

Adjusted Risk Difference Estimation: An Assessment of Convergence Problems with Application to Malaria Efficacy Studies

Mukaka Mavuto^{2,4,5*}, White SA³, Mwapasa V², Kalilani-Phiri L², Terlouw DJ^{1,2,3} and Faragher Brian E³

¹College of Medicine, University of Malawi, Malawi



²Department of Public Health, University of Malawi, Malawi

³Liverpool School of Tropical Medicine, UK

⁴Mahidol-Oxford Tropical Medicine Research Unit, Mahidol University, Thailand

⁵Nuffield Department of Medicine Research Building, University of Oxford, UK

***Corresponding author:** Mavuto Mukaka, Head of Statistics, Mahidol-Oxford Tropical Medicine Research Unit, Mahidol University, 60th Anniversary Chalermprakiat Building, 3rd Floor, 420/6 Ratchawithi Rd, Ratchathewi District, Bangkok, 10400, Thailand.

Submission:  November 10, 2017; **Published:**  December 22, 2017

Abstract

A common measure of treatment effect in malaria efficacy studies is the risk difference, which can be estimated using binomial regression models. These models can fail to provide estimates, however, due to model failure or model convergence problems. Such failure most commonly occurs when the rate is close to 0% or 100% (a "boundary problem") but can also occur occasionally even when the rate is not close to a boundary. This paper reports the findings from simulation studies performed to evaluate the factors that may contribute to model failure when using binomial regression to derive risk difference estimates.

Convergence rates were found to fall:

- i) As one or both efficacy rates moved towards a boundary value, irrespective of the number of covariates included in the model;
- ii) As the numbers of covariates in the model increased;
- iii) As the levels of correlation between covariates the covariates increased. In all circumstances, convergence was poor when the efficacy rate in either group was 90% or more.

Introduction

In Randomized Controlled Trials (RCTs), treatment effect for binary outcomes is usually measured using a risk ratio, an odds ratio or a risk difference [1]. Risk differences are becoming more widely reported in RCTs, especially in malaria efficacy studies [2-4]. These are now frequently used when the objective is to compare treatment effect between treatments as commonly used alternatives such as odds ratios and risk ratios can be difficult for general medical researchers to interpret [5]. A major concern with odds ratios is that these are prone to misinterpretation even when the efficacy of the outcome measure is common, easily leading to invalid inferences of treatment effect [6-9]. Likewise, risk ratios can be difficult to understand, especially in equivalence studies, as their interpretation depends on whether a failure or a success is being modelled; while failure and success rates are clearly complementary, the inference that can be drawn for each of these endpoints may differ in some cases [5,10].

Risk difference estimation is preferred by many researchers not only because it is easier to interpret but for the arguably more important reason that it is frequently the most relevant and biologically sensible measure to report [11] and provides a clear interpretation of the scientific question of interest. That is, it directly answers the question of whether there is a difference in risk and/or efficacy between two groups.

The binomial regression model with an identity link function is often the standard model used to estimate risk differences. Risk difference models, however, sometimes encounter the problem that the binomial regression algorithm fails to converge or to produce output [10]. The current literature suggests that this problem is most likely to occur when one or both rates are close to a boundary (0% or 100%) and hence when risk ratios approach either infinity or zero [12,13]. Experience working with risk difference modelling using the Stata software, however, shows that these models can

sometimes fail to converge or produce output even in instances where both rates are reasonably distant from boundary values. Furthermore, even in scenarios when one or both rates are close to a boundary, convergence can be achieved in some cases. These observations suggest that factors other than boundary problems may influence non-convergence and/or output failure.

