

Statistical Analyses of Mutually Exclusive Competing Risks in Neonatal Studies

C. Engel^{1,*} and A.R. Franz²

¹Center for Pediatric Clinical Studies, Biometry University Children's Hospital Tuebingen, Frondsbergstraße 23, 72070 Tübingen, Germany

²Center for Pediatric Clinical Studies, Methods University Children's Hospital Tuebingen, Frondsbergstraße 23, 72070 Tübingen, Germany

Abstract: Following the well-established approach on how to deal with competing risks in the situation of time-to-event endpoints, cumulative incidences have to be used to analyse each single category of an outcome in the situation of competing risks without a time-to-event structure as well. This can be easily done by applying a simple chi-square test.

Nevertheless, these categorial outcomes are usually combined to get a composed dichotomous outcome to face the problem on how to deal with a significant chi-square omnibus test in the situation of more than 1 df, i.e. > 2x2 tables.

The aim of this report is to question the practice of combined, i.e. composed dichotomized, endpoints because important information is lost and the real effect of interest in confirmatory phase III studies may only become apparent in explorative secondary analyses.

It is shown – by using hypothetical data and by recalculation of published phase III studies' results – how the use of a chi-square omnibus test and the scarcely known post-hoc testing answers the real question of interest within one primary confirmatory analysis. This method reveals insight into the actual effect of a new treatment or therapy on the event of interest in the presence of a mutually exclusive competing risk.

Keywords: Competing risk, randomized controlled trial, composite outcome, chi-square test, post-hoc testing.

INTRODUCTION

Competing risks reveal a special problem in many clinical studies. They may preclude or at least alter the probability of the event of interest. For the past twenty years or so, the problem of competing risks has been solved concerning time-to-event endpoints, e.g., in the context of a cancer disease, where time to relapse is the event of interest and the competing risk is death without relapse. This problem has been described and it was shown how to use cumulative incidence instead of Kaplan-Meier-estimates (Gooley [1], Cheng *et al.* [2]). Since then, different methods have been explained and demonstrated in applied and non-technical publications, e.g., by Dignam *et al.* [3], Chappel [4] or Schmoor *et al.* [5]. They are well-established in analysing data of clinical studies dealing with time-to-event data in the context of competing risks (e.g., Spratt *et al.* [6], Wulaningsih *et al.* [7], Turin *et al.* [8]).

New treatments to be evaluated in clinical studies in populations with a high mortality, e.g., very preterm infants, typically have the potential to prevent or improve a single serious condition such as bronchopulmonary dysplasia (BPD), necrotizing

enterocolitis (NEC), retinopathy of prematurity (ROP) or severe long-term neurodevelopmental impairment (NDI). Patients at high risk of such serious conditions, who are typically selected as study population for this type of interventional studies, are also at high risk to die before the diagnosis of the condition of interest can be made. Consequently, death is a competing risk to the real condition of interest.

To meet the concerns, that improvements or prevention of the condition of interest would possibly occur at the expense of an increased death rate, dichotomized composite outcomes are frequently used as primary outcome measures in such studies (e.g. Bassler *et al.* [9], Support Study Group [10, 11], Shankaran *et al.* [12], Azzopardi *et al.* [13]). The primary endpoint is then defined as “death or the occurrence of the condition of interest” versus “alive without occurrence of the condition of interest”. Each component of the primary endpoint (death versus alive and condition of interest versus alive) is then usually also analysed as secondary outcome, hence the actual condition of interest is only evaluated in the subgroup of surviving patients as conditional probability and assessed as secondary outcome only.

Additionally, analysis of composite endpoints for confirmatory analysis may fail to proof a given difference in the outcome of interest if there are trends

*Address correspondence to this author at the Center for Pediatric Clinical Studies, Biometry University Children's Hospital Tuebingen, Frondsbergstraße 23, 72070 Tübingen, Germany;
E-mail: corinna.engel@med.uni-tuebingen.de

in the opposite direction for the competing outcome. In this case, an existing treatment effect in a confirmative phase III study may only become visible in the secondary outcome analyses of the single components of the composed endpoint, which are no longer confirmative.

Finally, if the frequently used approach for analysis in interventional studies is not the most powerful one, more infants than necessary are subjected to the burden of clinical trials and potentially beneficial therapies were/are rejected by mistake because of pretended insufficient benefit.

In neonatal studies, outcomes are frequently evaluated when the preterm infant has reached a certain developmental stage, i.e. a postmenstrual age of 36 weeks, or term equivalent age.

