UNIVERSITA' DEGLI STUDI DI MILANO-BICOCCA
Corso di Dottorato in Epidemiologia e Biostatistica
XXVIII Ciclo

THE VALIDATION OF CANDIDATE SURROGATES FOR A TIME TO EVENT ENDPOINT IN CHILDHOOD LEUKEMIA

Tutor: Prof.ssa Stefania Galimberti

Tesi di dottorato di
Ausiliatrice Lucenti
Matricola - 715729

Anno Accademico 2014-2015

Acknowledgements

Maybe one page is not enough to fully express my gratitude to those people who have been fundamental in the course of my studies and in my life during these years.
I would like to thank Professor Maria Grazia Valsecchi for making it possible for me to start and conduct my studies. Thank you for your always positive and proactive attitude.
Thanks to my work team and to the other colleagues: Maria Chiara, Laura, Paola, Emanuela, Anita, Jessica, Gilda, Simona e Miriam. Each and every one of them has taught me and continues to teach me important lessons.
Thanks to Davide, a wonderful desk neighbour, colleague, statistical advisor, mediator and friend.
I would like to express my appreciation to Professor Burzykowski from Hasselt University for giving me the opportunity to gain a study experience abroad, in Belgium, for welcoming me, for following my work, for always being helpful and for his unforgettable sense of humor.
I would also like to mention all the friends who have always supported me during these years, even through difficult times: thanks to Alessandra G., Ana C., Elisa S. and her parents Carlo and Liliana, Raffaella C. and Arianna P.
Thanks to my family.
Last but not least, I would like to express my most deeply felt gratitude to Stefania Galimberti, the best tutor I could have ever asked for. Thank you for your exceptional perseverance, for all the precious things you taught me, not only from a scientific point of view, for your affection, for always being there for me and for all your useful noes. Thank you.

INDEX

List of Abbreviations **3**
1 INTRODUCTION **4**
2 METHODS **7**
2.1 SINGLE-TRIAL VALIDATION FRAMEWORK **8**
2.2 META-ANALYTIC VALIDATION FRAMEWORK **12**
2.2.1 Two continuous endpoints **13**
2.2.2 An ordinal (or binary) surrogate and a time to event endpoint **14**
2.2.3 Two failure-time endpoints **15**
2.2.4 A continuous surrogate endpoint versus a true failure-time endpoint **16**
2.2.4.1 Copulas and their likelihood **17**
2.2.4.2 The Clayton copula **19**
2.2.4.3 The Hougaard copula **20**
2.2.5 Choice of the trial units **21**
3 THE CASE-STUDY **23**
3.1 MINIMAL RESIDUAL DISEASE **23**
3.2 DESCRIPTION OF THE CLINICAL PROTOCOLS **25**
3.3 ENDPOINT DEFINITIONS **32**
3.4 CHOICE OF THE TRIAL UNITS **33**
4 RESULTS **39**
4.1 DESCRIPTION OF THE DATA **39**
4.2 VALIDATION OF MRD FOR EFS **42**
4.2.1 ONE-TRIAL APPROACH **46**
4.2.2 MULTITRIAL APPROACH **54**
4.2.2.1 Sensitivity analyses **56**
4.2.3 ANALYSIS O MRD IN CONTINUOUS **59**
4.3 VALIDATION OF EFS FOR OS **60**
4.3.1 Multi trial approach **61**
4.3.1.1 Sensitivity analysis **65**
5 DISCUSSION **68**
SUPPLEMENT **70**
Appendix A: Bivariate copula **71**
Appendix B – Clayton Copula **72**
Appendix C – Hougaard Copula **78**

REFERENCE **84**

List of Abbreviations

AIEOP	Associazione Italiana Ematologia e Oncologia Pediatrica
ALL	Acute Lymphoblastic Leukaemia
AR	MRD average risk in EORTC group
AUL	Acute Undifferentiated Leukaemia
BFM	Berlin-Frankfort-Münsterx
CNS	Central Nervous System
COG	Children's Oncology Group
CR	Complete Remission
DXM	Dexametasone treatment
EFS	Event Free Survival
EORTC	European Organization for Research and Treatment of Cancer
FCM	Flow cytometry
HR	Hazard Ratio
MRD	Minimal Residual Disease
MRD-HR	MRD High risk group for AIEOP-BFM
MRD-IR	MRD intermediate risk group for AIEOP-BFM
MRD-SR	MRD standard risk group for AIEOP-BFM
NLPNRR	Newton-Raphson algorithm available in the SAS routine
OR	Odds Ratio
OS	Overall Survival
PDN	Prendisone treatment
PE	Proportion Explained
RE	Relative Effect
RER	Rapid Early Responders in COG group
RQ-PCR	Polymerase Chain Reaction
SER	Slow Eearly Responders in COG group
TP1	Time point one in AIEOP-BFM
TP2	Time point two in AIEOP-BFM
VHR	MRD very high risk in EORTC group
VLR	MRD very low risk in EORTC group
ρZ	Adjusted association

1 INTRODUCTION

The selection of the primary "outcome measure" or "endpoint" is a crucial step in the design of a clinical trial, whose aim is to demonstrate the presence of the treatment effect on such endpoint. Two major criteria for endpoint selection should be followed: sensitivity to treatment effects and clinical relevance (Fleming 1996).

It is often evident, however, that it might be difficult in a clinical trial to use the "true" endpoint, or rather the most sensitive and relevant clinical endpoint. This might be due to several reasons, including the high costs of the true endpoint evaluation, the difficulty of measurement, a long follow-up time requirement, or a large sample size due to a low incidence of the event (Burzykowski T., Molenberghs G., Buyse M. 2005). In such cases, the use of a true endpoint increases the complexity and/or the duration of research. To overcome these problems, a seemingly attractive solution is to replace the true endpoint with another one, which is measured earlier, more conveniently or more frequently. Such "replacement" endpoint is termed "surrogate" and has the purpose of evaluating the effect of a specific treatment for a specific disease.

An endpoint should be characterized by precise properties to be defined as a potential surrogate. The first property is that the surrogate endpoint must be on the causal pathway of the disease process and not otherwise. The second property is that the surrogate endpoint must be associated to the true endpoint. Nevertheless, these properties are not sufficient to make it a surrogate endpoint, and it is clear from this definition that surrogacy is disease as well as treatment dependent (Buyse M. 2009). Once a candidate surrogate is identified, several formal methods are available for the validation of the surrogate endpoint, depending on the number of trials performed.

The first formal statistical approach dates back to 1989, when Prentice proposed a definition of surrogate endpoint and four criteria to validate it. The most important criterion among these is called "The Prentice's Criterion", which implies that "the full effect of treatment upon the true endpoint is captured by the surrogate" (Prentice 1989). Subsequently, the Prentice's approach was strengthened by Freedman, who introduced the proportion of treatment explained (PE), aimed at measuring the proportion of the treatment effect mediated by the surrogate (Freedman 1992). This proposal was important since it shifted the interest in the validation of surrogate endpoints from significance testing to estimation, but it is in itself surrounded with difficulties (Burzykowski T., Molenberghs G., Buyse M. 2005). Buyse and Molenberghs showed that PE can be decomposed in two different quantities: the Relative Effect (RE) and the adjusted association (ρ_Z). The first measure relates to the capability of the surrogate to predict the treatment effect on the clinical endpoint at a population level; the second measure describes its capability to predict the outcome

of the clinical endpoint by describing the subject-specific association between the surrogate and true endpoint.

Since earlier methods relied on data coming from a single trial, they lacked treatment effect replication; to solve this problem, a meta-analytic approach to the validation of a surrogate endpoint was proposed by a group of Dutch statistician: Buyse M., Burzykowski T. and Molenberghs G.. It consists in estimating associations at two different levels: the association between the surrogate and the clinical endpoint, called the "individual-level association", and the association between the effects of treatment on the surrogate and the clinical endpoint, called "trial-level association" (Burzykowski et al. 2005). A clinical endpoint can be reliably estimated from the biomarker in an individual patient if a strong individual-level association is present, whereas a strong trial-level association implies that the effect of treatment on the clinical endpoint can be reliably estimated from the effect of treatment on the biomarker. A good surrogate is one that has biologic plausibility and is showed, statistically, to have strong individual-level and trial-level associations with the final endpoint. This meta-analytic approach was first developed to deal with continuous endpoints, but it has now many extensions to cover situations where the candidate surrogate and the clinical endpoint are not continuous or they are of a different nature, for instance, when both are binary, when one of them is a time to event outcome and the other is categorical or when both endpoints are repeatedly measured over time, and so on.

The aim of my PhD thesis is the evaluation of Minimal Residual Disease (MRD) as surrogate endpoint in acute lymphoblastic leukaemia (ALL). In fact MRD has not yet been formally validated as a surrogate endpoint, whilst it is a well-established prognostic biomarker in ALL. The challenge has now evolved to the qualification of early MRD as an efficacy-response biomarker in the assessment of new drugs for the treatment of paediatric ALL. In addition, as methods on the validation of a continuous surrogate for a failure time endpoint are lacking, a proposal was made here. In line with the meta-analytic framework, a copula based approach was implemented and translated in a SAS macro dealing with two different copulas.

Specifically, the main goal of the present study is to assess whether MRD evaluated at the end of the induction treatment can be considered a surrogate for Event Free Survival (EFS) in childhood B-lineage ALL patients who were treated (after randomization) with Dexamethasone or Prednisone. A secondary aim is to explore if Event Free Survival can be considered a surrogate for Overall Survival (OS). For this purpose a very large database with data from three randomized clinical trial performed in different European countries and in the USA have been used.

This thesis is organized as follows: a brief overview of the statistical methods used in the clinical application is presented in chapter 2, together with the proposal developed specifically for the validation of a continuous surrogate for a failure time endpoint. The motivating clinical context is illustrated in chapter 3, where details on MRD and on the ALL childhood protocols are reported. The results of the analyses are descripted in chapter 4, which contains a section dedicated to the validation of MRD as surrogate of EFS, and a second section devoted to the validation of EFS as surrogate for OS. Some final remarks are given in chapter 5.

2 METHODS

A significant feature regarding a potential surrogate endpoint is that it must present some peculiar properties. First of all, the surrogate endpoint must be on the causal pathway of the disease process (Figure 2.1), and not otherwise, as shown in the four cases of Figure 2.2a-d. A surrogate endpoint might not be involved in the same pathophysiologic process that results in the clinical outcome (Figure 2.2a), and even when it does, it is likely that some disease pathways are causally related to the clinical outcome and not to the surrogate endpoint (Figure 2.2b-d). Of the disease pathways affecting the true clinical outcome, the intervention may only affect the pathway mediated through the surrogate endpoint (Figure 2.2b), or the pathway or pathways independent of the surrogate endpoint (Figure 2.2c). Most important, the intervention might also affect the true clinical outcome by unintended mechanisms of action that are independent of the disease process (Figure 2.2d) (Fleming and DeMets 1996). The surrogate endpoint must be associated to the true endpoint, and this is the second property which it presents. However, the existence of such properties is thought not to be sufficient for using the former as a surrogate. Also, we shall call surrogate a biomarker or endpoint that is able to replace a clinical endpoint for the purpose of evaluating the effect of a specific treatment for a specific disease. Note that this definition comprises both the disease dependency as well as the treatment dependency characterizing surrogacy (Buyse M. 2009). Once a candidate surrogate is identified, several formal methods for the validation of surrogate endpoints are available depending on the number of trials performed.

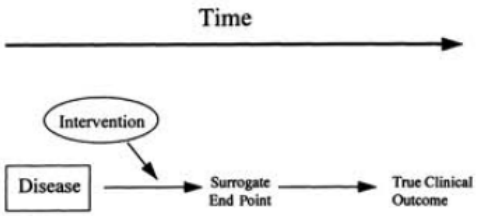


Figure 2.1: The setting that provides the greatest potential for the surrogate endpoint to be valid (Fleming and DeMets 1996)

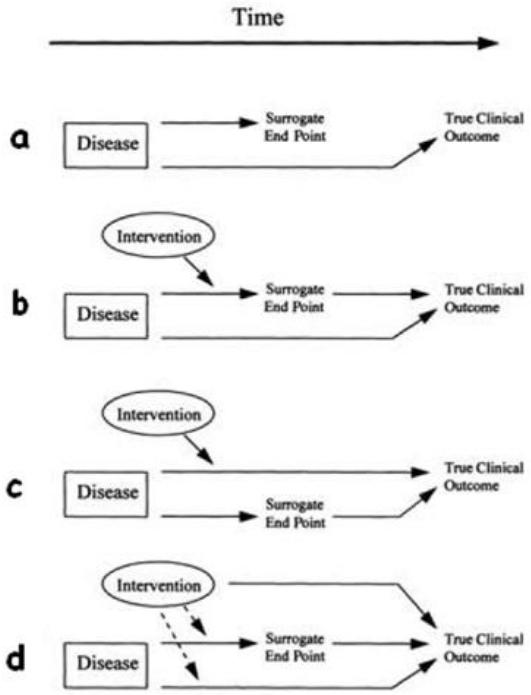


Figure 2.2a-d: Cases in which the surrogate isn't on the causal pathway of the disease process (Fleming and DeMets 1996)

The following notation will be adopted throughout this report: T and S are random variables denoting the true and surrogate endpoints, respectively, and Z is an indicator variable for treatment. This notation will be expanded by using two indices: i=1,..,N for trials/units and j=1,…,n_i for subjects within trials/units. The trial-specific effects of treatment Z on the two endpoints in trial *i* will be α_i and β_i.

2.1 SINGLE-TRIAL VALIDATION FRAMEWORK

It was in 1989 that Prentice published a paper in which the validation process for continuous endpoints was put within a statistical framework. Four operational criteria to check were proposed in his work that used data from a single trial. These criteria require:
(1) a significant impact of treatment on the surrogate endpoint (α);
(2) a significant impact of treatment on the true endpoint (β);
(3) a significant impact of the surrogate endpoint on the true endpoint (γ);
(4) the full effect of treatment upon the true endpoint is captured by the surrogate (β_S).
The latter criterion is also known as "The Prentice's Criterion".

Tests of significance on parameters α, β, γ can be used to verify the first three operational criteria in the following models:

$$S_j = \mu_S + \alpha Z_j + \varepsilon_{Sj} \quad \Leftrightarrow \quad f(S|Z) \neq f(S) \tag{1}$$

$$T_j = \mu_T + \beta Z_j + \varepsilon_{Tj} \quad \Leftrightarrow \quad f(T|Z) \neq f(T) \tag{2}$$

$$T_j = \mu + \gamma S_j + \varepsilon_j \quad \Leftrightarrow \quad f(T|S) \neq f(T) \tag{3}$$

$$T_j = \tilde{\mu}_T + \beta_S Z_j + \gamma_Z S_j + \tilde{\varepsilon}_{Tj} \quad \Leftrightarrow \quad f(T|S,Z) = f(T|S) \tag{4}$$

where $\beta_S = \beta - \sigma_{TS}\sigma_{SS}^{-1}\alpha$, $\gamma_Z = \sigma_{TS}\sigma_{SS}^{-1}$ and σ_{ST}, σ_{SS}, σ_{TT} are the elements of the variance-covariance matrix of the error terms of (1) and (2) (See Figure 2.3).

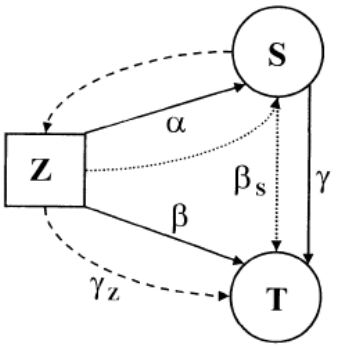


Figure 2.3: The associations between treatment (Z), a surrogate endpoint (S), and a true endpoint (T) are caracterized by the three parameters α, β and γ. Parameter β_S characterizes the effect of Z on T after adjustment for S, while γ_Z characterizes the effect of S on T after adjustment for Z (Buyse and Molenberghs, 1998).

These operational criteria are are informative and will tend to be fulfilled for valid surrogate endpoints, but they should not be regarded as strict criteria. They are necessary and sufficient to establish the validity of binary surrogate endpoint, but not for more complex surrogate endpoint (Burzykowski et al. 2005). Also, the criterion (4) might be useful to reject a poor surrogate endpoint, but it is inadequate to validate a good surrogate endpoint, for failing to reject the null hypothesis due merely to insufficient power (Freedman, L.S., Graubard, B.I., and Schatzkin, A., 1992).

Given that it cannot be proven that the effect of treatment upon the true endpoint is fully captured by the surrogate (4), a more direct approach for the estimation of the proportion of the exposure effect, that is explained by the surrogate endpoint, was designed by Freedman, Graubard, and Schatzkin (1992). A surrogate which explains a large proportion of that effect is to be considered a valid one, and for this reason it follows that the natural estimate of the "Proportion Explained" is

$$PE(T,S,Z) = 1 - \frac{\beta_S}{\beta} \quad (5)$$

where β and β_S are the estimates of the effect of Z on T, respectively, without and with adjustment for S.

Freedman's PE shares with the Prentice's Criteria the following criticism: if there is an interaction term between Z and S in (4), PE ceases to have a single interpretation and the validation process would have to stop (Freedman, Graubard, and Schatzkin (1992). Furthermore, if the number of observation is not large and the effect treatment upon the true endpoint is small, confidence limits tend to be wide and there will be substantial uncertainty about the proportion of the effect that is mediated by surrogate (Burzykowski et al. 2005). Even when large numbers of observations are available, however, the denominator of the proportion explained will be estimated with little precision, and the need for a surrogate endpoint would no longer exist. Therefore, the proportion explained will generally be too poorly estimated to be of much practical value. This conclusion has been supported by the results obtained by Freedman (2001) who reported that, to achieve 80% power for a test of the hypothesis that the surrogate explains more than 50% of treatment effect, the ratio $\beta/SE(\hat{\beta})$ should equal 5 or more. As noted by Freedman (2001), this requirement makes the use of PE practically infeasible. Moreover, the estimated proportion would be equal to 1 in the case of a perfect surrogate, but PE cannot be considered a true proportion ranging from 0 to 1 because it might assume values larger than 1 (Burzykowski et al. 2005).

Buyse and Molenberghs (1998) suggested to calculate two other quantities for the validation of a surrogate endpoint. The first quantity is the relative effect

$$RE(T,S,Z) = \frac{\beta}{\alpha} \quad (6)$$

Intuitively, RE can be interpreted as the slope of a regression line between β and α. If the multiplicative relation (6) could be assumed, and if RE were known exactly, it could be used to predict the effect of Z on T based on an observed effect of Z on S. In practice, RE will have to be estimated, and the precision of the estimation will be relevant for the precision of the prediction.

RE associates the effects of Z on T and on S averaged over all subjects and it will be equal to 1 if the effects of Z on T and on S are of identical magnitude. It will tend to be less than 1 if the true endpoint is more difficult to be affected than the surrogate endpoint. The second measure quantifies the association between S and T after adjustment for the treatment Z:

$$\rho_Z = \frac{\sigma_{ST}}{\sqrt{\sigma_{SS}\sigma_{TT}}} \quad (7)$$

where σ_{ST}, σ_{SS} and σ_{TT} are elements of variance-covariance matrix of the error terms of (1) and (2).

If ρ_Z is large (i.e. ρ_Z =1 for normal endpoint) it means that the surrogate and the true endpoints are very similar and there is a deterministic relationship between S and Z. The pair of measures ρ_Z and RE usefully complements the PE. Indeed, ρ_Z describes the subject-specific association between the surrogate and true endpoints, while RE links them at the population-averaged level. A perfect surrogate is one which has a large ρ_Z (the surrogate is perfect at the individual level) and RE =1 (the surrogate is perfect at the population level).

Buyse and Molenberghs (1998) noted that the use of RE and ρ_Z to validate surrogate endpoints is also complicated by a few problems. The two major limitations of RE are that its confidence limits may be too wide to permit clinically useful predictions and that its value may depend on the value of α. In other words, since RE is the slope of a regression line between α and β, the linearity of this regression may be questioned. RE might change with, the strength of the association between Z and the outcomes themselves. Also, estimate of RE is based on the strong assumption that the relationship between the treatment effects on the surrogate and true endpoints is multiplicative, an assumption that may be too strong to hold and unverifiable. This difficulty is more fundamental than the limited precision of RE that will typically be obtained in trials of small or moderate size (Buyse and Molenberghs 1998).

The methods presented in this section are summarized in Table 2.1.

Table 2.1: The quantities of interest for the validation of surrogate endpoint in a single trial (Buyse and Molenberghs, 1998).

Quantity of interest	Estimate	Test
Effect of treatment on true endpoint	β	H_0: $f(T \mid Z) = f(T)$
Effect of treatment on surrogate endpoint	α	H_0: $f(S \mid Z) = f(S)$
Effect of surrogate on true endpoint	γ	H_0: $f(T \mid S) = f(T)$
Proportion of treatment effect on true endpoint explained by surrogate	$PE(T,S,Z) = 1 - \frac{\beta_S}{\beta}$	
Effect of treatment on true endpoint relative to that on surrogate endpoint	$RE(T,S,Z) = \frac{\beta}{\alpha}$	
Adjusted effect of surrogate on true endpoint	ρ_Z	

2.2 META-ANALYTIC VALIDATION FRAMEWORK

The idea of validating surrogate endpoints through a meta-analytic approach has been first theoretically conceived and developed by Buyse et al. (2000a) by considering the case of two normally distributed endpoints in a multiple-trial setting, even if the general strategy was advocated by other authors earlier (Boissel et al. 1992, Freedman et al. 1992, Lin et al. 1997, Bycott et al. 1998).

This approach is essentially based on the estimation of associations at two levels: the association between the surrogate and the clinical endpoint, called the "individual-level association", and the association between the effects of treatment on the surrogate and the clinical endpoint, called "trial-level association". When a strong individual-level association is present, then the clinical endpoint can be reliably estimated from the biomarker in individual patient, whereas when a strong trial-level association exists, this implies that the effect of treatment on the clinical endpoint can be reliably estimated from the effect of treatment on the biomarker. A surrogate that has biologic plausibility and is showed, statistically, to have strong individual-level and trial-level associations with the final endpoint is to be regarded as a good one.

The specific methods used to validate a surrogate for a clinical endpoint will clearly depend on the nature of the variables involved in the problem at hand. Unlike for continuous outcomes, where the multivariate normal distribution and the linear mixed model provide a natural paradigm for model development, as shown in the first part of this section for contrast, other situations are addressed by bivariate models. In this perspective, the key aspect in the methods proposed by Burzykowski et al (2005), is the use of the copula, a general class of multivariate models that can be implemented starting by particular marginal models assumed for the surrogate and the true endpoint.

In the application on childhood acute lymphoblastic leukemia, the implemented methods are based on: a time to event endpoint and an ordinal or a continuous surrogate and two time to event endpoints. The sections that follow are devoted to a brief description of the methods available from the literature, while the proposal made on the evaluation of a continuous surrogate for a time to event endpoint will be more extensively illustrated.

All the analyses were done in SAS (version 9.2) and some of them were based on macros provided by Buyse et all. (Burzykowski T., Molenberghs G., Buyse M. 2005) and available on http://ibiostat.be/software/surrogate, while others were specifically implemented as part of this PhD project and reported in Appendix.

2.2.1 Two continuous endpoints

To evaluate two continuous endpoints that are assumed to be jointly normally distributed, a two level hierarchical model with trial level effects, either random or fixed, is postulated.
In the two-stage fixed-effects representation, models at both levels are fitted separately. The first stage is based on trial-specific models:

$$S_{ij} = \mu_{Si} + \alpha_i Z_{ij} + \varepsilon_{Sij} \quad (8)$$

$$T_{ij} = \mu_{Ti} + \beta_i Z_{ij} + \varepsilon_{Tij} \quad (9)$$

while at the second stage it is assumed that $(\mu_{Si}, \mu_{Ti}, \alpha_i, \beta_i)'$ follows a normal distribution with mean $(\mu_S, \mu_T, \alpha, \beta)'$ and with an unstructured covariance matrix that is:

$$D = \begin{pmatrix} d_{SS} & d_{ST} & d_{Sa} & d_{Sb} \\ & d_{TT} & d_{Ta} & d_{Tb} \\ & & d_{aa} & d_{ab} \\ & & & d_{bb} \end{pmatrix}. \quad (10)$$

The coefficient of determination R^2_{indiv} (11) that regards the distribution of T_{ij} conditional on S_{ij}, defines the association between the surrogate and the final endpoints, after adjustment for the effect of treatment. This coefficient is a measure of the precision with which we may predict the value of T_{ij} for an individual patient on the basis of the observed value of S_{ij} and the treatment assignment.

$$R^2_{indiv} = \frac{\sigma^2_{ST}}{\sigma_{SS}\sigma_{TT}} \quad (11)$$

The coefficient of determination, R^2_{trial} (12), is a natural quantity used to assess the quality of a surrogate at the trial level that pertains to the distribution of β_i conditional on μ_{Si} and α_i.

$$R^2_{trial} = \frac{\begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}^T \begin{pmatrix} d_{SS} & d_{Sa} \\ d_{Sa} & d_{aa} \end{pmatrix}^{-1} \begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}}{d_{bb}} \quad (12)$$

This coefficient measures how precisely we may predict the effect of treatment on the true endpoint on the basis of previous data and the observed treatment effect on the surrogate endpoint from a new trial. If $R^2_{trial} = 1$, then the treatment effect on the clinical endpoint can be predicted without error using the treatment effect on the surrogate, whereas $R^2_{trial} = 0$ implies that both treatment effects are independent and no meaningful prediction is possible (Burzykowski T., Molenberghs G., Buyse M. 2005).

2.2.2 An ordinal (or binary) surrogate and a time to event endpoint

To provide validation measures when the true endpoint is a failure-time random variable and the surrogate is a categorical variable with K ordered categories, Burzykowski et all. (2004) used at the first stage a bivariate copula model for the true endpoint and a latent variable underlying the surrogate endpoint ($\tilde{S}$)

$$F_{T_{ij},\tilde{s}_{ij}}(t,s;z) = C_\theta\left[F_{T_{ij}}(t;z), F_{\tilde{S}_{ij}}(s;z), \theta\right] \quad (13)$$

where $F_{\tilde{S}_{ij}}(s;z)$ and $F_{T_{ij}}(t;z)$ are the marginal cumulative distribution function of $\tilde{S}_{ij}$ and T_{ij} given $Z_{ij} = z$, respectively, and C_θ is a one parameter (θ) copula function (See Appendix A for details), i.e. a bivariate distribution function on $[0,1]^2$ with uniform margins, describing the association between T and $\tilde{S}$.

Specifically, the proportional odds model with K ordered categories is used to model S:

$$logit\{P(S_{ij} \le k|Z_{ij})\} = \eta_{ik} + \alpha_i Z_{ij} \quad (14)$$

and a proportional hazard model is used to model T:

$$\lambda(t|Z_{ij}) = \lambda_i(t)exp(\beta_i Z_{ij}) \quad (15)$$

where $\lambda_i(t)$ is the trial specific baseline hazard function.

At the first stage estimates of the parameter θ and of the trial specific treatment effects α_i and β_i are obtained by maximising the likelihood of (13). Noticeably, the estimation of the parameter in the proportional odds model (14) requires that in each trial unit all the K response levels are observed. In the case this assumption is not fulfilled, the trial units need to be modified or, in alternative, the model in (14) needs to be reparametrized as in (16):

$$logit\{P(S_{ij} \le k|Z_{ij})\} = \eta_k^0 + \eta_i + \alpha_i Z_{ij} \quad (16)$$

where the model assumes a fixed set of cutpoints $\eta_k^0, \dots, \eta_i$, but allows for trial-specific shifts η_i of the set.