Simulation studies were performed to evaluate the potential impact of convergence problems on the estimation of treatment effect using risk difference modelling when adjusting for potential confounders. This research was motivated by a malaria efficacy study that was conducted in Malawi between 2003 and 2005; full details of the methods and findings of this study are available in Bell DJ et al. [2] but to summarise briefly:

A four-group blinded randomized controlled study was performed to compare the efficacy of three Sulfadoxine-Pyrimethamine (SP)-Based Combination therapies (SP plus chloroquine (CQ), SP plus artesunate (ART) and SP plus amodiaquine (AQ)) plus a placebo (control) group. This control arm comprised of SP plus placebo as SP was the standard first line treatment for uncomplicated falciparum malaria in Malawi during the study period. The study was conducted in response to accumulating evidence indicating that some resistance to SP was developing, and because the WHO was recommending the use of combination therapies (especially the artemisinin combination therapies (ACTs)). Participants were children aged 1-5 years.

Treatment effect was modeled using risk differences. Covariates, denoted as X , included in the analyses

$$\text{were } X = \begin{bmatrix} \log_e(\text{age}) \\ hb \\ \log(wt) \\ \log(para) \end{bmatrix}$$

Adjusted risk difference modelling, however, encountered model failure problems.

Simulation studies were carried out to explore the impact of the following issues on model convergence rates: 1) the number of covariates in the model; 2) the level of correlation between the covariates; 3) the proximity to a boundary of one or more efficacies.

Methods

Datasets were simulated based on the methods and results of the randomized control trial described in the previous section Bell et al. [2]. These were used to compare the efficacy rates in two treatment groups ("active" and control) with adjustment for up to two covariates. Two patterns of correlation between the covariates were examined; no/negligible correlations and correlations similar to those found by Bell et al. [2]. All simulations used a sample size 200 (100 participants randomized to each group), again as used by Bell et al. [2]. Five thousand (5,000) datasets were simulated for each scenario considered. Independent (predictor) variables were generated using multivariate log-Normal distributions except for haemoglobin level (hb) which was generated using a Normal

distribution; these were used as covariates in a binomial model using an identity link to estimate the risk difference between the treatment and control arms.

The choice of covariates was based on expert knowledge from malaria researchers suggesting that age, weight (wt), haemoglobin (hb) level and parasite count (para) can confound the relationship between treatment and efficacy in malaria RCTs. Standard practice in such RCTs is to estimate adjusted treatment effects controlling for potential confounders.

Parameter values for simulation of covariates

The matrices of parameters and values for simulating the baseline covariates were as follows:

$$X = \begin{bmatrix} \log_e(\text{age}) \\ hb \\ \log(wt) \\ \log(para) \end{bmatrix}$$

$$\mu = \begin{pmatrix} 3.15 \\ 9.32 \\ 2.40 \\ 10.7 \end{pmatrix}$$

$$\sigma = \begin{pmatrix} 0.42 \\ 1.66 \\ 0.18 \\ 1.50 \end{pmatrix} \text{ and}$$

$$\rho = \begin{pmatrix} 1.00 & 0.09 & 0.16 & 0.02 \\ 0.09 & 1.00 & 0.4 & 0.2 \\ 0.16 & 0.4 & 1.00 & 0.05 \\ 0.02 & 0.2 & 0.05 & 1.00 \end{pmatrix}$$

where:

X is a vector of values for the four covariates: $\log(\text{age})$, $\log(wt)$, $\log(para)$ and haemoglobin (hb).

μ and σ are vectors of mean and standard deviation values respectively for (in order) $\log(\text{age})$, hb, $\log(wt)$ and $\log(para)$.

ρ is a matrix of correlations among the four covariates.

Randomization

The simulated pre-study covariate observations were randomly allocated to two treatment groups, A and B, in the ratio 1:1 using blocks of size 10.

Simulation of a binary outcome variable

After randomization of the simulated observations, the outcome measure (1=success, 0=failure) was generated for each of the two groups to achieve the required efficacy (treatment success) rates using a Bernoulli (π_i) distribution, where π_i is the mean proportion of subjects with treatment success (efficacy) in a group i ($i=A,B$). This resulted in a simulated binary outcome variable with group success (efficacy) and failure rates of π_i and $1-\pi_i$ respectively. The simulations were constrained to ensure that the specified rates were achieved exactly.