In this case, outcomes are categorical outcomes without a time-to-event structure. Although these outcomes may also be treated as time-to-event outcomes – with the same time span for each event occurrence – there are still dichotomized composite outcomes formed for primary analysis and conditional probabilities in secondary analyses used. The method of choice often is chi-square testing and calculation of risk ratios for the dichotomous composite primary outcome and, as secondary analyses, for the dichotomous single components, each analysis leading into a 2x2 table.

The aim of this report is to question this approach. Instead of calculating a condition's incidence as conditional probability in surviving infants, we propose that the more relevant outcome to be reported is the probability for an enrolled patient to either die, acquire the condition of interest or survive healthy based on the denominator of all patients included in a study. Consequently, statistical analyses should not only compare the conditional probability of the condition of interest in surviving infants between treatment groups in secondary analyses, but analyse the proportions of all possible categories (i.e. the event of interest and the competing risk and healthy survival) based on all infants included in the study simultaneously, to inform the clinician whether a new therapy or treatment may improve or worsen any of these ratios.

Besides using the complex SAS-macros [14] published to deal with competing risks in time-to-event outcomes, we propose an easier way for analysis of categorical, mutually exclusive competing outcomes, for

which time-to-event is not relevant: the calculation of cumulative incidences using well-known chi-square-testing with more than one degree of freedom (i.e., based on a contingency table with more than 2x2 fields), and its scarcely known post-hoc testing in case of a significant omnibus test.

This simple method gives insight into the real effect of a new treatment or therapy on each of the different components of a categorical outcome. It is limited to analyses without confounders or covariates as is the situation in confirmative randomized studies of phase III, for example.

MATERIAL AND METHODS

Description of the Conventional Approach for Handling Competing but Mutually Exclusive Outcomes without Time-to-Event Structure

Outcomes with more than two categories (i.e. event of interest, death or no event) are dichotomized to an outcome with two categories: event of interest or death versus no event. This is measured at a certain time point, i.e. in neonates at a postmenstrual age of 36 weeks.

In further analyses, the single components “event of interest versus no event” and “death versus no event” are analysed, whereas event of interest is only analysed as a conditional probability based on the patients that did not die.

To demonstrate the conventional approach unstratified Cochrane-Mantel-Haenzel-Test [15], was used.

Examples for this approach are given in Table 1 for hypothetical data and in Table 6 for five studies published in the New England Journal of Medicine within the last years.

Description of the Scientific Background of New Approach for Analysis of Competing but Mutually Exclusive Outcomes without Time-to-Event Structure

Chi-square testing with more than one degree of freedom (i.e., based on a contingency table with more than 2x2 fields) faces the problem of how to interpret a p-value below 0.05. Sharpe [16] and Franke *et al.* [17] addressed this problem by two different approaches.

Sharpe [16] describes adjusted standardized residuals as one of several possibilities how to decide

in a chi-square test with more than one degree of freedom and an omnibus test presenting a statistically significant result, which cells contributed mostly to this result.

He calculates adjusted standardized residuals for each cell by

$$Adj\ Residual = \frac{(O - E)}{\sqrt{E * \left(1 - \frac{RowMarginal}{n}\right) * \left(1 - \frac{ColMarginal}{n}\right)}}$$

with

- O = observed frequency
- E = expected frequency
- RowMarginal = row marginal for the cell
- ColMarginal = column marginal for the cell
- n = total number of cases across all cells

Agresti [18] described that an adjusted residual of more than +/- 2 if there are 'few' cells and of more than +/- 3 if there are 'many' cells, indicates a lack of fit of H₀ in that cell.

This is consistent with MacDonald and Gardner [19], who suggested a Bonferroni adjustment for the adjusted residual. According to them, an adjustment of the z critical of 1.96 should be done if the number of cells in the contingency table is large. For example a 3x5 contingency table consists of 15 cells. Therefore, alpha is adjusted to 0.05/15 = 0.003, which corresponds to a critical value of +/- 2.96 and is consistent with the above mentioned pragmatic proposal of Agresti [18].

