Statistical Methods in Medical Research
http://smm.sagepub.com/

A stepped wedge design for testing an effect of intranasal insulin on cognitive development of children with Phelan-McDermid syndrome: A comparison of different designs

Edwin R Van den Heuvel, Renée J Zwanenburg and Conny MA Van Ravenswaaij-Arts

Stat Methods Med Res published online 19 November 2014
DOI: 10.1177/0962280214558864

The online version of this article can be found at:
http://smm.sagepub.com/content/early/2014/11/17/0962280214558864

Published by:
SAGE
http://www.sagepublications.com

Review Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
0(0) 1–10

DOI: 10.1177/0962280214558864
smm.sagepub.com

SAGE

A stepped wedge design for testing an effect of intranasal insulin on cognitive development of children with Phelan-McDermid syndrome: A comparison of different designs

Edwin R Van den Heuvel,[1,2] Renée J Zwanenburg[3] and Conny MA Van Ravenswaaij-Arts[3]

Abstract

This paper compares the power of the parallel group design, the matched-pairs design, and several options for the stepped wedge and delayed start designs for testing a possible effect of intranasal insulin with respect to placebo on developmental growth of children with a rare disorder like Phelan-McDermid syndrome. A subject-specific linear mixed effects model for the primary outcome developmental age in a longitudinal setting with five time points was assumed. Monte Carlo simulation studies with small sample sizes were applied since the rare disorder prohibits large trials. The stepped wedge designs, which were initially preferred for ethical reasons, appear to be competitive in power to other designs and were in some settings even the best. The assumed statistical model also demonstrates that all of the designs can be viewed as a stepped wedge or delayed treatment design. Our results show that the stepped wedge design is an appropriate alternative for randomized controlled trials on developmental growth with small numbers of participants under the formulated statistical conditions.

Keywords

Developmental growth, Phelan-McDermid syndrome, random coefficient model, small trial designs, stepped wedge design, delayed start design

[1]Department of Mathematics and Computer Science, Eindhoven University of Technology, Eindhoven, The Netherlands
[2]Department of Epidemiology, University Medical Center Groningen, University of Groningen, Eindhoven, The Netherlands
[3]Department of Genetics, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands

Corresponding author:
Edwin R Van den Heuvel, Department of Mathematics and Computer Science, Eindhoven University of Technology, Den Dolech 2, Eindhoven, 5612 AZ, The Netherlands.
Email: e.r.v.d.heuvel@tue.nl

1 Introduction

Children with Phelan-McDermid syndrome (MIM# 606232) have a severe general developmental delay and autistic-like behavioral problems.[1] Phelan-McDermid syndrome is a rare chromosomal disorder with so far 50 diagnosed children and 25 diagnosed adults in the Netherlands. It is caused by a deletion of chromosome 22q13.3 and the neurological problems are thought to result from haploinsufficiency of the SHANK3 gene. Previous studies with intranasally administered insulin in healthy controls and dementia-related disorders showed a beneficial effect on cognitive function, declarative memory, and behavior.[2–7] In an explorative non-placebo controlled study with only six children with Phelan-McDermid syndrome, it was demonstrated that intranasal insulin might improve development in these children.[8] To confirm this, a double-blind randomized clinical trial (RCT) that compares intranasal insulin against placebo was needed to be designed that was suitable for an anticipated small number of participants.

The trial design should take into account that the cognitive abilities of children with Phelan-McDermid syndrome already develop or progress over time, although slower than in other children. This implies that the comparison of intranasal insulin with respect to placebo should focus on developmental growth. Thus, the RCT must entail a longitudinal study at which the development of the children is assessed at several time points. Development can be measured with the internationally accepted and validated Bayley-III assessment scales, resulting in an absolute score. This score can be transformed into a developmental age of the children using standardized tables.[9] Increase in developmental age of children between time points or developmental growth is then the primary outcome of the RCT and the average developmental growth may be compared between treatments.