At the second stage, it is assumed to use the trial level model:

$$\begin{pmatrix}\eta_i \\ \alpha_i \\ \beta_i\end{pmatrix} = \begin{pmatrix}\eta \\ \alpha \\ \beta\end{pmatrix} + \begin{pmatrix}e_i \\ a_i \\ b_i\end{pmatrix} \quad (17)$$

where $(\eta_i, \alpha_i, \beta_i)'$ follows a normal distribution with mean $(\eta, \alpha, \beta)'$ and with an unstructured covariance matrix.

We can appraise the quality of the surrogate at the individual trial level, which is a measure of the association between $\tilde{S}_{ij}$ and T_{ij}, on the basis of the copula parameter θ. When the bivariate Plackett copula is used (1965), θ takes the form of a (constant) global odds ratio:

$$\theta = \frac{P(T_{ij} > t, S_{ij} > k)\,P(T_{ij} \leqslant t, S_{ij} \leqslant k)}{P(T_{ij} > t, S_{ij} \leqslant k)\,P(T_{ij} \leqslant t, S_{ij} > k)} = \frac{P(T_{ij} > t|S_{ij} > k)}{P(T_{ij} \leqslant t|S_{ij} > k)} \left\{ \frac{P(T_{ij} > t|S_{ij} \leqslant k)}{P(T_{ij} \leqslant t|S_{ij} \leqslant k)} \right\}^{-1} \quad (18)$$

Thus, it can be interpreted as the (constant) ratio of the odds for surviving beyond time t given categories higher than k to the odds of surviving beyond time t given categories at most k. For a binary surrogate (k=2), it is just the odds ratio for a category versus the other, as the model in (14) reduces to a regression logistic model.

On the basis of the coefficient of determination R^2_{trial}, that pertains to the distribution of β_i conditional on the set of trial specific parameters including α_i and η_i, we can assess the quality of the surrogate at the trial level.

2.2.3 Two failure-time endpoints

Taking into consideration the case in which both the surrogate and the true endpoints are failure-time variables, Burzykowski et al. (2001) proposed a copula model that assumed the following joint survival function of (S_{ij}, T_{ij}):

$$F(s,t) = P(S_{ij} \geq s, T_{ij} \geq t) = C_\theta\{F_{Sij}(s), F_{Tij}(t)\} \quad s,t \geq 0 \quad (19)$$

where F_{Sij} and F_{Tij} denote the marginal survival functions and C_θ is a copula (Clayton 1978, Hougaard 1986, Plackett 1965). To model the effect of treatment on the marginal distributions of S_{ij} and T_{ij}, two proportional hazards models are introduced:

$$F_{Sij}(s) = exp\left\{-\int_0^s \lambda_{Si}(x)exp(\alpha_i Z_{ij})dx\right\} \quad (20)$$

$$F_{Tij}(t) = exp\left\{-\int_0^t \lambda_{Ti}(x)exp(\beta_i Z_{ij})dx\right\} \quad (21)$$

where λ_{Si} and λ_{Ti} are trial-specific marginal baseline hazard functions and α_i and β_i are trial-specific effects of treatment Z on the endpoints.

An remarkable feature of model (19) is that the margins do not depend on the choice of the copula function. Theorically, in model (19) any copula function can be used. For sake of simplicity, Burzykowski et al. (2001) considered primarily one-parameter families. In practical applications, they resorted to the Clayton (1978), the Hougaard (Gumbel, 1960) and the Plackett (1965) copula functions.

At the second stage it is assumed that:

$$\begin{pmatrix}\alpha_i\\ \beta_i\end{pmatrix} = \begin{pmatrix}\alpha\\ \beta\end{pmatrix} + \begin{pmatrix}a_i\\ b_i\end{pmatrix} \quad (22)$$

where $(\alpha_i, \beta_i)'$ follows a normal distribution with mean $(\alpha, \beta)'$ and with an unstructured covariance matrix.

The individual level association for two failure time endpoints is generally assessed by the Kendall's τ, an index that depends only on the copula function C_θ and it is independent of the marginal distributions of S_{ij} and T_{ij} (Burzykowski et al. 2005):

$$\tau = 4\int_0^1\int_0^1 C_\theta(u,v)C_\theta(du,dv) - 1 \quad (23)$$

The Kendall's τ represents the strength of association between the two endpoints remaining after adjustment, through the marginal models, for trial- and treatment effects.

The quality of the surrogate at the trial level is assessed on the basis of the coefficient of determination R^2_{trial}, obtained from the model at the second stage.

2.2.4 A continuous surrogate endpoint versus a true failure-time endpoint

Specific methods dealing with the validation of a continuous surrogate for a true failure-time endpoint are not available in the methodological literature. We propose here an approach that is mediated from the previous ones and is, thus, based on the copula models.

Similarly to the case of an ordinal surrogate, at the first stage we have a bivariate copula model:

$$F_{T_{ij},S_{ij}}(t,s;z) = C_\theta\left[F_{T_{ij}}(t;z), F_{S_{ij}}(s;z), \theta\right] \quad (24)$$

where $F_{S_{ij}}(s;z)$ and $F_{T_{ij}}(t;z)$ are the marginal cumulative distribution function of S_{ij} or T_{ij}, given $Z_{ij} = z$, respectively.

A linear model for the continuous surrogate S is assumed:

$$S_{ij} = \mu_{Si} + \alpha_i Z_{ij} + \varepsilon_{Sij} \quad (25)$$

and a proportional hazard model for the true time to event endpoint T:

$$\lambda(t|Z_{ij}) = \lambda_i(t)exp(\beta_i Z_{ij}). \quad (26)$$

At the second stage it is assumed that:

$$\begin{pmatrix}\mu_i\\ \alpha_i\\ \beta_i\end{pmatrix} = \begin{pmatrix}\mu\\ \alpha\\ \beta\end{pmatrix} + \begin{pmatrix}e_i\\ a_i\\ b_i\end{pmatrix} \quad (27)$$

where the vector $(\mu_i, \alpha_i, \beta_i)'$ follows a multivariate normal distribution with mean $(\mu, \alpha, \beta)'$ and with an unstructured covariance matrix:

$$D = \begin{pmatrix} d_{SS} & d_{Sa} & d_{Sb} \\ & d_{aa} & d_{ab} \\ & & d_{bb} \end{pmatrix}. \quad (28)$$

The quality of the surrogate at the trial level can be assesses on the basis of the coefficient of determination R^2_{trial}.

$$R^2_{trial} = \frac{\begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}^T \begin{pmatrix} d_{SS} & d_{Sa} \\ d_{Sa} & d_{aa} \end{pmatrix}^{-1} \begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}}{d_{bb}} \quad (29)$$

Since the correlation between S and T depends on the shape of the marginal function, we make use of a Copula function to estimate individual association at the first stage. For sake of simplicity, we first postulated a normal distribution for S, but different assumptions can actually be made. In the following sections the procedures based on Clayton and the Hougaard Copulas were described in details.

2.2.4.1 Copulas and their likelihood

The copula approach is a useful method for deriving a joint distribution given the marginal distributions and it is used to describe the dependence structure between two variables throught a proper association parameter. As the copula model splits the problem in two, with the marginal functions and an association parameter, a two-stage estimation procedure can be used by first jointly estimating the margins and then using the estimated margins to obtain the association parameter. Hougaard (1987) first suggested this two stage estimation procedure, which was also studied by Shih and Louis (1995) who examined the case in which each margin was modelled separately.

A remarkable feature of the copula is that the marginal models and the association model can be selected without constrains. Using the joint distribution function (24) with a proportional hazard model (26) and a fixed effect linear model (25) as marginal models, the corresponding likelihood function for the observed data can be identified for this specific situation, as described below.

The starting point is the general joint distribution copula function in (24), specified in terms of the cumulative distribution function (F) of the marginal models, that is now expressed in terms of the corresponding survival function: S=1-F:

$$P(S_{ij} > s, T_{ij} > t) = C_\theta(u, v) = C_\theta\left(S_s(s_{ij}), S_T(t_{ij})\right) \tag{30}$$

$$= \int_s^\infty \int_t^\infty f_{(S_{ij},T_{ij})}(u, v)\, du\, dv$$

where:

$$u: \qquad S(s) = 1 - \Phi(\mu, \alpha)_S \tag{31}$$

$$v: \qquad S(t) = e^{-(\lambda_{T_i} e^{\beta_{ij} z_{ij}} t)^{p_i}} \tag{32}$$

As the likelihood for a generic density f(T;λ) in the presence of censoring is expressed as

$$L(\lambda; t) = \prod_{i=1}^{n} [f(t_i)]^{\delta_i} [S(t_i)]^{1-\delta_i} \tag{33}$$

we have for f(S,T; $\mu_i, \alpha_i, \sigma_i^2, \lambda_i, \beta_i, \rho_i$) the following expression:

$$L(\mu_i, \alpha_i, \sigma_i^2, \lambda_i, \beta_i, \rho_i) = \prod_{i=1}^{c} \prod_{j=1}^{n_i} \left[f_{(S_{ij},T_{ij})}(S_{ij}, T_{ij})\right]^{\delta_T} \left[-\frac{\partial}{\partial s_{ij}} C_\theta\left(S_s(s_{ij}), S_T(t_{ij})\right)\right]^{1-\delta} \tag{34}$$

where the bivariate density is equal to:

$$f_{(S_{ij},T_{ij})}(S_{ij}, T_{ij}) = \frac{d^2}{ds_{ij} dt_{ij}} C_\theta\left(S_s(s_{ij}), S_T(t_{ij})\right) = \tag{35}$$

$$= \frac{d}{dt_{ij}}\left[\frac{\partial}{\partial s_{ij}} C_\theta\left(S_s(s_{ij}), S_T(t_{ij})\right)\right] = \frac{d}{dt_{ij}}\left[\frac{dC(u, v)}{du} \frac{dS_s(s_{ij})}{ds_{ij}}\right] =$$

$$= \frac{d^2 C_\theta(u, v)}{du\, dv} \frac{dS_T(t_{ij})}{dt_{ij}} \frac{dS_s(s_{ij})}{ds_{ij}}$$

Based on these results, we can write the likelihood as:

$$L(\mu_i, \alpha_i, \sigma_i^2, \lambda_i, \beta_i, \rho_i) = \prod_{i=1}^{c} \prod_{j=1}^{n_i} \left[\frac{d^2 C_\theta(u, v)}{ds_{ij} dt_{ij}}\right]^{\delta_T} \left[-\frac{dS_s(s_{ij})}{ds_{ij}}\right]^{1-\delta_T} \tag{36}$$

while the log-likelihood can be expressed as:

$$log L(\mu_i, \alpha_i, \sigma_i^2, \lambda_i, \beta_i, \rho_i) = \tag{37}$$

$$\sum_{i=1}^{c} \sum_{j=1}^{n_i} \left\{ \delta_T \log\left[\frac{d^2 C_\theta\left(S_s(s_{ij}), S_T(t_{ij})\right)}{ds_{ij} dt_{ij}}\right] + (1 - \delta_T) \log\left[-\frac{\partial C_\theta\left(S_s(s_{ij}), S_T(t_{ij})\right)}{\partial s_{ij}}\right] \right\}$$

Of the different copula functions that can be used, we concentrated on the Clayton and on Hougaard copulas. This was mainly motivated by the fact that in these two cases the association parameters are easy to interpret and to compute.

The estimation process of both the copulas we considered was implemented in a SAS IML macro. The log Likelihood estimates are obtained by using the Newton-Raphson algorithm available in

the SAS routine NLPNRR. Standard errors of the parameters were constructed via delta method, using the inverse of the Hessian Matrix, which was obtained using the SAS routine NLPFDD. The codes for the Clayton and Hougaard copulas are available in Appendix B and C, respectively.

2.2.4.2 The Clayton copula

The Clayton Copula (Clayton, 1978) takes the form

$$C_\theta(u,v) = (u^{1-\theta} + v^{1-\theta} - 1)^{\frac{1}{1-\theta}} \tag{38}$$

where *u* and *v* are the marginal cumulative distribution function and θ the dependence parameter (with θ>1). As θ approaches one, the marginal become independent. Replacing the marginal cumulative distribution function in the Clayton Copula, we obtain.

$$C_\theta(u,v) = ((1-\Phi(s))^{1-\theta} + S(t)^{1-\theta} - 1)^{\frac{1}{1-\theta}} \tag{39}$$

Based on (39), we can write the log likelihood function for the Clayton Copula by computing the following derivatives:

$$\frac{dC_\theta(u,v)}{du} = \frac{1}{1-\theta}\left(u^{1-\theta} + v^{1-\theta} - 1\right)^{\frac{1}{1-\theta}-1}(1-\theta)u^{-\theta} = \tag{40}$$

$$= \left(u^{1-\theta} + v^{1-\theta} - 1\right)^{\frac{\theta}{1-\theta}}u^{-\theta} = \{C_\theta(u,v)\}^\theta u^{-\theta}$$

$$\frac{d^2C_\theta(u,v)}{du\,dv} = \frac{1}{1-\theta}\left(u^{1-\theta} + v^{1-\theta} - 1\right)^{\frac{1}{1-\theta}-1}(1-\theta)v^{-\theta}u^{-\theta} = \tag{41}$$

$$= \theta\left(u^{1-\theta} + v^{1-\theta} - 1\right)^{\frac{2\theta-1}{1-\theta}}v^{-\theta}u^{-\theta} = \theta\{C_\theta(u,v)\}^{2\theta-1}v^{-\theta}u^{-\theta}$$

$$\frac{dS_T(t_{ij})}{dt_{ij}} = \frac{d\,(1-F(t))}{dt_{ij}} = -\frac{d\,F(t)}{dt_{ij}} = -f_T(t) \tag{42}$$

$$\frac{dS_s(s_{ij})}{ds_{ij}} = \frac{dC_\theta(u,v)}{du}\frac{dS_s(s_{ij})}{ds_{ij}} = \tag{43}$$

$$= \frac{dC_\theta(u,v)}{du}\,\frac{d\,(1-\Phi(s))}{ds_{ij}} = \frac{dC_\theta(u,v)}{du}(-\varphi(s))$$

therefore:

$$-\frac{dS_s(s_{ij})}{dS_{ij}} = \{C_\theta(u,v)\}^\theta u^{-\theta}\varphi(s)$$

$$\frac{d^2C_\theta(u,v)}{ds_{ij}dt_{ij}} = \theta\{C_\theta(u,v)\}^{2\theta-1}v^{-\theta}u^{-\theta}\big(-\varphi(s)\big)(-f_T(t)) \quad (44)$$

The log-likelihood with the **Clayton Copula** can thus be expressed as:

$$logL(\mu_i,\alpha_i,\sigma_i^2,\lambda_i,\beta_i,\rho_i) = \quad (45)$$

$$\sum_{i=1}^{c}\sum_{j=1}^{n_i}\{\delta_T \log[\theta\{C_\theta(u,v)\}^{2\theta-1}v^{-\theta}u^{-\theta}\big(-\varphi(s)\big)(-f_T(t))] + (1-\delta_T)\log[\{C_\theta(u,v)\}^{\theta}u^{-\theta}\varphi(s)]\}$$

that can be extended to:

$$logL(\mu_i,\alpha_i,\sigma_i^2,\lambda_i,\beta_i,\rho_i) = \quad (46)$$

$$\sum_{i=1}^{c}\sum_{j=1}^{n_i}\Big\{\delta_T\Big[\log\theta + \Big(\frac{2\theta-1}{1-\theta}\Big)log\big((S(s)^{1-\theta}+S(t)^{1-\theta}-1\big) - \theta\log S(s) - \theta logS(t) + log\varphi(s) + logf_T(t)\Big] +$$

$$+ (1-\delta_T)\Big[\frac{\theta}{1-\theta}log\big((S(s)^{1-\theta}+S(t)^{1-\theta}-1\big) - \theta log\, S(s) + log\varphi(s)\Big]\Big\}$$

2.2.4.3 The Hougaard copula

Tha Hougaard copula (Hougaard 1986) is given by

$$C_\theta(u,v) = exp\left\{-\left[(-\ln u)^{\frac{1}{\theta}} + (-\ln v)^{\frac{1}{\theta}}\right]^{\theta}\right\} \quad (47)$$

and to obtain the corresponding log-likelihood function, we need the following derivates

$$\frac{dC_\theta(u,v)}{du} = C_\theta(u,v)\left\{-\theta\left[(-ln\,u)^{\frac{1}{\theta}} + (-ln\,v)^{\frac{1}{\theta}}\right]\right\}^{\theta-1}\left[\frac{1}{\theta}(-ln\,u)^{\frac{1}{\theta}-1}\right]\left(-\frac{1}{u}\right) \quad (48)$$

$$\frac{d^2C_\theta(u,v)}{du\,dv} = \quad (49)$$

$$C_\theta(u,v)\left\{-\theta\left[(-ln\,u)^{\frac{1}{\theta}} + (-ln\,v)^{\frac{1}{\theta}}\right]\right\}^{\theta-1}\left[\frac{1}{\theta}(-ln\,v)^{\frac{1}{\theta}-1}\right]\left(-\frac{1}{v}\right)$$

$$\left\{-\theta\left[(-ln\,u)^{\frac{1}{\theta}} + (-ln\,v)^{\frac{1}{\theta}}\right]\right\}^{\theta-1}\left[\frac{1}{\theta}(-ln\,u)^{\frac{1}{\theta}-1}\right]\left(-\frac{1}{u}\right) +$$

$$+C_\theta(u,v)\left\{-\theta(\theta-1)\left[(-ln\,u)^{\frac{1}{\theta}} + (-ln\,v)^{\frac{1}{\theta}}\right]\right\}^{\theta-2}\left[\frac{1}{\theta}(-ln\,v)^{\frac{1}{\theta}-1}\right]\left(-\frac{1}{v}\right)$$

$$\left[\frac{1}{\theta}(-ln\,u)^{\frac{1}{\theta}-1}\right]\left(-\frac{1}{u}\right)$$

$$\frac{dS_T(t_{ij})}{dt_{ij}} = \frac{d\,(1-F(t))}{dt_{ij}} = -\frac{d\,F(t)}{dt_{ij}} = -f_T(t) \tag{50}$$

$$\frac{dS_s(s_{ij})}{ds_{ij}} = \frac{dC_\theta(u,v)}{du}\frac{dS_s(s_{ij})}{ds_{ij}} = \tag{51}$$

$$= \frac{dC_\theta(u,v)}{du}\,\frac{d\,(1-\Phi(s))}{ds_{ij}} = \frac{dC_\theta(u,v)}{du}(-\varphi(s))$$

The log-likelihood function can be written as:

$$logL(\mu_i, \alpha_i, \sigma_i^2, \lambda_i, \beta_i, \rho_i) = \tag{52}$$

$$\sum_{i=1}^{c}\sum_{j=1}^{n_i}\Bigg\{\delta_T\Bigg[log\{C_\theta(u,v)\} + (\theta-2)log\left((-\ln u)^{\frac{1}{\theta}} + (-\ln v)^{\frac{1}{\theta}}\right)$$

$$+ log\left(\left((-\ln u)^{\frac{1}{\theta}} + (-\ln v)^{\frac{1}{\theta}}\right)^{\theta} - \frac{\theta-1}{\theta}\right)$$

$$+ \frac{1}{\theta-1}\log(-\log S(s)) - \log S(s) + \frac{1}{\theta-1}log(-\log S(t)) - log\, S(t)$$

$$+ log\varphi(s) + logf_T(t)\Bigg]$$

$$+ (1-\delta_T)\Bigg[log\{C_\theta(u,v)\} + (\theta-1)log\left((-\ln u)^{\frac{1}{\theta}} + (-\ln v)^{\frac{1}{\theta}}\right)$$

$$+ \frac{1}{\theta-1}log(-\log S(s)) + log\varphi(s)\Bigg]\Bigg\}$$

2.2.5 Choice of the trial units

A fundamental step of the meta-analytic method is the choice of the units involved in the analysis, for example, trials, centers, or investigators. Practical considerations, such as the information available in the data, the experts consideration on the most appropriate unit for a specific problem, the amount of replication at a potential unit's level, and the number of patients per unit, may determine this choice. From a technical point of view, the optimal situation is the one in which the number of trials and the number of patients per unit are sufficiently large (Cortiñas et al. 2004).
The meta-analytic approach was originally formulated taking trials as the level of replications, but when the trials are few, this approach is not applicable, thus suggesting the use of countries,

treating centers or investigators, as the units of analysis. This effectively should imply the extension to a three (or higher) level model.

Omitting one lever of hierarchy in the analysis can induce different consequences on the estimation of the strength of association, depending on various factors (e.g., the sample size and the magnitude of the variability present at different levels). Cortiñas et al (2004) investigated through simulations the choice of replacing one level of replication with another one and the consequences of replacing a truly higher level model by a gross lower level structure, e.g. trial, center and patient levels replaced by center and patient levels.

The strategy of analysis we adopted for this particular project recognizes the hierarchy of the data and considers a structure based on two level of observation: the trial and the geographical area. As an example, in assessing a time to event endpoint, we considered a model that produces, for each geographical area, separated treatment effects for each endpoint, but the baseline hazards are assumed to be constant within trial.

3 THE CASE-STUDY

Over the past few decades a better understanding of the biological bases of ALL was made and Minimal Residual Disease has been accepted for routine clinical use worldwide. MRD is now considered so relevant that it has become essential to study whether graduate it to the status of clinical endpoint. We addressed this issue considering individual data from three large phase III studies performed in Europe and in the USA.

The clinical context that motivated this work was introduced in this chapter, first focusing on MRD and then to the synthesis of the protocols that were considered. Space has also been reserved to the description of the trial units choice.

3.1 MINIMAL RESIDUAL DISEASE

Minimal Residual Disease (MRD) defines small numbers of leukemic cells that circulate in the patient during the treatment period, or after treatment, even when the patient is classified in remission, e.g. with no symptoms or signs of disease. MRD is used to diagnose the disease, since it is characterized by a high sensitivity, and it is helpful to monitor the disease and to identify the presence of tumour cells after therapy or determine the best timing for a stem cells transplantation. Multiple reports suggest that the detection of MRD at an early time point of the treatment protocol (during/following Induction or Consolidation therapy) is a powerful and independent predictor of prolonged Event Free Survival in children and adults with acute lymphoblastic leukaemia (Reviewed in Campana, 2010; Cazzaniga et al., 2011 and Pui et al., 2015).

As a prognostic biomarker, MRD has had a profound impact world-wide on the design and conduct of clinical trials in ALL for all known risk groups of patients, currently defined by longstanding, accepted clinical and biological prognostic factors. It has being used to risk stratify patients on therapeutic trials designed to adapt treatment for individual patients relative to their risk for treatment failure (Campana, 2009). Conter et al. (2010) and Schrappe et al. (2011) have shown that MRD-based treatment strategies improve outcome in BCP-ALL and T-ALL childhood patients, and Vora et al. (2013) demonstrated that treatment can be reduced in MRD-based low-risk patient.

Sensitive methods (1 leukemic cell in 10,000 to 100,000 normal cells in a clinical sample) for the detection of MRD uses assays based on the real time quantitative polymerase chain reaction (RQ-PCR) and/or flow cytometry (FCM). In the United States, FCM is the most common technique,

while in Europe PCR is used. The two methods are quite standard, but each has its own advantages: FCM is less expensive, can often report quantitative results within a day and has a larger evidence base (having been used in most US based trials), while PCR is more sensitive.

These two methods for MRD evaluation were compared in childhood LLA by Basso et al. (2009) and Gaipa et al. (2012). Basso et al. (2009) concluded that FCM is more powerful to predict relapses at day 15, the concordance between methods was very high at day 15 and 78 of induction therapy, lower at day 33, at the end of induction therapy. Gaipa et al. (2012) compare the methods at day 15, 33 and 78 and they found that concordance rates between FCM and PCR largely depend on the time point of the analyses and most discordances occur at the lowest levels of MRD. In particular, the concordance rates were 87%, 72% and 89% for the three time points, respectively. Recently, PCR and FCM were also compared for MRD evaluation in Multiple Myeloma and Chronic Lymphocytic Leukemia also (Puing et al 2014; Raponi et al 2014) and they show that these methodologies are both valid, but PCR is favored by a slightly higher sensitivity, whereas FCM is significantly more applicable.

Nowadays, there is also an emerging technique that is the "next-generation sequencing". It has greater sensitivity than PCR, but it works on the same principle amplifying all possible sequences in the gene region of interest. At diagnosis, patients get a signature for that particular leukemia that can be screened for in subsequent tests. This methodology can become the future of MRD detection, but it is confined to the research areas and there are still hurdles to be overcome before it is used in the clinical practice.

MRD is considered a well-founded prognostic biomarker in ALL, but still not reliable as a clinical endpoint. For example, an open-label randomized trial comparing the effect of mitoxantrone with idarubicin in children with first relapse of ALL found that, although the mitoxantrone-treated patients had a lower relapse rate, there was no apparent difference in MRD between the two drugs in the intermediate-risk group. This finding induced researchers to believe that the decrease in relapse was unrelated to the kinetics of disease clearance therefore MRD is not a surrogate for efficacy (Parker, 2010). The hope was that in presence of no difference in MRD levels one month after treatment initiation, there would also be no difference in EFS, but evidence did not support this hypothesis. Another research (Bassan, 2009), which tends to agree that MRD can be an early indicator for Relapse in adult patient, found that MRD analysis during early post-remission therapy improve risk definitions and help improve risk-oriented treatment strategies. Patients who were MRD-negative had five-year overall survival and disease-free survival rates of 0.75 and 0.72, respectively, compared with rates of just 0.33 and 0.14 in MRD-positive patients. Presence of MRD, then, was the most significant risk factor for relapse, with a hazard ratio of 5.22.

MRD in childhood ALL has the potential to be a surrogate clinical endpoint, but no formal validation has been performed yet. The use of MRD as a clinical endpoint is recognized as a burning issue also by the regulatory agencies for the purpose of accelerated approval of new drugs.

3.2 DESCRIPTION OF THE CLINICAL PROTOCOLS

The clinical project was planned to include three randomized multicentre phase III studies that were performed in different European countries and in the USA, namely: AIEOP-BFM-ALL2000, EORTC and AALL0232. Among the objectives of these trials, the comparison between dexamethasone (DXM) and prednisone (PDN), used in the induction phase of the treatment of children with acute lymphoblastic leukaemia, was carried out in a randomized fashion.

These protocols had a complex structure and involved the great majority of the newly diagnosed cases of ALL observed in the study period in the different country (except for COG that is one of the few collaborative groups in the USA and is represented here with a protocol for high risk patients). They were different in many aspects and, in particular, they were not completely homogenous also in the process of MRD monitoring, but they had in common a time point of observation around one month from treatment start.

We included in the analysis all patients randomized either to DXM or PDN with diagnosis of B-cell immunophenotype (for the purpose of the secondary aim) and with complete info on MRD one month after randomization (for the primary aim).

A brief description of the protocols of the trials that we will analyse follows.