Franke *et al.* [17] proposed to use Scheffé post hoc values for pairwise comparisons because they present the most conservative approach:

$$\frac{\hat{\Psi}}{\sqrt{SE_{\hat{\Psi}}^2}} = \frac{w_1(p_1) - w_2(p_2)}{\sqrt{w_1^2 \left(\frac{p_1 * (1 - p_1)}{n_1}\right) + w_2^2 \left(\frac{p_2 * (1 - p_2)}{n_2}\right)}} = Z$$

With $\hat{\Psi}$ representing the linear combination of weights (W_k) and proportions (θ_k) of the contrast of interest:

$$\psi = W_1\theta_1 + W_2\theta_2 + \dots + W_k\theta_k$$

and

$$W_1 + W_2 + \dots + W_k = 0$$

whereas the numerator is defined as the square root of the weighted standard error of the contrast:

$$SE_{\theta}^2 = W_1^2 SE_{\theta_1}^2 + W_2^2 SE_{\theta_2}^2 + \dots + W_k^2 SE_{\theta_k}^2$$

$$= W_1^2 \left(\frac{p_1 q_1}{N_1}\right) + W_2^2 \left(\frac{p_2 q_2}{N_2}\right) + \dots + W_k^2 \left(\frac{p_k q_k}{N_k}\right)$$

The critical value for each contrast is given by

$$S^* = \sqrt{\chi_{v,1-\alpha}^2}$$

with

$\chi_{v,1-\alpha}^2$ = critical value for given degrees of freedom v and given α

Assessment of Relevance and Consequences Based on Previously Published Large Interventional Trials in the Very Preterm Population

A) A large study including about 2000 very preterm infants [20] found a mortality of 15% and a rate of events of interest of 35% in the placebo group. Baseline frequencies of death and a condition of interest in the placebo group were selected from this large interventional study in the population of extremely preterm infants resulting in 2x3 tables. Based on hypothetical treatment effects resulting in 5% (scenario 1), 6% (scenario 2), 7% (scenario 3), and 8% (scenario 4) absolute risk reduction for this event of interest, and a hypothetical sample size of n=1000, alpha-values were calculated to assess the sensitivity to confirm a given treatment effect by either the conventional approach (i.e. analysing as a dichotomized composed outcome and later on analyses of the single components, resulting in 2x2 tables each) or the new approach (inserting all components of the categorical outcome into one analysis of a 2x3 table).

The same analysis is then repeated for scenario 4 in a stratified fashion taking into account that treatment effects may be limited to subgroups.

Finally, the analysis is recalculated for a potential third treatment arm, resulting in a 3x3 table.

B) Five published large confirmatory phase III studies using dichotomized composite outcomes in the presence of competing risks in a situation without time-to-event outcomes were chosen for recalculation.

The results of these studies assessing new therapies for preterm neonates with high risk for the

concurrent but mutually exclusive risk death were recalculated using the methods described above for 2x3 tables. To facilitate the comparison of results, the data were also recalculated by Cochran-Mantel-Haenszel statistics in their original way using dichotomized composed endpoints and single components of these endpoints, resulting in 2x2 tables each. Nevertheless, original results in case of different strategies used by the authors (i.e. stratification, multivariate analyses, etc.) are also displayed.

RESULTS

Using basic proportions of 15% deaths and 35% events of interest, hypothetical frequencies for the 4

treatment effect scenarios and results for statistical analyses using dichotomized composed primary endpoint (2x2 tables) and single components (2x2 tables) as secondary endpoints versus chi-square testing with adjusted residuals and post-hoc testing (2x3 tables) are presented in Table 1: Analysing single components as secondary endpoints results in the above mentioned problem of conditional probabilities. Therefore, in scenario 2 with a population difference of 6% for the rate of the event of interest between new treatment A and treatment B (control), analysis of a secondary outcome “rate of event of interest in survivors” will actually assess a difference of 7.1%, due to assessment of conditional probabilities. Consequently, the rate of p-values below 0.05 will be