The risk difference model can be fitted in the Generalized Linear Model (GLM) framework. A GLM is constructed using three statistical functions: the linear function, the link function and the variance function [12,13]. The linear function may be parameterized as follows:

$$\phi_i = \beta_0 + \beta_1 x_{1i} + \dots + \beta_p x_{pi}$$

where x_{pi} is the p^{th} covariate or factor for individual i and $(\beta_0, \beta_1, \dots, \beta_p)$ is a vector of regression coefficients.

The link function is specified as: $g(E(Y_i)) = \phi_i$, that is, the link function describes how the expected value of the outcome variable $E(Y_i)$ depends on the linear predictor ϕ_i . Note that $g(E(Y_i))$ can be specified either as the identity function (no transformation on the $E(Y_i)$), or as a logit, log, log-binomial etc. along with a relevant family such as Poisson, Binomial, Gaussian and Gamma.

Finally, the variance function is given by $\text{var}(Y_i) = \phi^V(E(Y_i))$, where ϕ is a constant, commonly referred to as a dispersion parameter.

The outcome variable Y_i follows a binomial distribution denoted by $Y_i \sim \text{Binomial}(n_i, \pi_i)$, where n_i and π_i are the number of participants and treatment efficacy rate respectively in group i .

The main interest is to model the probability of a treatment success (efficacy) Y_i/n_i in the form of a risk difference model, where:

$$E(Y_i/n_i) = \pi_i$$

$$\text{var}(Y_i/n_i) = \frac{1}{n_i} \text{var}(Y_i)$$

$$= \frac{1}{n_i} \delta_i (1 - \delta_i)$$

$$= \frac{1}{n_i} \delta_i (1 - \delta_i)$$

So the variance function $V(\delta_i) = \frac{1}{n_i}$ which is a function of $E(Y_i/n_i)$, and the dispersion parameter is $\phi = 1/n_i$.

Estimates of the risk differences are obtained by specifying an identity link function to the binomial regression model. The coefficients in the Binomial regression model with identity link produce risk differences as follows:

Consider what happens when a variable X is increased by 1 unit.

$P(Y=1|X) = X\beta$, where $Y=1$ denote treatment success and $Y=0$ denote treatment failure.

$$P(Y=1|X+1) = (X+1)\beta$$

X can be thought of as a reference group and $X+1$ as a treatment group of interest. Taking the difference gives:

$P(Y=1|X+1) - P(Y=1|X) = (X+1)\beta - X\beta = \beta$, which is a risk difference.

In the multivariable model $P(Y=1|X_1, X_2, \dots, X_p) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_p X_{pi}$, the regression coefficients β_{pi} represent the risk difference resulting

Table 1: Convergence rates by efficacy rate and number of covariates in model (averaged over 5000 simulated datasets-correlated).

Covariates		Efficacy Rates			
Number	Names	60% vs. 70%	60% vs. 80%	60% vs. 85%	60% vs. 90%
1	age	99.7	97.9	94.6	85.5
1	hb	99.9	98.7	94.7	81.3

from a change (increase) of one unit in the explanatory variable X_{pi} . When using the identity link, fitted probabilities outside the probability range (0,1) are possible. Wacholder [11,13] suggested that when fitting a binomial regression model with identity link, the fitted probabilities should be checked for range conditions at each iteration. If found to be outside the permitted range, then they should be reset it to be within that range.

Risk difference models considered in this study included subsets of between one and three of the continuous covariates age, wt, para or hb, and a binary factor taking values of 1/0 indicating treatment given/not given.

