Design and analysis of multilevel intervention studies

Summary

In this thesis it was shown that the use of the multilevel model is necessary for the design and analysis of experiments when individuals are nested in clusters and the results from the study have to be generalizable to a population of clusters. Examples of such experiments are school-based smoking prevention interventions with pupils nested within schools, and multi-center clinical trials with patients nested within clinics. The use of traditional statistical techniques may lead to incorrect results of statistical tests of and confidence intervals for treatment effects, and to non-optimally designed experiments. Multilevel models are, however, more complicated than more traditional regression models and specialized software is necessary for data analysis when using multilevel models. Therefore a comparison was made in Chapter 2 between the multilevel model and three traditional models for continuous outcomes to show when and how these traditional models fail. The traditional models are the disaggregated data model, the aggregated data model, and the fixed effects model, which considers the clusters (i.e. schools, clinics) to be fixed. The results obtained with the fixed effects model only apply to the clusters in the study. The multilevel model, however, assumes that the clusters represent a random sample from some population, and so the results may be generalized to this population. The comparison in Chapter 2 was made for two levels of nesting (pupils within classes), continuous outcomes, and three different situations: randomization at the class level, and randomization at the pupil level with a treatment effect that is constant or that varies across classes, which correspond to a multilevel model with a fixed or random slope, respectively. Criteria for comparison were the estimator of the treatment effect and its variance. The results in this chapter show that traditional models generally lead to an under- or overestimation of the variance of the treatment effect estimator, which may result in type I or type II errors, respectively. Such errors especially occur when the dependency of outcomes within a class and/or the number of pupils per class is large. For balanced designs with a continuous outcome and no covariates several alternatives to the multilevel model are available: (a) the fixed effects model may be used when randomization is done at the pupil level and the treatment effect does not vary across classes, (b) the aggregated data model may be used when randomization is done at the pupil level and the treatment effect varies across classes, (c) or when randomization is done at the class level. When covariates are included in the model or the design is unbalanced, however, the multilevel model is

recommended as analysis model.

Two methods for the analysis of multilevel binary data were compared in Chapter 5: estimation by numerical integration, and Penalized Quasi Likelihood (PQL). PQL is based on linearization of the non-linear model using a Taylor series expansion, and appears in a first and a second order. The comparison was made for two levels of nesting, models without covariates, and the three situations that were also considered in Chapter 2. The methods were compared on the basis of their point estimator of the treatment effect, and the distribution of the test statistic for testing the treatment effect. The results show that first order PQL performs best in terms of Mean Squared Error of the point estimator of the treatment effect, although there is not much difference with the other methods. For this method, however, the distribution of the test statistic for testing the treatment effect does not have a standard normal distribution under the null hypothesis due to bias in the point estimator. The standard normal distribution may be used as a reference distribution under the null hypothesis for second order PQL. One may choose to use likelihood ratio tests instead of Wald tests when estimation by numerical integration is used. The latter estimation method at least as implemented in the current version of MIXOR has the disadvantage that non-convergence may occur quite often, or that the estimation process may converge to an evidently incorrect estimate. Both second order PQL and estimation by numerical integration perform well in terms of estimating the treatment effect since the relative bias is 5% or less. The bias may, however, be much higher for variance components estimation, especially with numerical integration with MIXOR, and therefore the development of other estimation methods for multilevel logistic models is necessary.

Chapters 3, 4, and 6 deal with the optimal design of experiments in a multilevel setting, for example school-based smoking prevention interventions with pupils nested within schools. When designing such experiments the optimal level of randomization, and the optimal sample sizes at each level of the data structure have to be established. This not only means determining the total number of pupils, but also the number of schools and the number of pupils per school. The number of schools and the number of pupils per school are restricted by financial limitations since the costs for enrolling schools and pupils may not exceed the budget. The costs are calculated as the sum of the total costs at the school level, which are equal to the costs per school multiplied by the number of schools, and the total costs at the pupil level, which are equal to the costs at the pupil level multiplied by the number of pupils. This means that the number of pupils per school decreases when the number of schools in the experiment increases and vice versa, and the optimal sample sizes at both levels have to be established. Furthermore the optimal sample sizes may be restricted by the actual number of schools or the actual number of pupils per school.