Table 1: Hypothetical Scenarios Based on Event Rates from Study Data

Results using combined endpoints (conventional approach)					
Scenario		1	2	3	4
Absolute Risk Reduction for Event of Interest		5%	6%	7%	8%
Typical Composite Primary Endpoint	Treatment A	225/500 (45%)	220/500 (44%)	215/500 (43%)	210/500 (42%)
	Treatment B	250/500 (50%)	250/500 (50%)	250/500 (50%)	250/500 (50%)
	p-value	0.114	0.058	0.027	0.011
Event of interest as conditional proportion as secondary endpoint	Treatment A	150/425 (35.3%)	145/425 (34.1%)	140/425 (32.9%)	135/425 (31.8%)
	Treatment B	175/425 (41.2%)	175/425 (41.2%)	175/425 (41.2%)	175/425 (41.2%)
	p-value	0.078	0.034	0.013	0.004
Mutually exclusive competing risk as secondary endpoint	Treatment A	75/500 (15%)	75/500 (15%)	75/500 (15%)	75/500 (15%)
	Treatment B	75/500 (15%)	75/500 (15%)	75/500 (15%)	75/500 (15%)
	p-value	1.0	1.0	1.0	1.0
Results considering for concurrent risk in definition of endpoint (new approach)					
Healthy	Treatment A	275/500 (55%)	280/500 (56%)	285/500 (57%)	290/500 (58%)
	Treatment B	250/500 (50%)	250/500 (50%)	250/500 (50%)	250/500 (50%)
	Adj. residual	/ *	/ *	+/- 2.22	+/- 2.54
	Scheffé			2.22	2.55
Event of interest	Treatment A	150/500 (30%)	145/500 (29%)	140/500 (28%)	135/500 (27%)
	Treatment B	175/500 (35%)	175/500 (35%)	175/500 (35%)	175/500 (35%)
	Adj. residual	/ *	/ *	+/- 2.38	+/- 2.73
	Scheffé			-2.39	-2.75
Mutually exclusive competing risk	Treatment A	75/500 (15%)	75/500 (15%)	75/500 (15%)	75/500 (15%)
	Treatment B	75/500 (15%)	75/500 (15%)	75/500 (15%)	75/500 (15%)
	Adj. residual	/ *	/ *	+/- 0.0	+/- 0.0
	Scheffé			0.0	0.0
Overall p-value		0.211	0.105	0.046	0.017

Treatment A = new hypothetical treatment with increasing treatment effect on event of interest from scenario 1 through scenario 4.
 Treatment B = hypothetical baseline / control / placebo with constant rate of event of interest and death.
 Typical composite primary endpoint = event of interest or death (as mutually exclusive competing risks).
 *not appropriate due to omnibus test resulting in p>0.05.
 Scheffé critical value for a table with two degrees of freedom is +/-2.45.

inappropriately high. In contrast, an analysis based on chi-square testing with more than 1 degree of freedom will be more conservative. This kind of analysis only gives a p-value below 0.05 in the third scenario where the difference in the event of interest is actually 7%.

Chi-Square omnibus testing followed by the calculation of adjusted residuals or Scheffé post-hoc critical values may also be applied to stratified analyses. Assuming that the scenarios in Table 1 were based on a population of preterm neonates stratified at randomisation by gestational age at birth between 26 and less than 28 (stratum 1) versus 24 and less than 26 weeks (stratum 2), respectively. Let us further suppose an analysis stratified for gestational age at birth and therapy for scenario four, which would reveal a much greater effect in the higher gestational age group, perhaps because neonates with lower gestational age suffer from too many comorbidities to profit from the new therapy. This results in the 4x3 table presented below (Table 2).

Chi-square tests controlled for gestational age group result in p-values of 0.009 for stratum 1 and 0.6 for stratum 2. These results are supported by the Scheffé-values for each contrast of interest (Table 3). To get confirmative results, these analyses may be

done in a hierarchical manner by first calculating the overall p-value and interpreting the results of the analyses controlled for gestational age according to the overall result.

Let us now regard studies with more than two therapies. Therefore the hypothetical study is rerun with a third study arm, for example with an observational arm to elucidate potential placebo effects. Due to the fact that now a 3x3-table is analysed, sample size has to be higher. Therefore, 750 hypothetical patients are included in each therapy group. The overall chi-square test reveals a p-value of 0.0004. As may be easily seen by the results for the different Scheffé-contrasts of interest, this p-value is caused by differences between treatment A and treatment B as well as between treatment A and treatment C concerning healthy outcome and event of interest. There is a small hypothetical difference between treatment B and treatment C which might be a small effect or a chance finding. But this small difference fails to reach significance (Tables 4 and 5).

Applying chi-square testing with adjusted residuals and post-hoc testing, we re-calculated results of interventional clinical studies in populations of preterm infants (Table 6). For clarity, calculations did not take

Table 2: Stratified Analysis for an Overall Absolute Risk Reduction for Treatment A of 8% (Scenario 4)

		Outcome			all
		Healthy	Event of interest	Mutually exclusive competing risk	
		N (%)	N (%)	N (%)	N (%)
Stratum	Therapy				
Stratum 1	Treatment A	170 (68%)	55 (22%)	25 (10%)	250 (100%)
	Treatment B	140 (56%)	85 (34%)	25 (10%)	250 (100%)
	All	310 (62%)	140 (28%)	50 (10%)	500 (100%)
Stratum 2	Treatment A	120 (48%)	80 (32%)	50 (20%)	250 (100%)
	Treatment B	110 (44%)	90 (36%)	50 (20%)	250 (100%)
	All	230 (46%)	170 (34%)	100 (20%)	500 (100%)

Overall p-value: p<0.0001.
 Chi-square-test stratum 1: p=0.009.
 Chi-square-test stratum 2: p=0.60.

