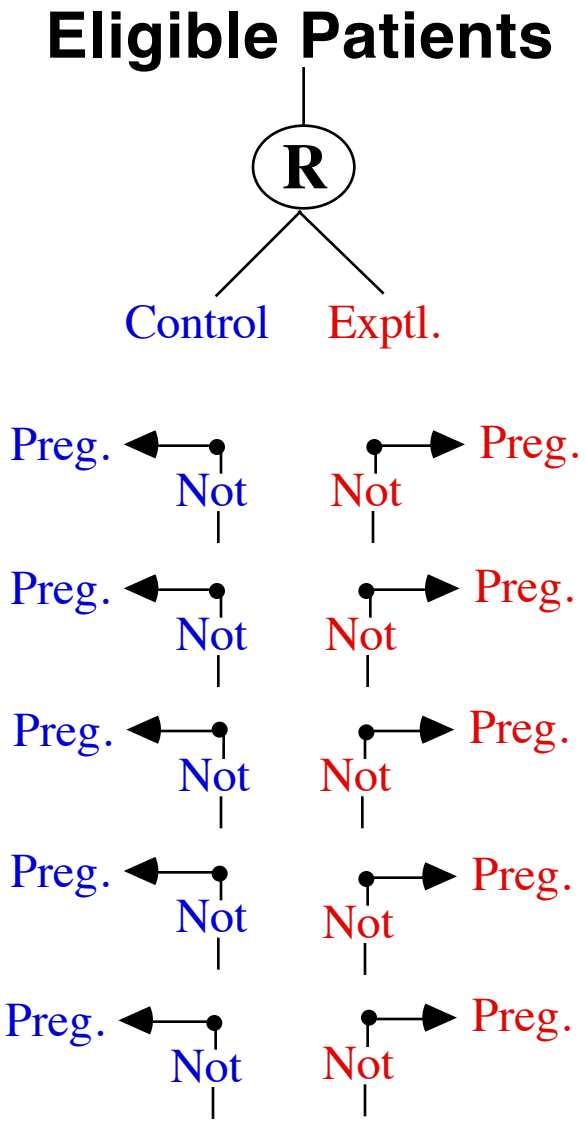
'Control' or 'standard' treatment for the same number of cycles to the remaining fraction.

'Alternating-sequence' design

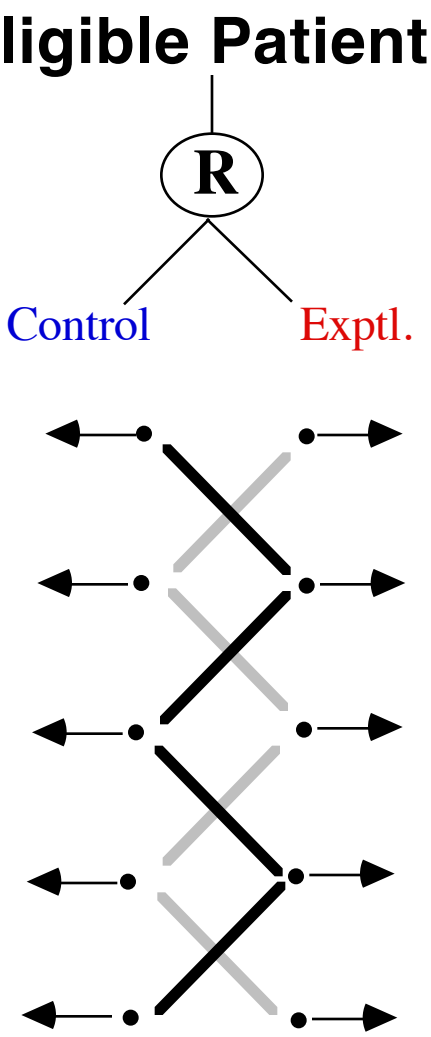
1/2 of women/couples randomized to receive standard, and the other 1/2 the experimental treatment in 1st cycle.

Those who do not become pregnant are crossed to the opposite treatment after each successive cycle.

"Parallel"



Alternating Sequence



Relative merits of 2 designs

Daya 1993; Khan et al. 1996; te Velde 1998; Cohlen et al. 1998; Daya 1999, Norman & Daya 2000

Arguments

- efficiency & sample sizes
- attractiveness of alternating design to couples
- possible biases in the resulting estimates of efficacy

  - Comparisons of results of trials that used different designs to evaluate same procedure [Khan et al., 1996]

    relative to those seen in parallel trials, effects of more effective treatment higher in -- i.e., overestimated by -- crossover trials

  - Monte Carlo evaluations [Cohlen et al., 1998]

    simulated patients from heterogeneous subfertile popln.

    - results from parallel trials appeared to be unbiased
    - alternating seq. design did indeed seem to slightly -- but in their opinion not materially -- overestimate the treatment efficacy.

    advice: "practical advantages ; more pregnancies achieved
    => crossover design should be first choice in infertility research"

Nature & extent of biases in estimates from 2 designs

Calculations/assumptions by Norman and Daya [2000].

- Heterogeneous population

... with the less effective treatment, at 1st (each) cycle ...

prob[becoming pregnant]	% of couples
10%	80%
40%	20%
---	
overall ave. 16%.	

- The more effective therapy had 'relative risk' of 2, i.e. at each cycle, probability[becoming pregnant] DOUBLED.
- *Expected* number of pregnancies at each cycle

adapted from Figure 2 and Table 2 of Norman and Daya ... .

Treatment Received in the Indicated Cycle

	Control				Experimental		
	Sub-population (fecundity)				Sub-population (fecundity)		
Cycle	High	Low	Combined	Ratio	Combined	High	Low
1	**200** (***80***)	**800** (***80***)	**1000** (***160***)		1000 (*320*)	200 (*160*)	800 (*160*)
			16%	2.00	32%		
2	40 (*16*)	640 (*64*)	680 (*80*)		**840** (***240***)	**120** (***96***)	**720** (***144***)
			11.8%	2.43	**28.6%**		
3	**24** (***9.6***)	**576** (***57.6***)	**600** (***67.2***)		600 (*134.4*)	24 (*19.2*)	576 (*115.2*)
			11.2%	2.00	22.4%		

4	4.8 (*1.9*)	460.8 (*46.1*)	465.6 (*48.0*) 10.3%	2.10	**532.8** **(*114.9*)** **21.6%**	**14.4** **(*11.5*)**	**518.4** **(*103.7*)**
5	**2.9** **(*1.2*)**	**414.7** **(*41.5*)**	**417.6** **(*42.6*)** **10.2%**	2.00	417.6 (*85.2*) 20.4%	2.9 (*2.3*)	414.7 (*82.9*)
1-5	271.7 (*108.7)*	2891.5 (*289.2*)	3163.2 (*397.8*) 12.6%	2.10*	3390.4* (*894.8*) 26.4%*	361.3 (*289.0*)	3029.1 (*605.8*)

Norman and Daya ... SUMMARY

Estimates of the efficacy of more effective treatment from ***both designs are biased*** (true relative risk = 2)

	Parallel	**Alternating seq.**
estimates	*under-*	*over-*
apparent relative risk:	1.83	2.10

bias limited to data from even-numbered cycles.

Despite the greater bias in the parallel design, Norman and Daya limited discussion of their concerns to -- and aimed their cautions at proponents of -- the alternating design.

Norman and Daya ... ADVICE

Compromise

"The objective of obtaining an accurate estimate of the effect of treatment, but also allowing all subjects to have the opportunity to receive the experimental treatment in at least one cycle, can now be achieved with the alternating-sequence design trial.

The proviso is that the trial should run for at least three cycles and all data from the even-numbered cycles would have to be excluded from the analysis, which would be restricted only to the odd-numbered cycles."

"When multiple cycles of treatment are undertaken to evaluate the efficacy of infertility therapy, the alternating-seq. design with restriction of the analysis to only the odd-numbered treatment cycles provides an unbiased estimation of the treatment effect".

Our Opinions / Questions / Suggestions / Plan

- Bias-avoiding strategy unlikely to be an acceptable to most investigators, patients and review committees.
- Must we discard 'biased' cycles and compensate for the decreased precision by increasing numbers of couples?
- If know what causes bias, why not remove it statistically?

2 approaches (use only the 'aggregated by cycle' data)

-1- no assumptions re form of heterogeneity;

strong assumptions re how a particular woman's fecundability with Exp'tl Tx is related to her fecundability with Standard Tx.

-2- specific parametric model for heterogeneity;

no connection b/w a particular woman's fecundability with Exp'tl Tx & that same woman's fecundability with Standard Tx

PRELIMINARIES

p_c a woman's average per-cycle probability of getting pregnant with Standard ("Control") Tx.

$f[p_c]$ distribution of p across source population

In Norman at al.'s example, $p_c \sim$ 2-point distribution

p_c *(per-cycle success prob.)*	$f[p_c]$ *(%)*
0.10	80
0.40	20
	summary
0.16	*<--- mean*
0.12	*<--- SD*

PRELIMINARIES ...

Suppose (as did Norman & Daya) that ...

Exp'tl Tx increases each p_C a constant-fold

i.e., by same multiple, θ ('Efficacy') in each fertility 'stratum'. *

$$p_E = \theta \times p_C \quad \text{..... for all values of } p_C$$

we drop the subscript C for next few slides..

---

* N+D later relaxed this assumption, making θ a function of p_C

difficult to imagine that a treatment which improves p 2-fold, from 0.1 to 0.2, with also raise the per-cycle success probability from 0.4 by same factor to 0.8.

Cycle-specific fecundability as fn. of θ and moments

For any distribution of p, the expected numerators, denominators, and success probabilities for each of the first 3 cycles, as a function of the efficacy, θ, and the (absolute) moments, p_1 to p_3 , of the distribution of $\mathbf{p_c}$ (**p** for short)

	Intervention	
Cycle	Control	Experimental
1	$\frac{\Sigma\ \mathbf{p\ .\ f[p]}}{\mathbf{1}}$ $\mathbf{\Rightarrow\ p_1}$	$\frac{\Sigma\ \theta\ .\ p\ .\ f[p]}{1}$ $\theta\ .\ p_1\ \Leftarrow$
2	$\frac{\Sigma\ p\ .\ (1\ -\ \theta\ .\ p).f[p]}{\Sigma\ \ (1\ -\ \theta\ .\ p).f[p]}$ $\Rightarrow \frac{p_1\ -\theta\ .\ p_2}{1\ -\theta\ .\ p_1}$	$\frac{\Sigma\ \theta\ \mathbf{.\ p\ .\ (1\ -\ p)\ .\ f[p]}}{\Sigma\ \ \mathbf{(1\ -\ p)\ .\ f[p]}}$ $\frac{\theta\ \mathbf{.\ p_1}\ \mathbf{-\theta\ .\ p_2}}{\mathbf{1\ -\ p_1}}$ **<=**
3	$\frac{\Sigma\ \mathbf{p\ .\ (1\ -\ p)\ .\ (1\ -}\ \theta\ \mathbf{.\ p)\ .\ f[p]}{\Sigma\ \ \mathbf{(1\ -\ p)\ .\ (1\ -}\ \theta\ \mathbf{.\ p)\ .\ f[p]}$ $\mathbf{\Rightarrow \frac{p_1\ -\ p_2\ -\ \theta\ .\ p_2\ +\ \theta\ .\ p_3}{1\ -\ p_1\ -\ \theta\ .\ p_1\ +\ \theta\ .\ p_2}}$	$\frac{\Sigma\ \theta\ .\ p\ .\ (1\ -\ p)\ .\ (1\ -\ \theta\ .\ p)\ .\ f[p]}{\Sigma\ \ \ \ (1\ -\ p)\ .\ (1\ -\ \theta\ .\ p)\ .\ f[p]}$ $\frac{\theta\ .\ p_1\ -\ \theta\ .\ p_2\ -\ \theta^2.\ p_2\ +\ \theta^2.\ p_3}{1\ -\ p_1\ -\ \theta\ .\ p_1\ +\ \theta\ .\ p_2}$ <=

It is assumed that either there are no dropouts [i.e. women/couples who haven't yet gotten pregnant do not abandon the study] or that dropouts are 'at random' and unrelated to their values of p.

Unspecified $f[p_C]$; $p_E = \theta \times p_C$ for all values of p_C

- Extension to cycles ≥ 4 is obvious, but algebra tedious!
- propn.'s in cycle 'k' involve 1st k moments of distrn. of p_C.
- Each addnl. cycle....

  => 2 new data points introduce 1 new parameter.

  => K cycles --> 2K datapoints; K+1 parameter model.

  => Thus, if K ≥ 3 --> remaining K-1 df to assess model fit.

- Unable to fit g.l.m. to 2K datapoints & K+1 parameters.
- Fit model numerically by maximizing Log Likelihood.

  Mathematica --> MLE --> I matrix --> SE for fitted E.

I. Fitting unspecified $f[p_C]$, with $p_E = \theta \times p_C \ \forall \ p_C$

MLE of θ (SE of *ln* estimate), as fn. of # data cycles used, and # moments fitted

	Method based on moments					Norman and Daya	
	Numbers of moments estimated					(odd-numbered cycles)	
Cycles used	1	2	3	4	5	*Cycles used*	Estimate (SE of *ln* est)
1	2.00 (0.086)					*1*	2.00 (0.086)
1, 2	2.33 (0.077)	2.00 (0.066)					
1, 2, 3	2.18 (0.066)	2.06 (0.060)	2.00 (0.060)			*1, 3*	2.00 (0.070)
1, 2, 3, 4	2.28 (0.064)	2.01 (0.055)	2.02 (0.055)	2.00 (0.056)			
1, 2, 3, 4, 5	2.18 (0.059)	2.07 (0.054)	2.01 (0.052)	2.00 (0.052)	2.00 (0.053)	*1, 3, 5*	2.00 (0.065)

Parametric model for heterogeneity

- Technical difficulties in fitting high-order nonlinear model.
- Large # of parameters rel. to # of observations.
- Instead... use a *known distributional form* for f[p].

Before we choose distrn. of (p_C,p_E) **across couples....**

Likelihood contribution: 1 couple; fixed but ? value of (p_C,p_E)

Cycle:	1	2	3	4	5
Treatment:	Exp'tl	Control	Exp'tl	Control	Exp'tl
Outcome:	–	–	–	–	+
Probability(+) :	p_E	p_C	p_E	p_C	p_E
Probability(Outcome):	$1 - p_E$	$1 - p_C$	$1 - p_E$	$1 - p_C$	p_E

- Likelihood: Π probs. of 5 observed outcomes (Bernoulli).
- Exploit equivalence of geometric and binomial likelihoods
  - rearrange the 5 cycle by cycle contributions above as:

    0 successes in 2 trials when prob[+] = p_C

    1 success in 3 trials when prob[+] = p_E.

II. Weinberg & Gladen's 'Beta-Geometric' GLM

- Beta distrn. for 'unexposed' population (e.g., non-smokers)

  2 parameters α and β ---> $E[p] = \alpha/(\alpha+\beta)$; $var[p] = \alpha b/((\alpha+\beta)^2(\alpha+\beta+1))$.

- **In (previously unsuccessful) couples who enter a particular cycle, the -- *now conditional* -- distrn. of p is again a beta distrn, but shifted towards zero by an amount that depends on spread of initial distrn.**

- Re-express parameters α and β in terms of 2 equivalent ones γ and δ

  $p_{[cycle]}$ = (conditional) mean p in those who enter this cycle

  $1 / p_{[cycle]} = \gamma + \delta$ (cycle - 1) .

  - γ : expected # cycles to become pregnant if per-cycle probability is $p_{[1]}$
  - δ: spread of the initial distribution of p

    $\delta = 0$ --> # cycles to pregnancy same (geometric) distrn for each.

II. Weinberg & Gladen's 'Beta-Geometric' Model

In those couples were previously unsuccessful in $u = k - 1$ previous cycles, the—*now conditional*—distribution of p at cycle k in this *selected sub*group is shifted towards the left i.e., towards zero, but *remains a*

***Beta distribution*, now with parameters $\{\alpha, \beta + u\}$.**

Expected probability of success among those who enter cycle k is related to number of previously unsuccessful cycles u via the simple reciprocal link:

$$1 / \mathrm{E}[\, p_{[k]} \,] = (\alpha+\beta+u)/\alpha = (\alpha+\beta)/\alpha + (1/\alpha) \times u$$

$$= \gamma + \delta \times u.$$

II. Fitting Weinberg & Gladen's 'Beta-Geometric' GLM
2 parallel arms: 'Control' Tx (t = 0) & Exp'tl Tx (t = 1)

Data structure:	# successes cycle *k* / # who undergo cycle *k*, each Tx
Model:	Binomial regression with inverse (i.e., power^{-1}) link.
2 groups	Usual indicator variables and product terms

Fit two *separate* Beta distributions for expected probabilities among smokers *(t =1)* & non-smokers(t = 0), via single equation:

$$1 / E[\, p_{[k,t]} \,] = (\gamma_0 + \delta_0 \times u \;) \times (1 - t) \; + \; (\gamma_1 + \delta_2 \times u \;) \times t$$

The alternating-sequence design...

- In source pop'lns: $p_C \sim \text{Beta}[\alpha_C, \beta_C]$ $p_E \sim \text{Beta}[\alpha_E, \beta_E]$.
- For distributions in cycle k (> 1) ...
  - need to be able to calculate the expected probability of success given the numbers of previous cycles, u_C and u_E respectively, where the standard and experimental treatments had been unsuccessful.
  - Unless one postulates the full bivariate form of the counterfactual probabilities for each woman, it is not possible to specify these subsequent probability distributions exactly.

Strategy used by Cohlen *et al*....

At cycle *k*, *odds* of success in group who were unsuccessful on opposite treatment at cycle *k*-1 is shifted up/down by odds ratio in 1st cycle.

	Odds		Reciprocal of Probability	
k	Control	Experimental	Control	Experimental
1	$\frac{\alpha_C}{\beta_C}$	$\frac{\alpha_E}{\beta_E}$	$\frac{\alpha_C + \beta_C}{\alpha_C}$	$\frac{\alpha_E + \beta_E}{\alpha_E}$
2	$\frac{\alpha_E}{\beta_E + 1} \times \frac{\alpha_C \beta_E}{\alpha_E \beta_C}$ $= \frac{\alpha_C}{\beta_C + \frac{\beta_C}{\beta_E}}$	$\frac{\alpha_C}{\beta_C + 1} \times \frac{\alpha_E \beta_C}{\alpha_C \beta_E}$ $= \frac{\alpha_E}{\beta_E + \frac{\beta_E}{\beta_C}}$	$\frac{\alpha_C + \beta_C}{\alpha_C}$ $+ \frac{\beta_C}{\alpha_C \beta_E} \times 1$	$\frac{\alpha_E + \beta_E}{\alpha_E}$ $+ \frac{\beta_E}{\alpha_E \beta_C} \times 1$
3	$\frac{\alpha_E}{\beta_E + \frac{\beta_E}{\beta_C} + 1} \times \frac{\alpha_C \beta_E}{\alpha_E \beta_C}$ $= \frac{\alpha_C}{\beta_C + \frac{\beta_C}{\beta_E} + 1}$	$\frac{\alpha_C}{\beta_C + \frac{\beta_C}{\beta_E} + 1} \times \frac{\alpha_E \beta_C}{\alpha_C \beta_E}$ $= \frac{\alpha_E}{\beta_E + \frac{\beta_E}{\beta_C} + 1}$	$\frac{\alpha_C + \beta_C}{\alpha_C}$ $+ \frac{\beta_C}{\alpha_C \beta_E} \times 1$ $+ \frac{1}{\alpha_C} \times 1$	$\frac{\alpha_E + \beta_E}{\alpha_E}$ $+ \frac{\beta_E}{\alpha_E \beta_C} \times 1$ $+ \frac{1}{\alpha_E} \times 1$

	Odds		Reciprocal of Probability	
k	Control	Experimental	Control	Experimental
4	$\frac{\alpha_E}{\beta_E + \frac{\beta_E}{\beta_C} + 2} \times \frac{\alpha_C \beta_E}{\alpha_E \beta_C}$ $= \frac{\alpha_C}{\beta_C + 2\frac{\beta_C}{\beta_E} + 1}$	$\frac{\alpha_C}{\beta_C + \frac{\beta_C}{\beta_E} + 2} \times \frac{\alpha_E \beta_C}{\alpha_C \beta_E}$ $= \frac{\alpha_E}{\beta_E + 2\frac{\beta_E}{\beta_C} + 1}$	$\frac{\alpha_C + \beta_C}{\alpha_C}$ $+ \frac{\beta_C}{\alpha_C \beta_E} \times 2$ $+ \frac{1}{\alpha_C} \times 1$	$\frac{\alpha_E + \beta_E}{\alpha_E}$ $+ \frac{\beta_E}{\alpha_E \beta_C} \times 2$ $+ \frac{1}{\alpha_E} \times 1$

Control Tx: equivalent to adding 1 to denominator of the odds for every previously unsuccessful cycle with this Tx, and adding an amount (β_C / β_E) for every previously unsuccessful cycle with Exp'tl Tx.

Exp'tl Tx: equivalent to adding 1 for every previously unsuccessful cycle with this treatment, and (β_E / β_C) for every previously unsuccessful cycle with the Control Tx.

Generalized linear model [Cohlen]:

Control Tx:

$$1 / E[\, p_{[k,C]} \,] = \frac{\alpha_C + \beta_C}{\alpha_C} + \frac{1}{\alpha_C} \times u_C + \frac{\beta_C}{\alpha_C \beta_E} \times u_E$$

$$= \gamma_0 + \delta_{0,C} \times u_C + \delta_{0,E} \times u_E$$

Exp'tl Tx:

.....

Modification factor that involves both α and β

For Control Tx:

Instead of $\phi = (\beta_C / \beta_E)$, use the initial *failure ratio*

$$\phi = [\beta_C/(\alpha_C + \beta_C)] / [\beta_E/(\alpha_E + \beta_E)],$$

to modify u_E before adding it to denominator of the odds

Control Tx	*Exp'tl Tx*
$\dfrac{\alpha_C}{\beta_C + u_C + (1/\phi) \times u_E}$	$\dfrac{\alpha_E}{\beta_E + u_E + \phi \times u_C}$

Generalized linear model [McGill version]:

$$1 / E[p_{[k,t]}] =$$

$$(\gamma_0 + \delta_{0,C} \times u_C + \delta_{0,E} \times u_E) \times (1 - t) + (\gamma_1 + \delta_{1,E} \times u_C + \delta_{1,C} \times u_E) \times t$$

Fitting...