An optimality criterion is necessary for calculating the optimal design. Generally the treatment effect is the parameter of most interest in experiments, and assuming (almost) unbiased point estimation, the variance of its estimator has to be minimized to obtain smallest confidence intervals for the treatment effect and maximal power of the test on treatment effect. This optimality criterion, together with the precondition that the costs may not exceed the budget,

were used in Chapters 3 and 6, in which optimal designs were established for linear and logistic multilevel models with two or three levels of nesting. The results in these chapters show that the pupil level is the optimal level of randomization, and randomization at this level is especially preferable when the variance components corresponding to the intercept at higher levels are relatively large and/or when the number of pupils per school is very large. The optimal sample sizes and the variance of the treatment effect estimator given the optimal sample sizes were derived, and of course this variance decreases when the budget increases. The optimal sample sizes at higher levels are equal to two when randomization is done at the pupil level and the treatment effect is constant across higher level units. One might wonder if such a small number of units in a random sample may give enough information about their population. A rule of thumb for selecting between the multilevel and fixed effect model is therefore necessary. Furthermore one may wonder if the variance components at higher levels can be estimated well with such small sample sizes, but these estimated variance components, however, are not necessary for calculating the variance of the treatment effect estimator. This problem, however, generally does not occur when the model contains a random slope since then the optimal numbers of higher level units are generally larger than two. The optimal design for multilevel logistic models can only be derived analytically for first order Marginal Quasi Likelihood (MQL), which is known to produce biased estimates. Therefore a simulation study was performed to 'translate' the analytical results on optimal designs obtained with first order MQL into results for PQL estimation or estimation by numerical integration.

Chapter 4 builds upon the results of Chapter 3 by including relevant covariates into the model, a multi-center clinical trial with patients nested within clinics was used as an example. In this chapter not only the variance of the treatment effect estimator was used as optimality criterion, D - and A - optimal designs were derived as well. For these three optimality criteria the patient level was again the optimal level of randomization. The optimal sample sizes did not depend on the optimality criterion for randomization at the clinic level. Furthermore, the effect of including or excluding covariates on the variance components and optimal designs was established in this chapter as well. It was shown that the variance of the treatment effect estimator generally increases when covariates are excluded from the model, at least assuming that the treatment factor is uncorrelated with the covariates due to the randomization procedure.

Chapter 7 illustrates the results on optimal designs from Chapters 3, 4, and 6. In this chapter a school-based smoking prevention intervention was optimally planned using results from a recent study. To plan the intervention as efficient as possible, the costs at both the pupil and school level and the budget have to be known in advance, as well as the drop-out rate at both levels, and the values of the model parameters. This leads to the following problem: intervention studies are developed and implemented to gain some knowledge about the values of model parameters, in particular the treatment effect, but the values of these parameters have to be known to plan an intervention study optimally. To solve this problem a prior estimate may be used, which may be obtained from theoretical knowledge about the smallest relevant treatment

effect, a pilot study, or a similar intervention study in which the same outcome variables were analyzed. In Chapter 7 several designs were compared and it was shown how to deal with limited school sizes and a limited number of schools. It was shown that an almost optimal design could be achieved with a considerably lower budget.

It has to be noted that this thesis mainly focused on the design and analysis of multilevel experimental data with one treatment factor with two levels. Future research may focus on designing optimal experimental designs for models with more than one treatment factor which may be randomized at either level, for models with one treatment factor with more than two levels, or for models with quantitative treatments. It is also necessary to develop optimal designs for multilevel logistic models with covariates.