Table 3: Contrasts for Stratified Analysis

Contrast	Scheffé-value	
	High GA	Low GA
Treatment A vs. Treatment B, healthy	2.79 ***	0.90
Treatment A vs. Treatment B, event of interest	-3.02 ***	-0.94
Treatment A vs. Treatment B, competing risk	0.0	0.0

***exceeds Scheffé-critical-value of 2.45.

Table 4: Analysis with Three Therapy Groups

	Outcome			all
	Healthy	Event of interest	Mutually exclusive competing risk	
	N (%)	N (%)	N (%)	N (%)
Therapy				
Treatment A	435 (58%)	203 (27.1%)	112 (14.9%)	750 (100%)
Treatment B	375 (50%)	263 (35.1%)	112 (14.9%)	750 (100%)
Treatment C	360 (48%)	278 (37.1%)	112 (14.9%)	750 (100%)
all	1170 (52%)	744 (33 %)	336 (14.9%)	2250 (100%)

Chi-square-test $p=0.0004$.

into account stratification, but original results were depicted with stratification for the sake of completeness where appropriate.

Table 5: Contrasts for Analysis with Three Therapy Groups

Contrast	Scheffé-value
Treatment A vs. Treatment B/Healthy	3.11841 ***
Treatment A vs. Treatment C/Healthy	3.89959 ***
Treatment B vs. Treatment C/Healthy	0.77491
Treatment A vs. Treatment B/Event	-3.36225 ***
Treatment A vs. Treatment C/Event	-4.17537 ***
Treatment B vs. Treatment C/Event	-0.80705
Treatment A vs. Treatment B/Risk	0.00000
Treatment A vs. Treatment C/Risk	0.00000
Treatment B vs. Treatment B/Risk	0.00000

***exceeds Scheffé-critical-value of 3.08 for 3x3-tables.

The proposed analyses offered the chance to proof differences in the effect of a new therapy on the event of interest in studies that failed to proof this for the combined primary endpoint used. Concerning the studies of the Neurosis Study group [9] and the NICHD Neonatal Research Network Support Study Group [11] the published analyses of the study data using dichotomized composed primary outcome failed to proof a difference for this endpoint (0.053 and 0.21, respectively) and found remarkable p-values only in the analysis of the single components as secondary outcomes (event of interest 0.004 and <0.0001, respectively). Analysing all components of the primary outcome in a 2x3 table by applying chi-square omnibus test and post-hoc testing would have revealed p-values of 0.006 and <0.0001, respectively, already in the confirmatory primary analyses (Table 6, columns "Neurosis Study group" and "NICHD Neonatal Research Network Support Study Group", respectively).

Due to the fact that different methods with different underlying algorithms are used, p-values differ slightly between different methods.

This may be critical in results close to p-values of 5%, as in the study of the NICHD Neonatal Research Network Childhood outcomes after hypothermia Study Group [12]. It may be useful to calculate Scheffé-values nevertheless, as in this example they reveal the difference being based on more deaths in one group, not in more or less successful treatment concerning IQ scores (Table 6, column "NICHD Neonatal Research Network Childhood outcomes after hypothermia Study Group"). The recalculation of the results of the Study of the Toby Study Group [13] demonstrates how the use of conditional probabilities regarding the single components may alter the overall result (Table 6, column "Toby Study Group").

Table 6 reveals only slight differences between the results using Cochrane-Mantel-Haenszel statistics and the original study results based on stratified data. These slight differences do not change the overall interpretation of the original study results. Therefore, the recalculation of the results using Cochrane-Mantel-Haenszel statistics to facilitate comparisons is justified.