Clearly, different designs can be selected. The most simple design is the classical parallel group design (PGD) in which a child is randomly allocated to one of the two treatment groups and then developmental growth is observed longitudinally and compared between treatments. An alternative approach is to match the participating children in pairs and then allocate the two treatments randomly within pairs. The matching may be conducted on developmental growth for a specified time period on the control treatment or before administration of any treatment. A third approach would be to use a two-by-two cross-over design. However, this option seems less realistic since it is indicated that the intranasal insulin may have long lasting effects[8] and this may introduce a carry-over effect that is different from placebo. A fourth option is to use a stepped wedge design (SWD).[10] In this design, each child starts with placebo, but at certain predefined time points one child or a portion of the children switches from placebo to the new treatment and remains on this treatment till the end of the study. The allocation of children to treatment switches is random. The SWD is a special type of cross-over design but with treatment switches in one direction.

The SWD is frequently considered more ethical than any of the other proposed designs, in particular when the treatment is considered to do more good than harm.[11–13] Indeed, it is considered unfair to withhold a truly beneficial treatment to a substantial number of participants in either the PGD or in the matched-pairs design (MPD) or to change to an inferior treatment in a cross-over trial. This ethical advantage of the SWD is an important motivation for parents to let their children with Phelan-McDermid syndrome participate in the clinical trial. However, the SWD has also received criticism and it has some serious drawbacks.[14,15] The most dominant disadvantages are that a SWD may take longer than the more traditional designs and that the required repeated measurements for the SWD could be a true burden. These arguments though, may not hold as strong for testing the effect of intranasal insulin on developmental growth of children with Phelan-McDermid syndrome, since the primary outcome would be developmental growth.

With regards to efficiency, the SWD has been compared primarily to cluster randomized designs in practice, i.e. designs where groups of participants, instead of participants themselves, are randomized to treatments. This comparison is reasonable, since the application of the SWD in practice has typically been focused on randomization of clusters of participants to the treatment switches.[11,12] Compared to cluster randomized trials, the SWD is somewhat more efficient.[10,13] When compared to more traditional designs, that do randomize individual patients to treatments, the performance of the SWD is unknown.

On the other hand, the SWD is similar in nature to the delayed start design (DSD) that is used in RCTs for Alzheimer and Parkinson's disease.[16,17] DSDs start as a PGD, but after a specific time period, part or all of the patients in the control group are changed to the new treatment. Besides demonstrating symptomatic efficacy between treatments, DSDs in Alzheimer and Parkinson studies are used to demonstrate that disease progression can be slowed by the new treatment (but this is not relevant in our trial). Thus, the DSD is typically characterized by just one switch, while the SWD would generally use multiple switches. Furthermore, in the SWD all patients typically start with the control treatment, while the DSD would allocate the new treatment to part of the patients at the start of the trial. Although the scientific scope in Alzheimer and Parkinson's disease is different for Phelan-McDermid syndrome, DSDs are suitable alternatives.

This paper compares the power of several designs that could be used to test intranasal insulin with respect to placebo on developmental growth under certain statistical assumptions. The designs involved are three SWDs, four DSDs, the classical PGD, and the MPD. The comparison is conducted with simulation studies, since the expected number of children that would participate in our trial in the Netherlands was determined at just 20 children. Asymptotic theory for small trials is not reasonable and is therefore disregarded. In our view, this is the first paper that investigates and discusses the SWD in comparison with classical and other designs for testing treatment effects on developmental growth of individuals and with small numbers of participants.

2 Methods

2.1A longitudinal design

For the longitudinal RCT, we decided to observe developmental age at five time points at intervals of six months, i.e. $t \in \{0,6,12,18,24\}$. Shorter time intervals are too much of a burden for the children, but more importantly, it is not very informative for developmental growth either and the potential test-retest effect is smaller with larger intervals. For the purpose of simulation and comparability reasons, we assume that all trials will start with administration of treatment (control or treatment) immediately at the beginning ($t = 0$). A visualization of the simulated trials, are provided in Figure 1. All trials use 20 subjects divided over different administration groups.

In the classical PGD, half of the children receive the new treatment and the other half receive placebo. In the MPD, the developmental growth in the first six months is used to create pairs of children with similar developmental growth. After six months, one child receives the treatment and the other child continues on placebo. For the SWD, the switches will typically coincide with the moments of measurements,[10,11] but the number of switches and the moment of switches must still be determined. Three SWDs were chosen. Two designs have only two switches (SWD-S2a and SWD-S2b) and one design has three switches (SWD-S3). The total treatment time on placebo and the new treatment was balanced in two designs (SWD-S2a and SWD-S3), but a shorter treatment period for placebo was selected in a third design (SWD-S2b). Four DSDs were selected