Correlations and number of covariates included in modelling

In the first set of simulations, the contributions to binomial regression model convergence rates due to the (i) the proximity of efficacy rates to boundary values and (ii) the number of covariates were assessed using correlation levels between the covariates reflecting those observed in the sample dataset. The efficacy rate in one group was allowed to approach a boundary (100% in this case) while that in the other group was held constant. For example, in one group efficacy was set at 60% and in the other group it was varied to take values 70%, 80%, 85% or 90%, and convergence rates were estimated in each scenario.

To assess the impact of level correlation between covariates on the ability of models to converge, the pair wise correlations between the covariates were reduced to zero and the same models fitted while monitoring convergence. In these simulations, for each pre-set risk difference, the number of covariates and the correlation levels were varied.

Statistical software

All simulations were performed using Stata 12.0, IC, StataCorp, 4905 Lakeway Drive, College Station, Texas 77845 USA. Graphs were constructed using Microsoft Excel 2007.

Simulation Results

Effect of proximity to boundary of efficacy levels and number of covariates on convergence of a binomial regression model

When using just a single predictor variable (covariate) and with the efficacy levels set at 60% and 70% for the two groups respectively (i.e. distant from the boundary of the parameter space), the percentage of datasets that converged when a binomial regression model was fitted was high (96.8% to 99.7%) (Table 1); however, while the corresponding non-convergence rates were very low, they were not negligibly so.

1	para	96.8	98.6	92	86.6
1	wt	99.9	98.1	94.2	82.1
2	hb, age	99.4	94.7	84.5	61.4
2	age, wt	99.3	94.1	84.7	61
2	age, para	96	91.1	82.6	66.1
2	hb, wt	99.4	95.1	84.7	59.8
2	hb, para	96.1	91.4	83.3	64.9
2	wt, para	96.4	92	83	64.6
3	age, hb, para	95.7	87.9	73.9	48.1
3	age, hb, wt	98.8	89.7	73.4	42.9
3	hb, wt, para	96.3	88.1	74.2	46.4
3	age, wt, para	95.9	87.6	74.5	47.7

Convergence rates remained high, ranging from 92.0% to 94.7%, even when the larger of the two efficacy levels were increased to 85%. However, when the larger efficacy level was increased to 90% (still some distance from the boundary), convergence levels dropped dramatically between 81.3% and 86.6%.

Adding additional covariates exacerbated this problem. With the larger of the two efficacy levels set at 90%, convergence was obtained in 59.8% to 66.1% of models involving two covariates, and in only 42.9% to 48.1% of models involving three covariates. The full extent of this trend is shown graphically in Figure 1.

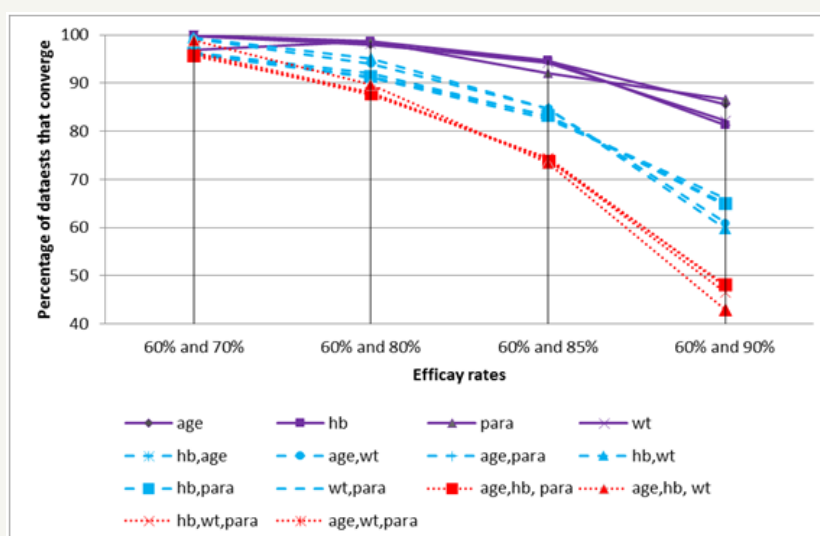


Figure 1: Percentage convergence by efficacy rates and correlation structure.