Only 4 free parameters..

$\delta_{0,E}$ and $\delta_{1,C}$ (non-linear) functions of other 4 parameters.

Eliminate 2 redundant ones... numerically maximize log Lik.

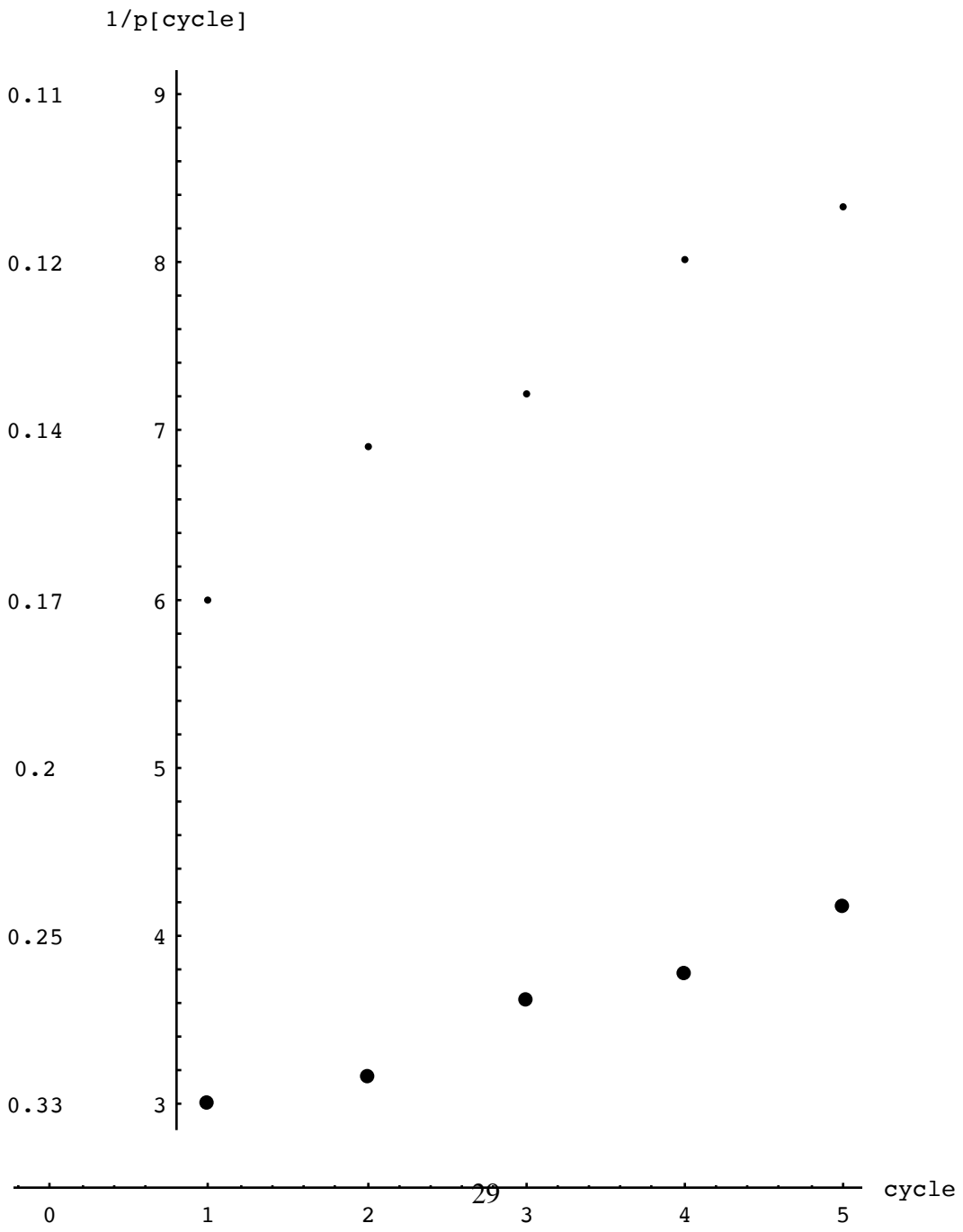
Simpler approach

Ignore redundancies;

Fit '6-parameter' model -- Binomial regression with μ^{-1} link.

IV. Adapting Weinberg & Gladen's Model

Reciprocals of success probabilities at each cycle, under alternating sequence design. Probabilities generated under (null) beta(3,15) distrn. of first-cycle probabilities, and an 'efficacy' of E=2. Cycle-specific ratios: 2, 2.18, 2, 2.13,2 .
{a=3,b=15} ==> mean[p] = 1/6 & SD[p] = 0.085; c=(a+b)/a = 6 & d = 1/a = 1/3.


1/p[cycle]
0.11
0.12
0.14
0.17
0.2
0.25
0.33
9
8
7
6
5
4
3
cycle
0
1
2
3
4
5

Pragmatic analytical model (GLM) for alt. seq. data...

$$1 / p_{[cycle,\ condn]} = c_1 + d_1 (cycle - 1) + g_1 z$$

$$(c_2 - c_1)\, I + (d_1 - d_2)(cycle - 1)\, I + g_2\, I\, z.$$

I =	1 if exptl. condn. 0 otherwise,	z =	1/(cycle-1) in even-# cycles 0 otherwise

θ estimated as ratio c_1 / c_2. [$c_1 = 1 / p_{[1]}$ & $c_2 = 1 / (E \times p_{[1]})$].

c_1: ave. # cycles to pregnancy under condn. 1
for couple with 'average' fertility.
c_2: ave. # cycles to pregnancy under condn. 2

Fit to N+D 'data':

$c_1 = 6.01$ cycles ; $c_2 = 3.01$ cycles; $E = (1/c_2) / (1/c_1) = 1.99$.

Parametric form for f[p] : Advantages

- Tx in specific cycle

  ==> outcome at that cycle, for each couple separately

- Cycle-specific covariates

  - intercourse frequency and timing

  - characteristics of the ova or the semen

  may modify couple's base per-cycle prob. of conception

  [Weinberg et al., 199x, Ecochard and Clayton, 2000].

Fitting parametric form for f[p] : Practical Issues

To model f[p] & 'constant-multiple of p' effect of exp'tl treatment, need:

(i) random-effects model for Bernoulli regression

(ii) log link (iii) $0 \leq p \leq 1$; $0 \leq E \times p \leq 1$.

- PROC NLMIXED in SAS
  - can constrain some parameters
  - only Gaussian distrn. currently avail. for f[p] -- in p, log[p] & logit[p] scales.
  - user must know g.l.m's and program p as function of linear predictor.
- (SAS) GLIMMIX macro:- full menu of links.
- Stata
- Gibbs Sampling

  direct & flexible

  (i) ability to use (among others) beta-distribution for p

  (ii) can ensure p and $E \times p$ stay in (0,1) interval

  (iii) interval estimates in most appropriate parameter scale

Fits to N+D 'data'

Estimates(SE) of efficacy parameter E .

Components of model	Software	Estimate of E (SE*)	Estimate(s) of other parameter(s)	Notes
Random Effects Models				
$\log[p] \sim N(0, \sigma^2)$	SAS PROC NLMIXED	2.42 (not calculated)	σ^2: 1.12	after 7 iterations
ditto	SAS Macro GLIMMIX	2.17 (0.04747)		after 14 iterations
$p \sim \text{Beta}(\alpha, \beta)$	WINBUGS	2.02 (0.11 *)	α: 1.7 ; β: 8.5	
Generalized Estimating Equations Approach				
Exchangeable correlation structure	SAS PROC GENMOD	1.96 (0.0468e96m)	ρ: 0.07	

III. Generalized Estimating Equations (GEE) Approach

Consider each couple as a separate 'cluster'.

- Less direct/Less transparent/Does not *explicitly* model heterogeneity in p. Instead, it models similarity of responses within same couple.
- pragmatic --> avoids high-dimensional design matrices; focus on correct SE.
- β's : not usual 'conditional on the cluster / covariate pattern' interpretation.
  - Limited range of between cluster variation --> distortion is small.
- binomial/log link applied to data in Table 1: E estimate = 1.96.

No explicit distributional form for f[p].

0.07 'correlation' of within-individual residuals (on 0/1 scale)

- *some* intra-individual variation in p
- difficult to convert r=0.07 into var[log[p]], or var[p]

EFFICACY ESTIMATES FROM ACTUAL TRIAL

Does 2nd generation protocol improve success rate with donor insemination that used frozen semen ? [Brown et al, 1988].

In 1984, same team had achieved a fecundability rate of 5.0 pregnancies/100 cycles with 1st-generation protocol, versus 18.9/100 using fresh semen.

For each woman, semen was from her matched donor for first six cycles; if pregnancy not achieved by then, the donor was changed.

Practical difficulties: "Cryopreserved semen was frequently substituted in a cycle scheduled to be fresh because the donor was not available".

Detailed information on which sequence was actually followed by each woman no longer available (S. Shapiro, personal communication, 2002). Thus, in order to allow a comparison of statistical analyses that use 'aggregated by cycle' versus 'individual' data, we constructed individual sequences to match, as closely as we could, the aggregated data in the 1988 report.

Pregnancies & Fecundability in Cycles of Insemination with Fresh & Frozen Semen. **Data adapted from Table 1 of Brown et al.**

	Fresh semen			Frozen semen		
	Number of			Number of		
Cycle	Patients	Pregnancies	Fecundability	Patients	Pregnancies	Fecundability
1	**163**	**55**	**0.337**	125	14	0.144
2	82	27	0.329	**137**	**13**	**0.095**
3	**70**	**17**	**0.243**	109	14	0.128
4	55	12	0.218	**93**	**9**	**0.097**
5	**60**	**16**	**0.267**	67	8	0.119
6	35	7	0.200	**68**	**6**	**0.088**
1-6	465	134	0.288	559	64	0.114

Efficacy, estimated from...	**Frozen / Fresh**
64/559 = 0.114 ÷ 134/465 = 0.288 ["crude"]	0.40(0.13)
odd-numbered cycles [36/301 ÷ 88/293]	0.40(0.18)
Mantel-Haenszel Risk Ratio	0.38(0.13)
Method I *(same ratio (to 2 dp), whether fit 1 or 2 moments)*	0.xx(0.xx)
Method II [Random effects model]	
- NLMIXED (f[l*n* p] ~ Normal)	0.35(0.13)
- GLIMMIX (f[l*n* p] ~ Normal)	0.36(0.14)
- WINBUGS (f[p] ~ Beta)	0.39(0.14)
Method III [GEE with exchangeable correlations]	0.40(0.13)[1]
Method IV [Generalized Linear Model*]	0.36(0.24)[2]

* see text for details. (SE) is the SE of the log of the ratio (in some cases, back-calculated from SE of ratio itself)

[1] model-based SE for *ln* ratio = 0.13; empirical SE = 0.12.

[2] ratio = 1/8.320 ÷ 1/2.994; SE[*ln* ratio] = { $(0.324/2.994)^2 + (1.766/8.320)^2$ }$^{1/2}$ = 0.24

*** Deviance / df = 0.69; Chi-square goodness of fit statistic = 7.3 (12 df).

Impressions

Estimates from random effects models based on a Gaussian distribution of *ln* p furthest from null.

Perhaps not surprisingly, given that the p's were generated from the same model used to fit the parameters, the estimate obtained by Gibbs sampling was close to the less parametric estimate from the GEE approach.

Both crude, and Mantel-Haenszel, summaries were very close to that given by the unbiased, but -- in terms of variance -- approximately half as precise, estimator advocated by Norman and Daya.

Imprecise estimate yielded by the Generalized Linear Model applied to the 'aggregated-by-cycle' data is probably a consequence of the large # of parameters (6) fitted to 12 datapoints.

IMPLICATIONS

Reassurance ...

to researchers who use alternating seq. design

to couples who generate the research data:

Not necessary to exclude even-numbered cycles

Statistical precision already low:

small n's: seldom > 25 pregnancies over all cycles.

WHY SMALL IMPACT OF MODEL CHOICE?

- N & D's advice based on

  - ***extreme*** 2-point distribution of p,

  - Tx efficacy ***doubled*** : 40% -> 80% in high fertility subgroup.

- **Bias** from using aggregated-by-cycle data

  << imprecision induced by small sample sizes used in practice.

- Freezing ***reduces*** average p from 33% to near 12%;

  0 <-- f [E × p]
  0 <-- f [p]

  ↓ impact of differential removal
  of most fertile at odd # cycles

OUR PREFERENCES

Random effects models

- accommodate woman-cycle outcomes using regression models.
- make use of *all of the individual level data* for each woman
  (baseline, cycle-by-cycle Tx indicators; other available cycle-dependent covariates).
- can include unexplained woman-woman heterogeneity

Software: WINBUGS.

- flexibility, transparency of estimation process.
- routines for binary data not standard/stable in main packages;
- GEE ?
  - sensible answers, but ...
  - ?? population-averaged vs. within-woman comparisons

CLARIFICATIONS

- Alternating sequence design not a full crossover study.

  Does **not** have statistical efficiency of paired-comparisons.

  Sample size /power calculations as per unmatched designs

- Cycle-specific results in same woman statistically independent ?

  - Condn'l on (unmeasurable) p (E × p) specific to a woman, cycles represent indep. Bernoulli trials with alternating prob's.

  - Random effects model allows p to vary from woman to woman.

OTHER ISSUES

- Choice of comparative parameter ?

  - unrestricted range of p, and of efficacy measure E

- p unlikely to have continuous distribution

  - subgroup where p=0 [Zhou and Weinberg]

Data-analysis options for comparisons of assisted reproductive technologies

James A. Hanley[1,2]
Nandini Dendukuri[1,3]
Robert Platt[1,4]
Marie-Hélène Mayrand[1]

Submitted to Statistics in Medicine, March 19, 2005

[1] Dept. of Epidemiology & Biostatistics, McGill University, Montreal, Quebec, Canada.
[2] Division of Clinical Epidemiology, Royal Victoria Hospital, Montreal.
[3] Technology Assessment Unit, McGill University Health Centre.
[4] Department of Pediatrics, Montreal Children's Hospital.

Correspondence:

Dr. J, Hanley, Department of Epidemiology and Biostatistics, McGill University,
1020 Pine Avenue West, Montreal, Quebec, Canada, H3A 1A2.

Telephone: +1 (514) 398-6270. Fax: +1 (514) 398-4503.
E-mail: James.Hanley@McGill.CA

Running Head: Trials of assisted reproductive technologies

SUMMARY

In the alternating-sequence design used to compare success rates with assisted reproductive technologies, women or couples are randomized to receive either the standard or experimental treatment in the first cycle, and— if they do not become pregnant—crossed between standard and experimental treatments after each successive cycle. Two authors have shown that, in the presence of heterogeneity of fecundability, and an effective treatment, the overall efficacy of the experimental treatment is overestimated by this design. These authors advised that in order to achieve an accurate estimate of efficacy, the trial should be run for at least three cycles and that all data from even-numbered cycles be excluded from the analysis, which should then be restricted only to odd-numbered cycles. In this paper, we describe approaches that make use of the data from all cycles. The methods are generalizations of those applicable to the constant-sequence design, where naive methods that do not take account of the heterogeneity produce underestimates of treatment efficacy.

Keywords

alternating sequence; fertility; experimental design; bias; precision; heterogeneity; generalized linear models

1. INTRODUCTION

Two experimental designs have been used to evaluate the efficacy of assisted reproductive technologies[1]. In the parallel-design, or constant-sequence randomized trial, the experimental treatment is administered for one or— if unsuccessful—more cycles to a fraction (usually one half, randomly chosen) of the eligible patients, and the control treatment for the same number of cycles to the remaining fraction. In the alternating-sequence design, some of the women or couples are randomized to receive the standard, and the others to receive the experimental treatment in the first cycle. Those who do not become pregnant are crossed to the opposite treatment after each successive cycle.

The relative merits of these two designs have been keenly debated [1-6]. Some arguments focus on efficiency and sample sizes: if the experimental therapy is effective, the alternating-design results in more pregnancies than the constant-sequence design, and is more attractive to couples. Others have to do with possible biases in the resulting estimates of efficacy. The first suggestion of bias came from comparisons of results of actual trials that used one or other of the two designs to evaluate the same procedure [2]. The authors noticed that, relative to those seen in parallel trials, treatment effects of the more effective treatment were higher in— i.e. overestimated by —crossover trials. Subsequent Monte Carlo evaluations [4], simulating patients from a heterogeneous subfertile population, indicated that while results from parallel trials appeared to slightly underestimate efficacy, the alternating-sequence design did indeed seem to slightly—but in their opinion not materially – overestimate it. Thus, they advised [4, p. 40] that "because of its practical advantages and because more pregnancies are achieved, a crossover design should be the first choice in infertility research."

The origin and nature of the biases in the estimates from these two designs can be readily understood by studying the worked example in Norman and Daya [6]. As shown in the first

row of Table 1, they assumed a heterogeneous population, where fecundability i.e., the per-cycle probability of getting pregnant, varied from couple to couple. To simplify matters, they assumed that with the less effective (Control) treatment, some 80% of couples had a $p_C = 10\%$, and the remaining 20% of couples a $p_C = 40\%$ fecundability, i.e.,

$$p_C = \begin{cases} 0.1 & \text{for } 80\% \text{ of couples,} \\ 0.4 & \text{for } 20\% \text{ of couples.} \end{cases}$$

Thus they assumed that the overall average fecundability is 16%, and the standard deviation is 12%.

They further assumed that the more effective (Experimental) treatment had a constant efficacy, θ, of 2, i.e., that at each cycle, a couple's probability of becoming pregnant was doubled. In Table 1 the course of the 1000 randomly allocated to undergo the 'control' treatment in the first cycle is tracked in bold. The entries at each cycle are the expected numbers of couples from the higher- and lower fecundity subpopulations who attempt to (and, in parentheses, the numbers who do) become pregnant. Using expected numbers of pregnancies at each cycle, Norman and Daya showed that estimates of efficacy that are based only on the total number of women and the total number of pregnancies are biased, irrespective of the design—the parallel design underestimates (apparent efficacy: θ =1.83, calculations not shown here but discussed later) and the alternating-sequence design overestimates (apparent efficacy $\theta = 2.10$, middle column Table 1). However, they noted that the bias in the alternating-sequence design is limited to the data from even-numbered cycles.

– Table 1 about here –

Despite the greater bias in the parallel design, Norman and Daya limited discussion of their concerns to—and aimed their cautions at prospective users of—the alternating sequence

design. They suggested [6, p. 323] a compromise between patient preference for this design and the statistical bias: "The objective of obtaining an accurate estimate of the effect of treatment, but also allowing all subjects to have the opportunity to receive the experimental treatment in at least one cycle, can now be achieved with the alternating-sequence design trial. The proviso is that the trial should run for at least three cycles and all data from the even-numbered cycles would have to be excluded from the analysis, which would be restricted only to the odd-numbered cycles." They concluded by advising [6, p. 310] that "when multiple cycles of treatment are undertaken to evaluate the efficacy of infertility therapy, the alternating-sequence design with restriction of the analysis to only the odd-numbered treatment cycles provides an unbiased estimation of the treatment effect."

This bias-avoiding strategy is unlikely to be an acceptable option for most investigators, patients and ethics review committees, and prompts the obvious questions: Must we discard 'biased' cycles and compensate for the decreased precision by increasing the numbers of couples enrolled? If we know the form of the bias, can we not remove it statistically using statistical models?

The purpose of this paper is to investigate this question, and several related ones. Under what model(s) is Norman's and Daya's approach really unbiased? If one can successfully eliminate the bias, at what price, in terms of increased imprecision, can this be achieved? Given the typically small sample sizes in this research area, can we afford this price, or might the overall mean squared error be smaller if we took a more naive approach? And, ultimately, if researchers use this design to collect their data, how should they analyze them, and how should they calculate the uncertainty in their estimates of efficacy? We restrict attention to models that use aggregated data for each treatment-cycle.

2. HOMOGENEOUS FECUNDABILITY

Let p_C denote a woman's fecundability i.e., her per-cycle probability of getting pregnant, with the less effective (control) treatment ($t = 0$). For now, assume no variation in p_C across women, i.e., that $\text{Var}[p_C] = 0$. Let p_E denote her fecundability with the experimental treatment ($t = 1$). Its efficacy with respect to the control treatment can be expressed in different ways using different forms for g in the generalized regression equation $g[p_E] = g[p_C] + \beta \times t$. For example, β is the absolute difference in fecundability if g is the identity function; $\exp[\beta]$ is the fecundability ratio θ if g is the ln function, or the fecundability odds ratio if g is the logit function. We will use the fecundability ratio to measure efficacy.

Suppose that one such woman, alternating from the experimental treatment in cycle 1, became pregnant on this treatment in the 5th cycle.

Cycle:	1	2	3	4	5
Treatment:	Exp'tl	Control	Exp'tl	Control	Exp'tl
Outcome:	—	—	—	—	+
Probability(+):	p_E	p_C	p_E	p_C	p_E
Probability(Outcome):	$1 - p_E$	$1 - p_C$	$1 - p_E$	$1 - p_C$	p_E

The observed data can be modeled as a sequence of independent Bernoulli trials with alternating probabilities of success. The likelihood based on this woman's data is the product of the probabilities of the 5 individual outcomes; it can also be re-arranged and written as a product of two geometric (but binomial-like) likelihoods, corresponding to $s_C = 0$ successful cycles, preceded by $u_C = 2$ unsuccessful ones, when the success probability was p_C ; and $s_E = 1$ successful cycle, preceded by $u_E = 2$ unsuccessful ones, when the success probability was p_E , i.e.,

$$L(u_C, u_E, s_C, s_E \mid p_C, p_E) \propto (1 - p_C)^2 \times (1 - p_E)^2 \times p_E$$

Since p_C and p_E are constant from woman to woman, so that all woman-cycles within the same treatment condition are exchangeable, the likelihood based on the data from several such women can again be written as the product of two binomial-like likelihoods

$$L(U_C, U_E, S_C, S_E \mid p_C, p_E) \propto (1-p_C)^{U_C} \times p_C^{S_C} \times (1-p_E)^{U_E} \times p_E^{S_E}$$

where $U_C = \Sigma u_C$ and U_E and S_C and S_E are the total numbers of unsuccessful and successful cycles when using C and E respectively, i.e., summed over all women and all cycles. The ML point estimator of the fecundability ratio is simply $(S_E/T_E)/(S_C/T_C)/$ where $T_E = S_E + U_S$ and $T_C = S_C + U_C$. A likelihood-based interval estimate is also easily calculated.