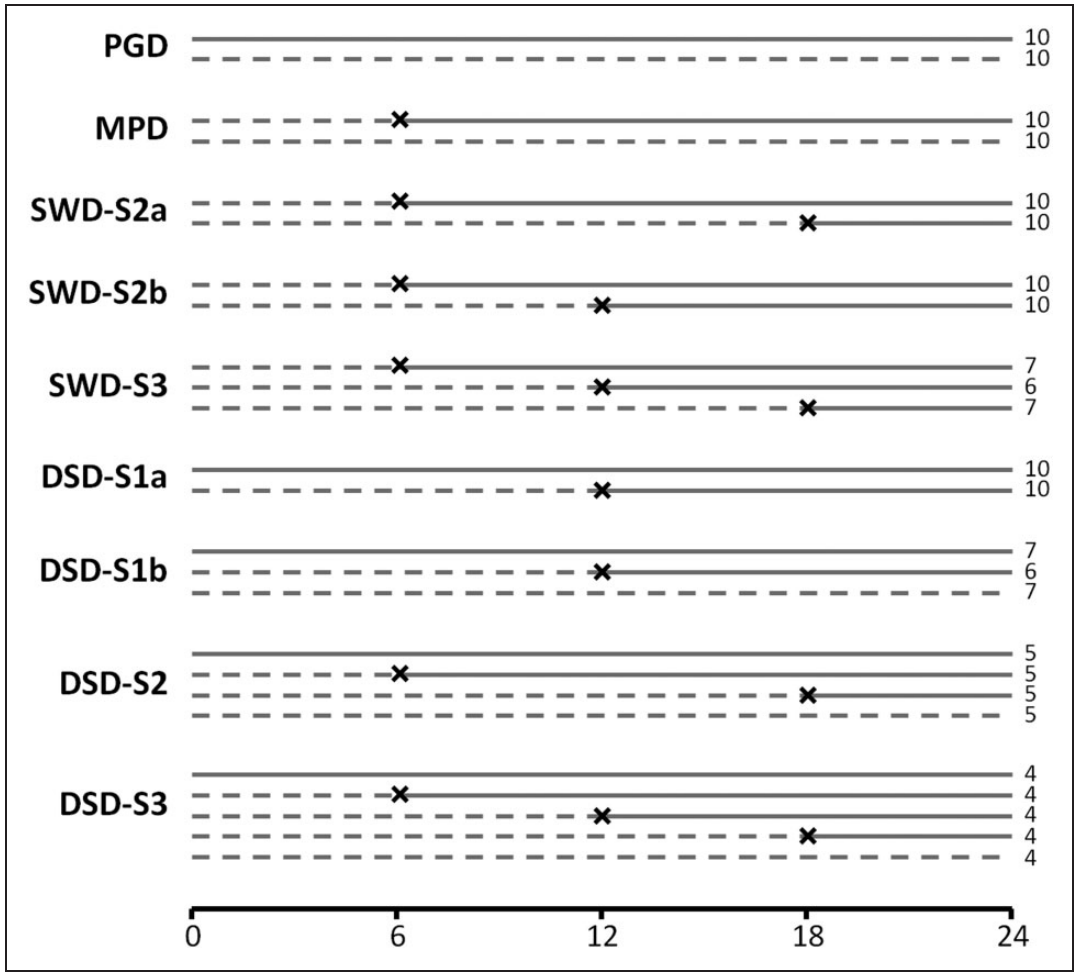

Figure 1. Visualization of the different trial designs with the number of children in each administration group. Dotted lines indicate placebo, solid lines indicate the new treatment, and crosses indicate a treatment switch. The numbers at the end of the lines refer to the number of children in this administration group.

(DSD-S1a, DSD-S1b, DSD-S2, DSD-S3), having one, two, or three switches. Two designs, which are common in practice, had just one switch, while two others had two or three switches. Again, we kept the designs balanced in the total time on placebo and the new treatment, except for DSD-S1a, which would have less treatment time on placebo. The allocation of children to the different administration groups would be determined by randomization.