The effect of correlations between covariates on model convergence

The influence of correlation levels on model convergence rates was examined for efficacy rates of 60% and 85% in the two groups

respectively (Table 2 and Figure 2). In these simulations, the impact of setting the correlations between each pair of covariates to zero was compared to that obtained using the correlation levels from the original RCT.

Table 2: Convergence rates in the presence and absence of correlation between covariates (averaged over 5000 simulated datasets): efficacy rates 60% and 85%.

Number of Covariates	Covariates	Original Correlations	Zero Correlations
1	age	94.6	97.2
1	hb	94.7	96.5
1	para	92	92.1
1	wt	94.2	96.3
2	hb, age	84.5	91
2	age, wt	84.7	91.1
2	age, para	82.6	85.7
2	hb, wt	84.7	91.1

2	hb, para	83.3	86.1
2	wt, para	83	85.8
3	age, hb, para	73.9	80
3	age, hb, wt	73.4	84.1
3	hb, wt, para	74.2	80.2
3	age, wt, para	74.5	80.4

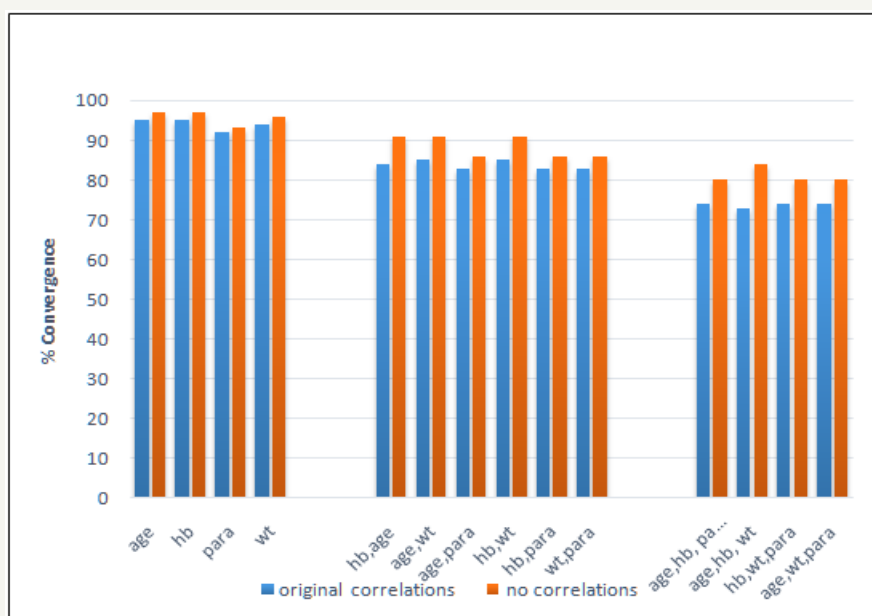


Figure 2: Effect of reducing correlation on convergence (efficacy rates: 60% and 85%).

For all models considered, the convergence percentage improved when the correlations between covariates were removed. The improvement in convergence rates increased as the number of covariates in the model was raised.

Discussion and Conclusion

For many researchers, risk difference is increasingly the preferred measure of treatment effect in RCTs and in epidemiological cohort studies where the primary outcome measure is dichotomous. Estimating risk differences adjusted for important covariates can be problematic, however, due to the problem of the binomial regression model failing to converge. This simulation study found that non-convergence rates tend to rise as the number of continuous covariates included in the model increases, as correlation levels between covariates increase, and (most severely) when efficacy rates in either treatment arms is close to a boundary (i.e. to 0% or 100%).