3. HETEROGENEOUS FECUNDABILITY

In reality, fecundability *does* vary among women, i.e., $\mathrm{Var}[p_C] > 0$ and $\mathrm{Var}[p_E] > 0$. We denote this variation by the general bivariate pdf $f(p_C, p_E)$, with marginal distribution $f(p_C)$. We present two data-analysis approaches which use aggregated data for each treatment in each cycle. The first makes no assumptions about the form of the marginal distribution $f(p_C)$, but strong ones about how a particular woman's fecundability with the experimental treatment is related to her fecundability with the standard one. In this approach, the observed data can be modeled either as (i) two sets of multinomial distributions, one for the numbers $\{S_{C_1}, S_{E_2}, \cdots\}$ who become pregnant in cycles $C_1, E_2, etc.$, the other for the numbers $\{S_{E_1}, S_{C_2}, \cdots\}$ who become pregnant in cycles $E_1, C_2, etc.$, or (ii) as cycle- and treatment-specific binomial random variables $\{\ S_{C_1}|T_{c_1}\}$, $\{S_{E_1}|T_{E_1}\}$, $\{S_{C_2}|T_{C_2}\}$, ... The second approach is based on a specific parametric form—Beta—for f, but does not 'connect' a particular woman's fecundability when undergoing experimental treatment with that same woman's fecundability with the standard treatment.

We evaluate these approaches using the data in Table 1 as well as data generated from continuous bivariate distribution for $\{p_C, p_E\}$. In the latter case, how a particular woman's fecundability, p_E, with the experimental treatment is related to her fecundability, p_C, with the standard one induced variability in the efficacy across women. In this second method, the observed data are modeled as cycle- and treatment-specific binomial random variables.

3.1 Unspecified-form for f; constant fecundability ratio

Consider an unspecified distribution $f(p_C)$ and let θ denote the constant ratio of p_E to p_C for all values of p_C. For a person with a specific value p_C, assigned to the sequence C, E, C, ... , the probabilities of becoming pregnant in cycle $1, 2, 3, ...$ are

$$p_C, (1-p_C) \times \theta \times p_C, (1-p_C) \times (1 - \theta \times p_C) \times p_C, ...$$

The probabilities, if that same person were assigned to the sequence E, C, E,. . . , are

$$\theta \times p_C, (1 - \theta \times p_C) \times p_C, (1 - \theta \times p_C) \times (1-p_C) \times \theta \times p_C, ...$$

Since p_C varies over persons, the multinomial proportions are the expectations of these probabilities, taken over the distribution, $f(p_C)$, of p_C. They form the numerators of the expressions given in Table 2. and represent the contribution to the (multinomial-based) likelihood of each person who becomes pregnant in that cycle. The extension beyond cycle 3 (not shown) is obvious, even if the algebra is tedious. Of note is the fact that the two likelihood contributions from cycle k involve the first k moments of the distribution of p_C. Others, e.g., [7,8,9], have noted this in the simpler constant-sequence design. Thus, each cycle adds two new data points and one new parameter; overall the $2K$ datapoints from K cycles are modeled by $K+1$ parameters. If $K \geq 2$, the remaining $K-1$ degrees of freedom can be used to assess model fit.

Table 2 about here

Maximum likelihood estimates for these $K+1$ parameters can be obtained from a non-linear modeling package, such as SAS PROC NLMIXED (see Appendix). Although our approach deals with heterogeneity, it does so without specifying a traditional random effects model: we used only the 'NL" portion of NLMIXED. We found that this *multinomial* approach is very sensitive to starting values, and have had more success by modeling the number who become pregnant on a specific treatment in a specific cycle—conditional on the number who used that specific treatment in that cycle—as a *binomial* random variable. These conditional probabilities are given as the quotients in Table 2. Again, they can be fitted using SAS PROC NLMIXED by (i) expressing the $2K$ binomial parameters in terms of the K moments and the parameter of interest θ, and (ii) for each of the $2K$ observed counts, modeling

$$Number_{pregnant} \sim Binomial(Number_{treated}, BinomialParameter).$$

For the data in Table 1, Maximum Likelihood estimates (and Standard Errors of their natural logarithms) for these parameters are shown in Table 3. This method correctly 'recovers' θ. Further, because the procedure uses data from all cycles, it produces smaller standard errors than those for the summary estimates from the odd-numbered cycles only. Moreover, one can achieve this increased precision, and only a slight inaccuracy, with fewer than the full K moments: one can omit i.e., set to zero in the likelihood, some of the higher order moments—those of order 3 or more in our example. This is because p_C is bounded by 0 and 1, so that the higher moments are of decreasing magnitudes, and thus increasingly negligible.

– Table 3 about here –

3.2 Beta-Geometric Model

That an experimental treatment would increase each p_C value a constant-fold, i.e., by the same multiple, θ for each woman, regardless of her value of p_C, is not realistic biologically. Whereas women whose natural fecundability is 10% might reasonably have it increased to 20% i.e., by a factor of $\theta = 2$, with experimental treatment, the treatment is unlikely to also raise other women's already high natural per-cycle success probability of 40%, say, by the same (multiplicative) factor of $\theta = 2$, i.e., to 80%. Moreover, if $p_C > 1/\theta$, this assumption of a constant θ is statistically impossible. In addition, there are practical technical difficulties in fitting such a high-order nonlinear model; the number of parameters (moments) relative to the numbers of observations is large. For these reasons, we turn to more natural parametric statistical models for p_C and p_E, ones with fewer constraints on how a particular woman's fecundability when undergoing experimental treatment relates to what might be (loosely) called its 'counterfactual' i.e., the same woman's fecundability with the standard treatment.

The Beta-Geometric (B-G) model has been used in demography [7]. More recently, Weinberg and Gladen [8] used it in a non-experimental study of the effect of smoking on fecundability. Since women were classified as smokers or non-smokers for the entire period of observation (up to 12 cycles), their model immediately applies to a parallel-sequence design [4]. The latter authors used a modified B-G model to *generate* data, but did not consider it for the *analysis* of their data. We extend these ideas to develop a beta-geometric model for data from the alternating-sequence design.

Before doing so, we review its use for the constant-sequence design. Weinberg and Gladen compared fecundability, measured over 12 cycles, in smokers relative to non-smokers. They modeled fecundability in the two source populations as Beta distributions, with their respective location and shape governed by the pairs of parameters $\{\alpha_C,\beta_C\}$ and $\{\alpha_E,\beta_E\}$.

Therefore, before the first cycle the probability density function of p_C is given by:

$$f[p_c] \propto {p_c}^{\alpha_C - 1}(1 - p_c)^{\beta_C - 1}$$

The mean and variance of the fecundability in the control group before the first cycle are $\mu_C = \alpha_C/(\alpha_C + \beta_C)$ and $\sigma_C^2 = \alpha_C\beta_C/((\alpha_C + \beta_C)^2(\alpha_C + \beta_C + 1))$ Weinberg and Gladen showed that in those couples who were unsuccessful in U previous cycles, the—*now conditional*—distribution of p_C at cycle $U + 1$ in this selected subgroup is shifted towards the left i.e., towards zero, but remains a Beta distribution, with probability density function:

$$f(p_c|U) \propto {p_c}^{\alpha - 1}(1 - p_c)^{\beta - 1 + U}.$$

The parameters of the fecundability distribution are now $\{\alpha_C,\ \beta_C + U\}$. Thus, after k cycles the mean fecundability is given by $\mu_C = \alpha_C/(\alpha_C + \beta_C + k)$. They further showed that the expected probability of success among those who enter cycle U+1 is related to the number of previously unsuccessful cycles U via the simple reciprocal link:

$$1/E[p_C|U] = (\alpha_C + \beta_C + U)/\alpha = (\alpha_C + \beta_C)/\alpha + (1/\alpha) \times U$$

$$= \gamma + \delta \times U$$

The parameter γ is the expected number of cycles to become pregnant if a couple had the average per-cycle probability under the control treatment i.e., $\gamma = 1/\mu_C$. The parameter δ reflects the spread of the initial distribution of p_C: under homogeneity, the number of cycles required to achieve pregnancy reduces to the same geometric random variable for each couple, i.e., $\delta = 0$. The parameters γ and δ can be fit using binomial regression with an inverse (i.e., power-1) link, e.g., using PROC GENMOD in the SAS, or glm in Stata. Since glm in R does not allow this link for the binomial, one needs to supply the variance function.

Weinberg and Gladen extended the model to effectively fit separate Beta distributions for the cycle-specific probabilities for smokers ($t = 1$) and non-smokers($t = 0$), via one equation:

$$1/E[p_C \mid U] = (\gamma_C + \delta_C \times U) \times (1 - t) + (\gamma_E + \delta_E \times U) \times t.$$

To extend it to the alternating sequence design, we model the expected probability of pregnancy for women who have already undergone U_E and U_C unsuccessful cycles on E and C, and are now about to receive (say) E. The initial Beta(α_E,β_E) distribution must be updated to reflect the U_E and U_C. The U_E is added to the β_E term as in the parallel sequence model, but a correction must be made to the U_C. If C were no more effective than E, and thus exchangeable with it, we would add the full U_C to the U_E to obtain the β_E term of $U_E + U_C$. But if C were less effective than E, then using the full $U_E + U_C$ in the β_E term would shift the fecundability distribution too far to the left. This is most easily seen if *no* pregnancies can occur with C. The greater the efficacy of E relative to C, (i.e., the smaller the 'failure ratio' $(1 - p_E)/(1 - p_C)$), the smaller should be the 'amalgam' of U_E and U_C. Thus, we suggest that the contribution of U_C be reduced, and propose the amalgam

$$U^* = U_E + FRR \times U_C$$

where FRR is the ratio of the probability of failure on the experimental and control treatments at cycle 1. This ratio is less than unity if E is more effective than C. Thus, the probability density function of p_E at cycle $u_C + u_E + 1$ is taken to be

$$f(p_E \mid u_E\ u_C) \propto {p_E}^{\alpha_E - 1}(1 - p_E)^{\beta_E - 1 + u_E + FRR \times u_C},$$

i.e., a Beta(α_E , $\beta_E + u_E + FRR \times u_C$) distribution. Thus the expected probability of success at this cycle is inversely proportional to a linear function of u_E and u_C, i.e .

$$E(p_E \mid u_C\ \ u_E) = \alpha_E/(\alpha_E + \beta_E + u_E + FRR \times u_C)$$

Similarly, we can show that

$$E(p_C \mid u_C \;\; u_E) = \alpha_C/(\alpha_C + \beta_C + u_C + (1/FRR) \times u_C)$$

We re-formulate it as a generalized linear model for binary data with an inverse link function:

$$\begin{array}{rlll} 1/E(p \mid u_C \;\; u_E) = & \gamma_{0E} \times t & + & \gamma_{0C} \times (1-t) \\ + & \gamma_{1E} \times (u_E \times t) & + & \gamma_{1C} \times (u_C \times (1-t)) \\ + & \gamma_{2E} \times (u_C \times t) & + & \gamma_{2C} \times (u_E \times (1-t)) \end{array} \quad (3)$$

where $t = 1$ under the experimental treatment and $t = 0$ under the control treatment. Constraints should be placed on γ_{2E}, which is a function of γ_{0E} and γ_{1E}, and on γ_{2C}. The model can be fit using PROC NLMIXED in SAS, which can accommodate constraints. The parameters of the original Beta distributions can be obtained by the following transformations.

$$\alpha_E = 1/\gamma_{1E};\;\; \beta_E = (\gamma_{0E} - 1)/\gamma_{1E};\;\; \alpha_C = 1/\gamma_{1C};\;\; \beta_C = (\gamma_{0C} - 1)/\gamma_{1C},$$

and efficacy is estimated by $\hat{\theta} = \hat{\gamma_{0C}}/\hat{\gamma_{0E}}$.

In a Bernoulli model, woman-level covariates could be added as linear terms in (3). Notice that at each cycle the fecundability distribution under treatment t is obtained by adding to the β parameter of its starting fecundability distribution, the number of unsuccessful cycles on t and a multiple of the number of those on the other treatment. While we have used the multiplying factors as FRR and $1/FRR$ above, we could use the generic expressions $\alpha_C/(\alpha_C + \beta_C + u_C + \phi_C \times u_C)$ and $\alpha_E/(\alpha_E + \beta_E + u_E + \phi_E \times u_C)$ for the respective mean fecundability after cycle $u_C + u_E$. ϕ_C and ϕ_E should be constrained to be > 1 and < 1 respectively, assuming E is more effective than C.

4. EVALUATION

We assessed the performance of these analysis models on 200 generated datasets. Following Cohlen et al. [4] we began with Beta distributions with means of 0.16 and 0.32, each

with a coefficient of variation of 75% (Cohlen et al. used 56%). We calculated the 9th, 18th, ..., 90th percentiles for each of these two initial distributions, and placed a point mass of 0.1 at each of these values, thereby creating two 10-point distributions for the first cycle. If a woman's fecundability was at say the 18th percentile when on C i.e. if $p_C = 5.2\%$, then her fecundability with E (if necessary) was the corresponding 18th percentile in that distribution, namely $p_E = 8.1\%$ (fecundability ratio = 1.56). Similarly, those just above the middle of the p_C distribution, i.e., a fecundability of $p_c = 15.2\%$, were considered to be just above the middle of the p_E distribution, namely $p_E = 31.6\%$ (fecundability ratio = 2.08), while in the 3rd highest subgroup, fecundabilities were $p_C = 22.6\%$ and $p_E = 48.2\%$ (ratio = 2.13). For each dataset, 1000 women were randomly assigned, using a multimomial distribution, to the 10 fecundability levels under C, and 1000 others to the 10 corresponding levels in the p_E distribution. At each cycle, these two sets of 10 subgroup frequencies were depleted using random numbers of pregnancies generated by the 20 corresponding binomial distributions. The numbers who were unsuccessful were switched to their corresponding level on the opposite distribution, before generating the pregnancies for the next cycle.

The estimates produced by our analysis models are summarized in Table 5. Leaving the form of f unspecified and estimating its first few moments is somewhat more efficient than using the Beta-Binomial model, where there is more of a tradeoff between bias and precision.

5. EFFICACY ESTIMATES FROM A CLINICAL TRIAL

The data in Table 6a are from (by today's standards) a very large clinical trial. It evaluated the performance of a second generation protocol for donor insemination with frozen semen [10]. Today, screening of semen for HIV infection precludes the use of fresh semen. Earlier, in the first use that we have found of the alternating sequence design, this group

achieved a fecundability rate of only 5.0 pregnancies per 100 cycles with a first-generation protocol, versus 18.9 per 100 using fresh semen [11].

The data for the first six cycles (during which, for each woman, semen was from her matched donor) are shown in Table 6a. However, as is obvious from the denominators, the same practical difficulties were encountered as those mentioned in the first study: "Cryopreserved semen was frequently substituted in a cycle scheduled to be fresh because the donor was not available." Unfortunately, information on the actual sequence for each woman is no longer available (S. Shapiro, personal communication, 2002).

Estimates of efficacy are given in Table 6b. Those produced by the methods in 3.1 and 3.2 are closest to the null, suggesting that they removed some of the bias induced when one ignores the heterogeneity. Possibly by chance, given its poor precision, the estimator advocated by Norman and Daya was furthest from the null, further than both the crude and the Mantel-Haenszel estimates. The Generalized Linear Model estimate was closest to the null, but had a high SE, possibly because of the large number of parameters (6) fitted to the 12 datapoints. The method of moments had the lowest SE. The fact that it was no different from that of the Mantel Haenszel estimator suggests that, in this study, heterogeneity does not substantially inflate or deflate the SE. Without individual-specific data, we are unable to assess how much heterogeneity could also be affected by women who dropped out.

6. DISCUSSION

The more attractive alternative sequence design also produces more pregnancies for the same number of cycles. We describe two approaches that allow clinical trials to use data from all of the cycles, while Norman and Daya's approach [6] squanders the statistical advantage of this design. Many contemporary trials generate fewer than 25 pregnancies in total. De-

creasing precision further by omitting even-numbered cycles, without trying to eliminate the biases by other methods, is difficult to defend. Moreover, with the sample sizes considered here reduced by a realistic factor of 25, sampling variability dominates analytic bias.

The bias caused by naively using all data is a function of the heterogeneity in p_C and the efficacy of the experimental treatment ; *both must be substantial in order to produce a serious bias*. Norman and Daya based their concerns on an extreme 2-point distribution of p_C, and a treatment that doubled the $p_C = 40\%$ in the high fertility subgroup to a 'biologically nearly impossible' $p_E = 80\%$. Even then, the bias was less than the sampling variability induced by the sample sizes used in practice. With the same $\theta = 2$, and a more realistic distribution where $f[0.025] = 0.8$ and $f[0.1] = 0.2$ so that $\text{mean}[p_C] = 0.04, SD[p_C] = 0.03$, the bias was much smaller Moreover, if, rather than increasing p, an new treatment— such as frozen semen—*reduces* it, the degree of bias in the naive estimate of θ is also less, because of the smaller impact of the differential success (removal) of the most fertile in odd-numbered cycles. For example, using $\theta = 0.5$ in Table 2, the data from cycle 2 yield $\hat{\theta} = 0.47$, a relative bias of only 6%. These 'low-bias' conditions would also apply in comparisons of contraceptive methods, where the probability of an unwanted pregnancy is already low.

The alternating sequence design is not a full *crossover* study. Nor does it carry the full statistical efficiency usually associated with self-matched comparisons. Thus, the sample size and power calculations/projections are best carried out by analogy with unmatched designs.

In practice, for example in the study by Brown et al. [10], there are unavoidable individual deviations from the alternating-sequence protocol. Some investigators may use variations of the alternating design: e.g., in Ecochard et al. [12], 1/2 patients received one treatment for the first two cycles and the competing treatment for the next two cycles, while the other 1/2 followed the opposite sequence. For these more complex designs, random effects models for

binary (Bernoulli) data allow one to model the cycle-by-cycle sequence of outcomes for each woman using a full regression approach that makes use of all of the individual level data for each woman (any baseline covariates, the cycle-by-cycle treatment indicators and any other available cycle-dependent covariates). Some quite complex 2-level hierarchical models have been used in such circumstances [13,14].

The approaches we have described are also applicable to simple data analyses for the parallel- or 'constant-sequence' randomized trial design. Since this competing design has the same data structure as Weinberg and Gladen's example (fecundability of smokers and non-smokers), their 'Beta-Geometric' Model is immediately applicable without modification. The method based on moments is also applicable. Applied to Norman and Daya's 'constant-sequence' example (and the 'data' in their Table 1), both of our methods produce estimates closer to the true $\theta = 2$, whereas a naive analysis produces an attenuated estimate of 1.83.

In an appendix, Norman and Daya [6, p324] claim that "the assumptions of a constant drug efficacy is not necessary" by considering an arbitrary distribution $f[p_C]$, and an arbitrary efficacy function, $\theta[p_C]$ They purport to show algebraically that "the outcome rates in the odd cycles in an alternating sequence are unbiased," i.e., that "the results will hold true regardless of the relationship between efficacy and fertility." In fact, the ratio from alternate cycles will not continue to be unbiased if the treatment effect is variable. This is illustrated in Table 7 by slightly perturbing Norman and Daya's simulated example so that the risk ratio in the low fecundability group is 2.5, while the risk ratio in the high fecundability group remains at 2. This is closer to what happens in reality where there is likely to be a greater relative shift in the low fecundability groups, compared to the high fecundability groups. The true average risk ratio across the population is thus $2.5 \times 0.8 + 2 \times 0.2 = 2.4$. However, from Table 7 we can see that this estimate is not obtained even in the first cycle: the ratio

of the expected probabilities of successes does not match the expectation of the ratios of the success probabilities, i.e.

$$\frac{\Sigma\ \theta[p_C] \times p_C \times f[p_C]}{\Sigma\ p_C \times f[p_C]} \neq \Sigma\ \theta[p_C] \times f[p_C] = \theta[p_C]$$

Further, it appears that the odd cycles underestimate the true risk ratio while the even cycles overestimate it. If the study continues to a point when only women in the low fecundability group remain, then the ratio approach the true ratio of 2.5 in both odd and even cycles

This contrary finding is an additional impetus to consider more general regression models that allow not just between-individual heterogeneity, and covariates at the woman-cycle level, but also more flexibility in the specification of the comparative parameter. We plan to investigate whether the amount of data from a typical alternating sequence design makes such models practical. Unlike traditional studies with multiple crossovers, the alternating sequence design involves at most one instance of Y=1 per subject, and such outcomes preclude further observations. This, the small sample sizes, and the small number of cycles usually used, may be a serious impediment to more complex modeling. The analyses we have presented, based on marginal distributions, may well be the appropriate ones for the amounts of clinical trial data generated by the alternating sequence design.

ACKNOWLEDGMENTS

Nandini Dendukuri is a chercheur-boursier of the Fonds de Recherche en Santé du Québec. Robert Platt is a career scientist of the Canadian Institutes of Health Research.This work was supported by individual operating grants from the Natural Sciences and Engineering Research Council of Canada and a team grant from Le Fonds Québécois de la recherche sur la nature et les technologies.

Code

xxxx

xxxx xxxx xxxxxxxx

xxxxx xxxxxx xxxxxx.