2.2 Statistical model

The primary outcome measure at the different moments of measurement is developmental age (in years), observed by the Dutch version of the Bayley III. The selected statistical model for analyzing this outcome is a subject-specific linear mixed model. We assume that developmental age tested at the five time points ($t \in \{0,6,12,18,24\}$) can be adequately described with a linear

profile. This assumption may be reasonable considering the fact that (i) children with Phelan-McDermid have a relatively slow cognitive development, (ii) the trial period is only two years, and (iii) a standardized measure (developmental age) is applied that would eliminate non-linear growth patterns in developmental scores. Both the intercept and the slope of the linear profile are assumed child specific. The intercept represents the patient's developmental age at the start of the trial ($t=0$) and the slope represents patient's developmental change in age (in years) expressed as fraction of biological age (in years). Indeed, for healthy children the average slope should be approximately equal to one, since developmental age of children would follow biological age perfectly, at least on average.

At the moment of intranasal insulin treatment, we will assume that the slope is increased with a fixed incremental value, an immediate response. If we let Y_{ij} be the j^{th} developmental age for the i^{th} child, the statistical model can be written as

$$Y_{ij} = Z_{0,i} + Z_{1,i} \cdot t_{ij} + \gamma \cdot \left(t_{ij} - x_i\right) \cdot \delta_{(x_i,\infty)}\left(t_{ij}\right) + \varepsilon_{ij} \tag{1}$$

with $i=1,2,\ldots,m$ children, $j=1,\ 2,\ldots,\ n$ measurements in time, $Z_{0,i} \sim N\left(\beta_0, \tau_0^2\right)$ a random intercept, $Z_{1,i} \sim N\left(\beta_1, \tau_1^2\right)$ a random slope, $\varepsilon_{ij} \sim N\left(0, \sigma_0^2\right)$ a residual, γ the effect of intranasal insulin on developmental growth with respect to the control treatment, x_i the switch moment of child i, $t_{ij} \in \{0,6,12,18,24\}$ the time point of measurement, and $\delta_A(t)$ an indicator variable equal to one when $t \in A$ and zero otherwise. Furthermore, the children are considered independent from each other, the residuals within a child are considered independently distributed and independent of the random coefficients, but the random coefficients may be correlated with $\mathrm{CORR}\left(Z_{0,i}, Z_{1,i}\right) = \rho$.

The model in equation (1) is different from earlier described models for SWDs.[10] Our model has implemented developmental growth using a linear function in time with a random slope for each child, while previous models did not assume any time trends, and therefore modeled time as a categorical variable, and they assumed only a random intercept for the clusters of patients. Thus differences are determined by an accepted time trend (per child separately) and by a change from clusters of patients to individual patients. One specific consequence of these differences is that generalized estimating equations, recommended for an analysis on cluster level,[10] cannot be applied in our setting due to the individual growth trajectories and the individual treatment switches. On the other hand, a similar model was used to estimate the impact of the critical period "menarche" on body fat of 162 girls.[18] This model assumed a random intercept and a random slope before and after the critical period, all possibly correlated. It is an extension of model (1), since it also treats γ random.

All trial designs in Figure 1 fit within the formulation of statistical model (1). For the classical PGD half of the children 'switches' immediately at time point zero ($x_i=0$) to the new treatment, while the remaining half will never switch or switches only after the trial has finished ($x_i \geq t_{in}$). Something similar holds true for the DSDs. One group starts immediately at the new treatment, other groups will switch during the trial, and possibly one group may never switch during the trial.

For the MPD, the statistical model in equation (1) will also hold for each child, but pairing children on developmental growth in the first period of the trial puts a restriction on the randomization procedure for allocating treatments to children. The randomization is related to the random slopes and the first two residuals in equation (1). Indeed, the monthly developmental growth of child i in the first six months is estimated by $V_{0,i} \equiv (Y_{i2} - Y_{i1})/(t_{i2} - t_{i1})$. The matching of children is then conducted by pairing children sequentially after the monthly developmental growths have been ordered in size. Under the assumptions in equation (1), the observed baseline growth is

equal to $Z_{1,i} + (\varepsilon_{i2} - \varepsilon_{i1})/(t_{i2} - t_{i1})$. Thus, if the residual variance $2\sigma_0^2/(t_{i2} - t_{i1})^2$ is substantially smaller than the variance τ_1^2, children with developmental growths $Z_{1,(2k-1)}$ and $Z_{1,(2k)}$ will most likely be matched, where $Z_{1,(i)}$ is the ordered developmental growth of child i. In that case, the pair of children may follow almost the exact same growth under placebo and this pair of children should be taken as the unit of experimentation in model (1) instead of the individual. However, when matching is not perfect or developmental growth within pairs is not identical, model (1) can better be applied to individuals, not utilizing the matching procedure. The advantage of the MPD is that children with similar developmental growths are randomly allocated to placebo and intranasal insulin. In other designs, an unlucky randomization may destroy such balances, particularly in small trials.