AIEOP-BFM-ALL2000

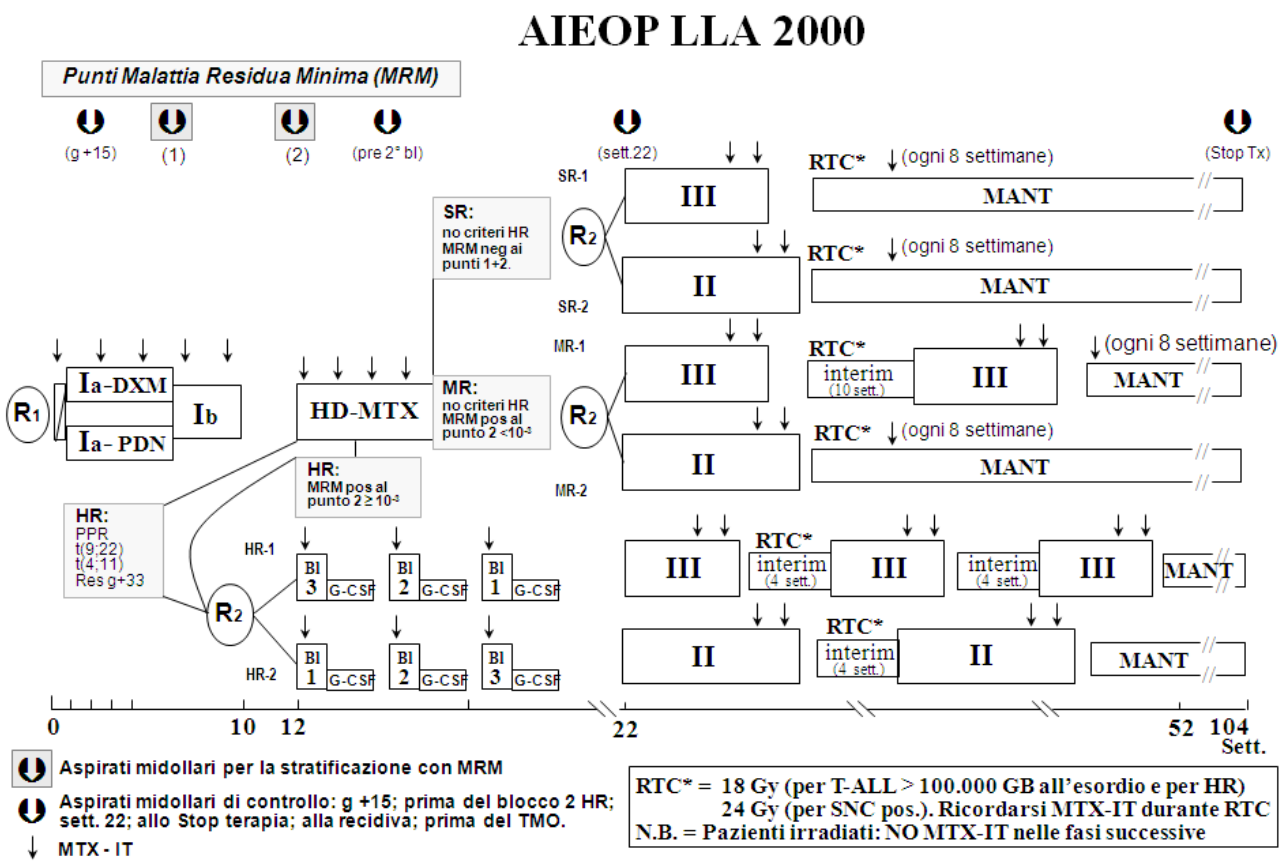
This is a collaborative trial, with the Italian (Associazione Italiana Ematologia e Oncologia Pediatrica – AIEOP) and the German (Berlin-Frankfort-Münsterx – BFM) groups sharing the same clinical protocols (Clin.Gov. registration codes are: NCT00613457 and NCT00430118 for the AIEOP and the BFM group, respectively). The AIEOP-BFM ALL 2000 study enrolled children between 1 and 18 years of age with Philadelphia negative ALL from September 1, 2000 to July 31, 2006. Randomization was stopped in 2005 for patients who were ≥10 years old because dexamethasone resulted more aggressive than PDN.

Diagnosis of ALL was performed using cytomorphology and cytochemistry when ≥ 25% of lymphoblastic cells were present in the bone marrow. All the enrolled patients entered the induction phase and were given a pre-phase of 7 day treatment, including steroid therapy

(prednisone) and one intrathecal dose of methotrexate, followed by induction protocol IA and induction consolidation protocol IB. On day 8, patients were randomized to continue steroid treatment with either prednisone (60 mg/m2 per day) or dexamethasone (10 mg/m2 per day) until day 28 with subsequent tapering doses. At the end of IA and IB phase, i.e. on day 33 (time point one – TP1) and 78 (time point two - TP2), MRD was evaluated with RQ-PCR method and the risk stratification was obtained. Patients were defined at MRD standard risk (MRD-SR) if MRD was found negative at both time points, using at least 2 molecular targets with a sensitivity of $\leq 10^{-4}$. If MRD levels differed between the 2 markers, the highest MRD level was chosen for the final MRD assessment. Patients were considered at intermediate risk (MRD-IR) when MRD was positive at one or both TPs, but at a level $<10^{-3}$ at TP2 with at least 2 markers. If MRD levels differed between the 2 markers, the highest MRD level was chosen for the final MRD classification, provided that the selected markers had a sensitivity of at least 10^{-3}. Patients with MRD $\geq 10^{-3}$ at TP2 were classified at MRD-HR, independently of the sensitivity and the number of markers (provided that at least 1 marker had a sensitivity of 10^{-3}). The treatment schemas by risk stratification based on MRD are summarized in Figure 3.1. MRD was measured through PCR, and available in continuous for AIEOP, while it was categorical for BFM. For the purpose of the analysis, we referred to MRD on day 33, that is the end of induction IA.

Other randomized questions were applied in the subsequent phase of this protocol, as shown in Figure 3.1. In short, 22 weeks after diagnosis, patients classified at Standard Risk were randomized to receive the standard reinduction protocol (II) or a less intensive treatment (III), while Medium Risk patients received a double reintensification with twice block III vs. one block II. The polychemiotherapy treatments in the standard and the medium risk groups are similar and less intense than the treatment in the high risk group. In fact, patients who did not achieve CR at the end of induction phase IA (or have translocations t(9,22)/t(4,11) or were PDN poor responders at day 7) were treated with phase IB of protocol I, and 3 subsequent high-risk (HR) blocks. Finally, specific indications to Bone Marrow Transplantations were set up in the protocol, which were also based on the MRD evaluated at TP1 or TP2.

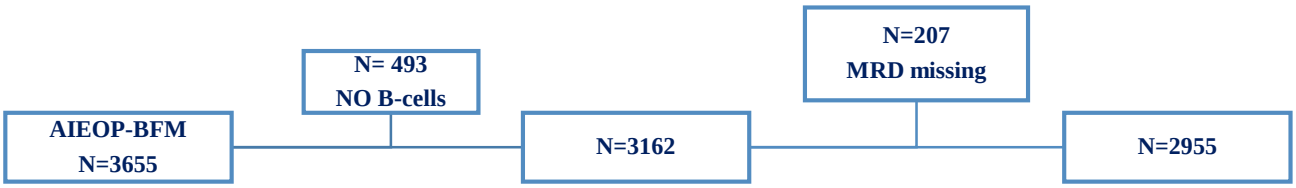
Figure 3.1: AIEOP-BFM experimental schema



The total number of AIEOP-BFM ALL patients randomized in 6 years of recruitment was 3655, 1192 were from AIEOP and 2463 from BFM. Excluding 493 not B-ALL precursor and 207 children with MRD missing at TP1, a final set of 2955 randomized patients was considered (Figure 2.2). Therefore, for the validation of MRD as surrogate for EFS, we analyzed 2955 patients for the AIEOP-BFM trial, while for the validation of EFS as surrogate of OS, 3162 patients were considered for the analysis.

Follow-up was uniformly updated as of September 2011.

Figure 3.2 AIEOP-BFM patients selection



AALL0232

AALL0232 is the Children's Oncology Group (COG) phase III study designed for NCI high risk ALL and B-precursor ALL patients from 1-30 years of age. The study utilized a 2 x 2 factorial design with an augmented intensity BFM backbone. Patients were randomized upfront to receive high dose methotrexate (5 gm/m) versus Capizzi escalating methotrexate during Interim Maintenance I. A second randomization compared dexamethasone 10 mg/m/day for 14 days versus prednisone 60 mg/m/day for 28 days during Induction. Based upon an increased rate of osteonecrosis (ON) observed in patients ≥10 years of age randomized to receive dexamethasone during Induction, an amendment restricted the Induction steroid randomization to patients 1-9 years of age, with older patients non-randomly assigned to prednisone during Induction therapy. MRD in peripheral blood was detect at day 29 (end of the induction) with Flow Cytometry and it was available as a continuous variable. Patients with negative MRD at day 29 and with M1 were classified as rapid early responders (RER) and received one Delayed Intensification course. Those with MRD at day 29 major of 1% (positivity status) and no M1 were classified as slow early responders (SER) and received two Delayed Intensification courses. The treatment protocols by risk stratification is summarized in Figure 3.3.

The AALL0232 protocol counts 2909 B-precursor patients enrolled in 7 years, from 2004 to 2011. We analysed 2023 patients because 782 children aged ≥10 years were not randomized (they received Prendisone by default) and 104 did not have the MRD value available (Figure 3.4). For the secondary aim, i.e. evaluate if EFS is a surrogate of OS, from 2127 subjects we excluded 34 patients who had data on Complete Remission missing, thus considering a total of 2093 patients.

Having AALL0232 a 2 x 2 factorial design with a quantitative interaction (results are not yet published), we considered patients treated with high dose methotrexate and Capizzi separately in our analysis.

Follow-up was uniformly updated to December 2014.

Figure 3.3 COG experimental schema

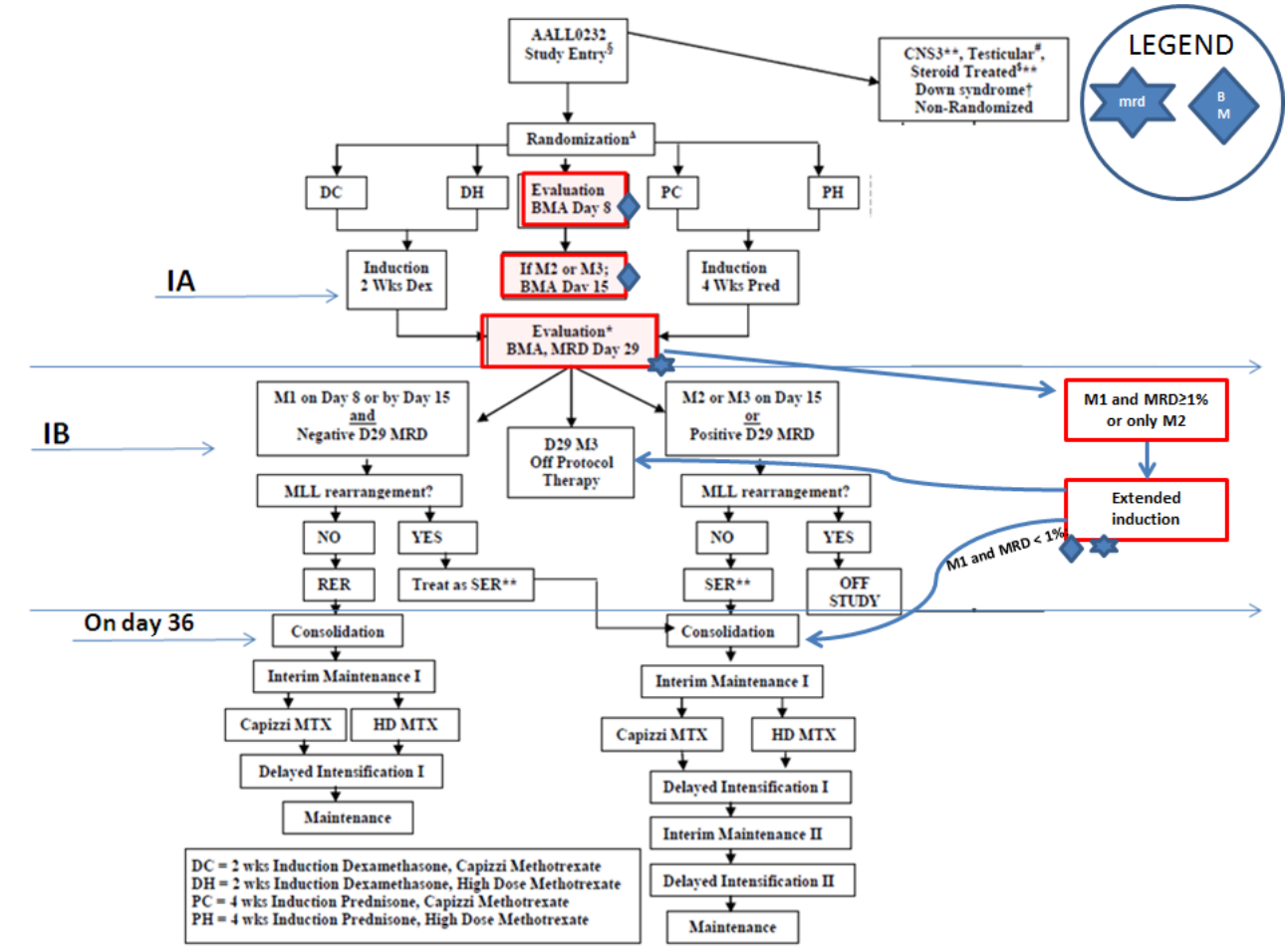
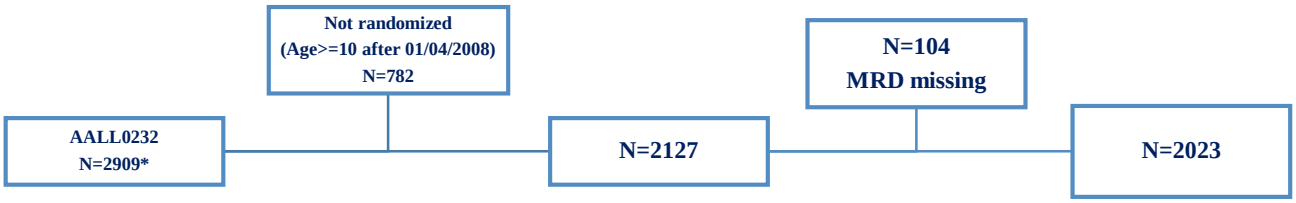


Figure 3.4 COG patients selection



*There was no exclusion related to the criteria on immunophenotype as the protocol enrolled exclusively B-ALL patients.

EORTC

The European Organization for Research and Treatment of Cancer (EORTC) is a non-profit Belgian research organization. In 1998, EORTC started a randomized trial in ALL where patients under 18 years of age with previously untreated ALL were eligible. Concerning the randomization

dexamethasone versus prednisolone, the patients were randomly assigned either before the beginning of the pre-phase (day 1), or at the beginning of protocol IA (day 8), at the investigator's discretion. In the latter case, prednisolone was used throughout the pre-phase. All patients had to receive dexamethasone (6 mg/m/day) or prednisolone (60 mg/m/day), orally, in two divided doses throughout the pre-phase (day 1 to day 7) and induction therapy (day 8 to day 35, including a tapering down period of 8 days). Minimal residual disease was evaluated at the end of induction (day 35) and was based on RQ-PCR method that returned data in categories.

Patients were assigned to different risk groups mostly depending on clinical features and also on MRD: very low risk (VLR), average risk (AR) and very high risk (VHR). VLR was defined as B-lineage ALL with no VHR criteria, with WBC counts below $10x10^9$/L, and with hyperdiploid karyotype or DNA index >1.16 and <1.5, in the absence of CNS and gonadal involvement. VHR criteria consisted of blast count in peripheral blood $\geq 1x10^9$/L at completion of the pre-phase (day 8), presence of t(9;22), t(4;11) or another MLL rearrangement, near-haploidy, acute undifferentiated leukaemia (AUL), MRD>10^{-2} (positivity cut off) at completion of induction (day 35) or failure to achieve complete remission (CR). AR patients had no VLR and VHR characteristics. Patients with B-cell lineage ALL, with WBC <$100x10^9$/L and without gonadal and CNS involvement were AR1. The others, including T-cell lineage, were classified as AR2. Patients with CNS-2 or with haemorrhagic cerebrospinal fluid becoming negative on day 4 of the pre-phase were included in AR1 group whereas non-equivocal CNS involvement at diagnosis or any CNS involvement on day 4 were considered as AR2. The treatment protocols by risk stratification is summarized in Figure 3.5 (Domenech et al. 2014).

One thousand and four hundred randomized patients from the 1947 initial subjects enrolled in 9 years from 1999 to 2008: 297 were excluded because they were not B-ALL precursors and 250 because MRD was missing (Figure 3.6). To analyze the surrogacy of EFS, we considered 1650 patients, while the surrogacy of MRD was evaluated based on 1400 patients.

Follow-up was uniformly updated to February 2011.

Figure 3.5: EORTC experimental schema

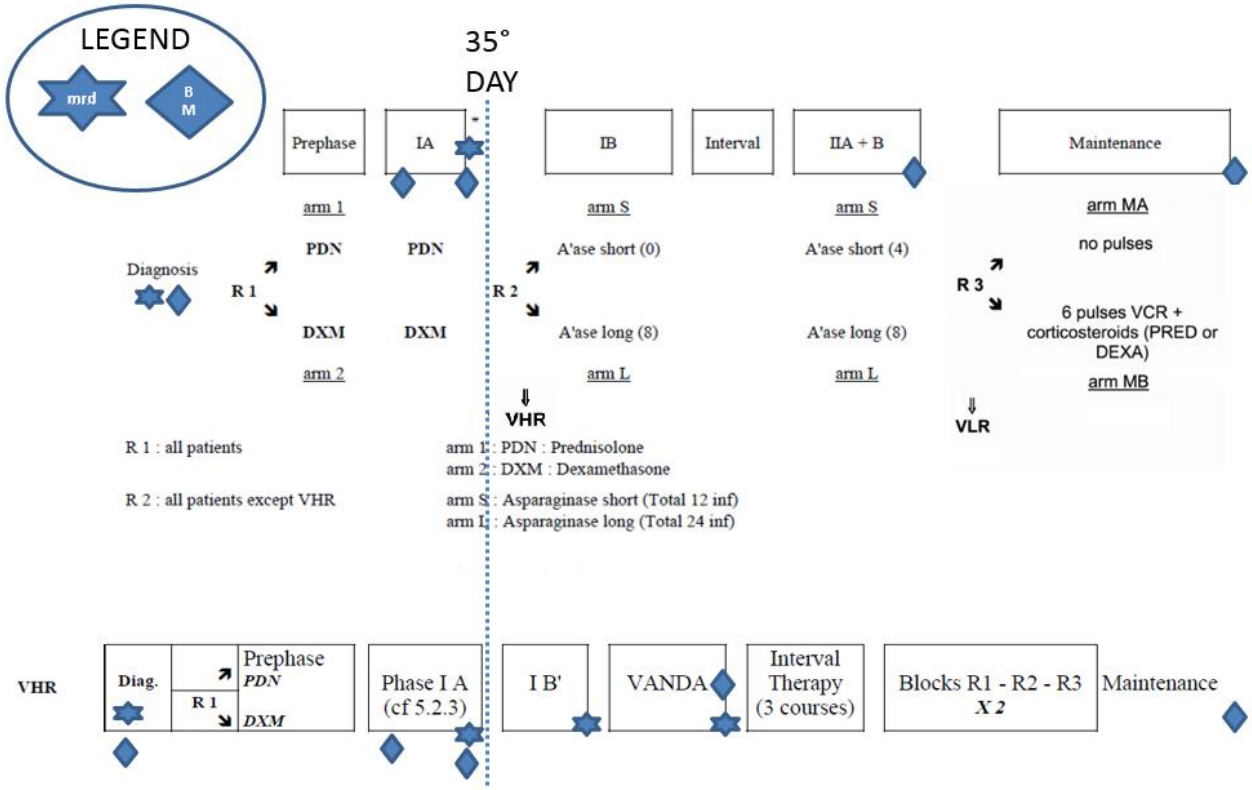
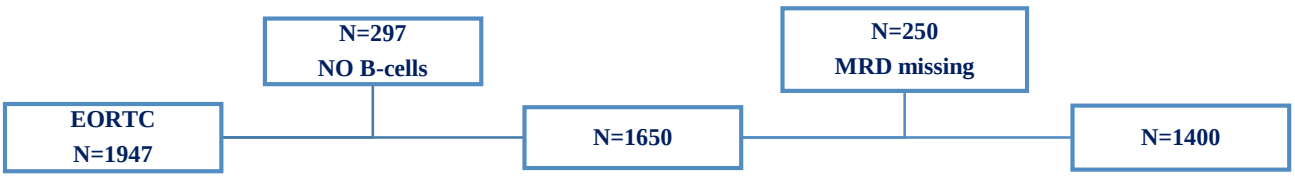


Figure 3.6: EORTC patients selection



The characteristics of the two treatments of interest and the main MRD features are summarized in Table 3.1. Dosing and duration are more homogeneous for PDN than DXM, while the time point for MRD evaluation differs in the three protocols. Another important difference is that in the COG only (NCI) high risk patients were randomized and MRD was detected with Flow Cytometry, not with PCR, as in AIEOP-BFM and EORTC.

Our point of view is that these differences are not a limitation, but represent an added value of the meta-analytic approach in terms of the generalizability of the results to future clinical trials and treatments.

Table 3.1: Characteristics of the two induction phase treatments and MRD features

TRIAL	TREATMENT				MRD	
	DXM		PDN		Timing (days from random)	Assay
	Dose	Duration (days)	Dose	Duration (days)		
AIEOP-BFM	10mg/mq/die	8-28	60mg/mq/die	8-28	+33	PCR
COG	10 mg/mq/die	1-14	60mg/mq/die	1-28	+29	FCM
EORTC	6mg/mq/die (bid)	1/8-35	60mg/mq/die (bid)	1/8-35	+35	PCR

3.3 ENDPOINT DEFINITIONS

The time to event endpoints of the study are defined as follows:

- Event Free Survival (EFS) is the time from randomization to first failure or last follow up, which over occurred first. The event we considered are: resistance, relapse at any site, development of a second malignant neoplasm (SMN) or death during remission.
- Overall Survival (OS) is the time from randomization to death from any cause or last follow-up, whichever occurred first.

As for MRD, we aimed at the evaluation of three ordered categories, with MRD full stratification as follows:

MRD class	MRD stratification
0	Negative
$0\text{-}5x10^{-4}$	Low positive
$\geq 5x10^{-4}$	Positive

In the Italian and the American clinical protocol, MRD was collected in continuous, while we had only partial information from the German and the Belgian clinical trials, since it was available categorized. Specifically, BFM used a 7 ordered categories classification: Negative (MRD=0), 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, where for example, the 10^{-3} category consist in the $5x10^{-3}<MRD<5x10^{-2}$ interval.

Unlike the BFM group, EORTC had a very different categorization of MRD (Table 3.2). We tried to harmonize the data considering two MRD cut-points for positivity, (i.e. $\geq 5x10^{-4}$ and $\geq 5x10^{-3}$). It resulted in a unsatisfactory classification that completely missed patients with MRD negativity.

Table 3.2: Description of the attempt to match EORTC and BFM MRD classification

EORTC classification MRD classes	Cut-points in agreement with BFM	
	Positivity: $\geq 5x10^{-4}$	Positivity: $\geq 5x10^{-3}$
10^{-3}-$5x10^{-3}$	NK	NK
$<5x10^{-3}$	Positive	Positive
$<10^{-3}$	NK	Not Positive
$5x10^{-3}$-10^{-2}	NK	Not Positive
$<10^{-2}$	Positive	Positive
$\geq 10^{-2}$	Positive	Not Positive

For this reason, we excluded EORTC from the validation of MRD while we used the data from this trial only to validate EFS for OS.

3.4 CHOICE OF THE TRIAL UNITS

Being this application on three trials only, the regions/states of the four participant Countries were considered as statistical units, as suggested in Burzykowski T. et al. (2005). We first considered the treating center, but since it is mandatory for the analysis that at least two categories of the ordinal surrogate and one event of the time to event endpoint should be present in each treatment group, centers were grouped according to geographical areas. Where the sample size was very limited, we aggregated regions/states based on proximity.

A total of 46 units were included in the analysis and these will be mentioned henceforth as "trial units". A description of the trial units within trial is reported in Table 3.3. We had 10-11 trial units in each trial except for EORTC, and heterogeneous sample sizes. The peculiarity of EORTC is related to the fact that we only had available the information on the countries involved, i.e. Belgium, France and Portugal.

Table .3.3: Description of the trial units within trials

		TRIAL					Overall
		AIEOP	BFM	COG-C*	COG-HD*	EORTC	
	N of trial-units	10	11	11	11	3	46
MRD vs EFS	Median pts/unit	109	143	92.5	91.4	349	104
	Min-Max pts/unit	76-222	54-343	46-147	52-158	93-958	46-958
EFS vs OS	Median pts/unit	115	148	96	95	407	114
	Min-Max pts/unit	80-240	58-369	47-156	57-167	121-1122	47-1122

* COG-C= COG-Capizzi COG-HD= COG-High Dose

The detailed list of the trial units is shown in Table 3.4, where it can be noticed that the units varied considerably in size, ranging from 46 (0.7%) to 958 (15%) (or 353 excluding the EORTC) patients. In Figures 3.7a-c the geographical distributions of the units in each country are represented.