REFERENCES

1. DAYA, S. Is there a place for the crossover design in infertility trials? *Fertility and Sterility* 1993; **59**: 67.
2. KHAN, K.S, DAYA S, COLLINS J.A., AND WALTER S.D. (1996). Empirical evidence of bias in infertility research: overestimation of treatment effect in crossover trials using pregnancy as the outcome measure. *Fertility and Sterility* **65**:939945.
3. TE VELDE, E.R., COHLEN B.J., LOOMAN C.W., AND HABBEMA, J,D. Crossover designs versus parallel studies in infertility research. [letter] *Fertility and Sterility* 1998; **69**:357-358.
4. COHLEN, B.J., TE VELDE, E.R., LOOMAN. C.W.N., EIJCHEMANS, R., AND HABBEMA, J.D.F. Crossover or parallel design in infertility trials? The discussion continues. *Fertility and Sterility* 1998; **70**:40-45.
5. DAYA, S. Differences Between Crossover and Parallel Study DesignsDebate? (letter) *Fertility and Sterility* 1999; **71**:771-772.
6. NORMAN, G.R. AND DAYA, S. The alternating-sequence design (or multiple-period crossover) trial for evaluating treatment efficacy in infertility. *Fertility and Sterility* 2000; **74**:319-324.
7. SHEPS, M.C., AND MENKEN, J.A. Mathematical models of conception. University of Chicago Press, 1973.
8. WEINBERG, C.R. AND GLADEN B.C. The beta-geometric distribution applied to comparative fecundability studies. *Biometrics* 1986; **42**:547-560.
9. LAU, T.S. On the heterogeneity of fecundability. *Lifetime Data Analysis* 1996; **2**:403-415.
10. BROWN, C.A., BOONE, W.R., AND SHAPIRO, S.S. Improved cryopreserved semen fecundability in an alternating fresh-frozen insemination program. *Fertility and Sterility* 1988; **50**:825-827.
11. RICHTER, M.A., HANING, R.V., AND SHAPIRO, S.S. Artificial donor insemination: fresh versus frozen semen: the patient as her own control. *Fertility and Sterility* 1984; **41**:277-280.
12. ECOCHARD, R, MATHIEU, C., ROYERE, D., BLACHE, G., RABILLOUD, M., AND CZYBA, J.C. A randomized prospective study comparing pregnancy rates after clomiphene citrate and human menopausal gonadotropin before intrauterine insemination. *Fertility and Sterility* 2000; **73**:90-93
13. ECOCHARD R, CLAYTON DG. Multi-level modelling of conception in artificial insemination by donor. *Statistics in Medicine* 1998; **17**(10):1137-1156.
14. ECOCHARD R, CLAYTON DG. Multivariate parametric random effect regression models for fecundability studies. *Biometrics* 2000; **56**(4):1023-1029.

Table 1: Cycle-specific ratios of expected pregnancy proportions if the alternating sequence design is applied to a population of 2000 which is heterogeneous with respect to spontaneous fecundity [20% with higher, 80% with lower fecundity].

	Treatment received in the indicated cycle						
	Control				Experimental		
	Sub-population (fecundability)				Sub-population (fecundability)		
Cycle	High	Low	All	Ratio	All	High	Low
1	**200**	**800**	1000		1000	200	800
	(80)	**(80)**	**(160)**		(320)	(160)	(160)
			16%	2.00	32%		
2	40	640	680		**840**	**120**	**720**
	(16)	(64)	(80)		**(240)**	**(96)**	**144**
			11.8%	2.43	**28.6%**		
3	**24**	**576**	600		600	24	576
	(9.6)	**(57.6)**	**(67.2)**		(134.4)	(19.2)	(115.2)
			11.2%	2.00	22.4%		
4	4.8	460.8	465.6		**532.8**	**14.4**	**518.4**
	(1.9)	(46.1)	(48.0)		**(114.9)**	**(11.5)**	**103.7**
			11.3%	2.10	**21.6%**		
5	**2.9**	**414.7**	417.6		417.6	2.9	414.7
	(1.2)	**(41.5)**	**(42.6)**		(85.2)	(2.3)	(82.9)
			10.2%	2.00	20.4%		
1-5	271.7	2891.5	3163.2		3390.4	361.3	3029.1
			(397.8)		(894.8)		
			(12.6%)	2.10*	(26.5%*)		

The course of the 1000 randomly allocated to undergo the 'control' treatment in the first cycle is tracked in bold. The entries at each cycle are the expected numbers of couples from the higher- and lower fecundity subpopulations who attempt to (and, in parentheses, the numbers who do) become pregnant. Table adapted from Figure 2 and Table 2 of Norman and Daya.

Table 2: Unconditional (multinomial) and conditional success probabilities for each of the first three cycles, as a function of the efficacy, θ, and the (absolute) moments of the unspecified distribution of p_C, the fecundability under the standard ["control" (C)] treatment.

Cycle	Control	Experimental
1	μ_1	$\theta \times \mu_1$
2	$\frac{\mu_1 - \theta \times \mu_2}{1 - \theta \times \mu_1}$	$\frac{\theta \times \mu_1 - \theta \times \mu_2}{1 - \mu_1}$
3	$\frac{\mu_1 - \mu_2 - \theta \times \mu_2 + \theta \times \mu_3}{1 - \mu_1 - \theta \times \mu_1 + \theta \times \mu_2}$	$\frac{\theta \times \mu_1 - \theta \times \mu_2 - \theta^2 \times \mu_2 + \theta^2}{1 - \mu_1 - \theta \times \mu_1 + \theta \times \mu_2}$

The numerators represent the unconditional probabilities of pregnancy in the indicated cycle for persons *entering* the study, while while the quotients represent conditional pregnancy probabilities for those who receive the indicated treatment in the *indicated* cycle. These probabilities are computed separately for those randomly allocated to the 'C to E to C' sequence, and conversely for their counterparts. Cycle 1 starts with denominators of 1 (100%) in each group; it is assumed that there are no dropouts [i.e. women/couples who have not yet become pregnant do not abandon the study] or that dropouts are 'at random' and unrelated to their values of p_C. The symbols μ_1 to μ_3 are the first 3 absolute moments of the distribution of p_C, the fecundability with standard treatment.

Table 3: MLEs of the efficacy parameter θ (SE of ln of estimate) as a function of number of data cycles used, and number of moments of unspecified distribution f estimated, compared with estimates obtained using approach of Norman and Daya.

	No. moments of f fitted					⋮	Norman and Daya	
Cycles	1	2	3	4	5	⋮	Cycles	$\hat{\theta}$
used						⋮	used	(SE*)
1	2.00					⋮	*1*	2.00
	(86)					⋮		(86)
1,2	2.33	2.00				⋮		
	(77)	(66)				⋮		
1,2,3	2.18	2.06	2.00			⋮	*1,3*	2.00
	(66)	(60)	(60)			⋮		(70)
1,2,3,4	2.28	2.01	2.02	2.00		⋮		
	(64)	(55)	(55)	(56)		⋮		
1,2,3,4,5	2.18	2.07	2.01	2.00	2.00	⋮	*1,3,5*	2.00
	(59)	(54)	(52)	(52)	(53)	⋮		(65)

'Data' from Table 1.

All SE's are multiplied by 1000.

* Mantel-Haenszel summary risk ratio, with SE of ln estimate back-calculated from test-based confidence interval.

Table 4: Fit of adapted beta-binomial generalized linear model to 5 cycles of Norman and Daya 'data.'

			Model	
Term	Parameter	Measure	$\phi = \frac{\beta_C/(\alpha_C+\beta_C)}{\beta_E/(\alpha_E+\beta_E)}$	No Constraint
t	γ_1		3.15(0.14)	3.14(0.14)
$1-t$	γ_0		6.31(0.45)	6.36(0.45)
		Ratio*	2.00	2.03
		ln Ratio	0.694(0.084**)	
$U_E \times t$	δ_{1E}		0.58(0.13)	0.36
$U_E \times (1-t)$	δ_{0E}		-	-0.00
$U_C \times t$	δ_{1C}		-	0.71
$U_C \times (1-t)$	δ_{0C}		0.93(0.31)	1.94

SE's shown in parentheses.

* The ratio estimate is 6.36/3.14.

** Since the covariance between the 6.36 and 3.14 is zero, the variance of the ln of the ratio, computed via the delta method, is $[(0.45/6.36)^2+(0.14/3.14)^2]^{1/2} = [1/499.7+1/196.6]^{1/2} = 0.084$.

Table 5: Estimates obtained by applying non-parametric and parametric methods to 200 datasets.*

	No. of moments fitted (f unspecified)					Beta-Binomial ϕ :		
Measure	1	2	3	4	5	$\frac{\beta_C}{\beta_E}$	$\frac{\beta_C/(\alpha_C+\beta_C)}{\beta_E/(\alpha_E+\beta_E)}$	N.C.**
Median$\{\hat{\theta}\}$	2.16	2.03	1.99	1.98	1.99	1.91	2.08	2.02
SD$\{\ln\hat{\theta}\}$	0.055	0.050	0.049	0.049	0.052	0.059	0.077	0.081
Mean{SE**}	0.058	0.054	0.052	0.052	0.054	0.062	0.078	0.083

* For details, see Section 4.

** N.C.: No constraint

** SE of $\ln\hat{\theta}$

Table 6a: Pregnancies and Fecundability in Cycles of Insemination with Fresh and Frozen Semen.

	Fresh semen			Frozen semen		
	Number of ...			Number of ...		
	Patients	Pregnancies	Rate	Patients	Pregnancies	Rate
1	**163**	**57**	**0.350**	125	18	0.144
2	69	18	0.261	**130**	**12**	**0.092**
3	**73**	**20**	**0.274**	87	8	0.092
4	59	12	0.203	**69**	**9**	**0.130**
5	**51**	**12**	**0.235**	50	1	0.020
6	51	12	0.235	**28**	**2**	**0.071**
1-6	466	131	0.281	489	50	0.102

The course of the patients who underwent insemination with fresh semen in the first cycle is tracked in bold. Data from Table 1 of Brown et al [10].

Table 6b: Estimates of Efficacy of Insemination with Frozen *vs.* Fresh Semen. SE of ln of estimate given in parentheses.

Cycles	Method/Model	Details ...	Efficacy
1-6	Brown *et al.*	$50/489 \div 131/466$	0.36(0.15)
1-6	M-H*	Summary Risk Ratio	0.37(0.15)
1, 3, 5	Norman & Daya	$27/262 \div 89/287$	0.35(0.19)
1, 3, 5	Norman & Daya, M-H*	Summary Risk Ratio	0.35(0.19)
1-6	Unspecified f	4 moments fitted	0.39(0.15)
1-6	Beta-binomial	Unconstrained	0.39(0.25)[1]

* Mantel-Haenszel summary risk ratio (summed over cycles), with SE of ln estimate back-calculated from test based-confidence interval.

[1] See footnote to Table 4. Deviance / df = 0.98; Chi-square goodness of fit statistic = 5.6 (6 df).

Table 7: Simulated example with varying risk ratio in each fecundability subpopulation; otherwise, same setup as in Table 1.

	Treatment received in the indicated cycle						
	Control				Experimental		
	Sub-population (fecundability)				Sub-population (fecundability)		
Cycle	High	Low	All	Ratio	All	High	Low
1	**200**	**800**	1000		1000	200	800
	(80)	**(80)**	**(160)**		(360)	(160)	(160)
			16%	2.25	36%		
2	40	640	640		**840**	**120**	**720**
	(16)	(60)	(76)		**(276)**	**(96)**	**180**
			11.8%	2.79	**32.9%**		
3	**24**	**540**	564		564	24	540
	(9.6)	**(54)**	**(63.6)**		(154.2)	(19.2)	(135)
			11.2%	2.43	27.3%		
4	4.8	405	409.8		**500.4**	**14.4**	**486**
	(1.9)	(40.5)	(42.4)		**(133)**	**(11.5)**	**121.5**
			10.3%	2.57	**26.6%**		
5	**2.9**	**364.5**	367.4		367.4	2.9	364.5
	(1.16)	**(36.45)**	**(37.61)**		(93.429)	(2.32)	(91.125)
			10.2%	2.48	25.4%		
1-5	271.7	2709.5	2981.2		3271.8	361.3	2910.5
	(108.7)	(270.6)	(379.3)		(1016.6)	(289.0)	(727.625)
			(12.7%)	2.45*	(31.1%*)		

The course of the patients who received the standard (control) treatment in the first cycle is tracked in bold.

Data-analysis options for comparisons of assisted reproductive technologies

James A. Hanley[1,2]
Nandini Dendukuri[1,3]
Robert Platt[1,4]
Marie-Hélène Mayrand[1]

Submitted to Statistical Methods in Medical Research, September 20, 2005

[1] Dept. of Epidemiology & Biostatistics, McGill University, Montreal, Quebec, Canada.
[2] Division of Clinical Epidemiology, Royal Victoria Hospital, Montreal.
[3] Technology Assessment Unit, McGill University Health Centre.
[4] Department of Pediatrics, Montreal Children's Hospital.

Correspondence:

Dr. J, Hanley, Department of Epidemiology and Biostatistics, McGill University,
1020 Pine Avenue West, Montreal, Quebec, Canada, H3A 1A2.

Telephone: +1 (514) 398-6270. Fax: +1 (514) 398-4503.
E-mail: James.Hanley@McGill.CA

Running Head: Trials of assisted reproductive technologies

SUMMARY

In the alternating-sequence design used to compare success rates with assisted reproductive technologies, women or couples are randomized to receive either the standard or experimental treatment in the first cycle, and— if they do not become pregnant—crossed between standard and experimental treatments after each successive cycle. Two authors have shown that, in the presence of heterogeneity of fecundability, and an effective treatment, the overall efficacy of the experimental treatment is overestimated by this design. These authors advised that in order to achieve an accurate estimate of efficacy, the trial should be run for at least three cycles and that all data from even-numbered cycles be excluded from the analysis, which should then be restricted only to odd-numbered cycles. In this paper, we describe approaches that make use of the data from all cycles. The methods are generalizations of those applicable to the constant-sequence design, where naive methods that do not take account of the heterogeneity produce underestimates of treatment efficacy.

Keywords

alternating sequence; fertility; experimental design; bias; precision; heterogeneity; generalized linear models

1. INTRODUCTION

Two experimental designs have been used to evaluate the efficacy of assisted reproductive technologies[1]. In the parallel-design, or constant-sequence randomized trial, the experimental treatment is administered for one or— if unsuccessful—more cycles to a fraction (usually one half, randomly chosen) of the eligible patients, and the control treatment for the same number of cycles to the remaining fraction. In the alternating-sequence design, some of the women or couples are randomized to receive the standard, and the others to receive the experimental treatment in the first cycle. Those who do not become pregnant are crossed to the opposite treatment after each successive cycle.

The relative merits of these two designs have been keenly debated [1-6]. Some arguments focus on efficiency and sample sizes: if the experimental therapy is effective, the alternating-design results in more pregnancies than the constant-sequence design, and is more attractive to couples. Others have to do with possible biases in the resulting estimates of efficacy. The first suggestion of bias came from comparisons of results of actual trials that used one or other of the two designs to evaluate the same procedure [2]. The authors noticed that, relative to those seen in parallel trials, treatment effects of the more effective treatment were higher in— i.e. overestimated by —crossover trials. Subsequent Monte Carlo evaluations [4], simulating patients from a heterogeneous subfertile population, indicated that while results from parallel trials appeared to slightly underestimate efficacy, the alternating-sequence design did indeed seem to slightly—but in their opinion not materially – overestimate it. Thus, they advised [4, p. 40] that "because of its practical advantages and because more pregnancies are achieved, a crossover design should be the first choice in infertility research."

The origin and nature of the biases in the estimates from these two designs can be readily understood by studying the worked example in Norman and Daya [6]. As shown in the first

row of Table 1, they assumed a heterogeneous population, where fecundability i.e., the per-cycle probability of getting pregnant, varied from couple to couple. To simplify matters, they assumed that with the less effective (Control) treatment, some 80% of couples had a $p_C = 10\%$, and the remaining 20% of couples a $p_C = 40\%$ fecundability, i.e.,

$$p_C = \begin{cases} 0.1 & \text{for } 80\% \text{ of couples,} \\ 0.4 & \text{for } 20\% \text{ of couples.} \end{cases}$$

Thus they assumed that the overall average fecundability is 16%, and the standard deviation is 12%.

They further assumed that the more effective (Experimental) treatment had a constant efficacy, θ, of 2, i.e., that at each cycle, a couple's probability of becoming pregnant was doubled. In Table 1 the course of the 1000 randomly allocated to undergo the 'control' treatment in the first cycle is tracked in bold. The entries at each cycle are the expected numbers of couples from the higher- and lower fecundity subpopulations who attempt to (and, in parentheses, the numbers who do) become pregnant. Using expected numbers of pregnancies at each cycle, Norman and Daya showed that estimates of efficacy that are based only on the total number of women and the total number of pregnancies are biased, irrespective of the design—the parallel design underestimates (apparent efficacy: θ =1.83, calculations not shown here but discussed later) and the alternating-sequence design overestimates (apparent efficacy $\theta = 2.10$, middle column Table 1). However, they noted that the bias in the alternating-sequence design is limited to the data from even-numbered cycles.

– Table 1 about here –

Despite the greater bias in the parallel design, Norman and Daya limited discussion of their concerns to—and aimed their cautions at prospective users of—the alternating sequence

design. They suggested [6, p. 323] a compromise between patient preference for this design and the statistical bias: "The objective of obtaining an accurate estimate of the effect of treatment, but also allowing all subjects to have the opportunity to receive the experimental treatment in at least one cycle, can now be achieved with the alternating-sequence design trial. The proviso is that the trial should run for at least three cycles and all data from the even-numbered cycles would have to be excluded from the analysis, which would be restricted only to the odd-numbered cycles." They concluded by advising [6, p. 310] that "when multiple cycles of treatment are undertaken to evaluate the efficacy of infertility therapy, the alternating-sequence design with restriction of the analysis to only the odd-numbered treatment cycles provides an unbiased estimation of the treatment effect."

This bias-avoiding strategy is unlikely to be an acceptable option for most investigators, patients and ethics review committees, and prompts the obvious questions: Must we discard 'biased' cycles and compensate for the decreased precision by increasing the numbers of couples enrolled? If we know the form of the bias, can we not remove it statistically using statistical models?

The purpose of this paper is to investigate this question, and several related ones. Under what model(s) is Norman's and Daya's approach really unbiased? If one can successfully eliminate the bias, at what price, in terms of increased imprecision, can this be achieved? Given the typically small sample sizes in this research area, can we afford this price, or might the overall mean squared error be smaller if we took a more naive approach? And, ultimately, if researchers use this design to collect their data, how should they analyze them, and how should they calculate the uncertainty in their estimates of efficacy? We restrict attention to models that use aggregated data for each treatment-cycle.

2. HOMOGENEOUS FECUNDABILITY

Let p_C denote a woman's fecundability i.e., her per-cycle probability of getting pregnant, with the less effective (control) treatment ($t = 0$). For now, assume no variation in p_C across women, i.e., that $\text{Var}[p_C] = 0$. Let p_E denote her fecundability with the experimental treatment ($t = 1$). Its efficacy with respect to the control treatment can be expressed in different ways using different forms for g in the generalized regression equation $g[p_E] = g[p_C] + \beta \times t$. For example, β is the absolute difference in fecundability if g is the identity function; $\exp[\beta]$ is the fecundability ratio θ if g is the ln function, or the fecundability odds ratio if g is the logit function. We will use the fecundability ratio to measure efficacy.

Suppose that one such woman, alternating from the experimental treatment in cycle 1, became pregnant on this treatment in the 5th cycle.

Cycle:	1	2	3	4	5
Treatment:	Exp'tl	Control	Exp'tl	Control	Exp'tl
Outcome:	—	—	—	—	+
Probability(+):	p_E	p_C	p_E	p_C	p_E
Probability(Outcome):	$1 - p_E$	$1 - p_C$	$1 - p_E$	$1 - p_C$	p_E

The observed data can be modeled as a sequence of independent Bernoulli trials with alternating probabilities of success. The likelihood based on this woman's data is the product of the probabilities of the 5 individual outcomes; it can also be re-arranged and written as a product of two geometric (but binomial-like) likelihoods, corresponding to $s_C = 0$ successful cycles, preceded by $u_C = 2$ unsuccessful ones, when the success probability was p_C ; and $s_E = 1$ successful cycle, preceded by $u_E = 2$ unsuccessful ones, when the success probability was p_E , i.e.,

$$L(u_C, u_E, s_C, s_E \mid p_C, p_E) \propto (1 - p_C)^2 \times (1 - p_E)^2 \times p_E$$

Since p_C and p_E are constant from woman to woman, so that all woman-cycles within the same treatment condition are exchangeable, the likelihood based on the data from several such women can again be written as the product of two binomial-like likelihoods

$$L(U_C, U_E, S_C, S_E \mid p_C, p_E) \propto (1-p_C)^{U_C} \times p_C^{S_C} \times (1-p_E)^{U_E} \times p_E^{S_E}$$

where $U_C = \Sigma u_C$ and U_E and S_C and S_E are the total numbers of unsuccessful and successful cycles when using C and E respectively, i.e., summed over all women and all cycles. The ML point estimator of the fecundability ratio is simply $(S_E/T_E)/(S_C/T_C)/$ where $T_E = S_E + U_S$ and $T_C = S_C + U_C$. A likelihood-based interval estimate is also easily calculated.

3. HETEROGENEOUS FECUNDABILITY

In reality, fecundability *does* vary among women, i.e., $\mathrm{Var}[p_C] > 0$ and $\mathrm{Var}[p_E] > 0$. We denote this variation by the general bivariate pdf $f(p_C, p_E)$, with marginal distribution $f(p_C)$. We present two data-analysis approaches which use aggregated data for each treatment in each cycle. The first makes no assumptions about the form of the marginal distribution $f(p_C)$, but strong ones about how a particular woman's fecundability with the experimental treatment is related to her fecundability with the standard one. In this approach, the observed data can be modeled either as (i) two sets of multinomial distributions, one for the numbers $\{S_{C_1}, S_{E_2}, \cdots\}$ who become pregnant in cycles $C_1, E_2, etc.$, the other for the numbers $\{S_{E_1}, S_{C_2}, \cdots\}$ who become pregnant in cycles $E_1, C_2, etc.$, or (ii) as cycle- and treatment-specific binomial random variables $\{\ S_{C_1}|T_{c_1}\}$, $\{S_{E_1}|T_{E_1}\}$, $\{S_{C_2}|T_{C_2}\}$, ... The second approach is based on a specific parametric form—Beta—for f, but does not 'connect' a particular woman's fecundability when undergoing experimental treatment with that same woman's fecundability with the standard treatment.