The goal of the proposed designs is to estimate the effect size γ of the intranasal insulin treatment and to test the null hypothesis $H_0 : \gamma = 0$ against the alternative $H_1 : \gamma \neq 0$. Simulation studies are performed to determine the size of the type I error rate (probability of incorrectly rejecting the null hypothesis) and the power (the probability of correctly rejecting the null hypothesis) of the selected designs under the statistical assumptions. Note that the analysis of the formulated statistical model in equation (1) can be performed with procedure MIXED of SAS, version 9.3, using restricted maximum likelihood and Satterthwaite's degrees of freedom.[18–20]

2.3 Simulation: Parameter settings

To investigate the power of the different designs, we simulated data according to model (1). To obtain accurate power values, we need realistic parameter settings. These settings were obtained from preliminary data at baseline and literature. For 18 children, we observed their developmental age with the Bayley-III assessment score approximately six months before the trial and for four patients we received their change in developmental and biological age. The test-retest reliability of the Bayley-III assessment score was determined from literature.[9]

The developmental ages of the 18 children gave an average developmental age of 20.7 months and a standard deviation of 7.61 month. From this result, we chose a baseline developmental age of $\beta_0 = 21$ months in the simulation study. Furthermore, the sum of the variances of the intercept and the residual $\sigma_0^2 + \tau_0^2$ was taken equal to approximately 57.9 ($\approx 7.61^2$). With the reported test-retest reliability of 0.80, the relationship between σ_0^2 and τ_0^2 could be obtained: $\sigma_0^2 \approx 0.25\tau_0^2$. For the simulation study, we chose $\sigma_0^2 = 10$ and $\tau_0^2 = 45$. Furthermore, the four patients with information on change in developmental age gave an average developmental growth of 0.28 developmental years per biological year and a standard deviation of 0.313 developmental years per biological year. Thus, children with Phelan-McDermid syndrome seem to develop at a rate of approximately 30% of what is normal. From these results, we chose $\beta_1 = 0.3$ and $\tau_1^2 = 0.1$.

The correlation between the random intercept and the random slope in model (1) could not be obtained from preliminary results. We therefore chose three different values $\rho \in \{0.25,0.50,0.75\}$. The treatment effect γ, which is also expressed in terms of percentage growth, was also selected at different values. We chose the values $\gamma \in \{0,0.15,0.30,0.45\}$. These choices will provide type I error rates ($\gamma = 0$) and powers at different effect sizes ($\gamma > 0$). Note that if the treatment effect of intranasal insulin is equal to 0.45, the treated children will develop in cognition at a rate that is on average at 75% of normal children. For all combinations of parameter choices we simulated 50,000 times model (1), to be able to detect at least 1% difference in power values between two designs with a type I error of 0.05 and a type II error of 0.20. For each simulation, a treatment effect was tested with the standard two-sided Wald test statistic at the significance level of $\alpha = 0.05$. The proportion of simulations for which an effect was significant is either the

simulated type I error rate or the power of the selected RCT, depending on the size of the treatment effect ($\gamma = 0$ or $\gamma > 0$, respectively).

3 Results

Fitting model (1) to the simulated data was not always successful. The numerical procedure was unable to maximize the likelihood function in at most five simulated data sets from 50,000 simulations. The compromised simulations were omitted from the summary results for each design separately, but we believe that this is only a minor issue in the comparisons between the different trial designs.

The selected designs were all capable of estimating the treatment effect $\gamma \in \{0,0.15,0.30,0.45\}$ in an unbiased way. The average value of the estimated treatment effects $\hat{\gamma}$ over all simulations were very close to the true value. The absolute bias that was observed from all simulation settings was not larger than 0.018. It was attained with DSD-S1a for a treatment effect of $\gamma = 0.45$ and a correlation coefficient of $\rho = 0.50$. All the other fixed model parameters were also estimated quite close to the input parameters, including the variance components.