When using maximum likelihood estimation, as is frequently the case with binomial regression models, software algorithms may fail to converge either because of poor starting values for the iterative computing process (e.g. starting values not in the restricted parameter space) or because the maximum likelihood solution occurs on the boundary of the parameter space [12]. The first of these causes can be fixed easily by specifying better starting values. In the latter situation, however, the derivative of the likelihood at its maximum may not be 0 (in which case standard software packages

that maximize the likelihood by finding the point at which the derivative is equal to 0 (namely Newton's method), may be unable to find a solution) or could be a result of having a flat likelihood distribution.

In summary, binomial model non-convergence rates were found to increase as efficacy rates in one or both treatment groups moved towards a boundary value irrespective of the number of covariates included in the model. Indeed, it is interesting to note that the problem worsened as the number of covariates in the model increased and as the efficacy rate in either treatment group moved towards a boundary value. That is, models with just one covariate had less convergence problems than models with two covariates, and in turn these had fewer convergence problems than models with three covariates.

For all scenarios examined, convergence was poor when the efficacy rate in either group was 90% or greater. Additionally, non-convergence rates increased with increasing levels of correlation between any covariates included in the model.

Acknowledgement

This work was supported by the European and Developing Countries Clinical Trials Partnership (EDCTP) grant number IP2007.31060.03; and the Johns Hopkins University Center for AIDS Research (Grant Number 1P30AI094189) from the National Institute of Allergy And Infectious Diseases.

References

1. Magder LS (2003) Simple approaches to assess the possible impact of missing outcome information on estimates of risk ratios, odds ratios, and risk differences. *Control Clin Trials* 24(4): 411-421.
2. Bell DJ, Nyirongo SK, Mukaka M, Zijlstra EE, Plowe CV, et al. (2008) Sulfadoxine-pyrimethamine-based combinations for malaria: a randomised blinded trial to compare efficacy, safety and selection of resistance in Malawi. *PLoS One* 3(2): 1578.
3. Arinaitwe E, Sandison TG, Wanzira H, Kakuru A, Homsy J, et al. (2009) Artemether-lumefantrine versus dihydroartemisinin-piperaquine for falciparum malaria: a longitudinal, randomized trial in young ugandan children. *Clin Infect Dis* 49(11): 1629-1637.
4. Mukaka M, White SA, Mwapasa V, Kalilani PL, Terlouw DJ, et al. (2016) Model choices to obtain adjusted risk difference estimates from a binomial regression model with convergence problems: An assessment of methods of adjusted risk difference estimation. *Journal of Medical Statistics and Informatics* 4: 5.
5. Mukaka M, White SA, Terlouw DJ, Mwapasa V, Kalilani PL, et al. (2016) Is using multiple imputation better than complete case analysis for estimating a prevalence (risk) difference in randomized controlled trials when binary outcome observations are missing? *Trials* 17: 341.
6. Case LD, Kimmick G, Paskett ED, Lohman K, Tucker R (2002) Interpreting measures of treatment effect in cancer clinical trials. *The Oncologist* 7(3): 181-187.
7. Davies HT, Crombie IK, Tavakoli M (1998) When can odds ratios mislead? *BMJ* 316(7136): 989-991.
8. Page J, Attia J (2003) Using Bayes' nomogram to help interpret odds ratios. *ACP J Club* 139(2): A11-A12.
9. McNutt LA, Wu C, Xue X, Hafner JP (2003) Estimating the relative risk in cohort studies and clinical trials of common outcomes. *Am J Epidemiol* 157(10): 940-943.
10. Cheung YB (2007) A modified least-squares regression approach to the estimation of risk difference. *Am J Epidemiol* 166(11): 1337-1344.
11. Walter SD (2000) Choice of effect measure for epidemiological data. *J Clin Epidemiol* 53(9): 931-939.
12. Deddens JA, Petersen MR (2008) Approaches for estimating prevalence ratios. *Occup Environ Med* 65(7): 481-506.
13. Wacholder S (1986) Binomial regression in GLIM: estimating risk ratios and risk differences. *Am J Epidemiol* 123(1): 174-184.