Table 3.4: Detailed list of the trial units within trials

Trial	*Center ID*	*Trial Units*	**MRD vs EFS**		**EFS vs OS**	
			n	***%***	***n***	***%***
AIEOP	**1**	Emilia Romagna	76	1.19	80	1.16
	2	Campania	138	2.16	152	2.20
	3	Lazio	114	1.79	122	1.77
	4	Liguria/toscana	128	2.01	133	1.93
	5	Lombardia	222	3.48	240	3.48
	6	Marche/Abruzzo/Umbria	89	1.40	93	1.35
	7	Piemonte	104	1.63	108	1.56
	8	Puglia/Calabria	93	1.46	96	1.39
	9	Sicilia/Sardegna	146	2.29	162	2.35
	10	Veneto/Trento	82	1.29	89	1.29
BFM	**11**	Austria	219	3.43	232	3.36
	12	Baden-Württemberg	262	4.11	279	4.04
	13	Bayern	197	3.09	205	2.97
	14	Brandenburg/Berlin	143	2.24	148	2.14
	15	Mecklenburg-Vor/Schleswig-Hols	58	0.91	62	0.90
	16	Niedersachsen	128	2.01	140	2.03
	17	Nordrhein-Westfalen	343	5.38	369	5.34
	18	Rheinland-Pf/Hessen/Saarland	201	3.15	215	3.11
	19	Sachsen	54	0.85	61	0.88
	20	Sachsen-Anhalt/Thüringen	54	0.85	58	0.84
	21	Switzerland	104	1.63	118	1.71
COG – HIGH D.	**22**	H CA	136	2.13	138	2.00
	23	H DE/MD/NJ/PA	89	1.40	93	1.35
	24	H FL/AL/MS/GA/SC	88	1.38	95	1.38
	25	H ID/NV/OR/WA/CANADA	114	1.79	119	1.72
	26	H IL/IN/OH/WI/MI	158	2.48	166	2.40
	27	H KS/MO/OK/AR/HI/TX/LA	115	1.80	118	1.71
	28	H Miscellanea	72	1.13	74	1.07
	29	H ND/SD/MN/IA/NE	54	0.85	56	0.81
	30	H NY/VT/NH/ME/MA/CT	61	0.96	62	0.90
	31	H TN/NC/VA/KY/WV	66	1.03	68	0.98
	32	H UT/CO/NM/AZ	52	0.82	54	0.78

Trial	*Center ID*	*Trial Units*	**MRD vs EFS**		**EFS vs OS**	
			n	***%***	***n***	***%***
COG - CAPIZZI	**33**	C CA	147	2.30	153	2.22
	34	C DE/MD/NJ/PA	92	1.44	94	1.36
	35	C FL/AL/MS/GA/SC	99	1.55	103	1.49
	36	C ID/NV/OR/WA/CANADA	106	1.66	110	1.59
	37	C IL/IN/OH/WI/MI	146	2.29	152	2.20
	38	C KS/MO/OK/AR/HI/TX/LA	123	1.93	125	1.81
	39	C Miscellanea	61	0.96	62	0.90
	40	C ND/SD/MN/IA/NE	50	0.78	51	0.74
	41	C NY/VT/NH/ME/MA/CT	64	1.00	67	0.97
	42	C TN/NC/VA/KY/WV	84	1.32	86	1.25
	43	C UT/CO/NM/AZ	46	0.72	47	0.68
EORTC	**44**	Belgium	349	5.47	407	5.89
	45	France	958	15.02	1122	16.25
	46	Portugal	93	1.46	121	1.75

Figure 3.7a: EORTC trial units

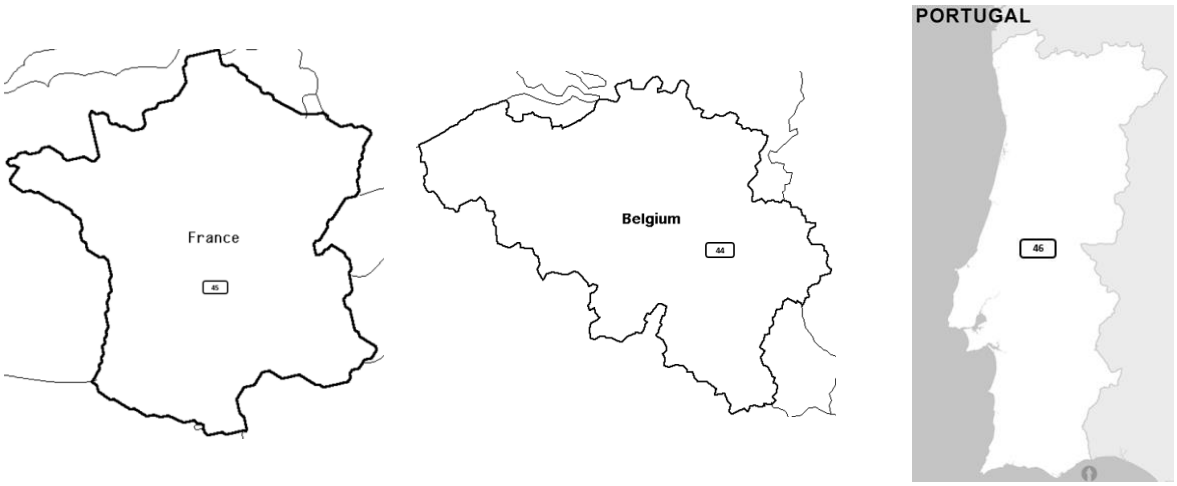


Figure 3.7b: COG trial units

CANADA

USA

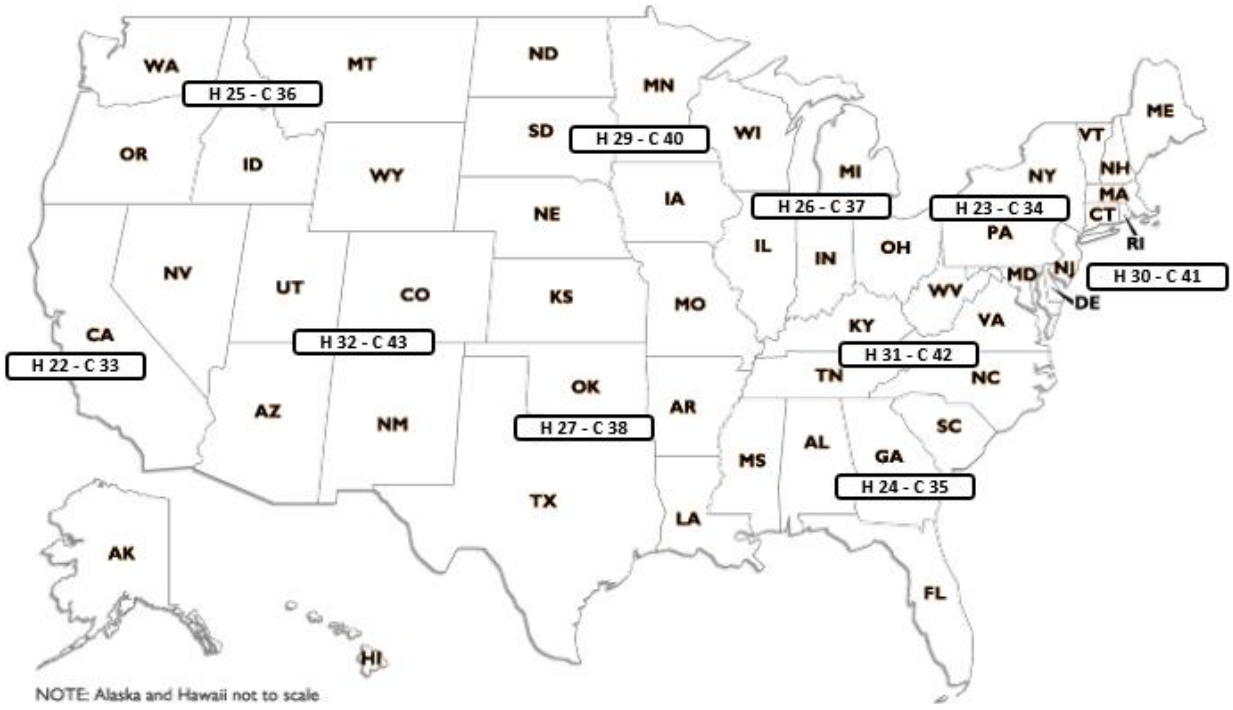
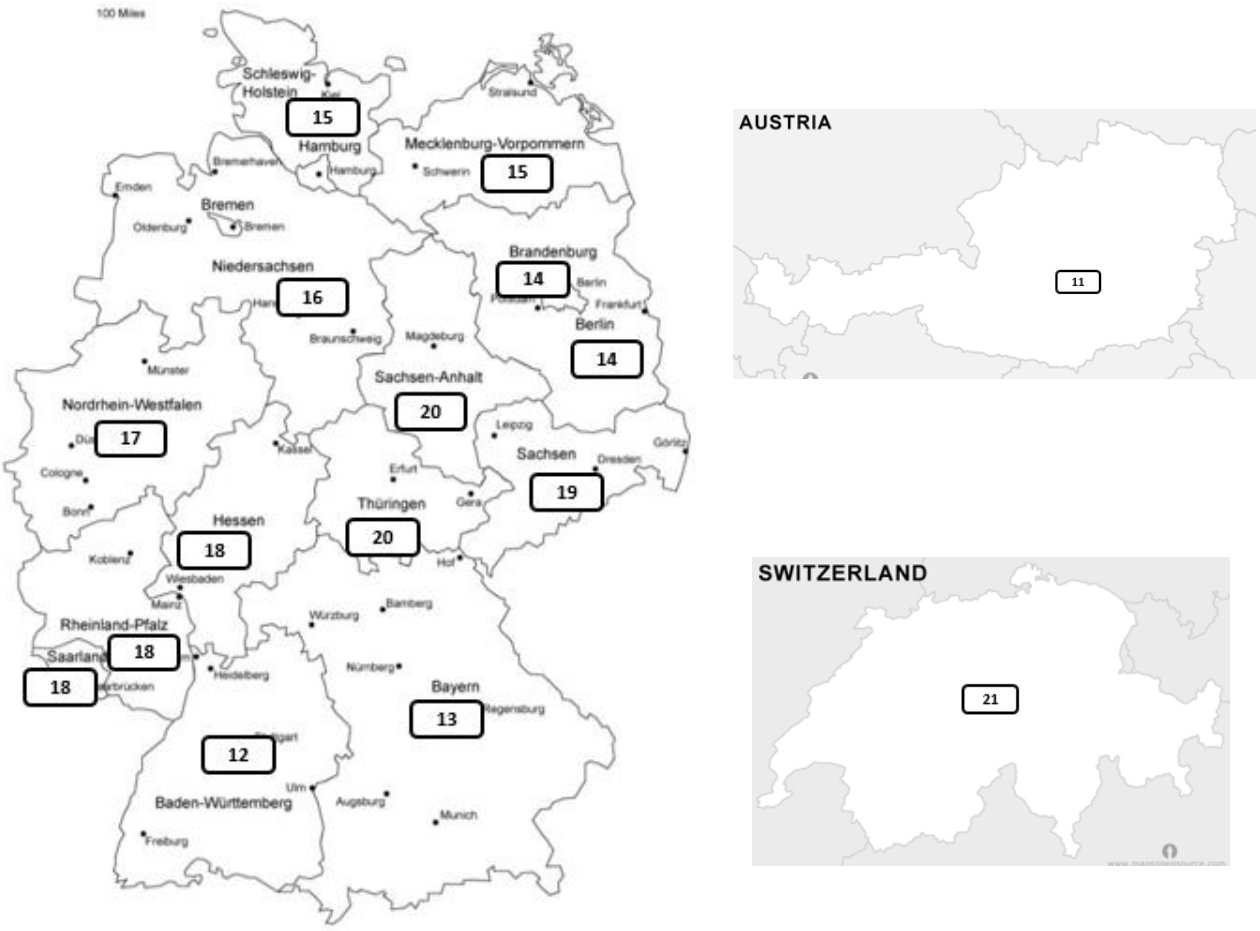


Figure 3.7. C: AIEOP-BFM trial units

GERMANY



ITALY

4 RESULTS

A brief description of the data involved in this project and the results of the validation process of both MRD and EFS as surrogate endpoints in the context of ALL are reported in this chapter. While the meta-analytic approach was used both for MRD and EFS, applying all the methods described in chapter 2, the traditional, yet controversial, assessment mainly based on Prentice Criteria was mainly considered as explorative for the validation of MRD.

4.1 DESCRIPTION OF THE DATA

The total number of patients screened from the selection process described in section 3.2 to be (potentially) included in the analyses are 6378 and 6905 for the primary and secondary objective of this thesis, respectively. The description that follows is shown by group and overall, and it is focused on the set of 6378 patients eligible for MRD validation. The results on the larger sample of 6905 patients eligible for EFS validation are not reported here, but are very similar.
From Table 4.1 it can be seen that the AIEOP-BFM group is the most represented and that DXM and PDN are uniformly distributed in each protocol. The global median follow-up time is 8.01 years, with a minimum observation of 6.77 years for EORTC to a maximum of 9.10 years for AIEOP-BFM.

Table 4.1: Distribution of patients and median Follow up within trials

Trial	N of patients			Median
	DXM	PDN	Total	Follow-up (y)
AIEOP-BFM	1460	1495	2955	9.10
COG	995	1028	2023	6.89
EORTC	699	701	1400	6.77
TOTAL	3154	3224	6378	8.01

The main clinical features reported in Table 4.2 are well balanced in the two treatment groups within each trial. Of note, AIEOP-BFM and EORTC patients are similar, while those from COG are older and with higher level of WBC, in line with the fact that the protocol was directed to NCI high risk ALL patients.

As anticipated in the previous chapter (see Section 3.2), we used individual data from different trials depending on the objective of the analysis. MRD validation was performed on 2 out of the 3 groups

we considered (i.e. AIEOP-BFM and COG), due to the impossibility to harmonize MRD data from EORTC, while EFS validation was done including also EORTC.
Ultimately, to assess whether MRD evaluated at the end of induction can be considered a surrogate for EFS, 4978 childhood B-lineage ALL patients, who were treated in induction with Dexamethasone or Prednisone, were considered, while, to explore if Event Free Survival can be used as a surrogate for Overall Survival, 6905 childhood B-lineage ALL patients were involved in the analysis.

Table 4.2: Clinical and biological features within trials

Characteristics	AIEOP-BFM N=2955				COG N=2023				EORTC N=1400				Total N=6378	
	DXM		PDN		DXM		PDN		DEXA		PDN			
	N	%	N	%	N	%	N	%	N	%	N	%	N	%
Total	1460	49.4	1495	50.6	995	49.2	1028	50.8	699	49.9	701	50.1		
Sex														
Male	804	55.1	757	50.6	534	53.7	554	53.9	371	50.2	368	49.8	3388	53.1
Female	656	44.9	738	49.4	461	46.3	474	46.1	328	49.6	333	50.4	2990	46.9
Age, y														
1-9	1256	86.0	1277	85.0	448	45.0	451	43.9	556	50.5	546	49.5	4534	71.1
>=10	204	14.0	218	15.0	547	55.0	577	56.1	143	48.0	155	52.0	1844	28.9
Age,y; mean (SD)	5.6	(3.7)	5.8	(3.8)	9.7	(5.8)	9.9	(5.7)	6.1	(4.1)	6.4	(4.2)	7.1	4.9
WBC count, x10^9/L														
Lower than 50	1252	85.8	1283	85.8	457	45.9	478	46.5	595	85.1	615	50.8	4680	73.4
50 or higher	208	14.2	212	14.2	538	54.1	550	53.5	104	14.9	86	45.3	1698	26.6
WBC count, x10^9/L median (min,max)	9.4	(0.4-567)	9.8	(0.1-875)	54.3	(0.3-1132)	53.8	(0.3-1306)	7.7	(0.4-424)	7.5	(0.5-480)	12.3	(0.1-1306)

4.2 VALIDATION OF MRD FOR EFS

Of the total sample of 4978 patients actually included in the analysis on MRD, 2455 (49%) and 2533 (51%) were randomized to DXM and PDN, respectively.

The distribution of the MRD stratification is shown in Table 4.3 by treatment within trial unit. When possible the description was done considering the trial units, otherwise a less detailed summary included the three groups/protocols, namely AIEOP-BFM and COG, with this latter separated in COG-Capizzi and COG-High Dose (see Section 3.2). The ID number in column 2 has been used in the graphical representation of the results to identify the trial units.

In general, a gradient seems to be present in the MRD distribution, with a negative MRD observed more frequently and a positive MRD less observed. This pattern can be consistently seen in all the trial units. Peculiarly, the Piemonte (ID=7) and Veneto/Trento (ID=10) units presented a very low percentage of MRD Positive patients in the DXM group, while in the Sachsen unit (ID=19) the percentage is zero. The MRD profiles in the two treatment groups are very heterogeneous.

Table 4.3: Decsription of MRD classification by treatment and trial unit

Trial	ID	Trial Unit	Treatment	Total	MRD		
					Negative	Low Positive	Positive
				N	%	%	%
AIEOP	1	Emilia R.	PDN	38	44.74	39.47	15.79
			DXM	38	44.74	36.84	18.42
	2	Campania	PDN	73	53.42	26.03	20.55
			DXM	65	50.77	33.85	15.38
	3	Lazio	PDN	72	45.83	25.00	29.17
			DXM	42	33.33	45.24	21.43
	4	Liguria/Toscana	PDN	62	50.00	33.87	16.13
			DXM	66	53.03	28.79	18.18
	5	Lombardia	PDN	103	40.78	33.01	26.21
			DXM	119	44.54	36.97	18.49
	6	Marche/Abruzzo/Umbria	PDN	45	46.67	35.56	17.78
			DXM	44	52.27	27.27	20.45
	7	Piemonte	PDN	56	37.50	39.29	23.21
			DXM	48	52.08	41.67	6.25
	8	Puglia/Calabria	PDN	51	43.14	27.45	29.41
			DXM	42	35.71	33.33	30.95
	9	Sicilia/Sardegna	PDN	74	43.24	40.54	16.22
			DXM	72	52.78	31.94	15.28
	10	Veneto/Trento	PDN	39	38.46	30.77	30.77
			DXM	43	69.77	27.91	2.33

Trial	ID	Trial Unit	Treatment	Total	MRD		
					Negative	Low Positive	Positive
				N	%	%	%
BFM	11	Austria	PDN	112	34.82	45.54	19.64
			DXM	107	44.86	41.12	14.02
	12	Baden-Württemberg	PDN	127	52.76	24.41	22.83
			DXM	135	54.07	28.89	17.04
	13	Bayern	PDN	87	55.17	20.69	24.14
			DXM	110	60.00	25.45	14.55
	14	Brandenburg/Berlin	PDN	69	44.93	30.43	24.64
			DXM	74	41.89	39.19	18.92
	15	Mecklenburg-Vor/Schleswig-Hols	PDN	28	46.43	28.57	25.00
			DXM	30	63.33	23.33	13.33
	16	Niedersachsen	PDN	67	49.25	29.85	20.90
			DXM	61	45.90	24.59	29.51
	17	Nordrhein-Westfalen	PDN	185	52.97	31.35	15.68
			DXM	158	55.06	31.01	13.92
	18	Rheinland-Pf/Hessen/Saarland	PDN	94	53.19	24.47	22.34
			DXM	107	53.27	36.45	10.28
	19	Sachsen	PDN	28	60.71	21.43	17.86
			DXM	26	69.23	30.77	.
	20	Sachsen-Anhalt/Thüringen	PDN	33	45.45	30.30	24.24
			DXM	21	42.86	33.33	23.81
	21	Switzerland	PDN	52	57.69	26.92	15.38
			DXM	52	51.92	28.85	19.23
COG HIGH D.	22	H CA	PDN	73	57.53	21.92	20.55
			DXM	63	74.60	9.52	15.87
	23	H DE/MD/NJ/PA	PDN	49	65.31	20.41	14.29
			DXM	40	60.00	15.00	25.00
	24	H FL/AL/MS/GA/SC	PDN	45	68.89	13.33	17.78
			DXM	43	53.49	25.58	20.93
	25	H ID/NV/OR/WA/CANADA	PDN	47	68.09	14.89	17.02
			DXM	67	64.18	16.42	19.40
	26	H IL/IN/OH/WI/MI	PDN	78	62.82	10.26	26.92
			DXM	80	71.25	11.25	17.50
	27	H KS/MO/OK/AR/HI/TX/LA	PDN	62	62.90	16.13	20.97
			DXM	53	71.70	9.43	18.87
	28	H Miscellanea	PDN	38	63.16	10.53	26.32
			DXM	34	73.53	14.71	11.76
	29	H ND/SD/MN/IA/NE	PDN	25	64.00	12.00	24.00
			DXM	29	55.17	20.69	24.14
	30	H NY/VT/NH/ME/MA/CT	PDN	37	78.38	16.22	5.41
			DXM	24	37.50	29.17	33.33
	31	H TN/NC/VA/KY/WV	PDN	31	64.52	22.58	12.90
			DXM	35	62.86	17.14	20.00
	32	H UT/CO/NM/AZ	PDN	27	66.67	14.81	18.52
			DXM	25	48.00	28.00	24.00

Trial	ID	Trial Unit	Treatment	Total	MRD		
					Negative	Low Positive	Positive
				N	%	%	%
COG Capizzi	33	C CA	PDN	79	64.56	13.92	21.52
			DXM	68	55.88	17.65	26.47
	34	C DE/MD/NJ/PA	PDN	45	62.22	22.22	15.56
			DXM	47	51.06	23.40	25.53
	35	C FL/AL/MS/GA/SC	PDN	56	55.36	21.43	23.21
			DXM	43	46.51	23.26	30.23
	36	C ID/NV/OR/WA/CANADA	PDN	54	61.11	14.81	24.07
			DXM	52	67.31	11.54	21.15
	37	C IL/IN/OH/WI/MI	PDN	73	69.86	12.33	17.81
			DXM	73	54.79	21.92	23.29
	38	C KS/MO/OK/AR/HI/TX/LA	PDN	60	55.00	21.67	23.33
			DXM	63	55.56	28.57	15.87
	39	C Miscellanea	PDN	35	65.71	11.43	22.86
			DXM	26	73.08	11.54	15.38
	40	C ND/SD/MN/IA/NE	PDN	23	82.61	8.70	8.70
			DXM	27	40.74	25.93	33.33
	41	C NY/VT/NH/ME/MA/CT	PDN	26	57.69	19.23	23.08
			DXM	38	57.89	23.68	18.42
	42	C TN/NC/VA/KY/WV	PDN	43	74.42	16.28	9.30
			DXM	41	48.78	24.39	26.83
	43	C UT/CO/NM/AZ	PDN	22	59.09	22.73	18.18
			DXM	24	58.33	25.00	16.67
		Total	PDN	1641	56.92	22.43	20.66
			DXM	1574	55.72	24.52	19.76

The distribution of the events that defines EFS are similar in the three protocols, as reported in table 4.4. DXM seems to better control the relapse occurrence, except for COG-Capizzi, where the rate of relapse is similar in the two treatment groups, with a slight advantage of PDN over DXM..

Table 4.4: Distribution of the events by trial and treatment

EFS events by treatment							
		DXM		PDN		Total	
Group	Event type	N	%	N	%	N	%
AIEOP-BFM	Resistant	13	0.9	16	1.1	29	1.0
	Relapses	159	10.9	239	16.0	398	13.5
	Deaths	17	1.2	20	1.3	37	1.3
	SMN	21	1.4	16	1.1	37	1.3
	Alive in CCR	1250	85.6	1204	80.5	2454	83.0
	Total	1460	100	1495	100	2955	100
COG - CAPIZZI	Resistant	9	1.8	12	2.3	21	2.1
	Relapses	96	19.1	85	16.5	181	17.8
	Deaths	14	2.8	22	4.3	36	3.5
	SMN	2	0.4	9	1.7	11	1.1
	Alive in CCR	381	75.9	388	75.2	769	75.5
	Total	502	100	516	100	1018	100
COG - HIGH D.	Resistant	10	2.0	11	2.1	21	2.1
	Relapses	61	12.4	89	17.4	150	14.9
	Deaths	14	2.8	17	3.3	31	3.1
	SMN	3	0.6	9	1.8	12	1.2
	Alive in CCR	405	82.2	386	75.4	791	78.7
	Total	493	100	512	100	1005	100

4.2.1 ONE-TRIAL APPROACH

We first explore here, in the first step for the evaluation of MRD surrogacy, if the four Prentice Criteria described in section 2.1 are satisfied. This kind of analysis is not a prerequisite for the meta-analytic approach. It is not a must and more importantly the meta-analytic approach can be applied regardless of the results of the analysis based on the Prentice Criteria (Burzykowski et al. 2015).
The results of the analyses on the 4 Prentice criteria below has been described point by point:

1. treatment affects the surrogate;
2. treatment affects the clinical end-point;
3. surrogate and clinical end-point are associated;
4. treatment effect disappears when adjusted by the surrogate.

1) TREATMENT AFFECTS THE SURROGATE

The first Prentice criteria is that there should be a significant impact of treatment on the surrogate endpoint (see model 1 in Section 2.1). We verified this criterion for each trial with a χ^2 test of association to assess whether the surrogate distribution differs in the two treatment groups and we quantified the level of association with a cumulative OR (higher vs lower level of MRD) obtained from a Proportional Odds model. In Table 4.5 is reported the MRD stratification by trial and treatment.

Table 4.5: MRD distribution by trial and treatment

Group			MRD			Proportional Odds model		χ2 test
			Negative	Low Positive	Positive	OR* (95% CI)	p-value	p-value
		N	%	%	%	DXM vs PDN		
AIEOP-BFM	Pdn	1495	47.8	30.8	21.4	0.83 (0.72-0.95)	0.006	0.001
	Dxm	1460	51.1	32.8	16.1			
COG	Pdn	1028	64.3	16.3	19.5	1.19 (1.003-1.42)	0.046	0.10
	Dxm	995	59.7	18.8	21.5			
COG Capizzi	Pdn	516	63.8	16.7	19.6	1.37 (1.07-1.74)	0.012	0.02
	Dxm	502	55.4	21.5	23.1			
COG High Dose	Pdn	512	64.8	15.8	19.3	1.03 (0.80-1.33)	0.790	0.97
	Dxm	493	64.1	16.0	19.9			

*Odds of Higher vs Lower MRD levels from a Proportional Odds model

In the AIEOP-BFM group the lower MRD categories are more represented (51.1% and 32.8 vs 47.8 and 30.8 for DXM and PDN, respectively) than the class with Positive MRD (16.1 vs 21.4). The test of association is significant and the OR was 0.83 (Table 4.5), indicating that DXM affect MRD positivity more than PDN. The opposite is seen in the COG group, where we have observed a higher percentage of lower MRD in PDN than in DXM (OR=1.19). When the analysis was performed in the Capizzi and High Dose separately, the difference in the MRD was driven by the Capizzi protocol (OR=1.37), while in the High Dose protocol the MRD distributions were very similar (OR=1). Thus, the first Prentice Criteria is satisfied for AIEOP-BFM and COG-Capizzi only, with the treatment effect pointing in two different directions.

2) TREATMENT AFFECTS THE CLINICAL END-POINT

The second Prentice Criteria, stating that a significant impact of treatment on the true endpoint is needed (see model 2 in section 2.1), was tested with a Log-rank test comparing the survival profile in the two treatment groups. Table 4.6 shows that in the AIEOP-BFM protocol, treatment has a significant effect on EFS with an HR of 0.71, indicating that DXM is better than PDN at least in preventing relapses (see also Table 4.4). This is illustrated in Figure 4.1, which represents the Kaplan-Meier (K-M) estimates of the EFS curves. In the COG-Capizzi trial, the two K-M curves are overlapping (Table 4.6 and Figure 4.2), whereas are significantly separated(p-values=0.01) in the COG-High Dose protocol, with a superiority of DXM. The second Criterion is thus satisfied for AIEOP-BFM and COG High Dose, but not for COG-Capizzi.