We evaluate these approaches using the data in Table 1 as well as data generated from continuous bivariate distribution for $\{p_C, p_E\}$. In the latter case, how a particular woman's fecundability, p_E, with the experimental treatment is related to her fecundability, p_C, with the standard one induced variability in the efficacy across women. In this second method, the observed data are modeled as cycle- and treatment-specific binomial random variables.

3.1 Unspecified-form for f; constant fecundability ratio

Consider an unspecified distribution $f(p_C)$ and let θ denote the constant ratio of p_E to p_C for all values of p_C. For a person with a specific value p_C, assigned to the sequence C, E, C, ... , the probabilities of becoming pregnant in cycle $1, 2, 3, ...$ are

$$p_C, (1-p_C) \times \theta \times p_C, (1-p_C) \times (1 - \theta \times p_C) \times p_C, ...$$

The probabilities, if that same person were assigned to the sequence E, C, E,. . . , are

$$\theta \times p_C, (1 - \theta \times p_C) \times p_C, (1 - \theta \times p_C) \times (1-p_C) \times \theta \times p_C, ...$$

Since p_C varies over persons, the multinomial proportions are the expectations of these probabilities, taken over the distribution, $f(p_C)$, of p_C. They form the numerators of the expressions given in Table 2. and represent the contribution to the (multinomial-based) likelihood of each person who becomes pregnant in that cycle. The extension beyond cycle 3 (not shown) is obvious, even if the algebra is tedious. Of note is the fact that the two likelihood contributions from cycle k involve the first k moments of the distribution of p_C. Others, e.g., [7,8,9], have noted this in the simpler constant-sequence design. Thus, each cycle adds two new data points and one new parameter; overall the $2K$ datapoints from K cycles are modeled by $K+1$ parameters. If $K \geq 2$, the remaining $K-1$ degrees of freedom can be used to assess model fit.

Table 2 about here

Maximum likelihood estimates for these $K+1$ parameters can be obtained from a non-linear modeling package, such as SAS PROC NLMIXED (see Appendix). Although our approach deals with heterogeneity, it does so without specifying a traditional random effects model: we used only the 'NL" portion of NLMIXED. We found that this *multinomial* approach is very sensitive to starting values, and have had more success by modeling the number who become pregnant on a specific treatment in a specific cycle—conditional on the number who used that specific treatment in that cycle—as a *binomial* random variable. These conditional probabilities are given as the quotients in Table 2. Again, they can be fitted using SAS PROC NLMIXED by (i) expressing the $2K$ binomial parameters in terms of the K moments and the parameter of interest θ, and (ii) for each of the $2K$ observed counts, modeling

$$Number_{pregnant} \sim Binomial(Number_{treated}, BinomialParameter).$$

For the data in Table 1, Maximum Likelihood estimates (and Standard Errors of their natural logarithms) for these parameters are shown in Table 3. This method correctly 'recovers' θ. Further, because the procedure uses data from all cycles, it produces smaller standard errors than those for the summary estimates from the odd-numbered cycles only. Moreover, one can achieve this increased precision, and only a slight inaccuracy, with fewer than the full K moments: one can omit i.e., set to zero in the likelihood, some of the higher order moments—those of order 3 or more in our example. This is because p_C is bounded by 0 and 1, so that the higher moments are of decreasing magnitudes, and thus increasingly negligible.

– Table 3 about here –

3.2 Beta-Geometric Model

That an experimental treatment would increase each p_C value a constant-fold, i.e., by the same multiple, θ for each woman, regardless of her value of p_C, is not realistic biologically. Whereas women whose natural fecundability is 10% might reasonably have it increased to 20% i.e., by a factor of $\theta = 2$, with experimental treatment, the treatment is unlikely to also raise other women's already high natural per-cycle success probability of 40%, say, by the same (multiplicative) factor of $\theta = 2$, i.e., to 80%. Moreover, if $p_C > 1/\theta$, this assumption of a constant θ is statistically impossible. In addition, there are practical technical difficulties in fitting such a high-order nonlinear model; the number of parameters (moments) relative to the numbers of observations is large. For these reasons, we turn to more natural parametric statistical models for p_C and p_E, ones with fewer constraints on how a particular woman's fecundability when undergoing experimental treatment relates to what might be (loosely) called its 'counterfactual' i.e., the same woman's fecundability with the standard treatment.

The Beta-Geometric (B-G) model has been used in demography [7]. More recently, Weinberg and Gladen [8] used it in a non-experimental study of the effect of smoking on fecundability. Since women were classified as smokers or non-smokers for the entire period of observation (up to 12 cycles), their model immediately applies to a parallel-sequence design [4]. The latter authors used a modified B-G model to *generate* data, but did not consider it for the *analysis* of their data. We extend these ideas to develop a beta-geometric model for data from the alternating-sequence design.

Before doing so, we review its use for the constant-sequence design. Weinberg and Gladen compared fecundability, measured over 12 cycles, in smokers relative to non-smokers. They modeled fecundability in the two source populations as Beta distributions, with their respective location and shape governed by the pairs of parameters $\{\alpha_C,\beta_C\}$ and $\{\alpha_E,\beta_E\}$.

Therefore, before the first cycle the probability density function of p_C is given by:

$$f[p_c] \propto {p_c}^{\alpha_C - 1}(1 - p_c)^{\beta_C - 1}$$

The mean and variance of the fecundability in the control group before the first cycle are $\mu_C = \alpha_C/(\alpha_C + \beta_C)$ and $\sigma_C^2 = \alpha_C\beta_C/((\alpha_C + \beta_C)^2(\alpha_C + \beta_C + 1))$ Weinberg and Gladen showed that in those couples who were unsuccessful in U previous cycles, the—*now conditional*—distribution of p_C at cycle $U + 1$ in this selected subgroup is shifted towards the left i.e., towards zero, but remains a Beta distribution, with probability density function:

$$f(p_c|U) \propto {p_c}^{\alpha - 1}(1 - p_c)^{\beta - 1 + U}.$$

The parameters of the fecundability distribution are now $\{\alpha_C,\ \beta_C + U\}$. Thus, after k cycles the mean fecundability is given by $\mu_C = \alpha_C/(\alpha_C + \beta_C + k)$. They further showed that the expected probability of success among those who enter cycle U+1 is related to the number of previously unsuccessful cycles U via the simple reciprocal link:

$$1/E[p_C|U] = (\alpha_C + \beta_C + U)/\alpha = (\alpha_C + \beta_C)/\alpha + (1/\alpha) \times U$$

$$= \gamma + \delta \times U$$

The parameter γ is the expected number of cycles to become pregnant if a couple had the average per-cycle probability under the control treatment i.e., $\gamma = 1/\mu_C$. The parameter δ reflects the spread of the initial distribution of p_C: under homogeneity, the number of cycles required to achieve pregnancy reduces to the same geometric random variable for each couple, i.e., $\delta = 0$. The parameters γ and δ can be fit using binomial regression with an inverse (i.e., power-1) link, e.g., using PROC GENMOD in the SAS, or glm in Stata. Since glm in R does not allow this link for the binomial, one needs to supply the variance function.

Weinberg and Gladen extended the model to effectively fit separate Beta distributions for the cycle-specific probabilities for smokers ($t = 1$) and non-smokers($t = 0$), via one equation:

$$1/E[p_C \mid U] = (\gamma_C + \delta_C \times U) \times (1 - t) + (\gamma_E + \delta_E \times U) \times t.$$

To extend it to the alternating sequence design, we model the expected probability of pregnancy for women who have already undergone U_E and U_C unsuccessful cycles on E and C, and are now about to receive (say) E. The initial Beta(α_E,β_E) distribution must be updated to reflect the U_E and U_C. The U_E is added to the β_E term as in the parallel sequence model, but a correction must be made to the U_C. If C were no more effective than E, and thus exchangeable with it, we would add the full U_C to the U_E to obtain the β_E term of $U_E + U_C$. But if C were less effective than E, then using the full $U_E + U_C$ in the β_E term would shift the fecundability distribution too far to the left. This is most easily seen if *no* pregnancies can occur with C. The greater the efficacy of E relative to C, (i.e., the smaller the 'failure ratio' $(1 - p_E)/(1 - p_C)$), the smaller should be the 'amalgam' of U_E and U_C. Thus, we suggest that the contribution of U_C be reduced, and propose the amalgam

$$U^* = U_E + FRR \times U_C$$

where FRR is the ratio of the probability of failure on the experimental and control treatments at cycle 1. This ratio is less than unity if E is more effective than C. Thus, the probability density function of p_E at cycle $u_C + u_E + 1$ is taken to be

$$f(p_E \mid u_E\ u_C) \propto {p_E}^{\alpha_E - 1}(1 - p_E)^{\beta_E - 1 + u_E + FRR \times u_C},$$

i.e., a Beta(α_E , $\beta_E + u_E + FRR \times u_C$) distribution. Thus the expected probability of success at this cycle is inversely proportional to a linear function of u_E and u_C, i.e .

$$E(p_E \mid u_C\ \ u_E) = \alpha_E/(\alpha_E + \beta_E + u_E + FRR \times u_C)$$

Similarly, we can show that

$$E(p_C \mid u_C \;\; u_E) = \alpha_C/(\alpha_C + \beta_C + u_C + (1/FRR) \times u_C)$$

We re-formulate it as a generalized linear model for binary data with an inverse link function:

$$\begin{array}{rlll} 1/E(p \mid u_C \;\; u_E) = & \gamma_{0E} \times t & + & \gamma_{0C} \times (1-t) \\ + & \gamma_{1E} \times (u_E \times t) & + & \gamma_{1C} \times (u_C \times (1-t)) \\ + & \gamma_{2E} \times (u_C \times t) & + & \gamma_{2C} \times (u_E \times (1-t)) \end{array} \quad (3)$$

where $t = 1$ under the experimental treatment and $t = 0$ under the control treatment. Constraints should be placed on γ_{2E}, which is a function of γ_{0E} and γ_{1E}, and on γ_{2C}. The model can be fit using PROC NLMIXED in SAS, which can accommodate constraints. The parameters of the original Beta distributions can be obtained by the following transformations.

$$\alpha_E = 1/\gamma_{1E};\;\; \beta_E = (\gamma_{0E} - 1)/\gamma_{1E};\;\; \alpha_C = 1/\gamma_{1C};\;\; \beta_C = (\gamma_{0C} - 1)/\gamma_{1C},$$

and efficacy is estimated by $\hat{\theta} = \hat{\gamma_{0C}}/\hat{\gamma_{0E}}$.

In a Bernoulli model, woman-level covariates could be added as linear terms in (3). Notice that at each cycle the fecundability distribution under treatment t is obtained by adding to the β parameter of its starting fecundability distribution, the number of unsuccessful cycles on t and a multiple of the number of those on the other treatment. While we have used the multiplying factors as FRR and $1/FRR$ above, we could use the generic expressions $\alpha_C/(\alpha_C + \beta_C + u_C + \phi_C \times u_C)$ and $\alpha_E/(\alpha_E + \beta_E + u_E + \phi_E \times u_C)$ for the respective mean fecundability after cycle $u_C + u_E$. ϕ_C and ϕ_E should be constrained to be > 1 and < 1 respectively, assuming E is more effective than C.

4. EVALUATION

We assessed the performance of these analysis models on 200 generated datasets. Following Cohlen et al. [4] we began with Beta distributions with means of 0.16 and 0.32, each

with a coefficient of variation of 75% (Cohlen et al. used 56%). We calculated the 9th, 18th, ..., 90th percentiles for each of these two initial distributions, and placed a point mass of 0.1 at each of these values, thereby creating two 10-point distributions for the first cycle. If a woman's fecundability was at say the 18th percentile when on C i.e. if $p_C = 5.2\%$, then her fecundability with E (if necessary) was the corresponding 18th percentile in that distribution, namely $p_E = 8.1\%$ (fecundability ratio = 1.56). Similarly, those just above the middle of the p_C distribution, i.e., a fecundability of $p_c = 15.2\%$, were considered to be just above the middle of the p_E distribution, namely $p_E = 31.6\%$ (fecundability ratio = 2.08), while in the 3rd highest subgroup, fecundabilities were $p_C = 22.6\%$ and $p_E = 48.2\%$ (ratio = 2.13). For each dataset, 1000 women were randomly assigned, using a multimomial distribution, to the 10 fecundability levels under C, and 1000 others to the 10 corresponding levels in the p_E distribution. At each cycle, these two sets of 10 subgroup frequencies were depleted using random numbers of pregnancies generated by the 20 corresponding binomial distributions. The numbers who were unsuccessful were switched to their corresponding level on the opposite distribution, before generating the pregnancies for the next cycle.

The estimates produced by our analysis models are summarized in Table 5. Leaving the form of f unspecified and estimating its first few moments is somewhat more efficient than using the Beta-Binomial model, where there is more of a tradeoff between bias and precision.

5. EFFICACY ESTIMATES FROM A CLINICAL TRIAL

The data in Table 6a are from (by today's standards) a very large clinical trial. It evaluated the performance of a second generation protocol for donor insemination with frozen semen [10]. Today, screening of semen for HIV infection precludes the use of fresh semen. Earlier, in the first use that we have found of the alternating sequence design, this group

achieved a fecundability rate of only 5.0 pregnancies per 100 cycles with a first-generation protocol, versus 18.9 per 100 using fresh semen [11].

The data for the first six cycles (during which, for each woman, semen was from her matched donor) are shown in Table 6a. However, as is obvious from the denominators, the same practical difficulties were encountered as those mentioned in the first study: "Cryopreserved semen was frequently substituted in a cycle scheduled to be fresh because the donor was not available." Unfortunately, information on the actual sequence for each woman is no longer available (S. Shapiro, personal communication, 2002).

Estimates of efficacy are given in Table 6b. Those produced by the methods in 3.1 and 3.2 are closest to the null, suggesting that they removed some of the bias induced when one ignores the heterogeneity. Possibly by chance, given its poor precision, the estimator advocated by Norman and Daya was furthest from the null, further than both the crude and the Mantel-Haenszel estimates. The Generalized Linear Model estimate was closest to the null, but had a high SE, possibly because of the large number of parameters (6) fitted to the 12 datapoints. The method of moments had the lowest SE. The fact that it was no different from that of the Mantel Haenszel estimator suggests that, in this study, heterogeneity does not substantially inflate or deflate the SE. Without individual-specific data, we are unable to assess how much heterogeneity could also be affected by women who dropped out.

6. DISCUSSION

The more attractive alternative sequence design also produces more pregnancies for the same number of cycles. We describe two approaches that allow clinical trials to use data from all of the cycles, while Norman and Daya's approach [6] squanders the statistical advantage of this design. Many contemporary trials generate fewer than 25 pregnancies in total. De-

creasing precision further by omitting even-numbered cycles, without trying to eliminate the biases by other methods, is difficult to defend. Moreover, with the sample sizes considered here reduced by a realistic factor of 25, sampling variability dominates analytic bias.

The bias caused by naively using all data is a function of the heterogeneity in p_C and the efficacy of the experimental treatment ; *both must be substantial in order to produce a serious bias*. Norman and Daya based their concerns on an extreme 2-point distribution of p_C, and a treatment that doubled the $p_C = 40\%$ in the high fertility subgroup to a 'biologically nearly impossible' $p_E = 80\%$. Even then, the bias was less than the sampling variability induced by the sample sizes used in practice. With the same $\theta = 2$, and a more realistic distribution where $f[0.025] = 0.8$ and $f[0.1] = 0.2$ so that $\text{mean}[p_C] = 0.04, SD[p_C] = 0.03$, the bias was much smaller Moreover, if, rather than increasing p, an new treatment— such as frozen semen—*reduces* it, the degree of bias in the naive estimate of θ is also less, because of the smaller impact of the differential success (removal) of the most fertile in odd-numbered cycles. For example, using $\theta = 0.5$ in Table 2, the data from cycle 2 yield $\hat{\theta} = 0.47$, a relative bias of only 6%. These 'low-bias' conditions would also apply in comparisons of contraceptive methods, where the probability of an unwanted pregnancy is already low.

The alternating sequence design is not a full *crossover* study. Nor does it carry the full statistical efficiency usually associated with self-matched comparisons. Thus, the sample size and power calculations/projections are best carried out by analogy with unmatched designs.

In practice, for example in the study by Brown et al. [10], there are unavoidable individual deviations from the alternating-sequence protocol. Some investigators may use variations of the alternating design: e.g., in Ecochard et al. [12], 1/2 patients received one treatment for the first two cycles and the competing treatment for the next two cycles, while the other 1/2 followed the opposite sequence. For these more complex designs, random effects models for

binary (Bernoulli) data allow one to model the cycle-by-cycle sequence of outcomes for each woman using a full regression approach that makes use of all of the individual level data for each woman (any baseline covariates, the cycle-by-cycle treatment indicators and any other available cycle-dependent covariates). Some quite complex 2-level hierarchical models have been used in such circumstances [13,14].

The approaches we have described are also applicable to simple data analyses for the parallel- or 'constant-sequence' randomized trial design. Since this competing design has the same data structure as Weinberg and Gladen's example (fecundability of smokers and non-smokers), their 'Beta-Geometric' Model is immediately applicable without modification. The method based on moments is also applicable. Applied to Norman and Daya's 'constant-sequence' example (and the 'data' in their Table 1), both of our methods produce estimates closer to the true $\theta = 2$, whereas a naive analysis produces an attenuated estimate of 1.83.

In an appendix, Norman and Daya [6, p324] claim that "the assumptions of a constant drug efficacy is not necessary" by considering an arbitrary distribution $f[p_C]$, and an arbitrary efficacy function, $\theta[p_C]$ They purport to show algebraically that "the outcome rates in the odd cycles in an alternating sequence are unbiased," i.e., that "the results will hold true regardless of the relationship between efficacy and fertility." In fact, the ratio from alternate cycles will not continue to be unbiased if the treatment effect is variable. This is illustrated in Table 7 by slightly perturbing Norman and Daya's simulated example so that the risk ratio in the low fecundability group is 2.5, while the risk ratio in the high fecundability group remains at 2. This is closer to what happens in reality where there is likely to be a greater relative shift in the low fecundability groups, compared to the high fecundability groups. The true average risk ratio across the population is thus $2.5 \times 0.8 + 2 \times 0.2 = 2.4$. However, from Table 7 we can see that this estimate is not obtained even in the first cycle: the ratio

of the expected probabilities of successes does not match the expectation of the ratios of the success probabilities, i.e.

$$\frac{\Sigma\ \theta[p_C] \times p_C \times f[p_C]}{\Sigma\ p_C \times f[p_C]} \neq \Sigma\ \theta[p_C] \times f[p_C] = \theta[p_C]$$

Further, it appears that the odd cycles underestimate the true risk ratio while the even cycles overestimate it. If the study continues to a point when only women in the low fecundability group remain, then the ratio approach the true ratio of 2.5 in both odd and even cycles

This contrary finding is an additional impetus to consider more general regression models that allow not just between-individual heterogeneity, and covariates at the woman-cycle level, but also more flexibility in the specification of the comparative parameter. We plan to investigate whether the amount of data from a typical alternating sequence design makes such models practical. Unlike traditional studies with multiple crossovers, the alternating sequence design involves at most one instance of Y=1 per subject, and such outcomes preclude further observations. This, the small sample sizes, and the small number of cycles usually used, may be a serious impediment to more complex modeling. The analyses we have presented, based on marginal distributions, may well be the appropriate ones for the amounts of clinical trial data generated by the alternating sequence design.

ACKNOWLEDGMENTS

Nandini Dendukuri is a chercheur-boursier of the Fonds de Recherche en Santé du Québec. Robert Platt is a career scientist of the Canadian Institutes of Health Research.This work was supported by individual operating grants from the Natural Sciences and Engineering Research Council of Canada and a team grant from Le Fonds Québécois de la recherche sur la nature et les technologies.

Code

xxxx

xxxx xxxx xxxxxxxx

xxxxx xxxxxx xxxxxx.

REFERENCES

1. DAYA, S. Is there a place for the crossover design in infertility trials? *Fertility and Sterility* 1993; **59**: 67.
2. KHAN, K.S, DAYA S, COLLINS J.A., AND WALTER S.D. (1996). Empirical evidence of bias in infertility research: overestimation of treatment effect in crossover trials using pregnancy as the outcome measure. *Fertility and Sterility* **65**:939945.
3. TE VELDE, E.R., COHLEN B.J., LOOMAN C.W., AND HABBEMA, J,D. Crossover designs versus parallel studies in infertility research. [letter] *Fertility and Sterility* 1998; **69**:357-358.
4. COHLEN, B.J., TE VELDE, E.R., LOOMAN. C.W.N., EIJCHEMANS, R., AND HABBEMA, J.D.F. Crossover or parallel design in infertility trials? The discussion continues. *Fertility and Sterility* 1998; **70**:40-45.
5. DAYA, S. Differences Between Crossover and Parallel Study DesignsDebate? (letter) *Fertility and Sterility* 1999; **71**:771-772.
6. NORMAN, G.R. AND DAYA, S. The alternating-sequence design (or multiple-period crossover) trial for evaluating treatment efficacy in infertility. *Fertility and Sterility* 2000; **74**:319-324.
7. SHEPS, M.C., AND MENKEN, J.A. Mathematical models of conception. University of Chicago Press, 1973.
8. WEINBERG, C.R. AND GLADEN B.C. The beta-geometric distribution applied to comparative fecundability studies. *Biometrics* 1986; **42**:547-560.
9. LAU, T.S. On the heterogeneity of fecundability. *Lifetime Data Analysis* 1996; **2**:403-415.
10. BROWN, C.A., BOONE, W.R., AND SHAPIRO, S.S. Improved cryopreserved semen fecundability in an alternating fresh-frozen insemination program. *Fertility and Sterility* 1988; **50**:825-827.
11. RICHTER, M.A., HANING, R.V., AND SHAPIRO, S.S. Artificial donor insemination: fresh versus frozen semen: the patient as her own control. *Fertility and Sterility* 1984; **41**:277-280.
12. ECOCHARD, R, MATHIEU, C., ROYERE, D., BLACHE, G., RABILLOUD, M., AND CZYBA, J.C. A randomized prospective study comparing pregnancy rates after clomiphene citrate and human menopausal gonadotropin before intrauterine insemination. *Fertility and Sterility* 2000; **73**:90-93
13. ECOCHARD R, CLAYTON DG. Multi-level modelling of conception in artificial insemination by donor. *Statistics in Medicine* 1998; **17**(10):1137-1156.
14. ECOCHARD R, CLAYTON DG. Multivariate parametric random effect regression models for fecundability studies. *Biometrics* 2000; **56**(4):1023-1029.