Samenvatting in het Nederlands (Dutch summary)

Uit dit proefschrift volgt dat het gebruik van het multi-niveau model noodzakelijk is voor het ontwerp en analyse van experimenten waarbij individuen genesteld zijn binnen clusters en wanneer de resultaten van het experiment generaliseerbaar moeten zijn naar een populatie van clusters. Voorbeelden van zulke experimenten zijn rook-preventie interventies met leerlingen genesteld binnen scholen, en multi-centre klinisch onderzoek met patiënten genesteld binnen klinieken. Het gebruik van traditionele statistische technieken kan leiden tot incorrecte resultaten van statistische toetsen op en betrouwbaarheidsintervallen voor het behandelingseffect, en tot niet-optimaal ontworpen experimenten. Multi-niveau modellen zijn echter complexer dan meer traditionele regressie modellen en specialistische computerprogrammatuur is nodig voor de data analyse wanneer het multi-niveau model gebruikt wordt. Daarom werd in Hoofdstuk 2 een vergelijking gemaakt tussen het multi-niveau model en drie traditionele modellen om aan te tonen wanneer en hoe de traditionele modellen tot incorrecte resultaten leiden. De traditionele modellen zijn het aggregatie model, het disaggregatie model, en het vaste effecten model, welke de clusters (bijvoorbeeld scholen of klinieken) als vaste effecten verondersteld. De resultaten van het vaste effecten model zijn alleen geldig voor die clusters die in het experiment betrokken zijn. Het multi-niveau model beschouwd de clusters echter als een toevallige steekproef uit een populatie, en de resultaten van het onderzoek mogen dan gegeneraliseerd worden naar deze populatie. De vergelijking in Hoofstuk 2 werd gemaakt voor twee niveaus van nesteling (leerlingen in klassen), continue uitkomsten, en drie verschillende situaties: randomisatie naar behandelingscondities op het klas niveau, randomisatie op het leerling niveau met een behandelingseffect dat constant is of dat varieert over de klassen. Dit correspondeert met een multi-niveau model met een vast respectievelijk een random hellingshoek voor het behandelingseffect. Criteria voor de vergelijking waren de schatter van het behandelingseffect en de variantie van deze schatter. De resultaten in dit hoofdstuk laten zien dat traditionele modellen in het algemeen tot een onder- of overschatting van deze variantie leiden. Dit kan resulteren in type I, respectievelijk type II fouten, en dit soort fouten komt vooral voor wanneer de afhankelijkheid van uitkomsten binnen eenzelfde klas en/of wanneer het aantal leerlingen per klas groot is. Voor gebalanceerde proefopzetten en modellen zonder covariaten zijn verschillende alternatieven voor het multi-niveau model beschikbaar: (a) het vaste effecten model mag gebruikt

worden wanneer op het leerling niveau gerandomiseerd wordt en het behandelingseffect varieert over klassen, (b) het aggregatie model mag gebuikt worden wanneer op het leerling niveau gerandomiseerd wordt en het behandelingseffect varieert over klassen, (c) of wanneer randomisatie is gedaan op het klas niveau. Het multi-niveau wordt echter aangeraden wanneer covariaten zijn opgenomen in het model of wanneer de proefopzet ongebalanceerd is.