Table 4.6: EFS estimated at 5 years from randomization by trial and treatment

Group	EFS at 5y (95% CI)		Log-rank test	HR (95% CI)
	DXM	PDN	p-value	DXM vs PDN
AIEOP-BFM	0.88 (0.86-0.89)	0.83 (0.81-0.84)	0.0002	0.71 (0.60-0.86)
COG	0.80 (0.77-0.82)	0.77 (0.74-0.79)	0.03	0.82 (0.70-1.98)
COG-Capizzi	0.77 (0.73-0.81)	0.77 (0.73-0.80)	0.65	0.95 (0.73-1.21)
COG-High Dose	0.83 (0.79-0.86)	0.77 (0.73-0.80)	0.01	0.70 (0.54-0.92)

Figure 4.1: AIEOP-BFM - Event Free Survival curves by treatment

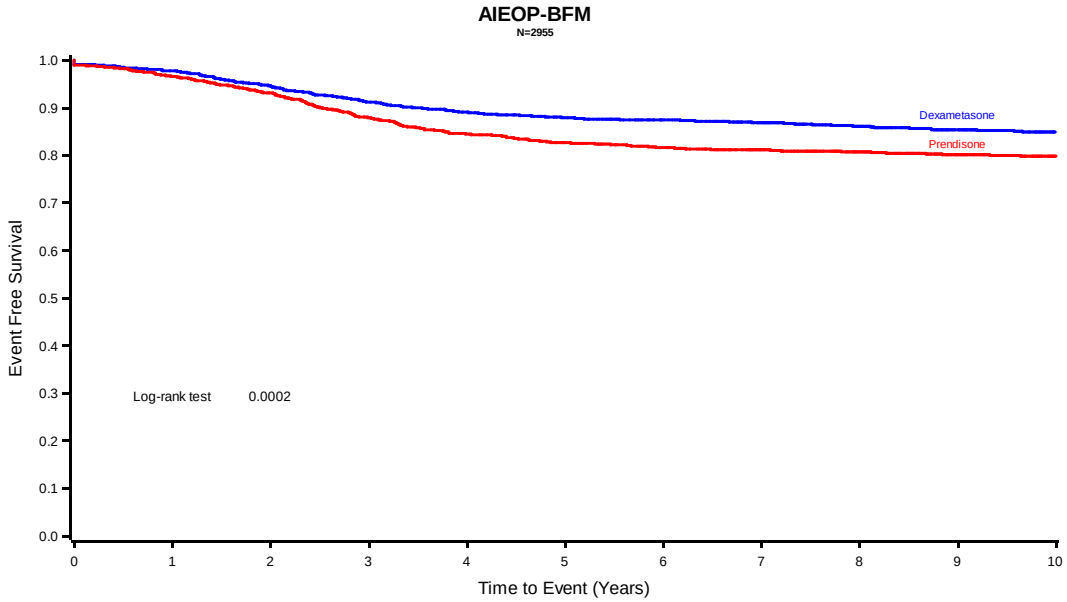


Figure 4.2: COG-Capizzi - Event Free Survival curves by treatment

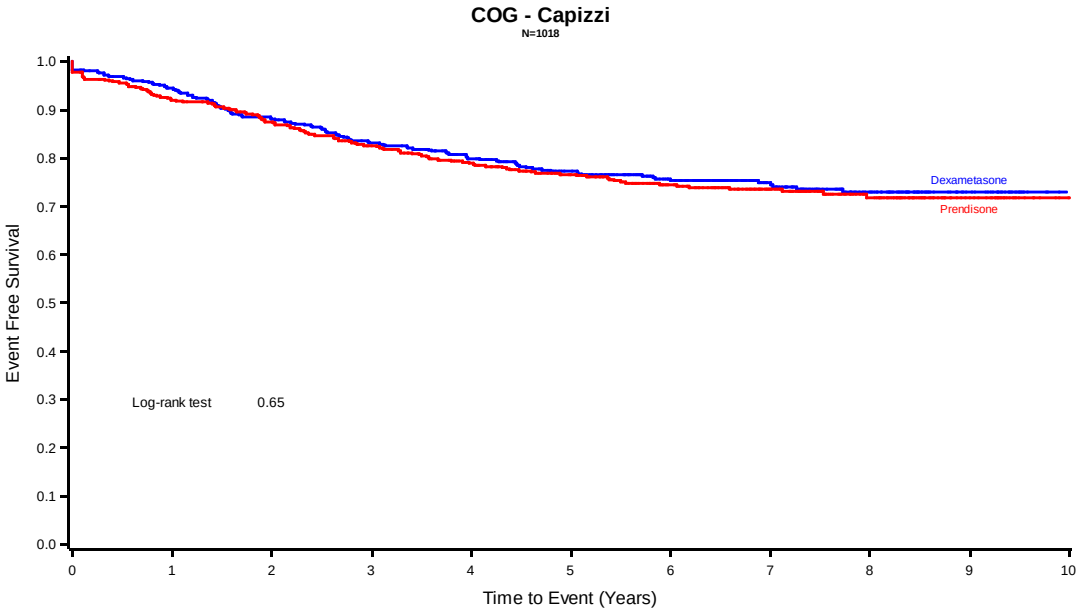
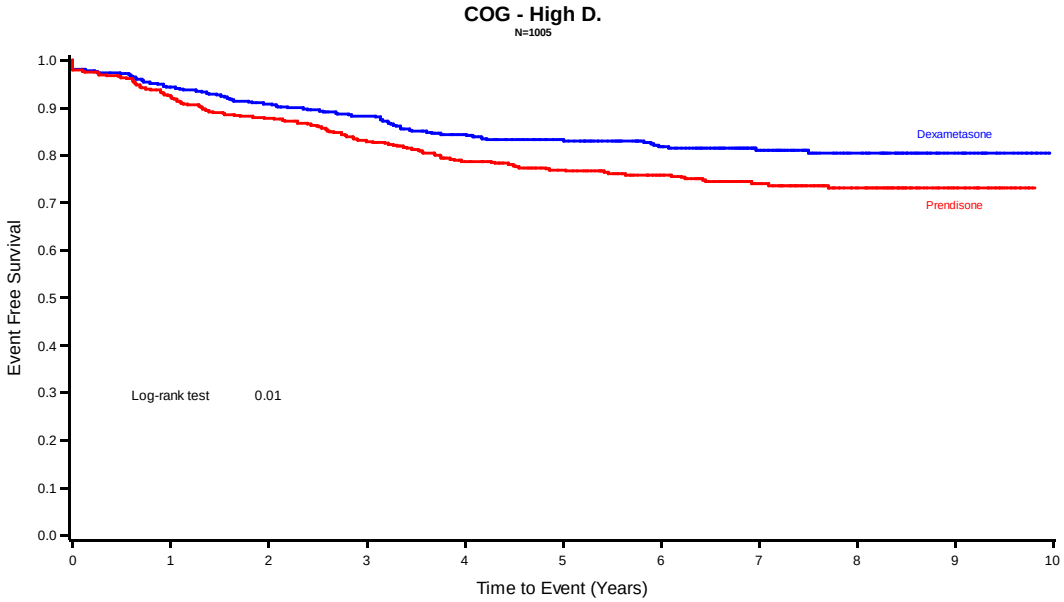


Figure 4.3: COG-High Dose - Event Free Survival curves by treatment



3) SURROGATE AND CLINICAL END-POINT ARE ASSOCIATED

The third Prentice Criterion, stating that a significant impact of the surrogate endpoint on the true endpoint (see model 3 in section 2.1), must be present, was verified by means of the Log-rank test (Table 4.7). The three Kaplan-Meier curves estimates of EFS reflecting MRD full stratification are represented for each protocol in Figures 4.4-4.6. Minimal Residual Disease significantly affected EFS in all the protocols, with a gradient that reflected the MRD pattern: patients with negative MRD had the best prognosis, while those with positive MRD had the worst prognosis.

Table 4.7: True endpoint (EFS) estimated at 5 years from diagnosis by trial and MRD classification

Group	Event Free Survival at 5y (95% CI)			Log-Rank test p-value	HR (95% CI) Low Positive vs Negative MRD	HR (95% CI) Positive vs Negative MRD
	Negative MRD	Low Positive MRD	Positive MRD			
AIEOP-BFM	0.93 (0.92-0.94)	0.85 (0.83-0.87)	0.64 (0.60-0.68)	<0.001	2.0 (1.6-2.5)	5.4 (4.3-6.7)
COG	0.87 (0.85-0.89)	0.77 (0.72-0.81)	0.54 (0.49-0.59)	<0.001	1.7 (1.3-2.3)	4.4 (3.4-5.4)
COG-Capizzi	0.86 (0.83-0.89)	0.75 (0.68-0.80)	0.54 (0.47-0.60)	<0.001	1.8 (1.3-2.6)	4.0 (3.1-5.3)
COG-High Dose	0.87 (0.85-0.90)	0.80 (0.73-0.86)	0.55 (0.48-0.62)	<0.001	1.6 (1.1-2.4)	4.8 (3.6-6.4)

Figure 4.4: AIEOP-BFM – Event Free Survival curves by MRD

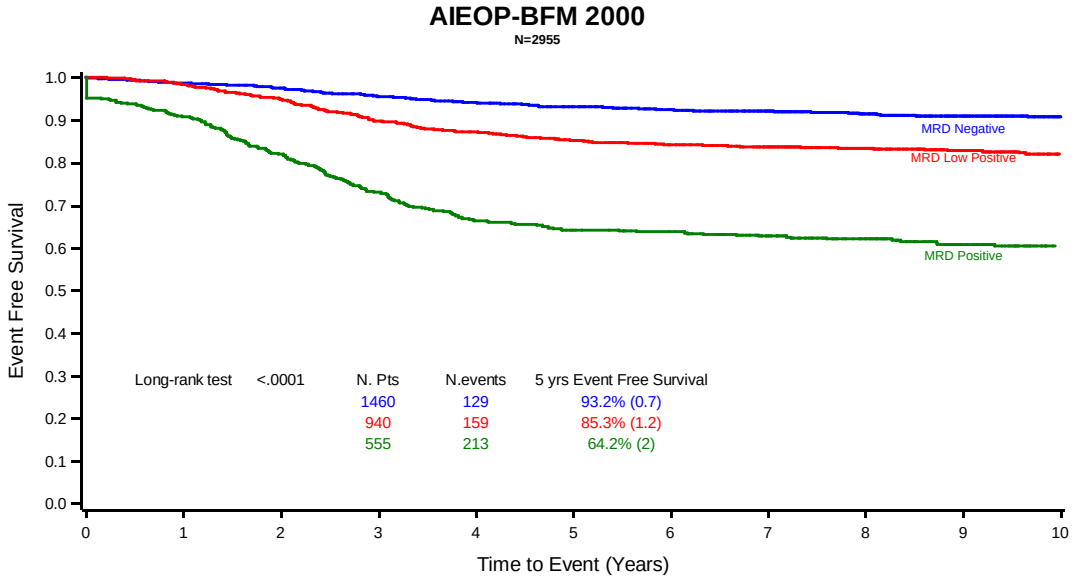


Figure 4.5: COG-Capizzi - Event Free Survival curves by MRD

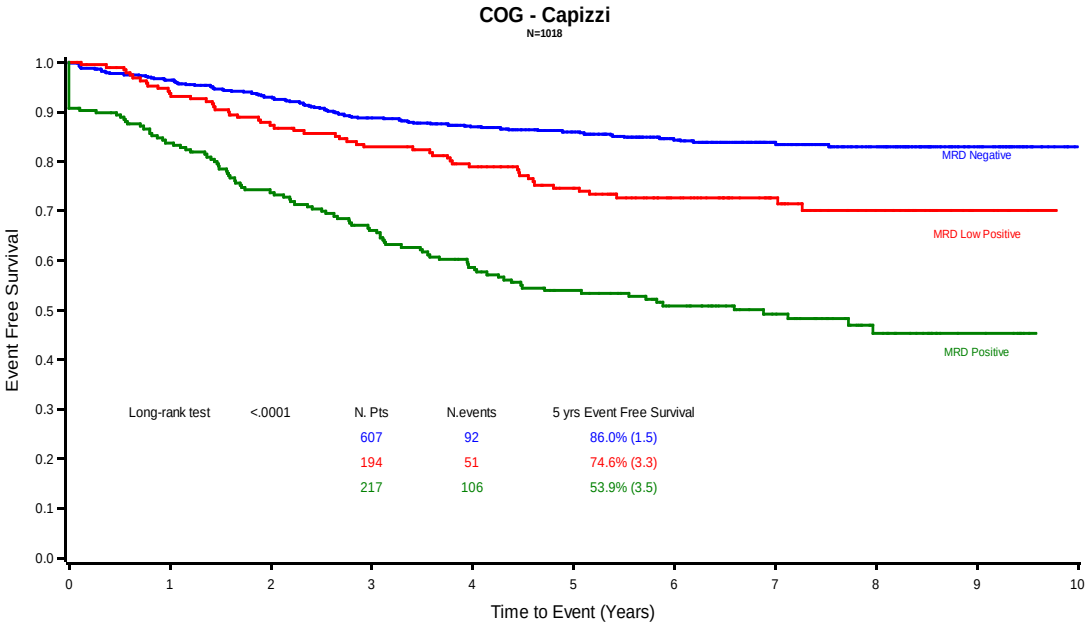
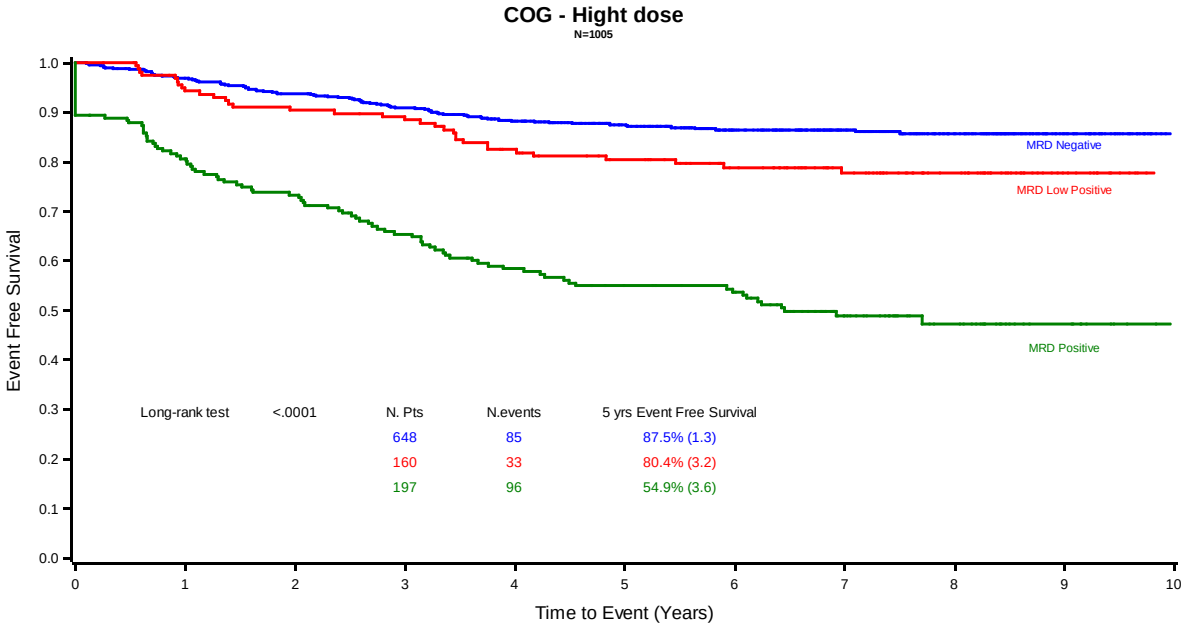


Figure 4.6: COG-High Dose - Event Free Survival curves by MRD



4) TREATMENT EFFECT DISAPPEARS WHEN ADJUSTED BY THE SURROGATE

The last and most important criterion states that the full effect of treatment upon the true endpoint has to be captured by the surrogate (see model 4 in section 2.1). To graphically check this statement the EFS curves by MRD stratification and treatment were depicted in Figures 4.7-4.9 and they were compared, within each level of MRD, with a Log-rank test.

In the AIEOP-BFM trial, only the Low Positive MRD captures the treatment effect, as the curves are very close and the Log-rank test is not significant. The same situation is depicted in the COG-High Dose group that is characterized by more marked differences in the curves. The only trial that satisfies the fourth criterion, based on the test of significance, is COG-Capizzi. However, the levels of the two curves and their distances in both the negative and positive MRD classes are similar to those observed in the AIEOP-BFM trial, but resulted non significant probably due to a lower sample size.

Figure 4.7: AIEOP-BFM - EFS curves by treatment and MRD stratification (thick line for DXM and thin line for PDN)

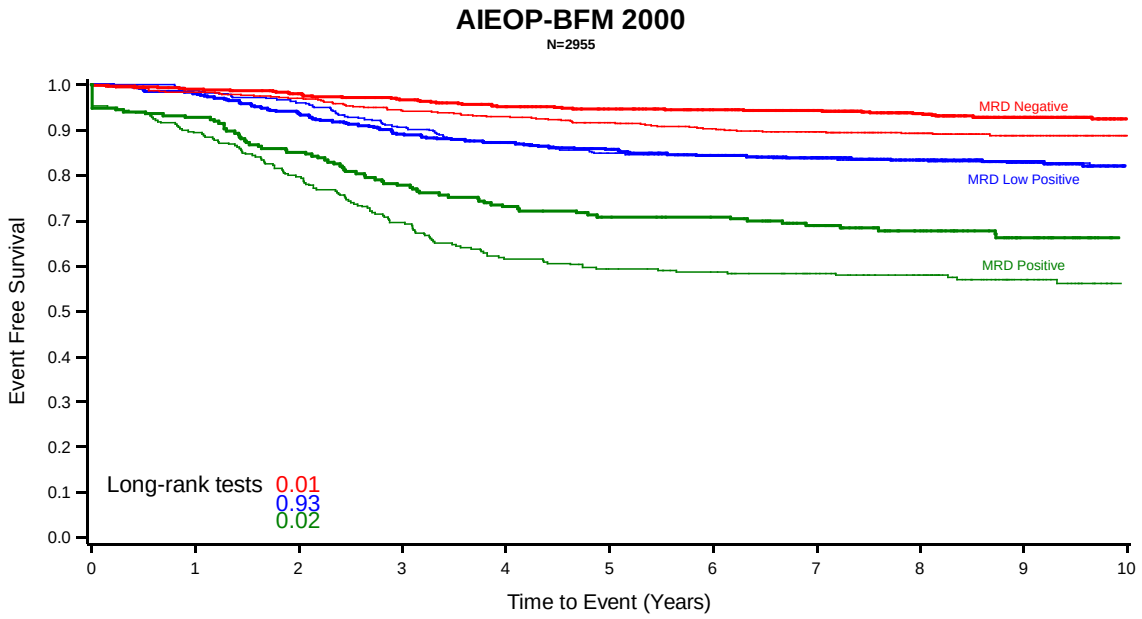


Figure 4.8: COG-Capizzi - EFS curves by treatment and MRD stratification (thick line for DXM and thin line for PDN)

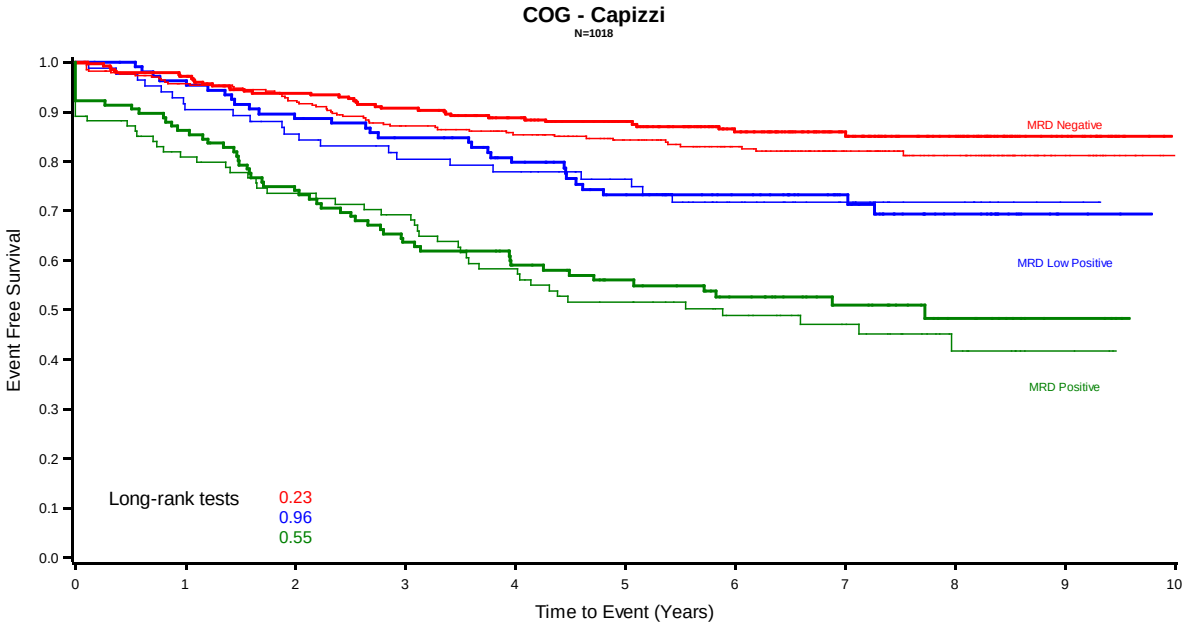
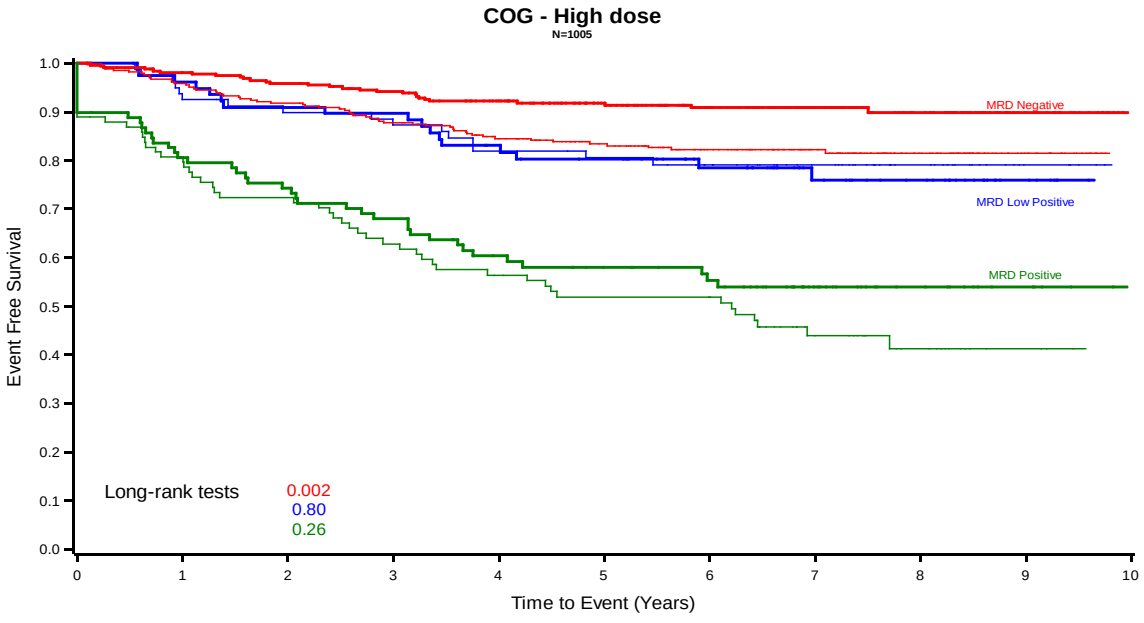


Figure 4.9: COG-High D - EFS curves by treatment and MRD stratification (thick line for DXM and thin line for PDN)



4.2.2 MULTI-TRIAL APPROACH

As indicated in section 2.2.2, the Cox proportional hazard model with a Weibull trial specific baseline hazard function was considered for the estimation of the trial unit specific treatment effects on the EFS (15), while the Proportional Odds model was used for MRD (14). The bivariate model was defined according to the Plackett copula (13).

Results considering both a model with no adjustment and after adjustment with age and WBC at diagnosis are reported in Table 4.8. The resulting estimate of the individual level association can be interpreted as a global (constant) odds ratio: for example the odds of surviving beyond time t given Negative or Low Positive MRD is at least 4.08 times higher than the odds for those with Positive MRD, or, in other hand, there is a 4.08-fold greater odds of surviving any specified time t for patients whose MRD is lower as compared to higher MRD levels. This suggests that there is a considerable association between MRD and the EFS time at individual level, after adjusting for treatment.

Table 4.8: Individual level association and Trial level association

	Patient-level Association		**Trial-level association**	
Model	θ^2_{indiv}	**95% CI**	R^2_{trial}	**95% CI**
No adjustment	4.08	(3.53-4.63)	0.15	(0-0.35)
With adjustment*	3.77	(3.26-4.28)	0.14	(0-0.34)

*age and WBC at diagnosis

The results of the second stage analysis in terms of trial level association are also reported in Table 4.8. The 95% confidence intervals for R^2_{trial} were obtained by finding values of these parameters, for which the corresponding estimates were equal to 2.5% and 97.5% quantiles of the cumulative distribution function of R^2 (Fisher 1928, Algina 1999). Contrary to the patient level evaluation, at the trial level the effects of the treatment on MRD and on EFS were poorly correlated (R^2_{trial}=0.15), meaning that MRD does not permit a reliable prediction of treatment effect on EFS.

The low association between the estimated trial-specific treatment effect for EFS and MRD can be observed in Figures 4.10-4.11, without and with adjustment, respectively. The circles represent the trial units and their size is proportional to their sample size, while the line is the prediction line from an estimated weighted regression analysis, with weights equal to the trial size. In the majority of the trial units DXM induced an advantage in EFS, but this seems not to be related with the benefit obtained on the MRD control (i.e. less MRD positivity), in the sense that only a part of the trial units presented a negative OR. Noticeably, almost all the trial units presenting with positive log(HR), which means that PDN performs better than DXM in terms of EFS, are from the COG-Capizzi protocol.

When the information on prognostic factors was taken into account, allowing for adjustment in the marginal models, the estimates of θ^2_{indiv} and R^2_{trial} changed at the individual level, but not at the trial level.

Figure 4.10: Treatment effects on MRD (categorical) vs treatment effect on EFS. Model with no adjustment

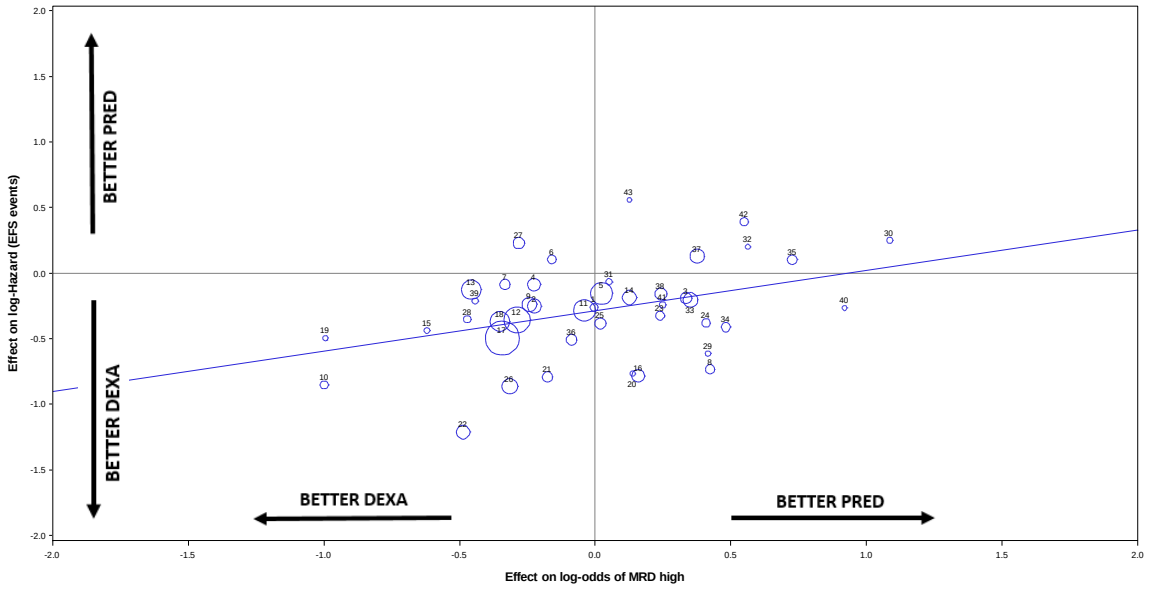
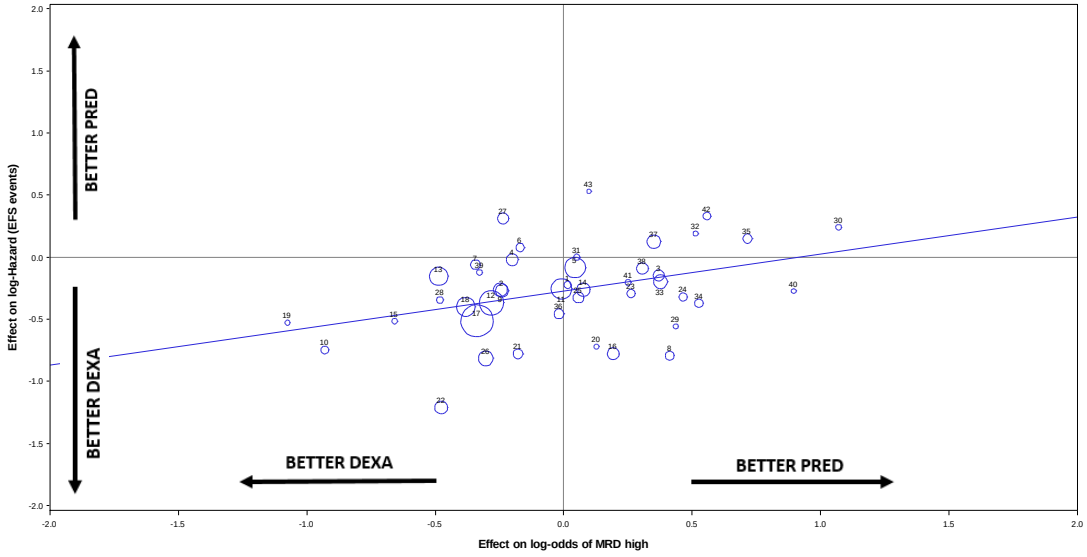


Figure 4.11: Treatment effects on MRD (categorical) vs treatment effect on EFS. Model with adjustment



4.2.2.1 Sensitivity analyses

In order to evaluate how sensitive are the results with respect to patients selection, we considered two different analysis excluding:

1) high risk patients
2) older patients.