Table 1: Cycle-specific ratios of expected pregnancy proportions if the alternating sequence design is applied to a population of 2000 which is heterogeneous with respect to spontaneous fecundity [20% with higher, 80% with lower fecundity].

	Treatment received in the indicated cycle						
	Control				Experimental		
	Sub-population (fecundability)				Sub-population (fecundability)		
Cycle	High	Low	All	Ratio	All	High	Low
1	**200**	**800**	1000		1000	200	800
	(80)	**(80)**	**(160)**		(320)	(160)	(160)
			16%	2.00	32%		
2	40	640	680		**840**	**120**	**720**
	(16)	(64)	(80)		**(240)**	**(96)**	**144**
			11.8%	2.43	**28.6%**		
3	**24**	**576**	600		600	24	576
	(9.6)	**(57.6)**	**(67.2)**		(134.4)	(19.2)	(115.2)
			11.2%	2.00	22.4%		
4	4.8	460.8	465.6		**532.8**	**14.4**	**518.4**
	(1.9)	(46.1)	(48.0)		**(114.9)**	**(11.5)**	**103.7**
			11.3%	2.10	**21.6%**		
5	**2.9**	**414.7**	417.6		417.6	2.9	414.7
	(1.2)	**(41.5)**	**(42.6)**		(85.2)	(2.3)	(82.9)
			10.2%	2.00	20.4%		
1-5	271.7	2891.5	3163.2		3390.4	361.3	3029.1
			(397.8)		(894.8)		
			(12.6%)	2.10*	(26.5%*)		

The course of the 1000 randomly allocated to undergo the 'control' treatment in the first cycle is tracked in bold. The entries at each cycle are the expected numbers of couples from the higher- and lower fecundity subpopulations who attempt to (and, in parentheses, the numbers who do) become pregnant. Table adapted from Figure 2 and Table 2 of Norman and Daya.

Table 2: Unconditional (multinomial) and conditional success probabilities for each of the first three cycles, as a function of the efficacy, θ, and the (absolute) moments of the unspecified distribution of p_C, the fecundability under the standard ["control" (C)] treatment.

Cycle	Control	Experimental
1	μ_1	$\theta \times \mu_1$
2	$\frac{\mu_1 - \theta \times \mu_2}{1 - \theta \times \mu_1}$	$\frac{\theta \times \mu_1 - \theta \times \mu_2}{1 - \mu_1}$
3	$\frac{\mu_1 - \mu_2 - \theta \times \mu_2 + \theta \times \mu_3}{1 - \mu_1 - \theta \times \mu_1 + \theta \times \mu_2}$	$\frac{\theta \times \mu_1 - \theta \times \mu_2 - \theta^2 \times \mu_2 + \theta^2}{1 - \mu_1 - \theta \times \mu_1 + \theta \times \mu_2}$

The numerators represent the unconditional probabilities of pregnancy in the indicated cycle for persons *entering* the study, while while the quotients represent conditional pregnancy probabilities for those who receive the indicated treatment in the *indicated* cycle. These probabilities are computed separately for those randomly allocated to the 'C to E to C' sequence, and conversely for their counterparts. Cycle 1 starts with denominators of 1 (100%) in each group; it is assumed that there are no dropouts [i.e. women/couples who have not yet become pregnant do not abandon the study] or that dropouts are 'at random' and unrelated to their values of p_C. The symbols μ_1 to μ_3 are the first 3 absolute moments of the distribution of p_C, the fecundability with standard treatment.

Table 3: MLEs of the efficacy parameter θ (SE of ln of estimate) as a function of number of data cycles used, and number of moments of unspecified distribution f estimated, compared with estimates obtained using approach of Norman and Daya.

	No. moments of f fitted					⋮	Norman and Daya	
Cycles	1	2	3	4	5	⋮	Cycles	$\hat{\theta}$
used						⋮	used	(SE*)
1	2.00					⋮	*1*	2.00
	(86)					⋮		(86)
1,2	2.33	2.00				⋮		
	(77)	(66)				⋮		
1,2,3	2.18	2.06	2.00			⋮	*1,3*	2.00
	(66)	(60)	(60)			⋮		(70)
1,2,3,4	2.28	2.01	2.02	2.00		⋮		
	(64)	(55)	(55)	(56)		⋮		
1,2,3,4,5	2.18	2.07	2.01	2.00	2.00	⋮	*1,3,5*	2.00
	(59)	(54)	(52)	(52)	(53)	⋮		(65)

'Data' from Table 1.

All SE's are multiplied by 1000.

* Mantel-Haenszel summary risk ratio, with SE of ln estimate back-calculated from test-based confidence interval.

Table 4: Fit of adapted beta-binomial generalized linear model to 5 cycles of Norman and Daya 'data.'

			Model	
Term	Parameter	Measure	$\phi = \frac{\beta_C/(\alpha_C+\beta_C)}{\beta_E/(\alpha_E+\beta_E)}$	No Constraint
t	γ_1		3.15(0.14)	3.14(0.14)
$1-t$	γ_0		6.31(0.45)	6.36(0.45)
		Ratio*	2.00	2.03
		ln Ratio	0.694(0.084**)	
$U_E \times t$	δ_{1E}		0.58(0.13)	0.36
$U_E \times (1-t)$	δ_{0E}		-	-0.00
$U_C \times t$	δ_{1C}		-	0.71
$U_C \times (1-t)$	δ_{0C}		0.93(0.31)	1.94

SE's shown in parentheses.

* The ratio estimate is 6.36/3.14.

** Since the covariance between the 6.36 and 3.14 is zero, the variance of the ln of the ratio, computed via the delta method, is $[(0.45/6.36)^2+(0.14/3.14)^2]^{1/2} = [1/499.7+1/196.6]^{1/2} = 0.084$.

Table 5: Estimates obtained by applying non-parametric and parametric methods to 200 datasets.*

	No. of moments fitted (f unspecified)					Beta-Binomial ϕ :		
Measure	1	2	3	4	5	$\frac{\beta_C}{\beta_E}$	$\frac{\beta_C/(\alpha_C+\beta_C)}{\beta_E/(\alpha_E+\beta_E)}$	N.C.**
Median$\{\hat{\theta}\}$	2.16	2.03	1.99	1.98	1.99	1.91	2.08	2.02
SD$\{\ln\hat{\theta}\}$	0.055	0.050	0.049	0.049	0.052	0.059	0.077	0.081
Mean{SE**}	0.058	0.054	0.052	0.052	0.054	0.062	0.078	0.083

* For details, see Section 4.

** N.C.: No constraint

** SE of $\ln\hat{\theta}$

Table 6a: Pregnancies and Fecundability in Cycles of Insemination with Fresh and Frozen Semen.

	Fresh semen			Frozen semen		
	Number of ...			Number of ...		
	Patients	Pregnancies	Rate	Patients	Pregnancies	Rate
1	**163**	**57**	**0.350**	125	18	0.144
2	69	18	0.261	**130**	**12**	**0.092**
3	**73**	**20**	**0.274**	87	8	0.092
4	59	12	0.203	**69**	**9**	**0.130**
5	**51**	**12**	**0.235**	50	1	0.020
6	51	12	0.235	**28**	**2**	**0.071**
1-6	466	131	0.281	489	50	0.102

The course of the patients who underwent insemination with fresh semen in the first cycle is tracked in bold. Data from Table 1 of Brown et al [10].

Table 6b: Estimates of Efficacy of Insemination with Frozen *vs.* Fresh Semen. SE of ln of estimate given in parentheses.

Cycles	Method/Model	Details ...	Efficacy
1-6	Brown *et al.*	$50/489 \div 131/466$	0.36(0.15)
1-6	M-H*	Summary Risk Ratio	0.37(0.15)
1, 3, 5	Norman & Daya	$27/262 \div 89/287$	0.35(0.19)
1, 3, 5	Norman & Daya, M-H*	Summary Risk Ratio	0.35(0.19)
1-6	Unspecified f	4 moments fitted	0.39(0.15)
1-6	Beta-binomial	Unconstrained	0.39(0.25)[1]

* Mantel-Haenszel summary risk ratio (summed over cycles), with SE of ln estimate back-calculated from test based-confidence interval.

[1] See footnote to Table 4. Deviance / df = 0.98; Chi-square goodness of fit statistic = 5.6 (6 df).

Table 7: Simulated example with varying risk ratio in each fecundability subpopulation; otherwise, same setup as in Table 1.

	Treatment received in the indicated cycle						
	Control				Experimental		
	Sub-population (fecundability)				Sub-population (fecundability)		
Cycle	High	Low	All	Ratio	All	High	Low
1	**200**	**800**	1000		1000	200	800
	(80)	**(80)**	**(160)**		(360)	(160)	(160)
			16%	2.25	36%		
2	40	640	640		**840**	**120**	**720**
	(16)	(60)	(76)		**(276)**	**(96)**	**180**
			11.8%	2.79	**32.9%**		
3	**24**	**540**	564		564	24	540
	(9.6)	**(54)**	**(63.6)**		(154.2)	(19.2)	(135)
			11.2%	2.43	27.3%		
4	4.8	405	409.8		**500.4**	**14.4**	**486**
	(1.9)	(40.5)	(42.4)		**(133)**	**(11.5)**	**121.5**
			10.3%	2.57	**26.6%**		
5	**2.9**	**364.5**	367.4		367.4	2.9	364.5
	(1.16)	**(36.45)**	**(37.61)**		(93.429)	(2.32)	(91.125)
			10.2%	2.48	25.4%		
1-5	271.7	2709.5	2981.2		3271.8	361.3	2910.5
	(108.7)	(270.6)	(379.3)		(1016.6)	(289.0)	(727.625)
			(12.7%)	2.45*	(31.1%*)		

The course of the patients who received the standard (control) treatment in the first cycle is tracked in bold.

Statistical models for data from the alternating sequence design.

Nandini Dendukuri[1,2,5], James A. Hanley[1,3], Robert Platt[1,4] and Marie-Hélène Mayrand[1]

[1] Dept. of Epidemiology, Biostatistics, and Occupational Health, McGill University, Montreal, Quebec, Canada. [2] Technology Assessment Unit, McGill University Health Centre. [3] Division of Clinical Epidemiology, Royal Victoria Hospital, Montreal. [4] Department of Pediatrics, Montreal Children's Hospital.

[5] to whom correspondence should be adressed at ..... E-mail: N.D@xxx.xx

BACKGROUND: The alternating-sequence design has been suggested as an attractive and statistically efficient way to compare pregnancy rates achievable with a new and an existing technology within a randomized clinical trial. However, some of the suggested statistical analysis strategies for data collected using this design either fail to make full use of the data, or use statistical models that ignore inter-patient variation. AIM / METHODS: This note describes – and provides methods for fitting – statistical models for data from the alternating-sequence design. RESULTS: These models accommodate patient heterogeneity and make full use of the data from the alternating-sequence design. They can also be used to more appropriately analyze data gathered in single-arm and parallel-sequence designs.

Keywords: Alternating sequence; fertility; experimental design; bias; precision; heterogeneity; generalized linear models

Running Head: Statistical models

[version May 12, 2008] ... To be submitted to Human Reproduction, 2008

Introduction

Two competing experimental designs have been used to evaluate the efficacy of assisted reproductive technologies[1]. In the parallel-design, or constant-sequence randomized trial, the experimental treatment is administered for one or— if unsuccessful—more cycles to a fraction (usually one half, randomly chosen) of the eligible patients, and the comparison ('control') treatment for the same number of cycles to the remaining fraction. In the alternating-sequence design, some of the women or couples are randomized to receive the standard, and the others to receive the experimental treatment in the first cycle. Those who do not become pregnant are crossed to the opposite treatment after each successive cycle.

The relative merits of these two designs have been vigorously debated [1-6]. Some arguments focus on efficiency and sample sizes: if the experimental therapy is effective, the alternating-design results in more pregnancies than the constant-sequence design, and is more attractive to couples. Others have to do with possible biases in the resulting estimates of efficacy. The first suggestion of bias came from comparisons of results of actual trials that used one or other of the two designs to evaluate the same procedure [2]. The authors noticed that, relative to those seen in parallel trials, treatment effects were higher in— i.e. were overestimated by —crossover trials. Subsequent Monte Carlo evaluations [4], simulating patients from a heterogeneous subfertile population, indicated that while results from parallel trials appeared

to slightly underestimate efficacy, the alternating-sequence design did indeed seem to slightly—but in their opinion not materially – overestimate it. Thus, they advised [4, p. 40] that "because of its practical advantages and because more pregnancies are achieved, a crossover design should be the first choice in infertility research."

Despite the greater bias in the parallel design, Norman and Daya limited discussion of their concerns to—and aimed their cautions at prospective users of—the alternating sequence design. Based on their numerical calculations (repeated here, in Appendix 1), they suggested [6, p. 323] a compromise between patient preference for this design and the statistical bias: "*The objective of obtaining an accurate estimate of the effect of treatment, but also allowing all subjects to have the opportunity to receive the experimental treatment in at least one cycle, can now be achieved with the alternating-sequence design trial. The proviso is that the trial should run for at least three cycles and all data from the even-numbered cycles would have to be excluded from the analysis, which would be restricted only to the odd-numbered cycles.*" They concluded by advising [6, p. 310] that "when multiple cycles of treatment are undertaken to evaluate the efficacy of infertility therapy, the alternating-sequence design with *restriction of the analysis to only the odd-numbered treatment cycles provides an unbiased estimation of the treatment effect.*"

This bias-avoiding strategy is unlikely to be an acceptable option for

most investigators, patients and ethics review committees, and prompts the obvious questions: *Must* we discard 'biased' cycles and compensate for the decreased precision by increasing the numbers of couples enrolled? *If we know the form of the bias, can we not remove it using statistical models while maintaining the full statistical efficiency*?

The most recent publication [7] on the topic, based on a regression model that uses the data from all cycles, appears to confirm the validity of the alternating-sequence design. However, the data-analysis methods in it, and in some earlier publications, rely on models that do not explicitly acknowledge inter-patient variation.

The purpose of this paper is to show that one can use biologically realistic statistical models to remove the bias demonstrated by Norman and Daya, without having to remove some of the data. In the process, we illustrate a number of statistical models that explicitly include between-patient heterogeneity, a phenomenon that all agree exists, but not all authors include in their analyses. In order to illustrate these models, many of which have been set forth in statistical rather than substantive journals, we begin with the simplest type of data, from a single-arm, 'constant-sequence', design. We then extend the presentation to include models for comparative studies, where the focus is on a comparative parameter, such as a success ratio, rather than on the heterogeneity parameters per se. The models are general, and can be used for both alternating and constant-sequence designs.

Modelling heterogeneity in one-arm studies

Consider the simplest type of data, from a single-arm 'constant-sequence' design, or a cohort study of persons with a common characteristic (in the first example, presented in Table 1, e.g. all were non-smokers). In the reporting of the results in such cohorts, it is common to see statements such as that "the probability of success was lower at each successive cycle." Thus, in the statistical modelling of the cycle-specific success rates, say by logistic regression, data-analysts will often add a term to reflect this 'decline.'

But does an *individual's* probability of success really *decrease* in each successive cycle? To see why this may not be the best (or only) interpretation, consider the cycle-specific 'success' rates one would observe in a large group of persons if the data-generation process followed the following statistical model: half are given a six-sided die, and asked to roll it once each cycle until they first "succeed" when a six shows upward; the others are given a coin, and asked to toss it until a specified face shows upwards. Persons are not allowed to switch between the die and the coin. The investigators do not observe which outcomes were generated by which subgroup – in statistical terminology, the subgroups and their associated differential success rates, are *latent*. As can easily be calculated, the *overall* cycle-specific success rates, shown below, would exhibit a *decreasing* trend, even though *each* person's probability of success remained *constant* (at either 1/6 or 1/2) from one cycle to the next.

Cycle:	1	2	3	4	5	6	7	8	9	10	11	12
% Success:	33	29	25	23	20	19	18	18	17	17	17	17

The reason for this trend is that those with a cycle-specific success probability of 50% tend to succeed sooner, so that by the later cycles, those remaining are predominately those whose cycle-specific success probability is 1/6th, or 17%. Had say only one-quarter been given the coin, and three quarters the die, the average success in the first cycle would have been 25%, and the decrease to the asymptote of 17% would have been more rapid; conversely, had say three-quarters been given the coin, and one quarter the die, the average success in the first cycle would have been 42%, and the decrease to the 17% would have been more gradual. The nature (center and spread) of the distribution (the 'mix' of probabilities) determines the observed sequence of success rates.

In reality, with small sample sizes, the observed sequence of success rates will not necessarily be monotonic. Moreover, in any clinical situation, a cohort will not consist of just two subtypes; rather there will be a continuum of probabilities, and some distribution of these probabilities. Figure 1 shows the infinite-sample and a finite-sample realization for a number of 'mixtures.'

Statisticians refer to statistical models for this heterogeneity by several names: a random effects model, a latent class model, a hierarchical model. If one has sufficient data, one can more precisely estimate the parameters of this distribution. Clearly, if in our simulation, we could ask each person to perform a large number of trials, and not stop at the first success, it would

be easier to estimate the distribution of success probabilities. Although the parameters that describe the success rates in a one-arm study are seldom of interest, it is instructive for the comparative study to be considered in the next section to how one statistically estimate them from a finite amount of persons, where observation terminates at the first success or after persons have tried for C cycles, whichever comes first.

Consider first the simplest case, where as before there are *just 2* probabilities of success, P_1 and P_2. Unlike the didactic example where we set them to 1/2 and 1/6 respectively, in practice P_1 and P_2 would be unknown, as would the relative frequencies F_1 and $F_2(= 1 - F_1)$ of the two classes of persons. Since there are 3 unknown parameters (P_1, P_2, F_1), we could use the observed success rates $\{S_1, S_2, S_3\}$ in the first $C = 3$ cycles in a *very large cohort* to estimate them, with very little sampling error, by solving the 3 estimating equations:

$$S_1 = \frac{F_1 P_1 + F_2 P_2}{1} \; ; \; S_2 = \frac{F_1 \bar{P}_1 P_1 + F_2 \bar{P}_2 P_2}{1 - S_1} \; ; \; S_3 = \frac{F_1 \bar{P}_1 \bar{P}_1 P_1 + F_2 \bar{P}_2 \bar{P}_2 P_2}{1 - (S_1 + S_2)},$$

where $\bar{P}_1$ and $\bar{P}_2$ are shorthand for $(1 - P_1)$ and $(1 - P_2)$, respectively. With a *smaller cohort*, the observed success rates $\{s_1, s_2, s_3\}$ in the first three cycles would be imperfect estimates of $\{S_1, S_2, S_3\}$. Thus the estimates of P_1, P_2 and F_1 derived from $\{s_1, s_2, s_3\}$[1], while seeming to yield a perfect fit, would contain some sampling error. The amount of estimation error could be reduced by using data from $C > 3$ cycles, just as the errors in

[1]Statisticians refer to this fitting technique as the *method of moments*.

the estimated parameters of a line could be made smaller by using more datapoints. Moreover, the discrepancies between the observed success rates $\{s_1, s_2, \ldots, s_C\}$ and the fitted ones $\{\hat{s_1}, \hat{s_2}, \ldots, \hat{s_C}\}$ could be used to gauge the statistical uncertainty of the estimates $\hat{P}_1, \hat{P}_2$ and $\hat{F}_1$.

Again, in reality, there is a *continuum* of success probabilities, and these would have a frequency distribution, with a shape described by a (frequency, or density) function, $f(P)$, say. If a very large number of persons each performed a large number of trials and did not stop at the first success, the resulting individual success rates would – as with batting averages of professional baseball players over one or more years – yield a precise estimate of the location and shape of the $f(P)$ distribution. However, with data on say a maximum of $C = 12$ trials on a finite number of persons, one is forced to assume a functional form of the distribution, one that is governed by just a few parameters. These can then be estimated from a series of observed success proportions $\{s_1, s_2, \ldots, s_C\}$. One of the most common forms is the Beta distribution: it is governed by two parameters, α and β, the mean is $\mu = \alpha/(\alpha + \beta)$, the standard deviation is $\sigma = \{\mu(1-\mu)/(\alpha + \beta + 1)\}^{1/2}$, and the upper tail is longer than the lower one when $\mu < 0.5$. Another 2-parameter form, that can more easily accommodate covariates, is the *logit-Normal* distribution, where $logit(P) = \log(P/[1-P])$ has a Normal (Gaussian) distribution with mean μ_{logit} and standard deviation σ_{logit}.

These two hierarchical models – the *Beta-geometric*[2] and the *logit-*

[2]If there is no variation in P, the beta-geometric distribution becomes the familiar

Normal – are easily fit by widely available software for hierarchical models. in Appendix 1 we apply them to the data from a 'one-arm' observational study (all non-smokers, with no known fertility problems).

geometric distribution, used to describe the probabilities of achieving say a first success in a series of trials.