The type I error rates and the power values for the different designs are provided in Table 1 and Table 2, respectively. Each row in the table is essentially based on the same simulated data, while different rows are based on different simulated data. The simulated data in one row contains the

Table 1. Type I error rates (expressed as percentage) for the different designs.

			Stepped wedge			Delayed start			
Correlation	Parallel group	Matched Pairs	S2a	S2b	S3	S1a	S1b	S2	S3
$\rho = 0.25$	6.34	6.36	6.22	5.27	5.84	5.85	6.80	6.77	6.68
$\rho = 0.50$	6.33	6.64	6.12	5.45	5.89	5.91	6.61	6.68	6.71
$\rho = 0.75$	6.09	6.13	6.18	5.31	5.93	5.85	6.56	6.39	6.40

Table 2. Power values (expressed as percentage) for the different treatment effects in the different designs.

				Stepped wedge			Delayed start			
Treatment Effect	Correlation	Parallel group	Matched pairs	S2a	S2b	S3	S1a	S1b	S2	S3
$\gamma = 0.15$	$\rho = 0.25$	16.31	17.90	18.51	16.95	18.72	15.14	18.97	18.37	19.92
	$\rho = 0.50$	17.17	18.55	18.83	16.85	18.97	14.78	19.24	18.50	19.11
	$\rho = 0.75$	19.12	20.68	19.66	17.20	19.49	15.04	20.34	20.23	20.18
$\gamma = 0.30$	$\rho = 0.25$	45.23	48.98	51.53	50.05	53.00	41.83	51.46	49.96	51.99
	$\rho = 0.50$	48.18	51.91	53.19	50.49	53.99	42.20	53.03	52.03	53.57
	$\rho = 0.75$	55.42	58.39	56.59	51.37	56.76	42.60	57.85	57.27	57.80
$\gamma = 0.45$	$\rho = 0.25$	77.39	81.50	83.82	83.12	85.40	73.81	83.10	81.96	83.98
	$\rho = 0.50$	80.22	84.52	85.35	83.51	86.33	73.43	84.74	83.88	85.22
	$\rho = 0.75$	86.98	89.33	88.03	84.26	88.18	74.42	88.79	88.30	88.78

same random intercepts, slopes, and residuals for a child, but responses for children in different designs may still be different across designs since children will fall in different administration groups. The difference is determined by the systematic effect of treatment $\gamma \cdot (t_{ij} - x_i) \cdot \delta_{(x_i,\infty)}(t_{ij})$.

All designs demonstrate somewhat inflated type I error rates. Stepped wedge designs, SWD-S2b and SWD-S3, and delayed start design DSD-S1a provided the lowest inflated type I error rates. The correlation coefficient does not seem to affect the type I error rate clearly. Inflated type I error rates for linear mixed models on repeated measures with small sample sizes have been discussed in literature.[21] The reason is that the asymptotic chi-square approximation to the null-distribution of the test statistic is not fully appropriate for small sample sizes. We believe that more sophisticated approaches[22] to improve the approximation of the null-distribution are not needed since the inflation is only limited in our settings.

Delayed start design DSD-S1a has power values lower than the other designs. The reason is probably that this design is highly imbalanced in total amount of treatment time between placebo and the new treatment. The new treatment is administered in total for 360 months while placebo is only administered for 120 months, a ratio of 3:1. An imbalance is also present in SWD-S2b and the MPD, although less dramatic, the ratios being 5:3 and 3:5, respectively. SWD-S2b performs less than the other SWDs, but the MPD has the largest power whenever the correlation between the intercept and slope is large ($\rho = 0.75$). At other correlation coefficients though, the MPD is less powerful than the (best performing) stepped wedge and delayed start designs, in particular for low correlation coefficients. Although matching children with the same developmental growth seems to work best when the correlation coefficient is high, the ratio of the variation in developmental growths of children and the residual error at the first six months was roughly 1:5. Indeed, the variation in growth was set to $\tau_1^2 = 0.1$ and the residual variation for the difference in growth at the first six months was equal to $2\sigma_0^2/(t_{i2} - t_{i1})^2 \approx 0.56$. Thus, our matching procedure was probably not fully optimal, which may explain why the MPD is not better than the other designs for all settings, but it did support our decision to still apply model (1) on individuals and not on pairs.