1) NO HIGH RISK PATIENTS

Of the 607 patients that we did not consider in this analysis, 293 were from AIEOP-BFM, 105 from COG-Capizzi and 105 (21%) from COG-High Dose. In addition, a trial unit of 104 patients was excluded because there was no EFS event in one treatment group (Campania from AIEOP-BFM).
The results obtained on 4371 patients are reported in Table 4.9 and Figures 4.12-4.13 and indicated that the level of both the individual and the trial association decreased as compared to the global analysis. The final message on MRD surrogacy does not change much, also considering a more homogeneous group of ALL patients that excludes subjects with negative clinical feature at diagnosis (or based on MRD). A wider heterogeneity is observed in the treatment effects on MRD estimated by means of the copula model, with trial-units 40 (C ND/SD/MN/IA/NE) and 30 (H NY/VT/NH/ME/MA/CT) much separated.

Table 4.9: Individual level association and Trial level association

	Patient-Level Association		**Trial-Level association**	
Model	θ^2_{indiv}	**95% CI**	R^2_{trial}	**95% CI**
No adjustment	2.96	(2.51-3.40)	0.08	(0-0.23)
With adjustment*	2.83	(2.40-3.25)	0.08	(0-0.23)

*age and WBC at diagnosis

Figure 4.12: Treatment effects on MRD (categorical) vs treatment effect on EFS. Model with no adjustment

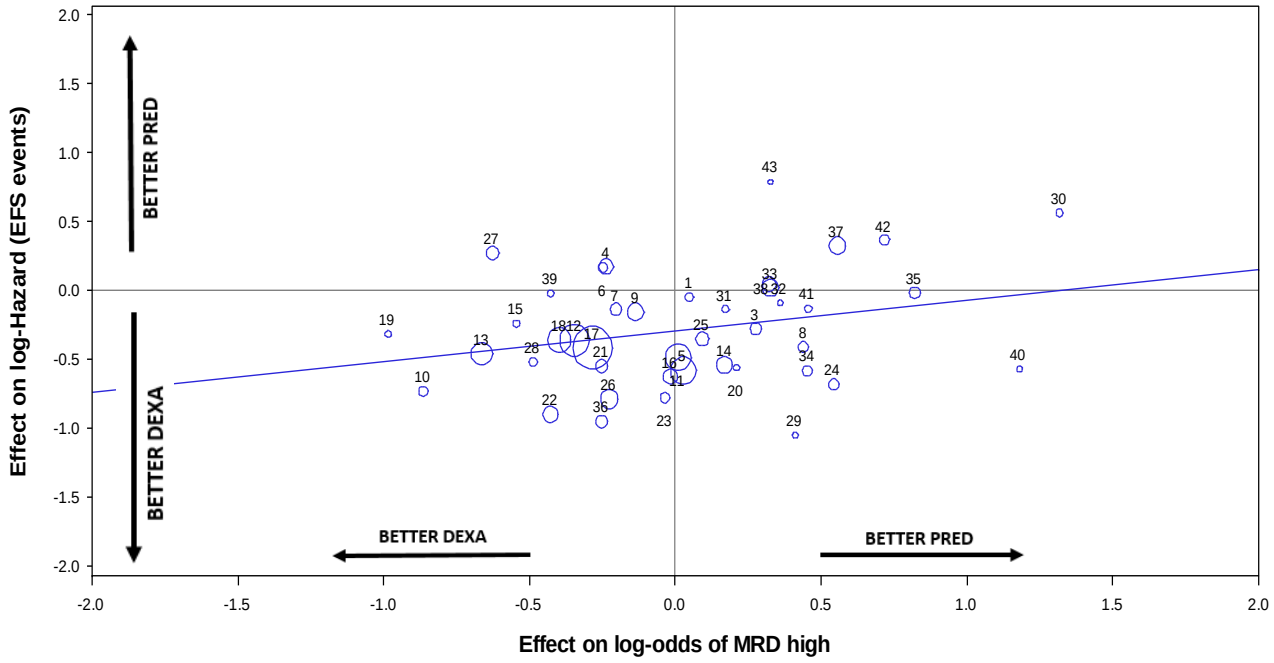
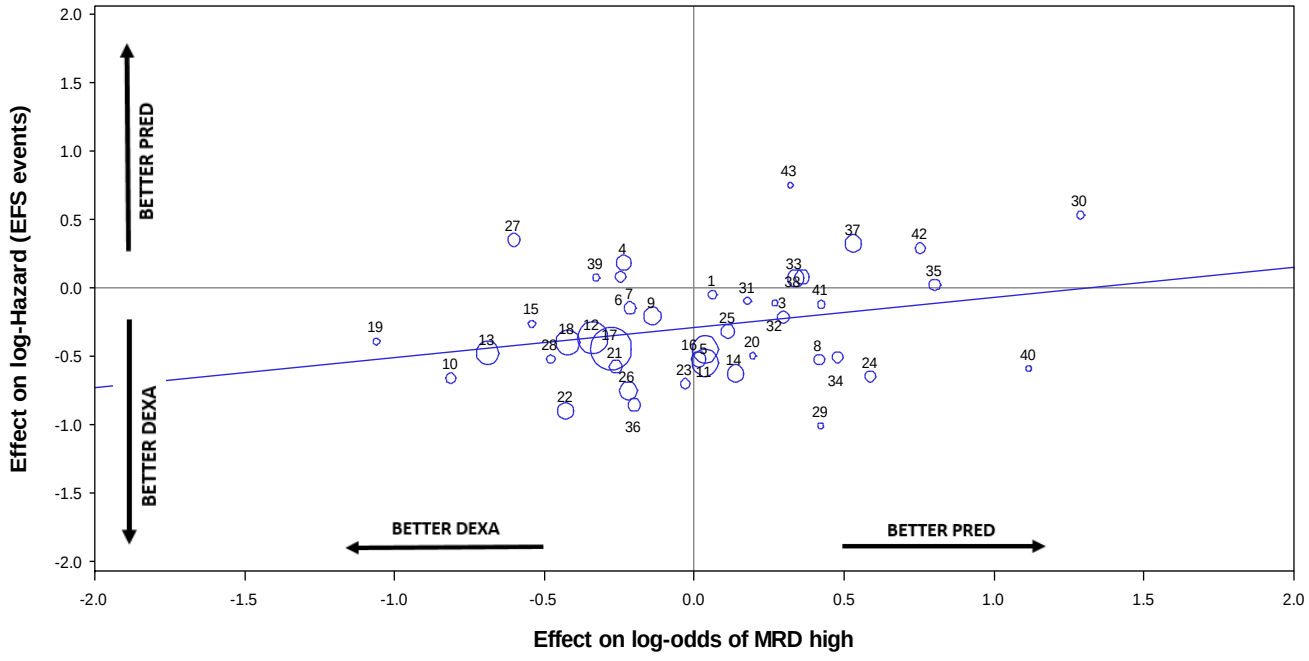


Figure 4.13: Treatment effects on MRD (categorical) vs treatment effect on EFS. Model with adjustment



2) NO OLDER PATIENTS

The second exploratory analysis excluded 1621 patients: 1546 due to the age indication (>10 years) and 75 patients from three trial units without at least one EFS event for each treatment arm (C NY/VT/NH/ME/MA/CT, H NY/VT/NH/ME/MA/CT, H Miscellanea). In the analysis on 3357 patients, while the patient level association increased to 4.28 as compared to the global analysis, the trial level association was approximately zero (Table 4.10). This absence of correlation is evident in Figure 4.14 where we can also observe a more wide variability of the estimated treatment effects, both in terms of HR and OR.

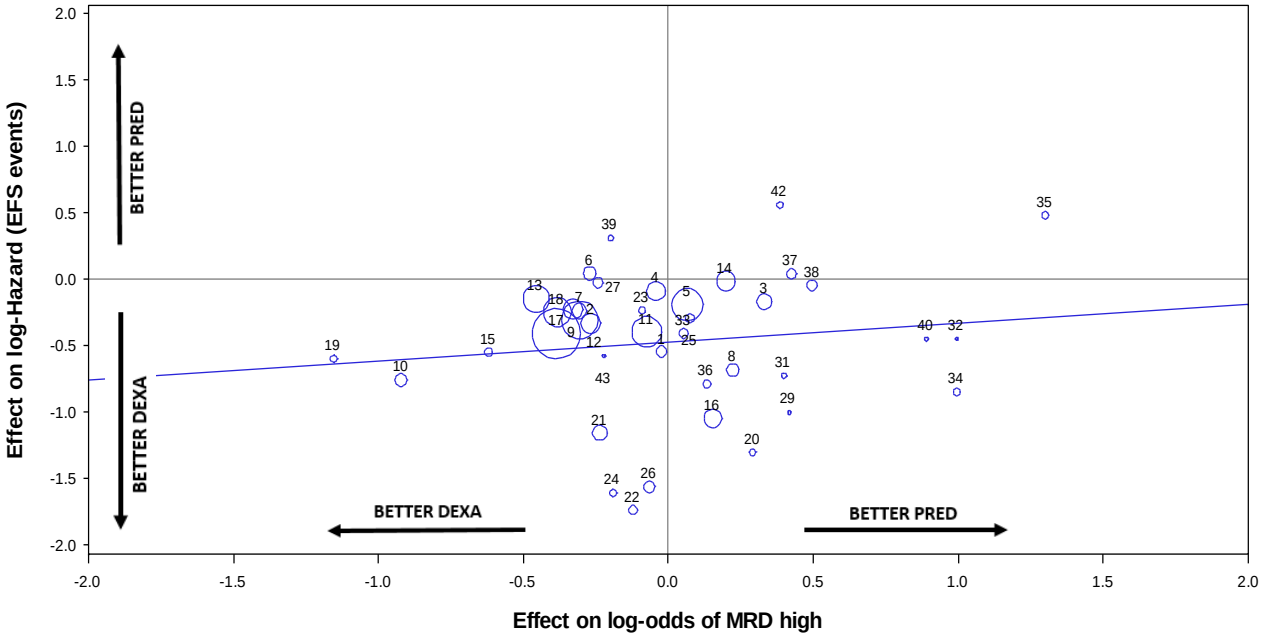
The model with adjustment is not reported because it did not converge.

Table 4.10: Individual level association and Trial level association

	Patient-Level Association		Trial-Level association	
Model	θ^2_{indiv}	**95% CI**	R^2_{trial}	**95% CI**
No adjustment	4.28	(3.53-5.03)	0.02	(0-0.10)
With adjustment*				

*WBC at diagnosis

Figure 4.14: Treatment effects on EFS versus effect on category MRD – No older patient - Model with no adjustment



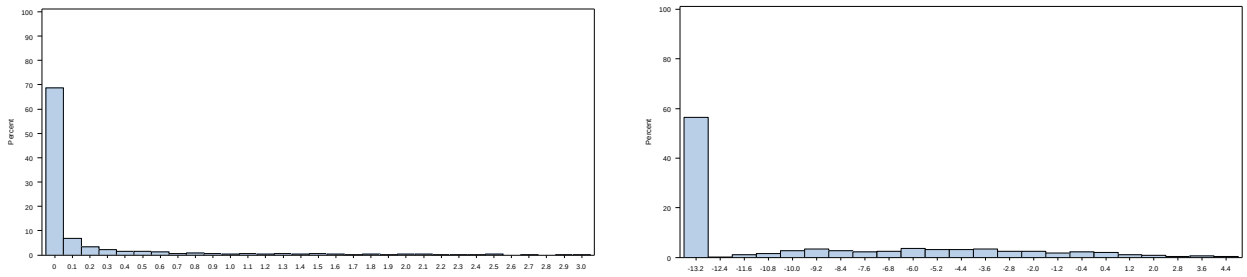
4.2.3 ANALYSIS ON MRD IN CONTINUOUS

We made an attempt to evaluate MRD in continuous as a surrogate of EFS with a meta-analytic approach that was developed ad-hoc for this clinical application. The overall distribution of the MRD on the natural scale is shown in Figure 4.15a, where a skewed behaviour is evident with a marked peak on zero. The horizontal axis was set at 3,thus excluding a certain number of extreme observations.

Different transformations were tested in order to make the distribution more regular and the best one was the logarithm. The log-transformed MRD measures are shown in Figure 4.15b (from the observed minimum and maximum), with zero values of MRD that were set by default equal to 0.0000021 (minimum observed). This graphic clearly suggest a mixture of a point distribution at zero and a continuous (normal) distribution on the right side. For simplicity, we analysed the log transformed MRD assuming a gaussian distribution; other situations will be evaluated in the future by means of different Copula models based on particular marginal distributions.

Figure 4.15a-b: MRD distribution on the original scale (a) and on the logarithm scale (b)

a b



The analysis was conducted on the 3215 patients from the AIEOP and COG groups who had the measure of MRD available, using the macro implemented in SAS for the Clayton and Hougaard Copulas. The estimates of the individual level association θ is near 1 for both the Clayton and the Hougaard Copulas, indicating that MRD and EFS are not associated at the individual level. The performance is not good also at the trial level as a R^2_{trial}=0.17 was obtained.

The lack of association at the individual level is not in line with the results obtained so far, and this can be justified by the fact that we made a very strong assumption that does not capture the peak at 0. More complex models should be developed for this at the cost of considerable computational challenges.

4.3 VALIDATION OF EFS FOR OS

The second aim of this study is to analyse if EFS can be considered a surrogate for OS in LLA. As anticipated in section 3.2, individual data on 6905 patients from different trials were evaluated, including also the 1400 EORTC patients in the three big trial units that identify: Belgium, Portugal and France.

The distribution of the EFS events by trial/group and treatment group is reported in Table 4.11. No marked difference with respect to the patients used for the primary analysis on MRD was observed in this augmented sample. The pattern of events in EORTC is very similar to AIEOP-BFM childhood patients.

Table 4.11: Distribution of the EFS events by Trial and Treatment

		Treatment					
		DXM		PDN		Total	
Group	Events type	N	%	N	%	N	%
AIEOP-BFM	Resistant	46	2.9	33	2.1	79	2.5
	Relapses	176	11.2	256	16.1	432	13.7
	Deaths	22	1.4	21	1.3	43	1.4
	SMN	21	1.3	17	1.1	38	1.2
	Alive in C-CR	1306	83.2	1264	79.5	2570	81.3
	Total	1571	100	1591	100	3162	100
COG - CAPIZZI	Resistant	11	2.1	13	2.5	24	2.3
	Relapses	99	18.9	88	16.7	187	17.8
	Deaths	14	2.7	22	4.2	36	3.4
	SMN	2	0.4	9	1.7	11	1.0
	Alive in CCR	398	75.9	394	74.9	792	75.4
	Total	524	100	526	100	1050	100
COG - HIGH D.	Resistant	10	1.9	11	2.1	21	2.0
	Relapses	72	14.0	90	17.1	162	15.5
	Deaths	14	2.7	19	3.6	33	3.2
	SMN	3	0.6	9	1.7	12	1.2
	Alive in CCR	417	80.8	398	75.5	815	78.1
	Total	516	100	527	100	1043	100
EORTC	Resistant	16	1.9	11	1.3	27	1.6
	Relapses	99	12.0	118	14.3	217	13.2
	Deaths	11	1.3	12	1.5	23	1.4
	SMN	3	0.4	3	0.4	6	0.4
	Alive in CCR	697	84.4	680	82.5	1377	83.5
	Total	826	100	824	100	1650	100

4.3.1 MULTI-TRIAL APPROACH

For EFS to be considered as a valid surrogate for OS with respect to DXM and PDN, two preliminary conditions must be met. The first condition requires that EFS and OS are correlated, while the second condition requires that the treatment effects evaluated on both the time to event endpoints are also correlated. The strength of the correlations reflects the quality of the surrogate, with a perfect surrogate that is expected to induce a correlation coefficient equal to 1.

The first condition was tested by performing a weighted linear regression analysis (WLRA) between trial unit specific Kaplan-Meier estimates of OS versus EFS evaluated in different time points. Data for each trial unit were weighted by the effective sample size at the time point considered for estimation. WLRA was performed with EFS at 3 years versus OS at 5 years (Figure 4.16 and Table 4.13) and EFS at 5 years versus OS at 7 years (Figure 4.17 and Table 4.13). For the trial specific Kaplan-Meier estimates of OS_{5y} and EFS_{3y}, the WLRA equation was $OS_{5y}= 0.32 + 0.68*EFS_{3y}$ with a coefficient of determination of $R^2=0.58$, indicating that about more than half of the variance could be explained by the linear regression. Similarly, for the trial specific Kaplan-Meier estimates of OS at 7 years versus EFS at 5 years, the WLRA equation was $OS_{7y} = 0.26 + 0.77*EFS_{5y}$ with $R^2=0.59$. These results suggest that condition one can be considered reasonably satisfied. In Figures 4.16-4.17, each trial unit is represented with two circles, one for each treatment: orange indicates Dexamethasone, while green Prednisone.

Table 4.13: Results of the Weighted Linear Regression

Equation	R^2
$OS_{5y}= 0.32 + 0.68*EFS_{3y}$	0.58
$OS_{7y} = 0.26 + 0.77*EFS_{5y}$	0.59

Also the second condition was tested by fitting a weighted linear regression model on the treatment effects in terms of Hazard Ratio (log-transformed) estimated on both EFS and OS by means of Cox proportional hazard models. The WLRA fitted equation was $\log(HR_{OS}) = 0.21+0.96*\log(HR_{EFS})$ with a coefficient of determination $R^2=0.44$, indicating that 44% of the variance could be explained by the linear regression (Figure 4.18).

Figure 4.16: Kaplan Meier estimates of EFS (3y) versus OS (5y) by treatment (in orange DXM, in green PDN)

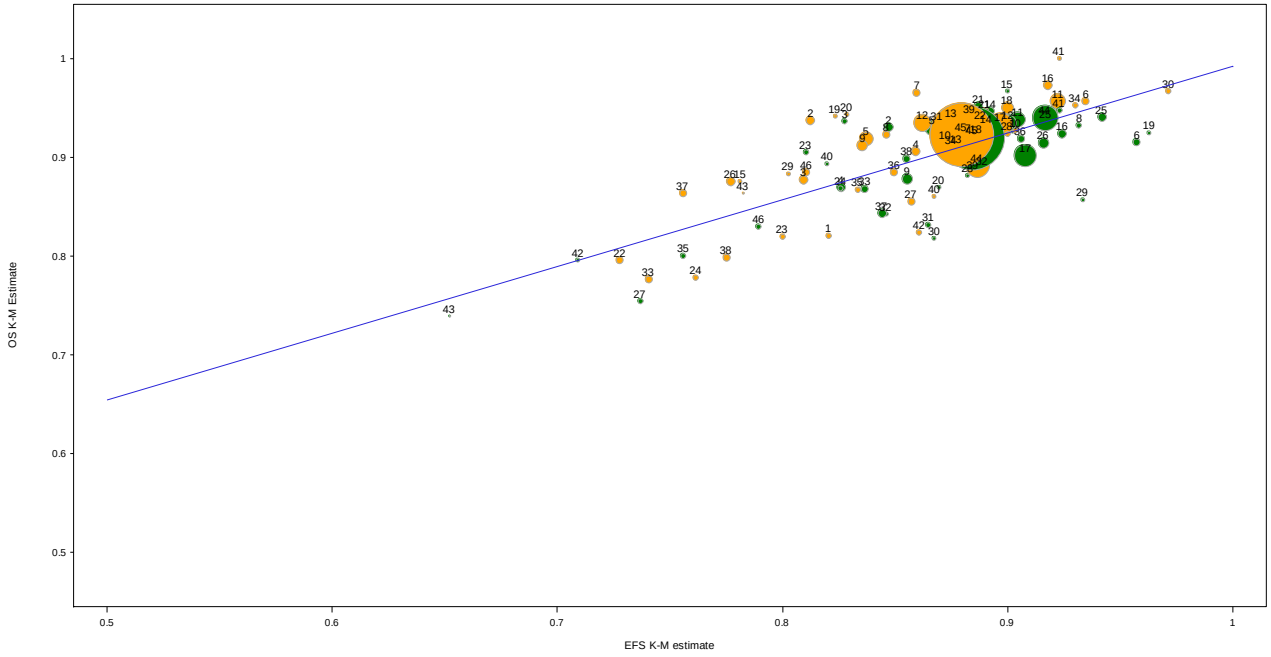


Figure 4.17: Kaplan Meier estimates of EFS (3y) versus OS (5y) by treatment (in orange DXM, in green PDN)

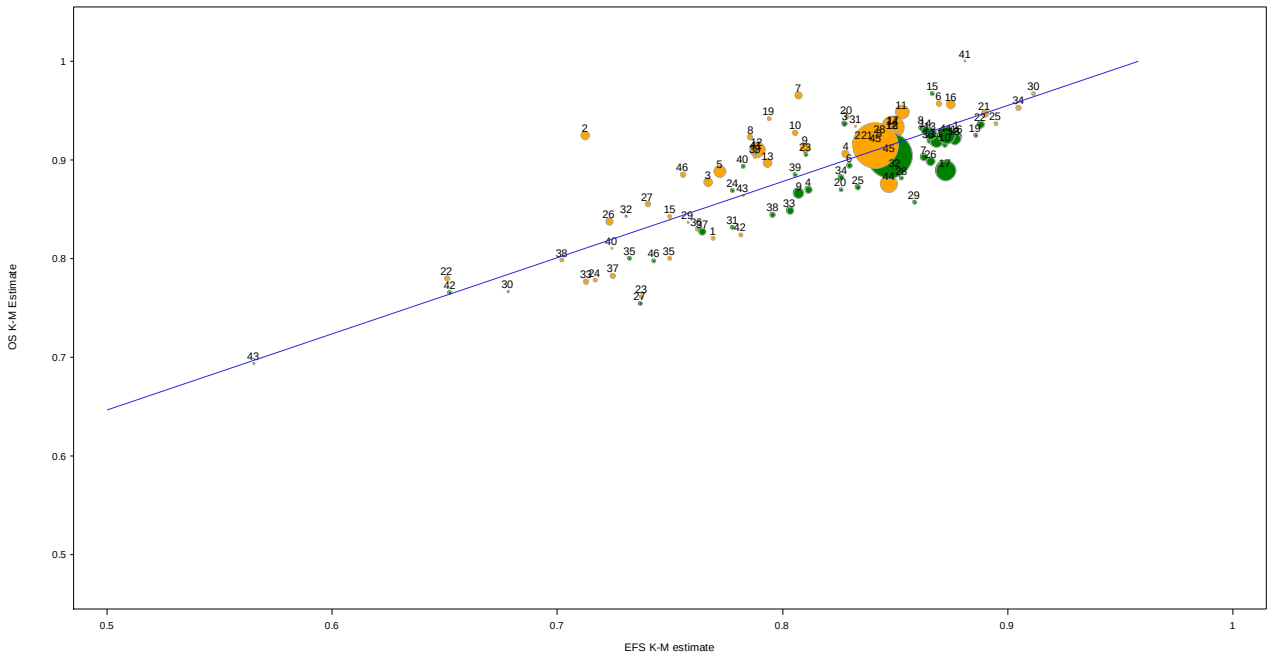
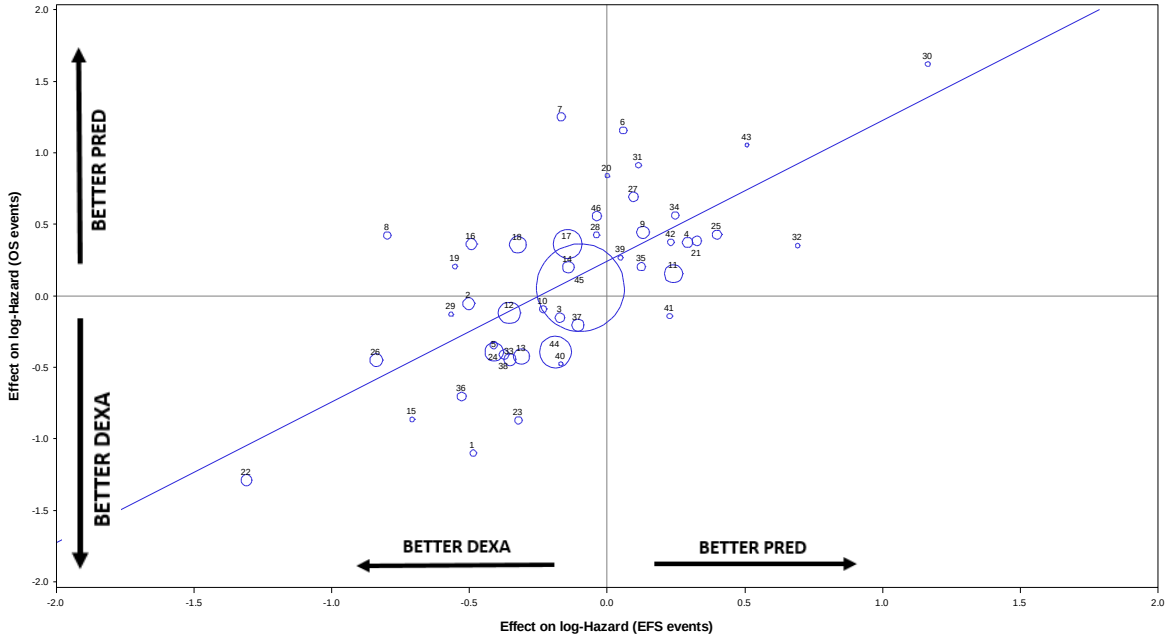


Figure 4.18: Treatment effect on OS versus effect on EFS.



The results of the meta-analytic approach to validation in terms of the degree of association at individual and trial level is reported in Table 4.14.

Table 4.14: Individual level association and Trial level association

	Patient-Level Association		Trial-Level association	
Model	**τ**	**95% CI**	**R^2_{trial}**	**95% CI**
No adjustment	0.94	(0.93-0.95)	0.62	(0.44-0.79)
With adjustment*	0.93	(0.928-0.94)	0.65	(0.48-0.81)

*age and WBC at diagnosis

The Kendall concordance coefficient **τ,** which is the copula association parameter, is 0.94 and 0.93 in the non-adjusted and adjusted analysis, respectively. These values indicate that the degree of association between EFS and OS, at the level of the individual patient, is very high: each patient there is an approximately 94% chance to observe a short OS given a short EFS. The R^2_{trial} ranges from 0.62 to 0.65 for the adjusted and non-adjusted analysis, as it is possible to observe in Figures 4.19-4.20, showing the estimated treatment effects on EFS (log HR) and OS (log HR) (see also Table 4.15 for the estimated WLRM).

Table 4.15: Results of the Weighted Linear Regression

Model	Equation
No adjustment	Log HR(OS)= 0.08 + 1.18* Log HR(EFS)
With adjustment*	Log HR(OS)= 0.07 + 1.20* Log HR(EFS)

The log(HR) estimated from the Cox model applied to each single trial unit in Figure 4.18 and those obtained from the proportional hazard model in the copula (Figure 4.19) are quite different due to the fact that the latter analysis accounts for the hierarchical structure of the data that consider the trial units within each protocol/group.

Of note is the behavior of the trial unit 40 (C ND/SD/MN/IA/NE) that, while presenting with a slight effect of PDN on EFS, it is characterized by a marked benefit on survival that points in favor of DXM. However, this seems not to influence much the analysis.

In conclusion, EFS can be considered a valid surrogate for OS .