DISCUSSION

We demonstrated how to use cumulative incidences in the situation of dichotomous outcomes of interest and the presence of mutually exclusive competing risks by using the well-known chi-square test for > 1df and its scarcely known post-hoc testing. Even if this seems to be a quite simple approach, it offers many advantages. In the situation of the birth of a very preterm baby, parents and caretakers want to know the probability for this baby either to die, to acquire a condition of concern (e.g., to suffer from chronic lung

Table 6: Recalculation of Study Results with Chi-Square Testing and Adjusted Residuals and Scheffé Post-Hoc Test, Respectively

Description of the studies	Publication		Neurosis Study Group [9]	NICHD Neonatal Research Network SUPPORT Study Group [10]	NICHD Neonatal Research Network SUPPORT Study Group [11]	NICHD Neonatal Research Network Childhood outcomes after hypothermia Study Group [12]	Toby Study Group[13]**
	Event of interest		BPD	BPD	ROP	IQ score < 70	IQ score ≤ 85
	Concurrent risk		Death	Death	Death	Death	Death
	Definition of PE		BPD or Death at 36 weeks GA	BPD or Death	ROP or Death	IQ score < 70 or Death	IQ score ≤ 85 or Death
	Treatment A		Budesonide (n=437)	CPAP (n=663)	Lower O ₂ -Sat. (n=605)	Hypothermia (n=97)	Hypothermia (n=145)
	Treatment B		Placebo (n=419)	Intubation + surfactant (n=653)	Higher O ₂ -Sat. (n=616)	Control, i.e. usual care (n=93)	Control, i.e. standard care (n=132)
Conventional Approach	Composite Primary Endpoint	Treatment A	175 (40%)	317 (47.8%)	171 (28.3)	46 (47%)	70 (48%)
		Treatment B	194 (46.3%)	333 (51%)	198 (32.1)	58 (62%)	80 (61%)
		p-value	0.065	0.249	0.14	0.04	0.04
	Event of interest	Treatment A	101/363 (27.8%)	223/569 (39.2%)	41/474 (8.6%)	19/70 (27%)	23/98 (24%)
		Treatment B	137/362 (37.8%)	219/539 (40.6%)	91/509 (17.9%)	17/52 (32%)	31/83 (37%)
		p-value	0.004	0.625	<0.0001	0.51	0.04
	Concurrent risk	Treatment A	74 (16.9%)	94 (14.2%)	130 (21.5)	27 (28%)	47 (32%)
		Treatment B	57 (13.6%)	114 (17.5%)	107 (17.4)	41 (44%)	49 (37%)
		p-value	0.176	0.103	0.069	0.02	0.41
New Approach	Healthy	Treatment A	262 (60%)	346 (52.2%)	434 (71.7%)	51 (53%)	75 (52%)
		Treatment B	225 (53.7%)	320 (49%)	418 (67.9)	35 (42%)	52 (39%)
		Adj. residual	+/- 1.85	/*	+/- 1.48	+/- 2.03	/*
		Scheffé	1.85	/*	1.48	2.09	/*
	Event of interest	Treatment A	101 (23.1%)	223 (33.6%)	41 (6.8)	19 (20%)	23 (16%)
		Treatment B	137 (32.7%)	219 (33.5%)	91 (14.8)	17 (18%)	31 (23%)
		Adj. residual	+/- 3.13	/*	+/- 4.5	+/- 0.23	/*
		Scheffé	-3.14***	/*	-4.55***	0.23	/*
	Concurrent risk	Treatment A	74 (16.9%)	94 (14.2%)	130 (21.5)	27 (28%)	47 (32%)
		Treatment B	57 (13.6%)	114 (17.5%)	107 (17.4)	41 (44%)	49 (37%)
		Adj. residual	+/- 1.35	/*	+/- 1.82	+/- 2.34	/*
		Scheffé	1.60	/*	2.26	-2.57***	/*
Overall p-value			0.0064	0.235	<0.0001	0.053	0.09
Published results in original publication							
Stratified for		GA	GA, Center, familial clustering	GA, Center, familial clustering	Center	/	
PE	p-value	0.053	0.3	0.21	0.06	0.04	
Event of interest	p-value	0.004	0.92	<0.001	0.51	0.05	
Concurrent risk	p-value	0.17	0.09	0.04	0.04	0.81	

*not appropriate due to omnibus test resulting in p>0.05.

**original endpoint has been "survivors with IQ score ≥ 85". Nominators in original publication are not consistent.

***exceeds Scheffé-critical-value of 2.45.

disease) or to be alive and healthy. The common approach to analyse combined endpoints (e.g., death or BPD) and perform secondary analysis of conditional probabilities for the events of interest does not provide the correct answer to this question. Additionally, this approach includes two methodical problems:

The combined endpoint may not lead to a significant result due to changes in the opposite direction for the event of interest and the competing risk. Even small deviations from the expected ratio within accepted random ranges for the competing risk may mask the effect of a new therapy on the event of interest. Because the single event of interest is then only analysed as a secondary outcome, the result of this analysis cannot be considered confirmatory. Additionally, the analysis of the rate of the event of interest is then based on the conditional probability (i.e., the rate of this event in survivors) which leads to inflation of p-values < 0.05 due to the smaller denominator for the calculation of the ratios.