Comparing the power values of the remaining designs demonstrates that there is not much difference in power between the stepped wedge and delayed start designs. The SWD-S3 has often the highest power, but the delayed start design with three steps also performs very well. At lower treatment effects ($\gamma = 0.15$) DSD-S1b is somewhat better than DSD-S3. The PGD, however, does not perform as good as the other designs. It is systematically worse than the MPD and all balanced stepped wedge and delayed start designs. Note that for cluster-randomized designs, the SWDs also have better power.[10,13]

A power of approximately 80–90% is only attained at a treatment effect of $\gamma = 0.45$. Lower treatment effects will not be detected as easily. Whether this effect size for the treatment with intranasal insulin is realistic is unknown, since literature has not reported any quantitative effect sizes of developmental growth (expressed in developmental age). It seems that the correlation coefficient does not change the power a lot, although a higher correlation seems to provide a little higher power. A positive correlation is to be expected, since a relative higher developmental age for a particular biological age would also imply a larger developmental growth.

4 Discussion

This paper discussed several options for a (double-blind) RCT that would be able to test an effect of intranasal insulin against placebo on developmental growth in children with Phelan-McDermid syndrome. The treatment effect was unbiasedly estimated in all combinations of designs and

analyses. All designs had a somewhat inflated type I error rate due to an imperfect approximation of the null-distribution of our test statistic in linear mixed models for small numbers of subjects. Three designs (PGD, SWD-S2b, and DSD-S1a) showed lower power values, thus including the parallel group design, but the remaining designs (MPD, SWD-S2a, SWD-S3, DSD-S1b, DSD-S2, and DSD-S3) provided similar power values. When the stepped wedge and delayed start designs are unbalanced in total treatment time between the actual treatment and placebo, power seems to be reduced. The MPD provided the highest power when the random intercept is highly correlated with the random slope of the random coefficients model. For other settings though, the stepped wedge and delayed start design with three steps seem to provide the highest powers. A positive effect of the number of steps on the power was however limited, which may indicate that the number of steps can be chosen on the basis of logistic or other practical reasons.

One concern about the different designs is the limited power when only an expected number of 20 children will be included in the trial. The developmental growth of children with Phelan-McDermid syndrome must then increase with 0.45 to an average growth of approximately 0.75 developmental years per biological year to obtain a power of more than 80%. An average growth of 0.75 means that these children will mentally develop on average at 75% of their biological age. Although the studies would benefit from an increased number of participants, we believe that if intranasal insulin would be truly effective, there is still a reasonable chance that the positive effect will be detected even if only 20 children participate in the RCT.

The SWD under the assumed statistical conditions is competitive with the more traditional designs and with the DSDs (typically used in Alzheimer and Parkinson's disease). In SWDs, all participants will receive the new treatment during the trial, which is considered an ethical advantage and a reason for them to participate in the trial. An extension phase to a PGD could be considered another alternative, but this resembles in a way the somewhat less powerful DSD with only one switch and where everyone switches to the new treatment (DSD-S1a). A disadvantage of any design that uses switch moments is the increased vulnerability to temporal changes unrelated to treatment. Temporal changes are less of a problem in PGDs and MPDs. The MPD though, which had the highest power under certain settings, is logistically more difficult to execute.

Although each design has its own advantages and disadvantages for a given study project, we selected the SWD with three steps (SWD-S3) to study the effect of intranasal insulin on developmental growth of Phelan-McDermid children. Multiple steps could reduce a possible placebo effect that may occur at the only switch moment of a one-step SWD, when the switch moment is known to participants. However, we expect no or hardly any difference between natural growth and placebo treatment, since children have a limited mental capacity and are not aware of the effect of medication on their development. This provided us the option to combine our trial with another planned six months follow-up study on natural developmental growth in the same Phelan-McDermid children. In our stepped wedge design, administration of intranasal insulin and placebo started only at six months and no treatment was administered at the first six months. This choice resulted in a lower burden of daily intranasal administration compared to the simulated SWD with three steps. Our trial design has been approved by our institutional medical ethical committee.

Funding

As part of the 'Priority Medicine for Children' program, the Netherlands Organization for Health Research and Development awarded C.M.A. Van Ravenswaaij-Arts a grant (ZonMw grant nr. 113202009).

Acknowledgements

We would like to thank S.A.J. Ruiter and E. De Kievit for contributing the developmental information that was used for the simulation study.