Figure 4.19: Treatment effect on OS versus effect on EFS. Model with no adjustment

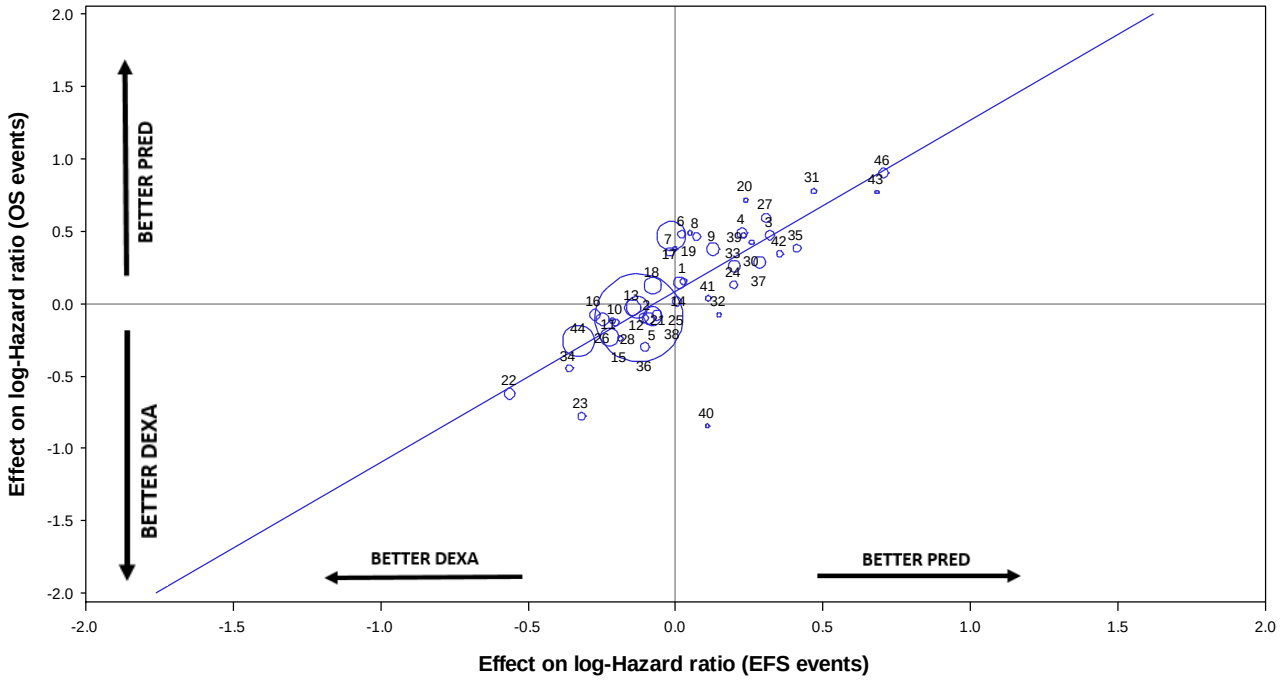
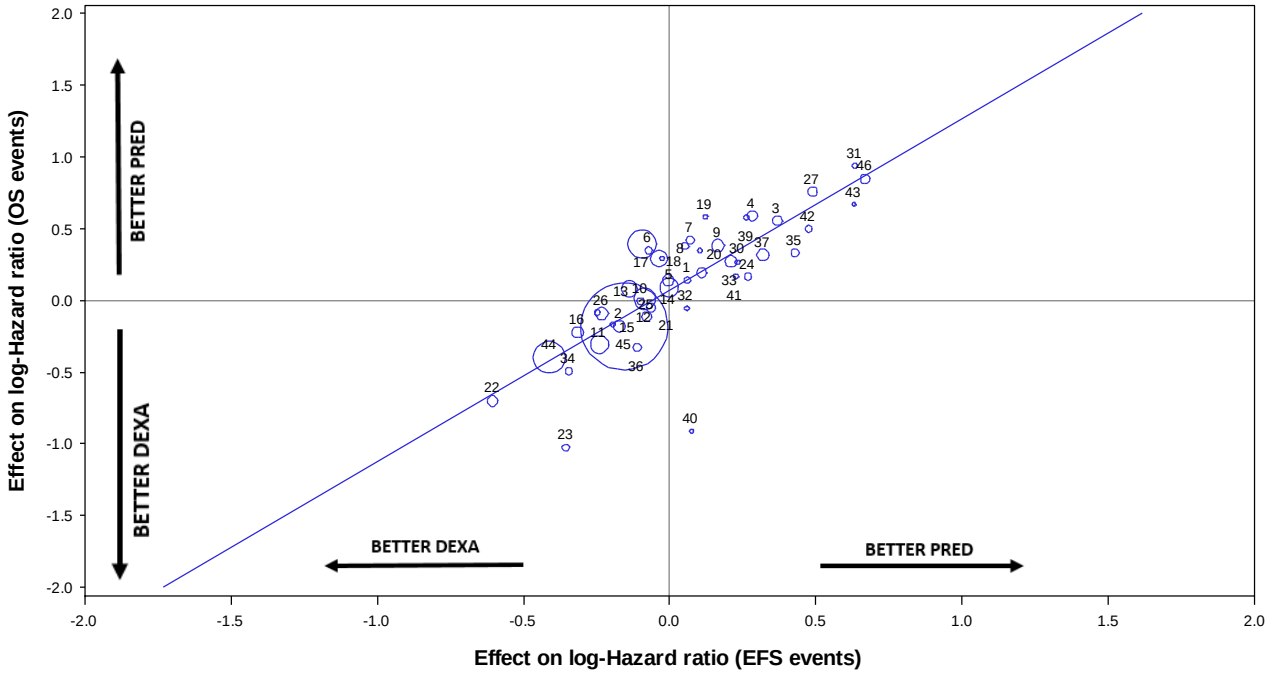


Figure 4.20: Treatment effect on OS versus effect on EFS. Model with adjustment



4.3.1.1 Sensitivity analysis

Two sensitivity analyses were also performed in this context on a selection of more homogeneous patients, thus excluding 1) high risk and 2) older patients.

1) *NO HIGH RISK PATIENTS*

Considering the 6152 patients presenting without negative clinical features at diagnosis, we obtained a marginal improvement both in τ and R^2_{trial} (Table 4.16 and Figure 4.21-4.22), despite the presence on the trial unit 40 (C ND/SD/MN/IA/NE), which seems to be more influential here than in the global analysis of the previous section. Of note, in this trial unit there is a change in the direction of the treatment effect on EFS, which now became coherent with the one observed on OS. In general, the estimated treatment effect in this subpopulation of patients are more marked as compared to the ones combined in the general population.

Table 4.16: Individual level association and Trial level association

	Patient-Level Association		Trial-Level association	
Model	**τ**	**95% CI**	**R^2_{trial}**	**95% CI**
No adjustment	0.96	(0.96-0.97)	0.67	(0.51-0.83)
With adjustment*	0.96	(0.96-0.96)	0.67	(0.51-0.82)

*age and WBC at diagnosis

Figure 4.21: Treatment effect on OS versus effect on EFS – No HR. Model with no adjustment

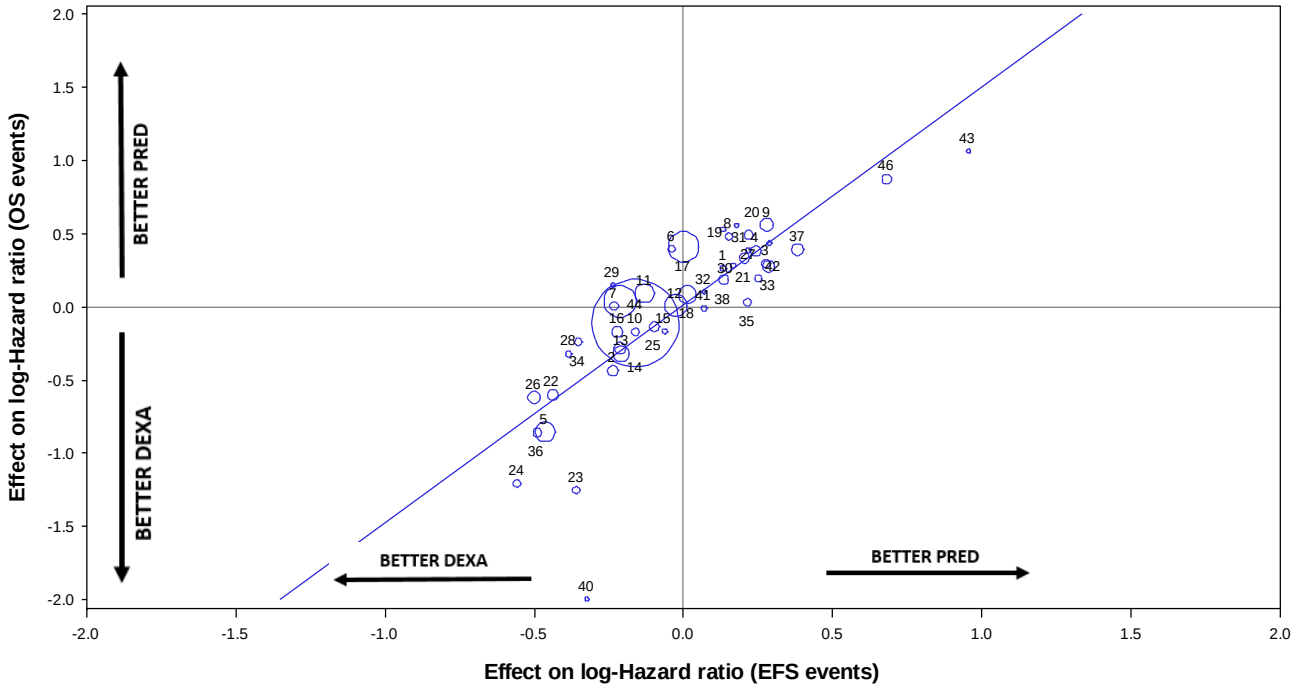
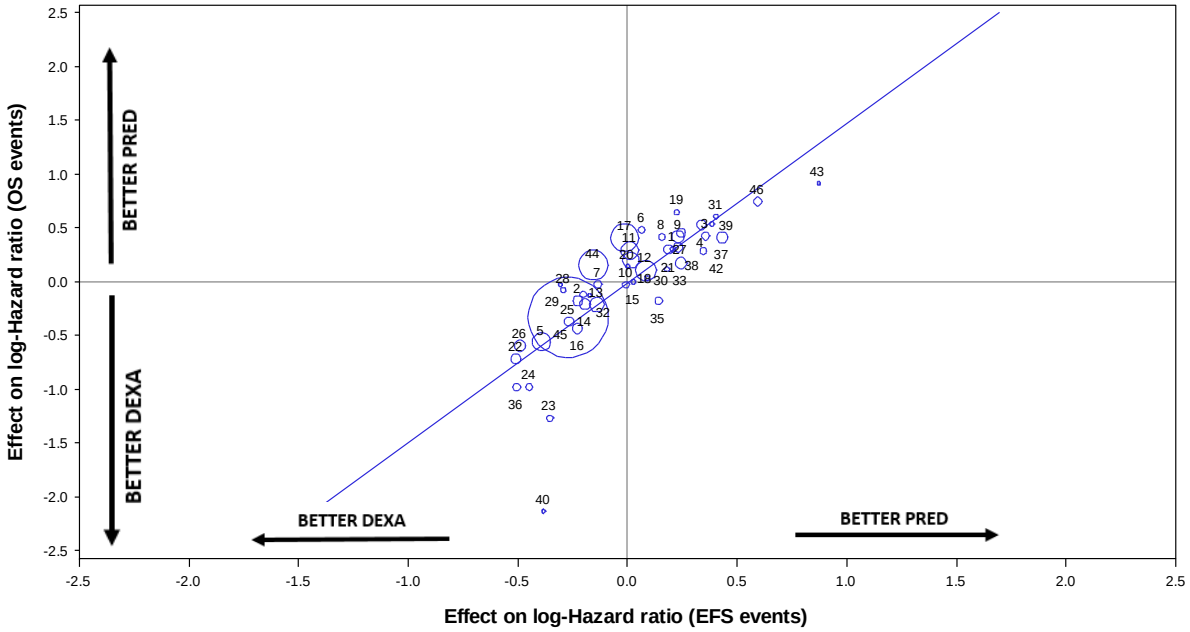


Figure 4.22: Treatment effect on OS versus effect on EFS – No HR. Model with adjustment



2) *NO OLDER PATIENTS*

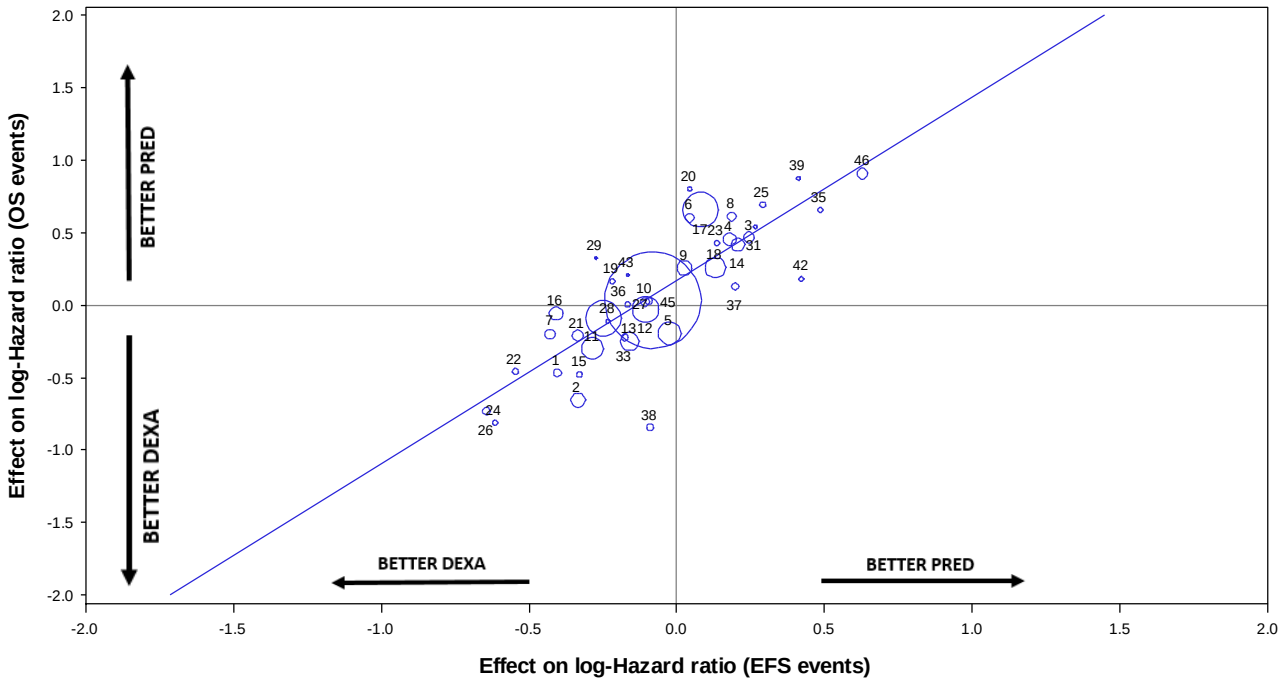
Considering a more restrictive condition on age, we selected 4899 ALL patients with age <10y that mainly involved patients from the COG-groups. The analysis confirmed a high level of individual association, with τ similar to that obtained in the global assessment (Table 4.17), while the performance at the trial level strongly worsened, as the R^2_{trial} halved. This can be confirmed by the graphical representation of the Copula based estimates in Figure 4.23: the points around the regression line are much more spread than those obtained in the global analysis. The model with adjustment is not reported because it did not converge.

Table 4.17: Individual level association and Trial level association

	Patient-Level Association		Trial-Level association	
Model	**τ**	**95% CI**	**R^2_{trial}**	**95% CI**
No adjustment	0.95	(0.945-0.958)	0.35	(0.12-0.57)
With adjustment*	-	-	-	-

*WBC at diagnosis

Figure 4.23: Treatment effect on OS versus effect on EFS – No older patients. Model with no adjustment



5 DISCUSSION

The monitoring of minimal residual disease (MRD) has become a routine clinical practice in frontline treatment of childhood ALL and it is crucial in refining the individual risk class. MRD evaluated early in the course of the disease has proven to be one of the strongest prognostic factor in ALL and this has encouraged clinicians to define it also as surrogate endpoint. Indeed, the two concept are very different, a prognostic factor is a measurement that is associated with the clinical outcome in the absence of therapy or with the application of a standard therapy that patients are likely to receive (Clark 2006); a surrogate endpoint is a biomarker that is intended to substitute a clinical endpoint when specific treatments are applied. A surrogate endpoint is expected to predict the clinical outcome. In this work, we have assessed if MRD, measured at the end of the induction, can be qualified as surrogate for EFS in childhood B-precursor ALL patients treated with DXM and PDN in the induction phase. To do this we have considered data from three of the most important groups in the world that conduct innovative research in this field, and a statistical meta-analytic approach to validation that is now considered the gold standard. An important issue related to this approach is the minimal number of trials involved in the multi-trial analysis. One common solution to the problem of too few similar trials consists in performing trial-level surrogacy analyses on trial sub-units (e.g., centers within trials), thereby artificially increasing the trial-level sample size. In our case, we had three trials at disposal, which were not sufficient to implement the meta-analytic approach and for this reason the treating centers were aggregated according to geographical areas.

The included trials are different in some aspects, e.g. enrolled patients, characteristics of the induction treatment, post induction therapies and schedules. The presence of these diversities, which might be seen as additional sources of heterogeneity in the analysis, is indeed a strength that can be defended on the ground of generalizability of the results of the validation process to future clinical trials and treatments. More importantly, the methods used for the measurement of MRD are different (PCR in Europe and FCM in the USA) and also the way in which MRD is collected (continuous for AIEOP and COG; categorical in the BFM and EORTC groups, with classes that cannot be harmonized). We thus performed the analysis considering MRD as an ordinal surrogate endpoint, defining the three MRD ordered classes as closely to the clinical practice (i.e. MRD Negative, Low Positive and Positive). We also focused on the analysis of the subset of MRD data in continuous, and for this purpose we developed an ad-hoc method (implemented in a SAS macro) as no specific approach was available in the statistical literature to validate a continuous surrogate for a time to event endpoint. In this extension, the explicit use of the copula models, which is one of the hallmark of the meta-analytic approach, allows the joint modelling of different types of endpoints and this broaden the range of

possible models that can be formulated. An additional advantage is that the meta-analytic method fully captures the two dimensions of the validation of a proposed surrogate, exploring both the individual and the trial level association.

We implemented the meta-analytic approach, using the analyses based on the Prentice's Criteria as explorative, and we found that MRD, no matter how it is measured (in categories or in continuous), is a poor surrogate for EFS at the trial level and this does not allow reliable prediction of the treatment effects on EFS. In contrast, there is a strong association between MRD and EFS time for individual patients, after adjusting for treatment. The results are confirmed also in the subgroup of the patients that excluded high risk patients and those < 10 years old. This seemingly contradictory findings can coexist as they refers to different areas of competence. The individual level surrogacy quantify the attitude of the two measurements to co-vary in the same subject, whereas the trial level surrogacy is related to the joint behaviour of the subjects within a trial, based on their treatment allocation. The clinical trialist or the statistician will primarily be interested in the trial level surrogacy, while the treating clinician will consider the individual level surrogacy as the more relevant quantity because he will be interested in predicting the behaviour of a given patient.

In this analysis we have not completely addressed an important issue related to the complex interaction between the potential surrogate and the true endpoint. In principle, MRD should lie on the causal pathway of the treatment, but in the reality, treatment decisions that might affect the true endpoint are made after observing the surrogate. MRD is observed after the induction phase of the treatment, while the true endpoint is observed under subsequent therapies that may confound the effect of DXM or PDN on the true endpoint. What we have assumed here is that the randomized post-induction treatments have limited impact on the analyses since they are similar in the final outcome, based on the results of the different study protocols (Conter et al. 2000, Domenech et al. 2014, Borowitz MJ et al. 2015).

Finally, the results on the appropriateness of EFS as surrogate endpoint for OS are more promising, even if the results of the sensitivity analysis suggested that EFS is not deemed acceptable as surrogate in the subset of younger patients, probably due to the imprecision of the trial-units estimates.

In conclusion, in a trialist perspective, MRD (detected early) cannot be used as a clinical endpoint that replaces EFS in trials involving traditional induction therapies in B-ALL children. On the contrary, we can maintain EFS as primary endpoint in substitution of OS that is considered the hard endpoint in cancer.

SUPPLEMENT

Appendix A: Bivariate copula

Some general definitions and properties of the bivariate copula models that were used in section 2.2 are reported here. An introduction to this topic that discuss also the more general multivariate copula can be found in Nelsen (1999).

Definition 1: *A bivariate copula C is a function from [0,1]×[0,1] into [0,1] such that:*

1. *For every u, v in [0,1], C(u,0) = C(0,v) = 0, C(u,1) = u and C(1,v) = v.*
2. *For every $u_1 \leq u_2$ and $v_1 \leq v_2$ in [0,1],*
   $C(u_2,v_2) - C(u_2,v_1) - C(u_1,v_2) + C(u_1,v_1) \geq 0$

It follows from this definition that a copula is a bivariate distribution function with uniform margins.
In case of independence of the two margins, the so-called product copula is obtained:
$C_P(u,v)=uv$

Theorem 1: *Let X_1 and X_2 be random variables with F, F_1 and F_2 the joint distribution function and the marginal respectively. There exists a copula function Cx_1x_2 such that*
$F(x_1,x_2) = Cx_1x_2\{F_1(x_1), F_2(x_2)\}$.
If F_1 and F_2 are continuous, then Cx_1x_2 is unique, otherwise Cx_1x_2 is uniquely determined on Ran(F1) × Ran (F2), as before.
We can interpret a copula as a function that establishes a particular dependence structure on two given random margins. The following theorem formalizes this fact.

Theorem 2: : *Let X_1 and X_2 be continuous random variables with margins F_1 and F_2 ; respectively. The variables X_1 and X_2 are independent if and only if the corresponding copula Cx_1x_2 equals the product copula $Cx_1x_2 \equiv C_p$.*

We present some examples of copula functions used in this work, i.e. Clayton, Placket and Hougaard.

- Clayton Copula

$$C_\theta(u,v) = (u^{1-\theta} + v^{1-\theta} - 1)^{\frac{1}{1-\theta}}$$

with
$\theta \in [-1, +\infty)$ and $\theta \neq 0$
$C_0(u,v)=uv$
$C_{-1}= max(u+v-1,0)$

$C_{+\infty}$= *min(u,v)*

- Placket Copula

$$C_\theta(u,v) = \frac{[1+(\theta-1)(u+v)] - \sqrt{[1+(\theta-1)(u+v)]^2 - 4uv\theta(\theta-1)}}{2(\theta-1)}$$

with

θ ϵ (0, +∞) and θ≠1

C_1*(u,v)=uv*

C_0*= max(u+v-1,0)*

$C_{+\infty}$*= min(u,v)*

- Gumbel-Hougaard Copula

$$C_\theta(u,v) = exp\left\{-\left[(-\ln u)^{\frac{1}{\theta}} + (-\ln v)^{\frac{1}{\theta}}\right]^\theta\right\}$$