Regarding the sample size needed to proof a given difference by combined endpoints or cumulative incidences of each single endpoint, there may even be a benefit in using cumulative incidences in slight smaller sample sizes needed (data not shown).

In summary, we propose that the analysis of cumulative incidences ought to be the method of choice for outcome assessment of dichotomous outcomes with mutually exclusive competing risk(s).

To get an effect size which is more informative than just event ratios within the n*m-tables of the chi-square testing, risk ratios may be easily calculated as usual by

breaking down the n*m-table (as described in Table 7a) into several 2x2-tables (refer to Table 7b as an example):

Risk Ratios for 'healthy versus event of interest and death' and 'death versus event of interest and healthy' may be calculated respectively.

The proposed method based on cumulative incidences by using well-known chi-square testing on n*m-tables with its scarcely known post-hoc testing reveals an easy method to deal with mutually competing risks in the situation of competing dichotomous outcomes excluding each other and without a time-to-event structure. In this particular setting, there is no need to use combined primary endpoints and conditional proportions as secondary endpoints, which, additionally, may be misleading.

Whereas the proposed analysis is suitable for mutually exclusive competing outcomes, it cannot be applied in situations of non-mutually-exclusive events. As an example from the context of neonatal clinical trials, the composite outcome of death or neurodevelopmental impairment (NDI) is often used as an endpoint in neonatal studies, too. NDI in this context is considered to be present if at least one of several adverse outcomes is observed (i.e. hearing or visual impairment or cognitive delay or cerebral palsy). Whereas death before reaching the time point of assessment excludes the observation of the other conditions, each of the components of the diagnosis NDI may occur as a single handicap or in all possible combinations with the other components/handicaps. Obviously, the child's outcome is much more impaired

Table 7a: Chi-Square-Table

	Healthy	Event of interest	Death	
Group A	n_1m_1	n_2m_1	n_3m_1	M1
Group B	n_1m_2	n_2m_2	n_3m_2	M2
	N1	N2	N3	

Table 7b: Table to Calculate Risk Ratios for "Event of Interest"

	Event of interest	Healthy + Death	
Group A	n_2m_1	$n_1m_1 + n_3m_1$	M1
Group B	n_2m_2	$n_1m_2 + n_3m_2$	M2
	N2	N1 + N3	

if there are multiple handicaps – in comparison for example with just the need for a hearing aid. To analyse all these different situations just within one combined dichotomous endpoint (NDI present or not present) reveals a very imprecise and unsatisfactory picture of the real situation. Efforts should be made to develop strategies for analysing these non-mutually-exclusive endpoints in a more informative way.

CONCLUSION

The new approach by applying well-known chi-square test with > 1df and its scarcely known post-hoc testing gives better insight in the effect of a treatment or therapy on the effect of interest in the presence of a mutually exclusive competing risk than the conventional method of dichotomizing the several categories of the outcome. It has been shown that in situations, where - by applying the conventional method - the effect of interest is only been seen in a secondary explorative analysis, this effect may already be shown in the primary analysis of confirmatory phase III studies.