References

1. Phelan K and McDermid HE. The 22q13.3 Deletion Syndrome (Phelan-McDermid Syndrome). *Mol Syndromol* 2012; **2**: 186–201.
2. Benedict C, Hallschmid M, Hatke A, et al. Intranasal insulin improves memory in humans. *Psychoneuroendocrinology* 2004; **29**: 1326–1334.
3. Benedict C, Kern W, Schultes B, et al. Differential sensitivity of men and women to anorexigenic and memory-improving effects of intranasal insulin. *J Clin Endocrinol Metabol* 2008; **93**: 1339–1344.
4. Reger MA, Watson GS, Frey 2nd WH2nd, et al. Effects of intranasal insulin on cognition in memory-impaired older adults: modulation by APOE genotype. *Neurobiol Aging* 2006; **27**: 451–458.
5. Reger MA, Watson GS, Green PS, et al. Intranasal insulin administration dose-dependently modulates verbal memory and plasma amyloid-beta in memory-impaired older adults. *J Alzheimers Dis* 2008; **13**: 323–331.
6. Krug R, Benedict C, Born J, et al. Comparable sensitivity of postmenopausal and young women to the effects of intranasal insulin on food intake and working memory. *J Clin Endocrinol Metabol* 2010; **95**: E468–E72.
7. Craft S, Baker LD, Montine TJ, et al. Intranasal insulin therapy for Alzheimer disease and amnestic mild cognitive impairment: a pilot clinical trial. *Arch Neurol* 2012; **69**: 29–38.
8. Schmidt H, Kern W, Giese R, et al. Intranasal insulin to improve developmental delay in children with 22q13 deletion syndrome: an exploratory clinical trial. *J Med Genet* 2009; **46**: 217–222.
9. Bayley N. *Bayley scales of infant and toddler development*. San Antonio: Hartcourt Assessment, 2006.
10. Hussey MA and Hughes JP. Design and analysis of stepped wedge cluster randomized trials. *Contemp Clin Trials* 2007; **28**: 182–191.
11. Brown CA and Lilford RJ. The stepped wedge design: a systematic review. *BMC Med Res Meth* 2006; **5**: 1–9.
12. Mdege ND, Man M-S, Taylor CA, et al. Systematic review of stepped wedge cluster randomized trials shows that design is particularly used to evaluate interventions during routine implementation. *J Clin Epidemiol* 2011; **64**: 936–948.
13. Woertman W, De Hoop E, Moerbeel M, et al. Stepped wedge designs could reduce the required sample size in cluster randomized trials. *J Clin Epidemiol* 2013; **66**: 752–758.
14. Zhan ZZ, Van den Heuvel ER, Doornbos PM, et al. Strength and weaknesses of a stepped wedge cluster randomized design: its application in a colorectal cance follow-up study. *J Clin Epidemiol* 2014; **67**: 454–461.
15. Kotz D, Spigt M, Arts ICW, et al. Use of the stepped wedge design cannot be recommended: a critical appraisal and comparison with the classic cluster randomized controlled trial design. *J Clin Epidemiol* 2012; **65**: 1249–1252.
16. McDermott MP, Hall WJ, Oakes D, et al. Design and analysis of two-period studies potentially disease modifying treatments. *Contr Clin Trials* 2002; **23**: 635–649.
17. Zhang RY, Leon AC, Chuang-Stein C, et al. A new proposal for randomized start design to investigate disease-modifying therapies for Alzheimer disease. *Clin Trials* 2011; **8**: 5–14.
18. Naumova EN, Must A and Laird NM. Tutorial in biostatistics: evaluating the impach of'critical periods' in longitudinal studies of growth using piecewise mixed effects models. *Int J Epidemiol* 2001; **30**: 1332–1341.
19. Software Manual, SAS Institute Inc., *SAS/STAT® 9.3 User's Guide*, Cary, NC: SAS Institute Inc, 2011.
20. Fitzmaurice G, Davidian M, Verbeke G, et al. *Longitudinal data analysis*. Boca Raton, FL: Chapman and Hall, 2008.
21. Manor O and Zucker DM. Small sample inference for the fixed effects in the mixed linear model. *Comput Stat Data Anal* 2004; **46**: 801–817.
22. Skene SS and Kenward MG. The analysis of very small samples of repeated measurements I: an adjusted sandwich estimator. *Stat Med* 2010; **29**: 2825–2837.