Twee methoden voor de analyse van multi-niveau binaire data werden vergeleken in Hoofdstuk 5: schatting door middel van numeriek integratie, en *Penalized Quasi Likelihood* (PQL). PQL is gebaseerd op linearisatie van het niet-lineaire model door middel van een Taylor reeks, waarbij een eerste of tweede orde benadering wordt gebruikt. De vergelijking werd gemaakt voor twee niveaus van nesteling, modellen zonder covariaten, en de drie situaties die ook in Hoofdstuk 2 werden behandeld. De methoden werden vergeleken op basis van hun puntschatter van het behandelingseffect, en de verdeling van de toetsingsgrootheid voor de toets op het behandelingseffect. De resultaten laten zien dat eerste orde PQL de beste parameter schattingen geeft aangezien deze methode de kleinste *Mean Squared Error* oplevert, alhoewel er weinig verschil is met de andere methoden. Voor eerste orde PQL is de toetsingsgrootheid voor het toetsen van het behandelingseffect niet standaard normaal verdeeld onder de nul hypothese, wat een gevolg is van de onzuiverheid in de puntschatter. De standaard normale verdeling mag echter wel gebruikt worden als referentieverdeling onder de nul hypothese voor tweede orde PQL. Men kan ervoor kiezen *likelihood ratio* toetsen te gebruiken in plaats van Wald toetsen wanneer schatting door middel van numerieke integratie wordt gedaan. Deze methode heeft echter het nadeel dat de schattingsprocedure vaak niet convergeert, of dat convergentie naar een duidelijk incorrecte schatting optreedt. Zowel tweede orde PQL als schatting via numerieke integratie schatten het behandelingseffect vrij goed aangezien de relatieve onzuiverheid 5% of minder is. De onzuiverheid kan echter veel hoger zijn voor het schatten van variantiecomponenten, en de ontwikkeling van andere schattingsmethoden voor multi-niveau logistische modellen is daarom noodzakelijk.

De Hoofdstukken 3, 4, en 6 behandelen het optimaal ontwerpen van experimenten in multi-niveau situaties, bijvoorbeeld rook-preventie interventies met leerlingen genesteld binnen scholen. Bij het ontwerpen van dit soort experimenten moeten het optimale niveau van randomisatie, en de optimale steekproefgrootte op elk niveau bepaald worden. Dit betekent dat niet alleen het totaal aantal leerlingen, maar ook het aantal scholen en het aantal leerlingen per school bepaald moet worden. Het aantal scholen en het aantal leerlingen per school wordt beperkt door financiële restricties aangezien de kosten voor het opnemen van leerlingen en scholen in het onderzoek het budget dat daarvoor aanwezig is niet mogen overschrijden. De kosten van het onderzoek worden berekend als de som van de totale kosten op het school niveau, welke gelijk zijn aan de kosten per school vermenigvuldigd met het aantal scholen, en de totale kosten op het leerling niveau, welke gelijk zijn aan de kosten per leerling vermenigvuldigd met het aantal leerlingen. Dit betekent dat het aantal leerlingen per school afneemt wanneer het aantal scholen in het experiment toeneemt en vice versa, en de optimale steekproefgroottes op beide niveaus

moeten bepaald worden. Deze optimale steekproefgroottes kunnen niet groter zijn dan het werkelijke aantal scholen en het aantal leerlingen per school.

Een optimaliteitscriterium is noodzakelijk voor de berekening van de optimale proefopzet. Het behandelingseffect is over het algemeen de parameter waarin men het meest geïnteresseerd is, en de variantie van de schatter van deze parameter moet geminimaliseerd worden om zo smal mogelijke betrouwbaarheidsintervallen voor het behandelingseffect te verkrijgen, en een maximaal onderscheidingsvermogen voor de toets op het behandelingseffect. Dit optimaliteitscriterium, in combinatie met de randvoorwaarde dat het budget niet overschreden mag worden door de kosten, werd gebuikt in Hoofdstukken 3 and 6, waarin optimale proefopzetten werden afgeleid voor lineaire en logistische multi-niveau modellen met twee of drie niveaus van nesteling. De resultaten in deze hoofdstukken laten zien dat het leerling niveau het optimale niveau van randomisatie is, en dat randomisatie op dit niveau vooral de voorkeur heeft wanneer de variantiecomponenten op hogere niveaus relatief groot zijn en/of wanneer het aantal leerlingen per school groot is. De optimale steekproefgroottes en de variantie van de schatter van het behandelingseffect onder de optimale steekproefgroottes werden gegeven, en deze variantie neemt vanzelfsprekend af wanneer het budget toe neemt. Voor multi-niveau logistische modellen kunnen de optimale steekproefgroottes alleen analytisch afgeleid worden voor eerste orde *Marginal Quasi Likelihood* (MQL), welke onzuivere schattingen oplevert. Daarom werd een simulatie studie uitgevoerd om de analytische resultaten voor eerste orde MQL te vertalen in resultaten voor PQL of schattingen via numerieke integratie. De optimale steekproefgroottes op hogere niveaus zijn gelijk aan twee wanneer randomisatie is gedaan op het leerling niveau en het behandelingseffect constant is over hogere orde eenheden. Men kan zich afvragen of zo'n klein aantal eenheden in een toevallige steekproef genoeg informatie geeft over hun populatie. Een vuistregel voor het kiezen tussen het vaste en het random effecten model is daarom noodzakelijk. Tevens kan men zich afvragen of de variantie componenten op hogere niveaus wel goed geschat kunnen worden met zulke kleine steekproefgroottes. Dit probleem doet zich over het algemeen niet voor wanneer het model een random intercept bevat omdat dan de optimale steekproefgroottes op hogere niveaus over het algemeen groter zijn dan twee.