REFERENCES

- [1] Gooley TA, Leisenring W, Crowley J, Storer BE. Estimation of failure probabilities in the presence of competing risks: new representations of old estimators. *Stat Med* 1999; 18: 695-706.
[http://dx.doi.org/10.1002/\(SICI\)1097-0258\(19990330\)18:6<695::AID-SIM60>3.0.CO;2-O](http://dx.doi.org/10.1002/(SICI)1097-0258(19990330)18:6<695::AID-SIM60>3.0.CO;2-O)
- [2] Cheng SC, Fine JB, Wei LJ. Prediction of cumulative incidence function under the proportional hazards model. *Biometrics* 1998; 54: 219-228.
<http://dx.doi.org/10.2307/2534009>
- [3] Dignam JJ, Zhang Q, Kocherginsky M. The use and interpretation of competing risks regression models. *Clin Cancer Res* 2012; 18(8): 2301-2308.
<http://dx.doi.org/10.1158/1078-0432.CCR-11-2097>
- [4] Chappel R. Competing risk analyses: How are they different and why should you care? *Clin Cancer Res* 2012; 18(8): 2127-2129.
<http://dx.doi.org/10.1158/1078-0432.CCR-12-0455>
- [5] Schmoor C, Schumacher M, Finke J, Beyersmann J. Competing risks and multistate models. *Clin Cancer Res* 2013; 19(1): 12-21.
<http://dx.doi.org/10.1158/1078-0432.CCR-12-1619>
- [6] Spratt DE, Wu AJ, Adeseye V, *et al.* Recurrence patterns and second primary lung cancers after stereotactic body radiation therapy for early-stage non-small-cell lung cancer: Implications for surveillance. *Clin Lung Cancer* 2015; Oct 1. pii: S1525-7304(15)00238-7.
<http://dx.doi.org/10.1016/j.clcc.2015.09.006>
- [7] Wulaningsih W, Vahdaninia M, Rowley M, *et al.* Prediagnostic serum glucose and lipids in relation to survival in breast cancer patients: a competing risk analysis. *BMC Cancer* 2015; 15(913): 1-9.
<http://dx.doi.org/10.1186/s12885-015-1928-z>
- [8] Turin TC, Okamura T, Afzali AR, *et al.* Hypertension and lifetime risk of stroke. *J Hypertens* 2016; 34(1): 116-122.
<http://dx.doi.org/10.1097/HJH.0000000000000753>
- [9] Bassler D, Plavka R, Shinwell ES, *et al.* for the NEUROSIS Trial Group. Early inhaled budesonide for the prevention of bronchopulmonary dysplasia. *N Engl J Med* 2015; 373: 1497-1456.
<http://dx.doi.org/10.1056/NEJMoa1501917>
- [10] SUPPORT Study Group of the Eunice Kennedy shriver NICHD Neonatal Research Network. Early CPAP versus surfactant in extremely preterm infants. *N Engl J Med* 2010; 362: 1970-1979.
<http://dx.doi.org/10.1056/NEJMoa0911783>
- [11] SUPPORT Study Group of the Eunice Kennedy shriver NICHD Neonatal Research Network. Target ranges of oxygen saturation in extremely preterm infants. *N Engl J Med* 2010; 362: 1959-1969.
<http://dx.doi.org/10.1056/NEJMoa0911781>
- [12] Shankaran S, Pappas A, McDonald SA, *et al.* for the Eunice Kennedy shriver NICHD Neonatal Research Network. Childhood outcomes after hypothermia for neonatal encephalopathy. *N Engl J Med* 2012; 366: 2085-2092.
<http://dx.doi.org/10.1056/NEJMoa1112066>
- [13] Azzopardi D, Strohm B, Marlow N, *et al.* for the TOBY Study Group. Effects of hypothermia for perinatal asphyxia on childhood outcomes. *N Engl J Med* 2014; 371: 140-149.
<http://dx.doi.org/10.1056/NEJMoa1315788>
- [14] Mayo Clinic [homepage on the internet]. Locally written SAS macros (cited 2016 January 14): available from <http://www.mayo.edu/research/departments-divisions/department-health-sciences-research/division-biomedical-statistics-informatics/software/locally-written-sas-macros>
- [15] Mantel N, Haenszel W. Statistical Aspects of the Analysis of Data From Restrospective Studies of Disease. *JNCI – J Natl Cancer Institute* 1959; 22(4): 719-748.
- [16] Sharpe D. Your Chi-Square Test is Statistically Significant: Now What? Practical Assessment, Research & Evaluation 2015; 20(8). Available online: <http://pareonline.net/getvn.asp?v=20&n=8>
- [17] Franke TM, Ho T, Christie ChA. The chi-square test: Often used and more often misinterpreted. *Am J Eval* 2012; 33(3): 448-458.
<http://dx.doi.org/10.1177/1098214011426594>
- [18] Agresti A. *Categorical data analysis* (3rd ed.). Hoboken NJ: Wiley 2013.
- [19] MacDonald PL, Gardner RC. Type I error rate comparisons of post hoc procedures for IxJ chi-square tables. *Educ Psychol Meas* 2000; 60(5): 735-754.
<http://dx.doi.org/10.1177/00131640021970871>
- [20] Schmidt B, Roberts RS, Davis P, *et al.* Caffeine therapy for apnea of prematurity. *NEJM* 2006; 354: 2112-2121.
<http://dx.doi.org/10.1056/NEJMoa054065>