2. HOMOGENEOUS FECUNDABILITY

Let p_C denote a woman's fecundability i.e., her per-cycle probability of getting pregnant, with the less effective (control) treatment ($t = 0$). For now, assume no variation in p_C across women, i.e., that $\text{Var}[p_C] = 0$. Let p_E denote her fecundability with the experimental treatment ($t = 1$). Its efficacy with respect to the control treatment can be expressed in different ways using different forms for g in the generalized regression equation $g[p_E] = g[p_C] + \beta \times t$. For example, β is the absolute difference in fecundability if g is the identity function; $\exp[\beta]$ is the fecundability ratio θ if g is the ln function, or the fecundability odds ratio if g is the logit function. We will use the fecundability ratio to measure efficacy.

Suppose that one such woman, alternating from the experimental treatment in cycle 1, became pregnant on this treatment in the 5th cycle.

Cycle:	1	2	3	4	5
Treatment:	Exp'tl	Control	Exp'tl	Control	Exp'tl
Outcome:	—	—	—	—	+
Probability(+):	p_E	p_C	p_E	p_C	p_E
Probability(Outcome):	$1 - p_E$	$1 - p_C$	$1 - p_E$	$1 - p_C$	p_E

The observed data can be modeled as a sequence of independent Bernoulli trials with alternating probabilities of success. The likelihood based on this woman's data is the product of the probabilities of the 5 individual outcomes; it can also be re-arranged and written as a product of two geometric

(but binomial-like) likelihoods, corresponding to $s_C = 0$ successful cycles, preceded by $u_C = 2$ unsuccessful ones, when the success probability was p_C ; and $s_E = 1$ successful cycle, preceded by $u_E = 2$ unsuccessful ones, when the success probability was p_E , i.e.,

$$L(u_C, u_E, s_C, s_E \mid p_C, p_E) \propto (1 - p_C)^2 \times (1 - p_E)^2 \times p_E$$

Since p_C and p_E are constant from woman to woman, so that all woman-cycles within the same treatment condition are exchangeable, the likelihood based on the data from several such women can again be written as the product of two binomial-like likelihoods

$$L(U_C, U_E, S_C, S_E \mid p_C, p_E) \propto (1 - p_C)^{U_C} \times p_C^{S_C} \times (1 - p_E)^{U_E} \times p_E^{S_E}$$

where $U_C = \Sigma u_C$ and U_E and S_C and S_E are the total numbers of unsuccessful and successful cycles when using C and E respectively, i.e., summed over all women and all cycles. The ML point estimator of the fecundability ratio is simply $(S_E/T_E)/(S_C/T_C)/$ where $T_E = S_E + U_S$ and $T_C = S_C + U_C$. A likelihood-based interval estimate is also easily calculated.

3. HETEROGENEOUS FECUNDABILITY

In reality, fecundability *does* vary among women, i.e., $\mathrm{Var}[p_C] > 0$ and $\mathrm{Var}[p_E] > 0$. We denote this variation by the general bivariate pdf $f(p_C, p_E)$, with marginal distribution $f(p_C)$. We present two data-analysis approaches which use aggregated data for each treatment in each cycle. The first makes

no assumptions about the form of the marginal distribution $f(p_C)$, but strong ones about how a particular woman's fecundability with the experimental treatment is related to her fecundability with the standard one. In this approach, the observed data can be modeled either as (i) two sets of multinomial distributions, one for the numbers $\{S_{C_1}, S_{E_2}, \cdots\}$ who become pregnant in cycles $C_1, E_2, etc.$, the other for the numbers $\{S_{E_1}, S_{C_2}, \cdots\}$ who become pregnant in cycles $E_1, C_2, etc.$, or (ii) as cycle- and treatment-specific binomial random variables $\{ S_{C_1}|T_{c_1}\}$, $\{S_{E_1}|T_{E_1}\}$, $\{S_{C_2}|T_{C_2}\}$, ... The second approach is based on a specific parametric form—Beta—for f, but does not 'connect' a particular woman's fecundability when undergoing experimental treatment with that same woman's fecundability with the standard treatment.

We evaluate these approaches using the data in Table 1 as well as data generated from continuous bivariate distribution for $\{p_C, p_E\}$. In the latter case, how a particular woman's fecundability, p_E, with the experimental treatment is related to her fecundability, p_C, with the standard one induced variability in the efficacy across women. In this second method, the observed data are modeled as cycle- and treatment-specific binomial random variables.

3.1 Unspecified-form for f; constant fecundability ratio

Consider an unspecified distribution $f(p_C)$ and let θ denote the constant ratio of p_E to p_C for all values of p_C. For a person with a specific value p_C, assigned to the sequence C, E, C, ... , the probabilities of becoming pregnant

in cycle $1, 2, 3, ...$ are

$$p_C, (1 - p_C) \times \theta \times p_C, (1 - p_C) \times (1 - \theta \times p_C) \times p_C, ...$$

The probabilities, if that same person were assigned to the sequence E, C, E,. . . , are

$$\theta \times p_C, (1 - \theta \times p_C) \times p_C, (1 - \theta \times p_C) \times (1 - p_C) \times \theta \times p_C, ...$$

Since p_C varies over persons, the multinomial proportions are the expectations of these probabilities, taken over the distribution, $f(p_C)$, of p_C. They form the numerators of the expressions given in Table 2. and represent the contribution to the (multinomial-based) likelihood of each person who becomes pregnant in that cycle. The extension beyond cycle 3 (not shown) is obvious, even if the algebra is tedious. Of note is the fact that the two likelihood contributions from cycle k involve the first k moments of the distribution of p_C. Others, e.g., [7,8,9], have noted this in the simpler constant-sequence design. Thus, each cycle adds two new data points and one new parameter; overall the $2K$ datapoints from K cycles are modeled by $K + 1$ parameters. If $K \geq 2$, the remaining $K - 1$ degrees of freedom can be used to assess model fit.

Table 2 about here

Maximum likelihood estimates for these $K + 1$ parameters can be obtained from a non-linear modeling package, such as SAS PROC NLMIXED

(see Appendix). Although our approach deals with heterogeneity, it does so without specifying a traditional random effects model: we used only the 'NL" portion of NLMIXED. We found that this *multinomial* approach is very sensitive to starting values, and have had more success by modeling the number who become pregnant on a specific treatment in a specific cycle—conditional on the number who used that specific treatment in that cycle—as a *binomial* random variable. These conditional probabilities are given as the quotients in Table 2. Again, they can be fitted using SAS PROC NLMIXED by (i) expressing the $2K$ binomial parameters in terms of the K moments and the parameter of interest θ, and (ii) for each of the $2K$ observed counts, modeling

$$Number_{pregnant} \sim Binomial(Number_{treated}, BinomialParameter).$$

For the data in Table 1, Maximum Likelihood estimates (and Standard Errors of their natural logarithms) for these parameters are shown in Table 3. This method correctly 'recovers' θ. Further, because the procedure uses data from all cycles, it produces smaller standard errors than those for the summary estimates from the odd-numbered cycles only. Moreover, one can achieve this increased precision, and only a slight inaccuracy, with fewer than the full K moments: one can omit i.e., set to zero in the likelihood, some of the higher order moments—those of order 3 or more in our example. This is because p_C is bounded by 0 and 1, so that the higher moments are of decreasing magnitudes, and thus increasingly negligible.

– Table 3 about here –

3.2 Beta-Geometric Model

That an experimental treatment would increase each p_C value a constant-fold, i.e., by the same multiple, θ for each woman, regardless of her value of p_C, is not realistic biologically. Whereas women whose natural fecundability is 10% might reasonably have it increased to 20% i.e., by a factor of $\theta = 2$, with experimental treatment, the treatment is unlikely to also raise other women's already high natural per-cycle success probability of 40%, say, by the same (multiplicative) factor of $\theta = 2$, i.e., to 80%. Moreover, if $p_C > 1/\theta$, this assumption of a constant θ is statistically impossible. In addition, there are practical technical difficulties in fitting such a high-order nonlinear model; the number of parameters (moments) relative to the numbers of observations is large. For these reasons, we turn to more natural parametric statistical models for p_C and p_E, ones with fewer constraints on how a particular woman's fecundability when undergoing experimental treatment relates to what might be (loosely) called its 'counterfactual' i.e., the same woman's fecundability with the standard treatment.

The Beta-Geometric (B-G) model has been used in demography [7]. More recently, Weinberg and Gladen [8] used it in a non-experimental study of the effect of smoking on fecundability. Since women were classified as smokers or non-smokers for the entire period of observation (up to 12 cycles), their model immediately applies to a parallel-sequence design [4]. The latter authors used a modified B-G model to *generate* data, but did not consider

it for the *analysis* of their data. We extend these ideas to develop a beta-geometric model for data from the alternating-sequence design.

Before doing so, we review its use for the constant-sequence design. Weinberg and Gladen compared fecundability, measured over 12 cycles, in smokers relative to non-smokers. They modeled fecundability in the two source populations as Beta distributions, with their respective location and shape governed by the pairs of parameters $\{\alpha_C,\beta_C\}$ and $\{\alpha_E,\beta_E\}$. Therefore, before the first cycle the probability density function of p_C is given by:

$$f[p_c] \propto {p_c}^{\alpha_C-1}(1-p_c)^{\beta_C-1}$$

The mean and variance of the fecundability in the control group before the first cycle are $\mu_C = \alpha_C/(\alpha_C+\beta_C)$ and $\sigma_C^2 = \alpha_C\beta_C/((\alpha_C+\beta_C)^2(\alpha_C+\beta_C+1))$ Weinberg and Gladen showed that in those couples who were unsuccessful in U previous cycles, the—*now conditional*—distribution of p_C at cycle $U+1$ in this selected subgroup is shifted towards the left i.e., towards zero, but remains a Beta distribution, with probability density function:

$$f(p_c|U) \propto {p_c}^{\alpha-1}(1-p_c)^{\beta-1+U}.$$

The parameters of the fecundability distribution are now $\{\alpha_C,\ \beta_C+U\}$. Thus, after k cycles the mean fecundability is given by $\mu_C = \alpha_C/(\alpha_C+\beta_C+k)$. They further showed that the expected probability of success among those who enter cycle U+1 is related to the number of previously unsuccessful

cycles U via the simple reciprocal link:

$$1/E[p_C|U] = (\alpha_C + \beta_C + U)/\alpha = (\alpha_C + \beta_C)/\alpha + (1/\alpha) \times U$$

$$= \gamma + \delta \times U$$

The parameter γ is the expected number of cycles to become pregnant if a couple had the average per-cycle probability under the control treatment i.e., $\gamma = 1/\mu_C$. The parameter δ reflects the spread of the initial distribution of p_C: under homogeneity, the number of cycles required to achieve pregnancy reduces to the same geometric random variable for each couple, i.e., $\delta = 0$. The parameters γ and δ can be fit using binomial regression with an inverse (i.e., power-1) link, e.g., using PROC GENMOD in the SAS, or glm in Stata. Since glm in R does not allow this link for the binomial, one needs to supply the variance function.

Weinberg and Gladen extended the model to effectively fit separate Beta distributions for the cycle-specific probabilities for smokers ($t = 1$) and non-smokers($t = 0$), via one equation:

$$1/E[p_C \mid U] = (\gamma_C + \delta_C \times U) \times (1 - t) + (\gamma_E + \delta_E \times U) \times t.$$

To extend it to the alternating sequence design, we model the expected probability of pregnancy for women who have already undergone U_E and U_C unsuccessful cycles on E and C, and are now about to receive (say) E. The initial Beta(α_E,β_E) distribution must be updated to reflect the U_E and U_C. The U_E is added to the β_E term as in the parallel sequence model, but a

correction must be made to the U_C. If C were no more effective than E, and thus exchangeable with it, we would add the full U_C to the U_E to obtain the β_E term of $U_E + U_C$. But if C were less effective than E, then using the full $U_E + U_C$ in the β_E term would shift the fecundability distribution too far to the left. This is most easily seen if *no* pregnancies can occur with C. The greater the efficacy of E relative to C, (i.e., the smaller the 'failure ratio' $(1-p_E)/(1-p_C)$), the smaller should be the 'amalgam' of U_E and U_C. Thus, we suggest that the contribution of U_C be reduced, and propose the amalgam

$$U^* = U_E + FRR \times U_C$$

where FRR is the ratio of the probability of failure on the experimental and control treatments at cycle 1. This ratio is less than unity if E is more effective than C. Thus, the probability density function of p_E at cycle $u_C + u_E + 1$ is taken to be

$$f(p_E \mid u_E\ u_C) \propto {p_E}^{\alpha_E - 1}(1 - p_E)^{\beta_E - 1 + u_E + FRR \times u_C},$$

i.e., a Beta(α_E , $\beta_E + u_E + FRR \times u_C$) distribution. Thus the expected probability of success at this cycle is inversely proportional to a linear function of u_E and u_C, i.e .

$$E(p_E \mid u_C\ \ u_E) = \alpha_E/(\alpha_E + \beta_E + u_E + FRR \times u_C)$$

Similarly, we can show that

$$E(p_C \mid u_C\ \ u_E) = \alpha_C/(\alpha_C + \beta_C + u_C + (1/FRR) \times u_C)$$

We re-formulate it as a generalized linear model for binary data with an inverse link function:

$$\begin{array}{rcccc} 1/E(p \mid u_C \;\; u_E) = & & \gamma_{0E} \times t & + & \gamma_{0C} \times (1-t) \\ & + & \gamma_{1E} \times (u_E \times t) & + & \gamma_{1C} \times (u_C \times (1-t)) \\ & + & \gamma_{2E} \times (u_C \times t) & + & \gamma_{2C} \times (u_E \times (1-t)) \end{array} \quad (3)$$

where $t = 1$ under the experimental treatment and $t = 0$ under the control treatment. Constraints should be placed on γ_{2E}, which is a function of γ_{0E} and γ_{1E}, and on γ_{2C}. The model can be fit using PROC NLMIXED in SAS, which can accommodate constraints. The parameters of the original Beta distributions can be obtained by the following transformations.

$$\alpha_E = 1/\gamma_{1E};\ \beta_E = (\gamma_{0E} - 1)/\gamma_{1E};\ \alpha_C = 1/\gamma_{1C};\ \beta_C = (\gamma_{0C} - 1)/\gamma_{1C},$$

and efficacy is estimated by $\hat{\theta} = \hat{\gamma_{0C}}/\hat{\gamma_{0E}}$.

In a Bernoulli model, woman-level covariates could be added as linear terms in (3). Notice that at each cycle the fecundability distribution under treatment t is obtained by adding to the β parameter of its starting fecundability distribution, the number of unsuccessful cycles on t and a multiple of the number of those on the other treatment. While we have used the multiplying factors as FRR and $1/FRR$ above, we could use the generic expressions $\alpha_C/(\alpha_C + \beta_C + u_C + \phi_C \times u_C)$ and $\alpha_E/(\alpha_E + \beta_E + u_E + \phi_E \times u_C)$ for the respective mean fecundability after cycle $u_C + u_E$. ϕ_C and ϕ_E should be constrained to be > 1 and < 1 respectively, assuming E is more effective than C.

4. EVALUATION

We assessed the performance of these analysis models on 200 generated datasets. Following Cohlen et al. [4] we began with Beta distributions with means of 0.16 and 0.32, each with a coefficient of variation of 75% (Cohlen et al. used 56%). We calculated the 9th, 18th, ..., 90th percentiles for each of these two initial distributions, and placed a point mass of 0.1 at each of these values, thereby creating two 10-point distributions for the first cycle. If a woman's fecundability was at say the 18th percentile when on C i.e. if $p_C = 5.2\%$, then her fecundability with E (if necessary) was the corresponding 18th percentile in that distribution, namely $p_E = 8.1\%$ (fecundability ratio = 1.56). Similarly, those just above the middle of the p_C distribution, i.e., a fecundability of $p_c = 15.2\%$, were considered to be just above the middle of the p_E distribution, namely $p_E = 31.6\%$ (fecundability ratio = 2.08), while in the 3rd highest subgroup, fecundabilities were $p_C = 22.6\%$ and $p_E = 48.2\%$ (ratio = 2.13). For each dataset, 1000 women were randomly assigned, using a multimomial distribution, to the 10 fecundability levels under C, and 1000 others to the 10 corresponding levels in the p_E distribution. At each cycle, these two sets of 10 subgroup frequencies were depleted using random numbers of pregnancies generated by the 20 corresponding binomial distributions. The numbers who were unsuccessful were switched to their corresponding level on the opposite distribution, before generating the pregnancies for the next cycle.

The estimates produced by our analysis models are summarized in Table

5. Leaving the form of f unspecified and estimating its first few moments is somewhat more efficient than using the Beta-Binomial model, where there is more of a tradeoff between bias and precision.

5. EFFICACY ESTIMATES FROM A CLINICAL TRIAL

The data in Table 6a are from (by today's standards) a very large clinical trial. It evaluated the performance of a second generation protocol for donor insemination with frozen semen [10]. Today, screening of semen for HIV infection precludes the use of fresh semen. Earlier, in the first use that we have found of the alternating sequence design, this group achieved a fecundability rate of only 5.0 pregnancies per 100 cycles with a first-generation protocol, versus 18.9 per 100 using fresh semen [11].

The data for the first six cycles (during which, for each woman, semen was from her matched donor) are shown in Table 6a. However, as is obvious from the denominators, the same practical difficulties were encountered as those mentioned in the first study: "Cryopreserved semen was frequently substituted in a cycle scheduled to be fresh because the donor was not available." Unfortunately, information on the actual sequence for each woman is no longer available (S. Shapiro, personal communication, 2002).

Estimates of efficacy are given in Table 6b. Those produced by the methods in 3.1 and 3.2 are closest to the null, suggesting that they removed some of the bias induced when one ignores the heterogeneity. Possibly by chance, given its poor precision, the estimator advocated by Norman and Daya was

furthest from the null, further than both the crude and the Mantel-Haenszel estimates. The Generalized Linear Model estimate was closest to the null, but had a high SE, possibly because of the large number of parameters (6) fitted to the 12 datapoints. The method of moments had the lowest SE. The fact that it was no different from that of the Mantel Haenszel estimator suggests that, in this study, heterogeneity does not substantially inflate or deflate the SE. Without individual-specific data, we are unable to assess how much heterogeneity could also be affected by women who dropped out.

6. DISCUSSION

The more attractive alternative sequence design also produces more pregnancies for the same number of cycles. We describe two approaches that allow clinical trials to use data from all of the cycles, while Norman and Daya's approach [6] squanders the statistical advantage of this design. Many contemporary trials generate fewer than 25 pregnancies in total. Decreasing precision further by omitting even-numbered cycles, without trying to eliminate the biases by other methods, is difficult to defend. Moreover, with the sample sizes considered here reduced by a realistic factor of 25, sampling variability dominates analytic bias.

The bias caused by naively using all data is a function of the heterogeneity in p_C and the efficacy of the experimental treatment ; *both must be substantial in order to produce a serious bias.* Norman and Daya based their concerns on an extreme 2-point distribution of p_C, and a treatment that

doubled the $p_C = 40\%$ in the high fertility subgroup to a 'biologically nearly impossible' $p_E = 80\%$. Even then, the bias was less than the sampling variability induced by the sample sizes used in practice. With the same $\theta = 2$, and a more realistic distribution where $f[0.025] = 0.8$ and $f[0.1] = 0.2$ so that mean$[p_C] = 0.04, SD[p_C] = 0.03$, the bias was much smaller Moreover, if, rather than increasing p, an new treatment— such as frozen semen— *reduces* it, the degree of bias in the naive estimate of θ is also less, because of the smaller impact of the differential success (removal) of the most fertile in odd-numbered cycles. For example, using $\theta = 0.5$ in Table 2, the data from cycle 2 yield $\hat{\theta} = 0.47$, a relative bias of only 6%. These 'low-bias' conditions would also apply in comparisons of contraceptive methods, where the probability of an unwanted pregnancy is already low.

The alternating sequence design is not a full *crossover* study. Nor does it carry the full statistical efficiency usually associated with self-matched comparisons. Thus, the sample size and power calculations/projections are best carried out by analogy with unmatched designs.

In practice, for example in the study by Brown et al. [10], there are unavoidable individual deviations from the alternating-sequence protocol. Some investigators may use variations of the alternating design: e.g., in Ecochard et al. [12], 1/2 patients received one treatment for the first two cycles and the competing treatment for the next two cycles, while the other 1/2 followed the opposite sequence. For these more complex designs, random effects models

for binary (Bernoulli) data allow one to model the cycle-by-cycle sequence of outcomes for each woman using a full regression approach that makes use of all of the individual level data for each woman (any baseline covariates, the cycle-by-cycle treatment indicators and any other available cycle-dependent covariates). Some quite complex 2-level hierarchical models have been used in such circumstances [13,14].

The approaches we have described are also applicable to simple data analyses for the parallel- or 'constant-sequence' randomized trial design. Since this competing design has the same data structure as Weinberg and Gladen's example (fecundability of smokers and non-smokers), their 'Beta-Geometric' Model is immediately applicable without modification. The method based on moments is also applicable. Applied to Norman and Daya's 'constant-sequence' example (and the 'data' in their Table 1), both of our methods produce estimates closer to the true $\theta = 2$, whereas a naive analysis produces an attenuated estimate of 1.83.

In an appendix, Norman and Daya [6, p324] claim that "the assumptions of a constant drug efficacy is not necessary" by considering an arbitrary distribution $f[p_C]$, and an arbitrary efficacy function, $\theta[p_C]$ They purport to show algebraically that "the outcome rates in the odd cycles in an alternating sequence are unbiased," i.e., that "the results will hold true regardless of the relationship between efficacy and fertility." In fact, the ratio from alternate cycles will not continue to be unbiased if the treatment effect is variable. This

is illustrated in Table 7 by slightly perturbing Norman and Daya's simulated example so that the risk ratio in the low fecundability group is 2.5, while the risk ratio in the high fecundability group remains at 2. This is closer to what happens in reality where there is likely to be a greater relative shift in the low fecundability groups, compared to the high fecundability groups. The true average risk ratio across the population is thus $2.5 \times 0.8 + 2 \times 0.2 = 2.4$. However, from Table 7 we can see that this estimate is not obtained even in the first cycle: the ratio of the expected probabilities of successes does not match the expectation of the ratios of the success probabilities, i.e.

$$\frac{\Sigma\ \theta[p_C] \times p_C \times f[p_C]}{\Sigma\ p_C \times f[p_C]} \neq \Sigma\ \theta[p_C] \times f[p_C] = \theta[p_C]$$

Further, it appears that the odd cycles underestimate the true risk ratio while the even cycles overestimate it. If the study continues to a point when only women in the low fecundability group remain, then the ratio approach the true ratio of 2.5 in both odd and even cycles

This contrary finding is an additional impetus to consider more general regression models that allow not just between-individual heterogeneity, and covariates at the woman-cycle level, but also more flexibility in the specification of the comparative parameter. We plan to investigate whether the amount of data from a typical alternating sequence design makes such models practical. Unlike traditional studies with multiple crossovers, the alternating sequence design involves at most one instance of Y=1 per subject, and such outcomes preclude further observations. This, the small sample sizes, and

the small number of cycles usually used, may be a serious impediment to more complex modeling. The analyses we have presented, based on marginal distributions, may well be the appropriate ones for the amounts of clinical trial data generated by the alternating sequence design.