with

θ ϵ [1, +∞) and θ≠1

C_1*(u,v)=uv*

C_1*= max(u+v-1,0)*

$C_{+\infty}$*= min(u,v)*

Appendix B –

SAS MACRO for the evaluation of a continuous surrogate for a time to event endpoint: Clayton Copula

```
/*These input are necessary:
- file "dataset" containing one record for each patient with measurements for both the true and the surrogate endpoint, with variables:
cens - measurement of the failure-time endpoint;
time - censoring indicator (1=event, 0=censoring);
surr - measurement of the continuous (normally distributed) surrogate;
treat - treatment indicator (0 or 1);
center - center/trial number (consecutive numbers: 1, 2, 3, ...).
- file "estab" containing estimated center/trial-specific coefficients for the linear regression model for the continuous surrogate, with variables:
intercept - estimated intercept;
effect      - estimated treatment effect (difference in means);
center - center/trial number (consecutive numbers: 1, 2, 3, ...).
- file "estt" containing estimated center/trial-specific coefficients effects for the Weibull model for the true endpoint (as
obtained by PROC LIFEREG), with variables:
intercept - estimated intercept;
treat     - estimated treatment effect (difference in means);
SCALE - estimated scale parameter ;
center - center/trial number (consecutive numbers: 1, 2, 3, ...).
*/

proc iml;
reset log;

/*1 - define loglikelihood function*/

start L_lik(param) /*define param*/
            global(time,cens,surr,treat,numcents,center);
lik=0;
      theta=exp(exp(param[,1])); /*exp(exp) - to constrain to be  >1 (from definition clayton copula)*/

do c=1 to numcents;

      sigma=exp(param[,2+(c-1)#6]);
      mu=param[,3+(c-1)#6];
      alpha=param[,4+(c-1)#6];
      lambda=exp(param[,5+(c-1)#6]); /*exp because the parameter can have only positive values */
      p=param[,6+(c-1)#6];
      beta=(param[,7+(c-1)#6]);

      t=time[loc(center=c),];
         delta=cens[loc(center=c),];
      s=surr[loc(center=c),];
      z=treat[loc(center=c),];

*      f1=logsdf('NORMAL',s,mu+alpha#z,sigma);/*log S(s)*/ /*I didn't use it*/
```

```
        f2=-lambda#exp(beta#z)#(t##p); /*log S(t)*/
        f6=exp(-lambda#exp(beta#z)#(t##p)); /*S(t)*/
        f4=p#lambda#exp(beta#z)#(t##(p-1))#f6; /*f(t)--> Surv#Hazard – SAS parameterization*/
        f3=pdf('NORMAL',s,mu+alpha#z,sigma); /*f(s)*/
        f5=sdf('NORMAL',s,mu+alpha#z,sigma); /*S(s)*/
*     print f2 f3 f4 f5 f6;

        /*define log-likelihood*/

        cop=(f5)##(1-theta)+(f6)##(1-theta)-1;
        a=log(theta)+((2#theta-1)/(1-theta))#log(cop)-theta#log(f5)-theta#log(f6)+log(f3)+log(f4);
        b=(theta/(1-theta))#log(cop)-theta#log(f5)+log(f3);
        likc=sum(delta#a+(1-delta)#b);
        lik=lik+likc;
end;

  return(lik);
finish L_lik;

/*2 - dataset*/
use dataset;
read all var{time} into time;
read all var{cens} into cens;
read all var{surr} into surr;
read all var{treat} into treat;
read all var{center} into center;
close dataset;

cents=unique(center)`;
numcents=nrow(cents);

print cents numcents;

use estsab;
read all var{intercept} into Intercepts;
read all var{effect} into effects;
read all var{center} into centers;
close estsab;
params=Intercepts||effects||centers;

print params;

use estt;
read all var{Intercept} into Interceptt;
read all var{treat} into treatt;
read all var{_SCALE_} into _SCALE_t;
read all var{center} into centert;
close estt;
```

```
paramt=Interceptt||treatt||_SCALE_t||centert;

print paramt;

use Covparm;
read all var{residual} into residual; /*Residual*/
close Covparm;

/*3 - subroutine*/

      x=J(1,1+6#numcents,.); /*J(nrow,ncol,value);*/

*       mattrib x colname={'Theta' 'Sigma' 'Mu' 'Alpha' 'Lambda' 'p' 'Beta'};

      x[,1]=-9;

do c=1 to numcents;

      x[,2+(c-1)#6]=0.5#log(residual[c,]); /*0.5#LOG(Residual)*/
      x[,3+(c-1)#6]=Intercepts[c,];/*intercept*/
      x[,4+(c-1)#6]=effects[c,]; /*Effect of treat on S*/
      x[,5+(c-1)#6]=-(Interceptt[c,]/_SCALE_t[c,]); /* -(intercept/scale)*/
      x[,6+(c-1)#6]=1/_SCALE_t[c,]; /*1/scale*/
      x[,7+(c-1)#6]=-(treatt[c,]/_SCALE_t[c,]); /* -(regression coeffincient/scale)*/

end;

      xopt=J(1,3,.);
      xopt[,1]=1; /*indicates whether the problem is minimization or maximization: 1=specifies a
maximization problem*/
      xopt[,2]=2; /*specifies the amount of printed output. 4=the approximate covariance matrix of parameter
esti- mates is printed if opt[3] is set*/
      xopt[,3]=0; /*Selects a scaling for the Hessian matrix, G: default for NLPNRR is 0= No scaling is done*/
      /*The other options are not for NLPNRR*/

      print x ;
      print xopt;
   maxiter=150;
   termin=maxiter||J(1,12,.); /*why 12?*/
   termin[1,4]=0;
   termin[1,6]=0.001;

call nlpnrr(rc,est,"L_lik",x,xopt,,termin);
/*CALL NLPNRR( rc, xr, "fun", x0 <,opt, blc, tc, par, "ptit", "grd", "hes">);*/
/*"Ridge" value should be zero, if not, the hessian is negative and the procedure don't converge.
      Solution? Change number of interation in subroutine, example in our case:"maxiter"*/

print est;
```

```
sigma=J(1,numcents,.);
mu=J(1,numcents,.);
alpha=J(1,numcents,.);
lambda=J(1,numcents,.);
p=J(1,numcents,.);
beta=J(1,numcents,.);

theta=exp(exp(est[,1]));

do c=1 to numcents;

      sigma[,c]=exp(est[,2+(c-1)#6]);
      mu[,c]=est[,3+(c-1)#6];
      alpha[,c]=est[,4+(c-1)#6];
      lambda[,c]=exp(est[,5+(c-1)#6]);
      p[,c]=est[,6+(c-1)#6];
      beta[,c]=est[,7+(c-1)#6];

end;

 print theta ;
 print sigma;
 print mu;
 print alpha;
 print lambda;
 print p;
 print beta;
 param_s=theta||sigma||mu||alpha||lambda||p||beta;

/* Standard Error - "Delta method" */

call nlpfdd(f,g,h,"L_lik",est);
/*CALL NLPFDD( f, g, h, "fun", x0, <,par, "grd">);*/

* print h; /*Hessian matrix*/

var=inv(-h);
se=t(sqrt(vecdiag(var))); /*standard error*/

* print var;
print (t(se));

se_est=J(1,1+6#numcents,.);

se_est[,1]=(exp(exp(est[,1]))#exp(est[,1]))#se[,1]; /*theta*/

do c=1 to numcents;
```

```
se_est[,2+(c-1)#6]=exp(est[,2+(c-1)#6])#se[,2+(c-1)#6]; /*sigma*/
se_est[,3+(c-1)#6]=se[,3+(c-1)#6]; /*mu*/
se_est[,4+(c-1)#6]=se[,4+(c-1)#6]; /*alpha*/
se_est[,5+(c-1)#6]=exp(est[,5+(c-1)#6])#se[,5+(c-1)#6]; /*lambda*/
se_est[,6+(c-1)#6]=se[,6+(c-1)#6]; /*p*/
se_est[,7+(c-1)#6]=se[,7+(c-1)#6]; /*beta*/

end;

print (t(param_s)) (t(se_est));

/*Confidence Intervals*/

lo_sigma=J(1,numcents,.);
up_sigma=J(1,numcents,.);
lo_mu=J(1,numcents,.);
up_mu=J(1,numcents,.);
lo_alpha=J(1,numcents,.);
up_alpha=J(1,numcents,.);
lo_lambda=J(1,numcents,.);
up_lambda=J(1,numcents,.);
lo_p=J(1,numcents,.);
up_p=J(1,numcents,.);
lo_beta=J(1,numcents,.);
up_beta=J(1,numcents,.);

lo_theta=exp(exp(est[,1]-2#se[,1]));
up_theta=exp(exp(est[,1]+2#se[,1]));

do c=1 to numcents;

lo_sigma[,c]=exp(est[,2+(c-1)#6]-2#se[,2+(c-1)#6]);
up_sigma[,c]=exp(est[,2+(c-1)#6]+2#se[,2+(c-1)#6]);
lo_mu[,c]=est[,3+(c-1)#6]-2#se,3+(c-1)#6];
up_mu[,c]=est[,3+(c-1)#6]+2#se[,3+(c-1)#6];
lo_alpha[,c]=est[,4+(c-1)#6]-2#se[,4+(c-1)#6];
up_alpha[,c]=est[,4+(c-1)#6]+2#se[,4+(c-1)#6];
lo_lambda[,c]=exp(est[,5+(c-1)#6]-2#se[,5+(c-1)#6]);
up_lambda[,c]=exp(est[,5+(c-1)#6]+2#se[,5+(c-1)#6]);
lo_p[,c]=est[,6+(c-1)#6]-2#se[,6+(c-1)#6];
up_p[,c]=est[,6+(c-1)#6]+2#se[,6+(c-1)#6];
lo_beta[,c]=est[,7+(c-1)#6]-2#se[,7+(c-1)#6];
up_beta[,c]=est[,7+(c-1)#6]+2#se[,7+(c-1)#6];

end;
```

```
ic=lo_theta||up_theta||lo_sigma||up_sigma||lo_mu||up_mu||lo_alpha||up_alpha||lo_lambda||up_lambda||lo_p||up
_p||lo_beta||up_beta;

print (t(lo_theta)) (t(up_theta));
print (t(lo_sigma)) (t(up_sigma));
print (t(lo_mu)) (t(up_mu));
print (t(lo_alpha)) (t(up_alpha));
print (t(lo_lambda)) (t(up_lambda));
print (t(lo_p)) (t(up_p));
print (t(lo_beta)) (t(up_beta));

quit;
```

Appendix C –
SAS MACRO for the evaluation of a continuous surrogate for a time to event endpoint: Hougaard Copula

```
/*These input are necessary:
- file "dataset" containing one record for each patient with measurements for both the true and the surrogate endpoint, with variables:
cens - measurement of the failure-time endpoint;
time - censoring indicator (1=event, 0=censoring);
surr - measurement of the continuous (normally distributed) surrogate;
treat - treatment indicator (0 or 1);
center - center/trial number (consecutive numbers: 1, 2, 3, ...).
- file "estab" containing estimated center/trial-specific coefficients for the linear regression model for the continuous surrogate, with variables:
intercept - estimated intercept;
effect      - estimated treatment effect (difference in means);
center - center/trial number (consecutive numbers: 1, 2, 3, ...).
- file "estt" containing estimated center/trial-specific coefficients effects for the Weibull model for the true endpoint (as
obtained by PROC LIFEREG), with variables:
intercept - estimated intercept;
treat     - estimated treatment effect (difference in means);
SCALE - estimated scale parameter ;
center - center/trial number (consecutive numbers: 1, 2, 3, ...).*/


proc iml;

reset log;

/*1 - define loglikelihood function*/

start L_lik(param) /*define param*/
            global(time,cens,surr,treat,numcents,center);
lik=0;
      theta=exp(param[,1])/(1+exp(param[,1])); /* Hougaard 0<theta<1 */

do c=1 to numcents;

      sigma=exp(param[,2+(c-1)#6]);
      mu=param[,3+(c-1)#6];
      alpha=param[,4+(c-1)#6];
      lambda=exp(param[,5+(c-1)#6]); /* exp because the parameter can have only positive values */
      p=param[,6+(c-1)#6];
      beta=(param[,7+(c-1)#6]);

      t=time[loc(center=c),];
         delta=cens[loc(center=c),];
      s=surr[loc(center=c),];
      z=treat[loc(center=c),];
```

```
*       f1=logsdf('NORMAL',s,mu+alpha#z,sigma);/*log S(s)*/ /*I didn't use it*/
        f2=-lambda#exp(beta#z)#(t##p); /*log S(t)*/
        f6=exp(-lambda#exp(beta#z)#(t##p)); /*S(t)*/
        f4=p#lambda#exp(beta#z)#(t##(p-1))#f6; /*f(t)--> Surv#Hazard – SAS parameterization */
        f3=pdf('NORMAL',s,mu+alpha#z,sigma); /*f(s)*/
        f5=sdf('NORMAL',s,mu+alpha#z,sigma); /*S(s)*/
*     print f2 f3 f4 f5 f6;

        /*define log-likelihood*/

        cop0=((-log(f5))##(1/theta))+((-log(f6))##(1/theta));
        cop=exp(-cop0##theta);

        b=log(cop)+(theta-1)#log(cop0)+(1/theta-1)#log(-log(f5))-log(f5)+log(f3);

        a=log(cop)+(theta-2)#log(cop0)+log(cop0##theta-(theta-1)/theta)+(1/theta-1)#log(-log(f5))-
log(f5)+(1/theta-1)#log(-log(f6))-log(f6)+log(f3)+log(f4);
        likc=sum(delta#a+(1-delta)#b);

        lik=lik+likc;
end;

  return(lik);

finish L_lik;

/*2 - dataset*/
use dataset;
read all var{time} into time;
read all var{cens} into cens;
read all var{surr} into surr;
read all var{treat} into treat;
read all var{center} into center;
close dataset;

cents=unique(center)`;
numcents=nrow(cents);

print cents numcents;

use estsab;
read all var{intercept} into Intercepts;
read all var{effect} into effects;
read all var{center} into centers;
close estsab;
params=Intercepts||effects||centers;
```

```
print params;

use estt;
read all var{Intercept} into Interceptt;
read all var{treat} into treatt;
read all var{_SCALE_} into _SCALE_t;
read all var{center} into centert;
close estt;
paramt=Interceptt||treatt||_SCALE_t||centert;

print paramt;

use Covparm;
read all var{residual} into residual; /*Residual*/
close Covparm;

/*3 - subroutine*/

      x=J(1,1+6#numcents,.); /*J(nrow,ncol,value);*/

*       mattrib x colname={'Theta' 'Sigma' 'Mu' 'Alpha' 'Lambda' 'p' 'Beta'};

      x[,1]=-1;

do c=1 to numcents;

      x[,2+(c-1)#6]=0.5#log(residual[c,]); /*0.5#LOG(Residual)*/
      x[,3+(c-1)#6]=Intercepts[c,];/*intercept*/
      x[,4+(c-1)#6]=effects[c,]; /*Effect of treat on S*/
      x[,5+(c-1)#6]=-(Interceptt[c,]/_SCALE_t[c,]); /* -(intercept/scale)*/
      x[,6+(c-1)#6]=1/_SCALE_t[c,]; /*1/scale*/
      x[,7+(c-1)#6]=-(treatt[c,]/_SCALE_t[c,]); /* -(regression coeffincient/scale)*/

end;

      xopt=J(1,3,.);
      xopt[,1]=1; /*indicates whether the problem is minimization or maximization: 1=specifies a maximization problem*/
      xopt[,2]=2; /*speci?es the amount of printed output. 4=the approximate covariance matrix of parameter esti- mates is printed if opt[3] is set*/
      xopt[,3]=0; /*Selects a scaling for the Hessian matrix, G: default for NLPNRR is 0= No scaling is done*/
      /*The other options are not for NLPNRR*/

      print x ;
      print xopt;
   maxiter=150;
   termin=maxiter||J(1,12,.); /*why 12?*/
   termin[1,4]=0;
```

```
   termin[1,6]=0.001;

call nlpnrr(rc,est,"L_lik",x,xopt,,termin);
/*CALL NLPNRR( rc, xr, "fun", x0 <,opt, blc, tc, par, "ptit", "grd", "hes">);*/
/*"Ridge" value should be zero, if not, the hessian is negative and the procedure don't converge.
      Solution? Change number of interation in subroutene, example in our case:"maxiter"*/

print est;

sigma=J(1,numcents,.);
mu=J(1,numcents,.);
alpha=J(1,numcents,.);
lambda=J(1,numcents,.);
p=J(1,numcents,.);
beta=J(1,numcents,.);

theta=exp(est[,1])/(1+exp(est[,1]));

do c=1 to numcents;

      sigma[,c]=exp(est[,2+(c-1)#6]);
      mu[,c]=est[,3+(c-1)#6];
      alpha[,c]=est[,4+(c-1)#6];
      lambda[,c]=exp(est[,5+(c-1)#6]);
      p[,c]=est[,6+(c-1)#6];
      beta[,c]=est[,7+(c-1)#6];

end;

 print theta;
 print sigma;
 print mu;
 print alpha;
 print lambda;
 print p;
 print beta;
 param_s=theta||sigma||mu||alpha||lambda||p||beta;

/*Standard Error*/

call nlpfdd(f,g,h,"L_lik",est);
/*CALL NLPFDD( f, g, h, "fun", x0, <,par, "grd">);*/

*print h; /* Hessian matrix*/

var=inv(-h);
se=sqrt(vecdiag(var)); /*standard error*/
```

```
*print var;
print se;

/*Delta Method*/

se_est=J(1,1+6#numcents,.);

se_est[,1]=sqrt(theta#(1-theta))#se[1,]; /*theta*/

do c=1 to numcents;

se_est[,2+(c-1)#6]=exp(est[,2+(c-1)#6])#se[2+(c-1)#6,]; /*sigma*/
se_est[,3+(c-1)#6]=se[3+(c-1)#6,]; /*mu*/
se_est[,4+(c-1)#6]=se[4+(c-1)#6,]; /*alpha*/
se_est[,5+(c-1)#6]=exp(est[,5+(c-1)#6])#se[5+(c-1)#6,]; /*lambda*/
se_est[,6+(c-1)#6]=se[6+(c-1)#6,]; /*p*/
se_est[,7+(c-1)#6]=se[7+(c-1)#6,]; /*beta*/

end;

print (t(param_s)) (t(se_est));

/*Confidence Intervals*/

lo_sigma=J(1,numcents,.);
up_sigma=J(1,numcents,.);
lo_mu=J(1,numcents,.);
up_mu=J(1,numcents,.);
lo_alpha=J(1,numcents,.);
up_alpha=J(1,numcents,.);
lo_lambda=J(1,numcents,.);
up_lambda=J(1,numcents,.);
lo_p=J(1,numcents,.);
up_p=J(1,numcents,.);
lo_beta=J(1,numcents,.);
up_beta=J(1,numcents,.);

lo_theta=exp(est[,1]-2#se[1,])/(1+exp(est[,1]-2#se[1,]));
up_theta=exp(est[,1]+2#se[1,])/(1+exp(est[,1]+2#se[1,]));

do c=1 to numcents;

lo_sigma[,c]=exp(est[,2+(c-1)#6]-2#se[2+(c-1)#6,]);
up_sigma[,c]=exp(est[,2+(c-1)#6]+2#se[2+(c-1)#6,]);
lo_mu[,c]=est[,3+(c-1)#6]-2#se[3+(c-1)#6,];
up_mu[,c]=est[,3+(c-1)#6]+2#se[3+(c-1)#6,];
lo_alpha[,c]=est[,4+(c-1)#6]-2#se[4+(c-1)#6,];
up_alpha[,c]=est[,4+(c-1)#6]+2#se[4+(c-1)#6,];
```

```
lo_lambda[,c]=exp(est[,5+(c-1)#6]-2#se[5+(c-1)#6,]);
up_lambda[,c]=exp(est[,5+(c-1)#6]+2#se[5+(c-1)#6,]);
lo_p[,c]=est[,6+(c-1)#6]-2#se[6+(c-1)#6,];
up_p[,c]=est[,6+(c-1)#6]+2#se[6+(c-1)#6,];
lo_beta[,c]=est[,7+(c-1)#6]-2#se[7+(c-1)#6,];
up_beta[,c]=est[,7+(c-1)#6]+2#se[7+(c-1)#6,];

end;

ic=lo_theta||up_theta||lo_sigma||up_sigma||lo_mu||up_mu||lo_alpha||up_alpha||lo_lambda||up_lambda||lo_p||up
_p||lo_beta||up_beta;

print (t(lo_theta)) (t(up_theta));
print (t(lo_sigma)) (t(up_sigma));
print  (t(lo_mu)) (t(up_mu));
print (t(lo_alpha)) (t(up_alpha));
print (t(lo_lambda)) (t(up_lambda));
print (t(lo_p)) (t(up_p));
print (t(lo_beta)) (t(up_beta));

quit;
```

REFERENCE

Algina, J. (1999) A comparison of methods for constructing confidence intervals for the squared multiple correlation coefficient. Multivariate Behavioral Research, 34, 494–504.

Alonso A, Elst W.V., Molenberghs G., Buyse M., Burzykowski T. (2014). On the relationship between the Causal-Inference and Meta-Analytic Paradigms for the Validation of Surrogate Endpoints. Biometrics, doi: 10.1111/biom.12245

Amemiya, T. (1984), "Tobit Models: A Survey," Journal of Econometrics, 24, 3–61.

Baker, S.G. and Kramer, B.S. (2003). A perfect correlate does not make a surrogate. BioMed Central Medical Reseach Methodology, 3, 16.

Basso et al. (2009). Risk of relapse of childhood acute lymphoblastic leukemia is predicted by flow cytometric measurement of residual disease on day 15 bone marrow. J Clin Oncol. 2009 Nov 1;27(31):5168-74.

Bessan R. et al. (2009). Improved risk classification for risk-specific therapy based on the molecular study of minimal residual disease (MRD) in adult acute lymphoblastic leukemia (ALL). Blood. 2009 Apr 30;113(18):4153-62.

Biomarkers Definitions Working Group (2001). Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. Clinical Pharmacology and Therapy, 69, 89–95.

Boissel JP, Collet J-P, Moleur P, Haugh M. Surrogate endpoints: a basis for a rational approach. European Journal of Clinical Pharmacology 1992; 43:235 –244.

Boone CW, Kelloff GJ. (1993). Intraepithelial neoplasia, surrogate endpoint biomarkers, and cancer chemoprevention. J Cell Biochem Suppl; 17F: 37–48.

Borowitz MJ et al.(2015). Prognostic significance of minimal residual disease in high risk B-ALL: a report from Children's Oncology Group study AALL0232. Blood. 2015 Aug 20;126(8):964-71. doi: 10.1182/blood-2015-03-633685. Epub 2015 Jun 29.

Buyse, M. and Molenberghs, G. (1998). The validation of surrogate endpoints in randomized experiments. Biometrics, 54, 1014–1029.

Buyse, M., Molenberghs, G., Burzykowski, T., Renard, D., and Geys, H. (2000a). The validation of surrogate endpoints in meta-analysis of randomized experiments. Biostatistics, 1, 49–67.

Buyse, M., Molenberghs, G., Burzykowski, T., Renard, D., and Geys, H. (2000b). Statistical validation of surrogate endpoints: problems and proposals. Drug Information Journal, 34, 447–454.

Buyse, M. (2009). Use of meta-analysis for the validation of surrogate endpoints and biomarkers in cancer trials. Cancer Journal, 15, 421–425.

Burzykowski, T., Molenberghs, G., Buyse, M., Renard, D., and Geys, H. (2001) Validation of surrogate endpoints in multiple randomized clinical trials with failure-time endpoints. Applied Statistics, 50, 405–422.

Burzykowski, T., Molenberghs, G., and Buyse, M. (2004) The validation of surrogate endpoints by using data from randomized clinical trials: a case study in advanced colorectal cancer. Journal of the Royal Statistical Society, Series A, 167, 103–124.

Burzykowski T.,Molenberghs G., Buyse M. (2005). The evaluation of surrogate endpoint. New York. Springer.

Bycott PW, Taylor JMG. An evaluation of a measure of the proportion of the treatment eect explained by a surrogate marker. Controlled Clinical Trials 1998; 19:555–568.

Brotman B, Prince AM. (1988). Gamma-glutamyltransferase as a potential surrogate marker for detection of the non-A, non-B carrier state. Vox Sang; 54: 144–7.

Campana, D. (2009) Role of minimal residual disease monitoring in adult and pediatric acute lymphoblastic leukemia. Hematol/Oncol Clin North America, 23, 1083-1098

Campana, D. (2010) Minimal residual disease in acute lymphoblastic leukemia. The Education Program of the American Society of Hematology, 2010, 7-12.

Cazzaniga, G., Valsecchi, M.G., Gaipa, G. Et al. (2011) Defining the correct role of minimal residual disease tests in the management of acute lymphoblastic leukemia. Br. J. Haematol., 155, 45-52.

Chen H., Geng Z., Jia J. (2007). Criteria for surrogate endpoint. J.R. Stat. Soc. Ser. B69, 919-932.

G.M. Clark, D.M. Zborowski, J.L. Culbertson et al. (2006). Clinical utility of epidermal growth factor receptor expression for selecting patients with advanced non-small cell lung cancer for treatment with erlotinib. J. Thorac. Oncol., 1, pp. 837–846

Clayton, D.G. (1978) A model for association in bivariate life Tables and its application in epidemiological studies of familial tendency in chronic disease incidence. Biometrika, 65, 141–151.

Conter, V., Bartram, C.R., Valsecchi, M.G. et al. (2010) Molecular response to treatment redefines all prognostic factors in children and adolescents with B-cell precursor acute lymphoblastic leukemia: results in 3184 children in the AEIOP-BFM ALL 2000 study. Blood, 115, 3206-3214.

Cortinas Abrahantes, J., Molenberghs, G., Burzykowski, T., Shkedy, Z., Alonso Abad, A., and Renard, D. (2004) Choice of units of analysis and modeling strategies in multilevel hierarchical models. Computational Statistics and Data Analysis, 47, 537–563.

Domenech C et al. (2014). Dexamethasone (6 mg/m2/day) and prednisolone (60 mg/m2/day) were equally effective as induction therapy for childhood acute lymphoblastic leukemia in the EORTC CLG 58951 randomized trial. Haematologica. 2014 Jul;99(7):1220-7.

Fisher, R.A. (1928) The general sampling distribution of the multiple correlation coefficient. Proceedings of the Royal Society, 121, 654–673.

Fleming T.R. (1996). Surrogate endpoints in clinical trials. Drug Information Journal, 30, 545–551.

Fleming, T.R. and DeMets, D.L. (1996). Surrogate endpoints in clinical trials: are we being misled? Annals of Internal Medicine, 125, 605– 613.

Food and Drug Administration (2012). Minimal Residual Disease (MRD) as a Surrogate Endpoint in Acute Lymphoblastic Leukemia (ALL) Workshop. FDA Briefing Document.

Frangakis C.E., Rubin D.B.(2002). Principal stratication in causal inference. Biometrics, 58, 21-29.

Freedman, L.S., Graubard, B.I., and Schatzkin, A. (1992). Statistical validation of intermediate endpoints for chronic diseases. Statistics in Medicine, 11, 167–178.

Freedman, L.S. (2001). Confidence intervals and statistical power of the 'Validation' ratio for surrogate or intermediate endpoints. Journal of Statistical Planning and Inference, 96, 143–153.

Gail, M.H., Pfeiffer, R., van Houwelingen, H.C., and Carroll, R.J. (2000).On meta-analytic assessment of surrogate outcomes. Biostatistics, 1, 231–246.

Gaipa G. et al. (2012), Time point-dependent concordance of flow cytometry and real-time quantitative polymerase chain reaction for minimal residual disease detection in childhood acute lymphoblastic leukemia. Haematologica. 2012 Oct;97(10):1582-93.

Glidden DV. (2000), A Two-Stage Estimator of the Dependence Parameter for the Clayton-Oakes Model. Lifetime Data Anal. Jun;6(2):141-56.

Gumbel, E.J. (1960) Bivariate exponential distributions. Journal of the American Statistical Association, 55, 698–707.

Hougaard, P. (1986) Survival models for heterogeneous populations derived from stable distributions. Biometrika, 73, 387–396.

Hougaard, P. (1987) Modelling multivariate survival. Scandinavian Journal of Statistics, 14, 291–304.

Karol SE et al. (2015), Prognostic factors in children with acute myeloid leukaemia and excellent response to remission induction therapy. Br J Haematol. 2015 Jan;168(1):94-101

Lin DY, Fleming TR, De Gruttola V. Estimating the proportion of treatment eect explained by a surrogate marker. Statistics in Medicine 1997; 16:1515 –1527.

Molenberghs, G., Buyse, M., Geys, H., Renard, D., Burzykowski, T., and Alonso, A. (2002). Statistical challenges in the evaluation of surrogate endpoints in randomized trials. Controlled Clinical Trials, 23, 607–625.

Nelsen, Roger B. (1999), An Introduction to Copulas, New York: Springer, ISBN 0-387-98623-5

Paone JF, Waalkes TP, Baker RR, Shaper JH. (1980). Serum UDP-galactosyl transferase as a potential biomarker for breast carcinoma. J Surg Oncol; 15: 59–66

Parker C. et al.(2010). Effect of mitoxantrone on outcome of children with first relapse of acute lymphoblastic leukaemia (ALL R3): an open-label randomised trial. Lancet. 2010 Dec 11;376(9757):2009-17.

Plackett, R.L. (1965) A class of bivariate distributions. Journal of the American Statistical Association, 60, 516–522.

Prentice, R.L. (1989). Surrogate endpoints in clinical trials: definitions and operational criteria. Statistics in Medicine, 8, 431–440.

Pui C.H. (2015); Clinical utility of sequential minimal residual disease measurements in the context of risk-based therapy in childhood acute lymphoblastic leukaemia: a prospective study. Lancet Oncol. 2015 Apr;16(4):465-74.

Puing N. (2014), Critical evaluation of ASO RQ-PCR for minimal residual disease evaluation in multiple myeloma. A comparative analysis with flow cytometry. Leukemia. 2014 Feb;28(2):391-7.

Raponi et al (2014), Minimal residual disease monitoring in chronic lymphocytic leukaemia patients. A comparative analysis of flow cytometry and ASO IgH RQ-PCR. Br J Haematol. 2014 Aug;166(3):360-8.

Renfro L.A., Shi Q., Xue Y, Li J., Shang H., Sargent D.J. (2014). Center-within-trial versus trial-level evaluation of surrogate endpoint. Computationl statistic & Data Analysis, 78, 1-20.

Schrappe M, Valsecchi MG, Bartram CR, Schrauder A, Panzer-Grümayer R, Möricke A, Parasole R, Zimmermann M, Dworzak M, Buldini B, Reiter A, Basso G, Klingebiel T, Messina C, Ratei R, Cazzaniga G, Koehler R, Locatelli F, Schäfer BW, Aricò M, Welte K, J.M. van Dongen J, Gadner H, Biondi A, Conter V (2011). Late MRD response determines relapse risk overall and in subsets of childhood T-cell ALL: results of the AIEOP-BFM-ALL 2000 study. Blood, 118: 2077-2084.

Shih, J.H. and Louis, T.A. (1995a) Inferences on association parameter in copula models for bivariate survival data. Biometrics, 51, 1384–1399

Tibaldi, F.S., Corti˜nas Abrahantes, J., Molenberghs, G., Renard, D., Burzykowski, T., Buyse, M., Parmar, M., Stijnen, T., and Wolfinger, R. (2003) Simplified hierarchical linear models for the evaluation of surrogate endpoints. Journal of Statistical Computing and Simulation, 73, 643–658.

Tobin J. Estimation of relationships for limited dependent variables. Econometrica. 1958;26:24–36

Van Dongen J.J.M. et al. (2015). Minimal residual disease diagnostics in acute lymphoblastic leukemia: need for sensitive, fast, and standardized technologies. Blood. 2015 Jun 25;125(26):3996-4009.

Van Houwelingen, J.C., Arends, L.A., and Stijnen, T. (2002) Advanced methods in meta-analysis: multivariate approach and meta-regression. Statistics in Medicine, 21, 589-624.

Vora et al. (2013).Treatment reduction for children and young adults with low-risk acute lymphoblastic leukaemia defined by minimal residual disease (UKALL 2003): a randomised controlled trial. Lancet Oncol. 2013 Mar;14(3):199-209.