Hoofdstuk 4 bouwt voort op Hoofdstuk 3 door relevante covariaten in het model op te nemen, een multi-center klinisch onderzoek met patiënten genesteld binnen klinieken werd gebruikt als voorbeeld. In dit hoofdstuk werd niet alleen de variantie van de schatter van het behandelingseffect als optimaliteitscriterium gebruikt, D - en A - optimale proefopzetten werden ook afgeleid. Voor deze drie criteria was het patiënt niveau het optimale niveau van randomisatie. De optimale steekproefgroottes hingen niet af van het optimaliteitscriterium wanneer randomisatie werd gedaan op het kliniek niveau. Tevens werd het effect van toevoegen of verwijderen van covariaten op optimale proefopzetten bepaald. De variantie van de schatter van het behandelingseffect neemt over het algemeen toe wanneer covariaten uit het model verwijderd worden, maar in sommige situaties blijft deze variantie constant wanneer covariaten worden verwijderd of toegevoegd.

Hoofdstuk 7 illustreert de resultaten over optimale proefopzetten uit de Hoofdstukken 3, 4, en 6. In dit hoofdstuk werd een rook-preventie interventie optimaal gepland waarbij gebruik werd gemaakt van resultaten van een recent onderzoek. Om de interventie zo optimaal mogelijk te kunnen plannen moeten de kosten op zowel het leerling als het school niveau en het budget van te voren bekend zijn, evenals de uitvalpercentages op beide niveaus, en de waarden van de parameters in het model. Dit leidt tot het volgende probleem: interventie studies worden ontwikkeld en uitgevoerd om enig idee te krijgen over de waarden van de model parameters, in het bijzonder het behandelingseffect, maar de waarden van deze parameters moeten van te voren bekend zijn om een interventie onderzoek zo optimaal mogelijk te kunnen plannen. Om dit probleem op te lossen kan een *prior* schatting gebruikt worden, welke verkregen kan worden uit theoretische kennis over het kleinste relevante behandelingseffect, een vooronderzoek, of een vergelijkbare interventie studie waarin dezelfde uitkomstmaten werden geanalyseerd. In Hoofdstuk 7 werden verschillende proefopzetten vergeleken en het werd uitgelegd hoe men om moet gaan met beperkte schoolgroottes en een beperkt aantal scholen. Er werd ook getoond dat een vrijwel optimale proefopzet kan worden bereikt men een aanzienlijk lager budget.

Het moet opgemerkt worden dat in dit proefschrift voornamelijk aandacht besteed werd aan multi-niveau experimentele data met één behandelingsfactor met twee niveaus. Toekomstig onderzoek kan zich richten op het ontwerp van optimale proefopzetten voor modellen met meer dan één behandelingsfactor die op elk niveau gerandomiseerd kunnen worden, voor modellen met één behandelingsfactor met meer dan twee niveaus, of voor modellen met kwantitatieve behandelingsfactoren. Het is ook noodzakelijk om optimale proefopzetten te ontwikkelen voor multi-niveau logistische modellen met covariaten.