ACKNOWLEDGMENTS

Nandini Dendukuri is a chercheur-boursier of the Fonds de Recherche en Santé du Québec. Robert Platt is a career scientist of the Canadian Institutes of Health Research.This work was supported by individual operating grants from the Natural Sciences and Engineering Research Council of Canada and a team grant from Le Fonds Québécois de la recherche sur la nature et les technologies.

Code

Appendix 1: Why heterogeneity creates biased efficacy estimates in the alternating-sequence design

The origin and nature of the biases in the estimates from these two designs – as well as the concept of inter-patient heterogeneity – can be readily understood by studying the worked example in Norman and Daya [6]. As shown in the first row of Table 1, for didactic purposes, they assumed a heterogeneous population, where fecundability i.e., the per-cycle probability of getting pregnant, varied from couple to couple. For didactic purposes, they assumed the simplest possible model of heterogeneity – with just two classes. In this simplistic model, with the less effective (Control) treatment, some 80% of couples had a per-cycle probability $p_C = 10\%$, and the remaining 20% of couples a $p_C = 40\%$ fecundability, i.e.,

$$p_C = \begin{cases} 0.1 & \text{for 80\% of couples,} \\ 0.4 & \text{for 20\% of couples.} \end{cases}$$

Thus they assumed that the overall average fecundability is 16%, and the standard deviation is 12% (in reality fecundability would vary along a continuum, and there might also be an sub-populationt with zero fecundability).

They further assumed that the more effective (Experimental) treatment had a constant efficacy, θ, of 2, i.e., that at each cycle, a couple's probability of becoming pregnant was doubled. In Table 1 the course of the 1000 randomly allocated to undergo the 'control' treatment in the first cycle is tracked in bold. The entries at each cycle are the expected numbers

of couples from the higher- and lower fecundity subpopulations who attempt to (and, in parentheses, the numbers who do) become pregnant. Using expected numbers of pregnancies at each cycle, Norman and Daya showed that estimates of efficacy that are based only on the total number of women and the total number of pregnancies are biased, irrespective of the design—the parallel design *under*estimates (apparent efficacy: θ =1.83, calculations not shown here but discussed later) and the alternating-sequence design *over*estimates (apparent efficacy $\theta = 2.10$, middle column Table 1). However, they noted that the bias in the alternating-sequence design is limited to the data from even-numbered cycles.

Appendix Table 1: Cycle-specific ratios of expected pregnancy proportions if the alternating sequence design is applied to a population of 2000 which is heterogeneous with respect to spontaneous fecundity [20% with higher, 80% with lower fecundity].

	Treatment received in the indicated cycle						
	Control				Experimental		
	Sub-population (fecundability)				Sub-population (fecundability)		
Cycle	High	Low	All	Ratio	All	High	Low
1	**200**	**800**	1000		1000	200	800
	(80)	**(80)**	**(160)**		(320)	(160)	(160)
			16%	2.00	32%		
2	40	640	680		**840**	**120**	**720**
	(16)	(64)	(80)		**(240)**	**(96)**	**144**
			11.8%	2.43	**28.6%**		
3	**24**	**576**	600		600	24	576
	(9.6)	**(57.6)**	**(67.2)**		(134.4)	(19.2)	(115.2)
			11.2%	2.00	22.4%		
4	4.8	460.8	465.6		**532.8**	**14.4**	**518.4**
	(1.9)	(46.1)	(48.0)		**(114.9)**	**(11.5)**	**103.7**
			11.3%	2.10	**21.6%**		
5	**2.9**	**414.7**	417.6		417.6	2.9	414.7
	(1.2)	**(41.5)**	**(42.6)**		(85.2)	(2.3)	(82.9)
			10.2%	2.00	20.4%		
1-5	271.7	2891.5	3163.2		3390.4	361.3	3029.1
			(397.8)		(894.8)		
			(12.6%)	2.10*	(26.5%*)		

The course of the 1000 randomly allocated to undergo the 'control' treatment in the first cycle is tracked in bold. The entries at each cycle are the expected numbers of couples from the higher- and lower fecundity subpopulations who attempt to (and, in parentheses, the numbers who do) become pregnant. Table adapted from Figure 2 and Table 2 of Norman and Daya.

REFERENCES

McDonnell J Angelique J.Goverde AJ and Vermeiden J.P.W. The place of the crossover design in infertility trials: a maximum likelihood approach Human Reproduction Vol.19, No.11 pp. 25372544, 2004 doi:10.1093/humrep/deh475 Advance Access publication September 30, 2004

1. DAYA, S. Is there a place for the crossover design in infertility trials? *Fertility and Sterility* 1993; **59**: 67.
2. KHAN, K.S, DAYA S, COLLINS J.A., AND WALTER S.D. (1996). Empirical evidence of bias in infertility research: overestimation of treatment effect in crossover trials using pregnancy as the outcome measure. *Fertility and Sterility* **65**:939945.
3. TE VELDE, E.R., COHLEN B.J., LOOMAN C.W., AND HABBEMA, J,D. Crossover designs versus parallel studies in infertility research. [letter] *Fertility and Sterility* 1998; **69**:357-358.
4. COHLEN, B.J., TE VELDE, E.R., LOOMAN. C.W.N., EIJCHEMANS, R., AND HABBEMA, J.D.F. Crossover or parallel design in infertility trials? The discussion continues. *Fertility and Sterility* 1998; **70**:40-45.
5. DAYA, S. Differences Between Crossover and Parallel Study DesignsDebate? (letter) *Fertility and Sterility* 1999; **71**:771-772.
6. NORMAN, G.R. AND DAYA, S. The alternating-sequence design (or multiple-period crossover) trial for evaluating treatment efficacy in infertility. *Fertility and Sterility* 2000; **74**:319-324.
7. McDonnell J Angelique J.Goverde AJ and Vermeiden J.P.W. The place of the crossover design in infertility trials: a maximum likelihood approach Human Reproduction Vol.19, No.11 pp. 25372544, 2004 doi:10.1093/humrep/deh475 Advance Access publication September 30, 2004
7. SHEPS, M.C., AND MENKEN, J.A. Mathematical models of conception. University of Chicago Press, 1973.
8. WEINBERG, C.R. AND GLADEN B.C. The beta-geometric distribution applied to comparative fecundability studies. *Biometrics* 1986; **42**:547-560.
9. LAU, T.S. On the heterogeneity of fecundability. *Lifetime Data Analysis* 1996; **2**:403-415.
10. BROWN, C.A., BOONE, W.R., AND SHAPIRO, S.S. Improved cryopreserved semen fecundability in an alternating fresh-frozen insemination program. *Fertility and Sterility* 1988; **50**:825-827.

11. RICHTER, M.A., HANING, R.V., AND SHAPIRO, S.S. Artificial donor insemination: fresh versus frozen semen: the patient as her own control. *Fertility and Sterility* 1984; **41**:277-280.

12. ECOCHARD, R, MATHIEU, C., ROYERE, D., BLACHE, G., RABILLOUD, M., AND CZYBA, J.C. A randomized prospective study comparing pregnancy rates after clomiphene citrate and human menopausal gonadotropin before intrauterine insemination. *Fertility and Sterility* 2000; **73**:90-93

13. ECOCHARD R, CLAYTON DG. Multi-level modelling of conception in artificial insemination by donor. *Statistics in Medicine* 1998; **17**(10):1137-1156.

14. ECOCHARD R, CLAYTON DG. Multivariate parametric random effect regression models for fecundability studies. *Biometrics* 2000; **56**(4):1023-1029.

Figure 1: Random effects models fitted to pregnancy rates in (fertile) non-smokers (data from Gladen and Weinberg). *Lower portion:* n: number attempting to become pregnant; s: number successful; × = success rate (%); • fitted rate from Beta-geometric model; • fitted rate from logit-normal model. *Upper portion:* + cumulative success rate; • fitted rate from Beta-geometric model;

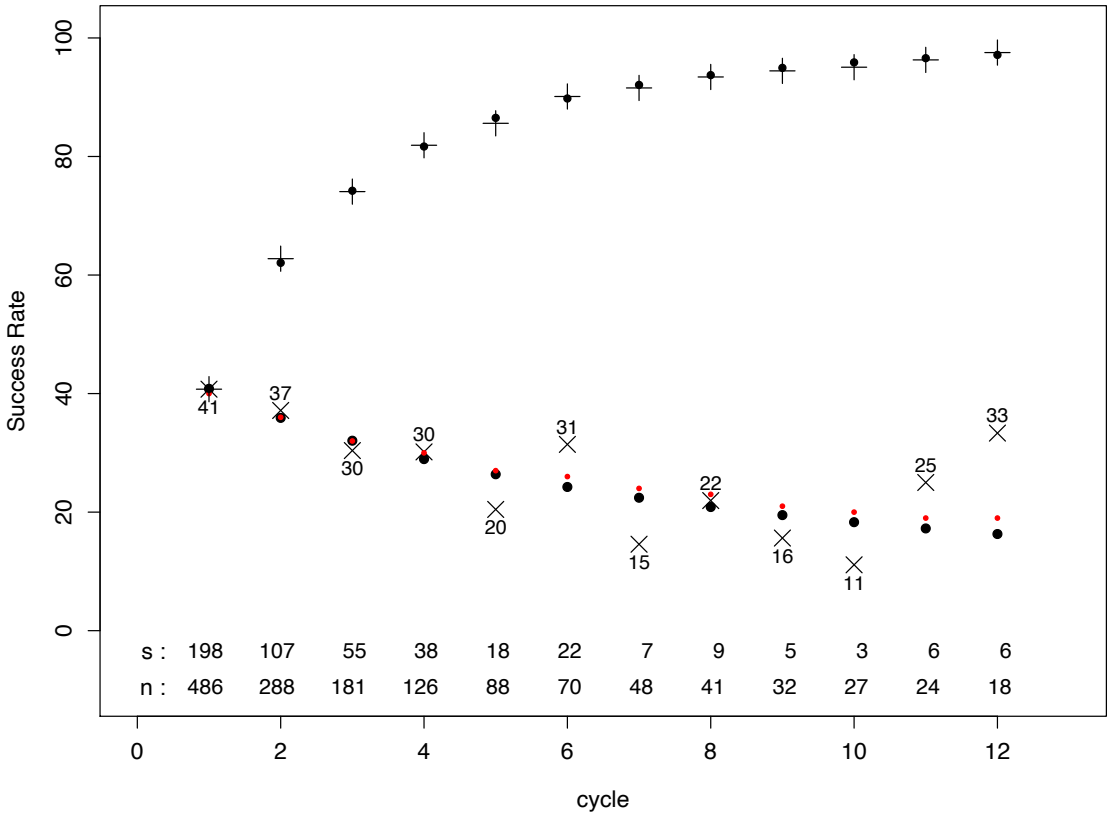


Figure 2: Distribution of individual success probabilities estimated from random effects models fitted to pregnancy rates in (fertile) non-smokers. Data, from Gladen and Weinberg, are given in Figure 1. Black curve: Beta-geometric model; Grey curve: logit-Normal model.

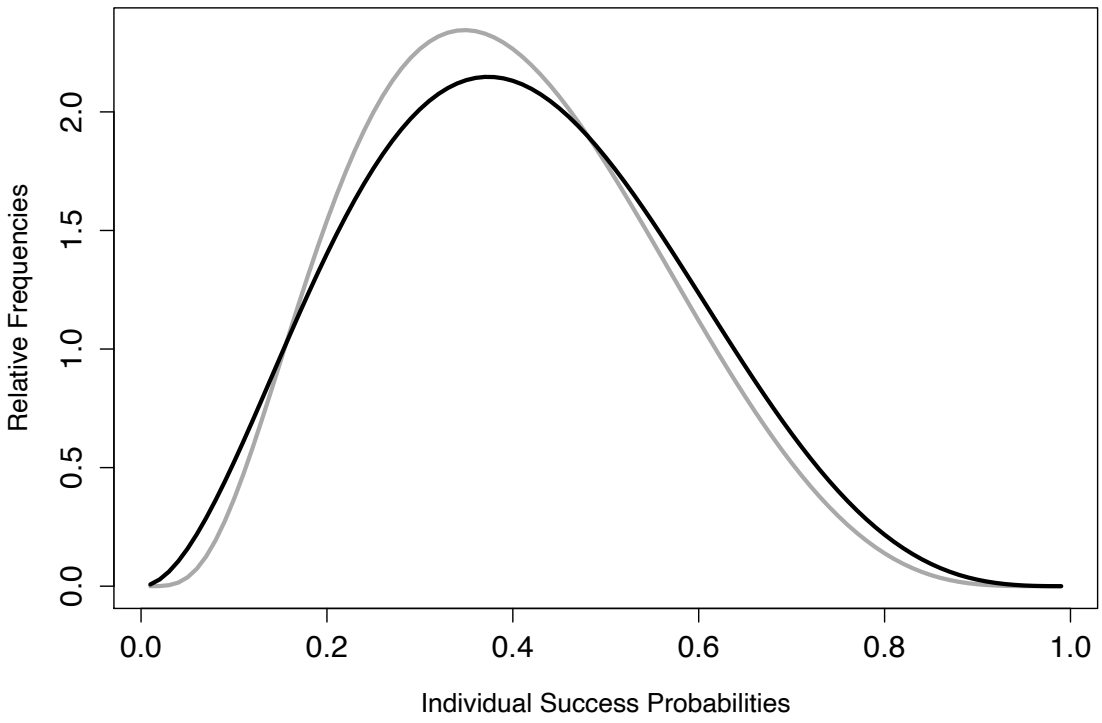


Table 2: Unconditional (multinomial) and conditional success probabilities for each of the first three cycles, as a function of the efficacy, θ, and the (absolute) moments of the unspecified distribution of p_C, the fecundability under the standard ["control" (C)] treatment.

Cycle	Control	Experimental
1	μ_1	$\theta \times \mu_1$
2	$\frac{\mu_1 - \theta \times \mu_2}{1 - \theta \times \mu_1}$	$\frac{\theta \times \mu_1 - \theta \times \mu_2}{1 - \mu_1}$
3	$\frac{\mu_1 - \mu_2 - \theta \times \mu_2 + \theta \times \mu_3}{1 - \mu_1 - \theta \times \mu_1 + \theta \times \mu_2}$	$\frac{\theta \times \mu_1 - \theta \times \mu_2 - \theta^2 \times \mu_2 + \theta^2}{1 - \mu_1 - \theta \times \mu_1 + \theta \times \mu_2}$

The numerators represent the unconditional probabilities of pregnancy in the indicated cycle for persons *entering* the study, while while the quotients represent conditional pregnancy probabilities for those who receive the indicated treatment in the *indicated* cycle. These probabilities are computed separately for those randomly allocated to the 'C to E to C' sequence, and conversely for their counterparts. Cycle 1 starts with denominators of 1 (100%) in each group; it is assumed that there are no dropouts [i.e. women/couples who have not yet become pregnant do not abandon the study] or that dropouts are 'at random' and unrelated to their values of p_C. The symbols μ_1 to μ_3 are the first 3 absolute moments of the distribution of p_C, the fecundability with standard treatment.

Table 3: MLEs of the efficacy parameter θ (SE of ln of estimate) as a function of number of data cycles used, and number of moments of unspecified distribution f estimated, compared with estimates obtained using approach of Norman and Daya.

	No. moments of f fitted					Norman and Daya	
Cycles used	1	2	3	4	5	Cycles used	$\hat{\theta}$ (SE*)
1	2.00					*1*	2.00
	(86)						(86)
1,2	2.33	2.00					
	(77)	(66)					
1,2,3	2.18	2.06	2.00			*1,3*	2.00
	(66)	(60)	(60)				(70)
1,2,3,4	2.28	2.01	2.02	2.00			
	(64)	(55)	(55)	(56)			
1,2,3,4,5	2.18	2.07	2.01	2.00	2.00	*1,3,5*	2.00
	(59)	(54)	(52)	(52)	(53)		(65)

'Data' from Table 1.

All SE's are multiplied by 1000.

* Mantel-Haenszel summary risk ratio, with SE of ln estimate back-calculated from test-based confidence interval.

Table 4: Fit of adapted beta-binomial generalized linear model to 5 cycles of Norman and Daya 'data.'

			Model	
Term	Parameter	Measure	$\phi = \frac{\beta_C/(\alpha_C+\beta_C)}{\beta_E/(\alpha_E+\beta_E)}$	No Constraint
t	γ_1		3.15(0.14)	3.14(0.14)
$1-t$	γ_0		6.31(0.45)	6.36(0.45)
		Ratio*	2.00	2.03
		ln Ratio	0.694(0.084**)	
$U_E \times t$	δ_{1E}		0.58(0.13)	0.36
$U_E \times (1-t)$	δ_{0E}		-	-0.00
$U_C \times t$	δ_{1C}		-	0.71
$U_C \times (1-t)$	δ_{0C}		0.93(0.31)	1.94

SE's shown in parentheses.

* The ratio estimate is 6.36/3.14.

** Since the covariance between the 6.36 and 3.14 is zero, the variance of the ln of the ratio, computed via the delta method, is $[(0.45/6.36)^2 + (0.14/3.14)^2]^{1/2} = [1/499.7 + 1/196.6]^{1/2} = 0.084$.

Table 5: Estimates obtained by applying non-parametric and parametric methods to 200 datasets.*

	No. of moments fitted (f unspecified)					Beta-Binomial ϕ :		
Measure	1	2	3	4	5	$\frac{\beta_C}{\beta_E}$	$\frac{\beta_C/(\alpha_C+\beta_C)}{\beta_E/(\alpha_E+\beta_E)}$	N.C.**
Median$\{\hat{\theta}\}$	2.16	2.03	1.99	1.98	1.99	1.91	2.08	2.02
SD$\{\ln\hat{\theta}\}$	0.055	0.050	0.049	0.049	0.052	0.059	0.077	0.081
Mean{SE**}	0.058	0.054	0.052	0.052	0.054	0.062	0.078	0.083

* For details, see Section 4.

** N.C.: No constraint

** SE of $\ln\hat{\theta}$

Table 6a: Pregnancies and Fecundability in Cycles of Insemination with Fresh and Frozen Semen.

	Fresh semen			Frozen semen		
	Number of ...			Number of ...		
	Patients	Pregnancies	Rate	Patients	Pregnancies	Rate
1	**163**	**57**	**0.350**	125	18	0.144
2	69	18	0.261	**130**	**12**	**0.092**
3	**73**	**20**	**0.274**	87	8	0.092
4	59	12	0.203	**69**	**9**	**0.130**
5	**51**	**12**	**0.235**	50	1	0.020
6	51	12	0.235	**28**	**2**	**0.071**
1-6	466	131	0.281	489	50	0.102

The course of the patients who underwent insemination with fresh semen in the first cycle is tracked in bold. Data from Table 1 of Brown et al [10].

Table 6b: Estimates of Efficacy of Insemination with Frozen *vs.* Fresh Semen. SE of *ln* of estimate given in parentheses.

Cycles	Method/Model	Details ...	Efficacy
1-6	Brown *et al.*	$50/489 \div 131/466$	0.36(0.15)
1-6	M-H*	Summary Risk Ratio	0.37(0.15)
1, 3, 5	Norman & Daya	$27/262 \div 89/287$	0.35(0.19)
1, 3, 5	Norman & Daya, M-H*	Summary Risk Ratio	0.35(0.19)
1-6	Unspecified f	4 moments fitted	0.39(0.15)
1-6	Beta-binomial	Unconstrained	0.39(0.25)[1]

* Mantel-Haenszel summary risk ratio (summed over cycles), with SE of *ln* estimate back-calculated from test based-confidence interval.

[1] See footnote to Table 4. Deviance / df = 0.98; Chi-square goodness of fit statistic = 5.6 (6 df).

Table 7: Simulated example with varying risk ratio in each fecundability subpopulation; otherwise, same setup as in Table 1.

	Treatment received in the indicated cycle						
	Control				Experimental		
	Sub-population (fecundability)				Sub-population (fecundability)		
Cycle	High	Low	All	Ratio	All	High	Low
1	**200**	**800**	1000		1000	200	800
	(80)	**(80)**	**(160)**		(360)	(160)	(160)
			16%	2.25	36%		
2	40	640	640		**840**	**120**	**720**
	(16)	(60)	(76)		**(276)**	**(96)**	**180**
			11.8%	2.79	**32.9%**		
3	**24**	**540**	564		564	24	540
	(9.6)	**(54)**	**(63.6)**		(154.2)	(19.2)	(135)
			11.2%	2.43	27.3%		
4	4.8	405	409.8		**500.4**	**14.4**	**486**
	(1.9)	(40.5)	(42.4)		**(133)**	**(11.5)**	**121.5**
			10.3%	2.57	**26.6%**		
5	**2.9**	**364.5**	367.4		367.4	2.9	364.5
	(1.16)	**(36.45)**	**(37.61)**		(93.429)	(2.32)	(91.125)
			10.2%	2.48	25.4%		
1-5	271.7	2709.5	2981.2		3271.8	361.3	2910.5
	(108.7)	(270.6)	(379.3)		(1016.6)	(289.0)	(727.625)
			(12.7%)	2.45*	(31.1%*)		

The course of the patients who received the standard (control) treatment in the first cycle is tracked in bold.