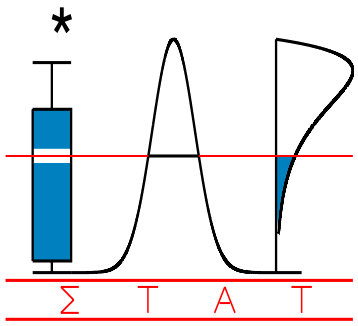
T E C H N I C A L

R E P O R T

0405

ASSESSING HETEROGENEITY IN OUTCOME BETWEEN INSTITUTIONS IN RANDOMIZED CLINICAL TRIALS WITH TIME TO EVENT ENDPOINTS

Catherine LEGRAND, Luc DUCHATEAU, Richard SYLVESTER,
Paul JANSSEN and Patrick THERASSE



I A P S T A T I S T I C S

N E T W O R K

INTERUNIVERSITY ATTRACTION POLE

http://www.stat.ucl.ac.be/IAP

Assessing heterogeneity in outcome between institutions in randomized clinical trials with time to event endpoints

Catherine Legrand, MSc[a,*], Luc Duchateau, PhD[b], Richard Sylvester, ScD[a], Paul Janssen, PhD[c], Patrick Therasse, MD[a].

[a]European Organization for Research and Treatment of Cancer, Brussels, Belgium.

[b]Faculty of Veterinary Medicine, Ghent University, Merelbeke, Belgium.

[c]Limburgs Universitair Centrum, Center for Statistics, Diepenbeek, Belgium.

[*]Corresponding author: C. Legrand, EORTC, Av. E. Mounier 83 box 11 - 1200 Brussels - Belgium, Email: cle@eortc.be

Running title: Heterogeneity in outcome over institutions

Abstract

Treatment outcome research investigates the heterogeneity in outcome between patients according to factors such as country or institution. Most treatment outcome studies are performed on data from registries. However, data from multi-center clinical trials may also be relevant for treatment outcome research. The treatment and outcome measures are more standardized and more detailed information on prognostic factors, measured in a standardized way, is available. In this paper, we investigate the heterogeneity between institutions in disease-free survival in an international multi-center randomized breast cancer trial with 2793 patients from 14 institutions. To explain heterogeneity, factors inherent to the institution and factors inherent to the patients, e.g. patient baseline characteristics, are considered. Most often descriptive statistics are used but these lead to highly subjective conclusions and can thus be misleading. We demonstrate how the frailty model can be used to investigate heterogeneity between institutions in an objective way. In particular we show that although descriptive statistics suggest that type of surgery explains heterogeneity in disease free survival, frailty modeling reveals that the geographical area is in fact the only factor which explains differences in outcome between institutions in this study.

Keywords: frailty model, treatment outcome, multicenter clinical trial, heterogeneity.

1 Introduction

While the main objective of randomized controlled clinical trials is to compare treatments for a specific disease, the objective of treatment outcome studies is to investigate the heterogeneity in outcome between patients according to factors such as country, institution or physician. Such studies can lead to improvements in the quality of care by pointing out which factors are associated with a better outcome. There is no unique way of conducting, analyzing and drawing conclusions from a treatment outcome study. A broad set of questions can be investigated and a wide range of approaches can be used in an attempt to answer these questions [1].

Most of the published treatment outcome studies have been performed using cancer registry data. In this paper we will study the heterogeneity between institutions in cancer clinical trials with time to event endpoints (e.g. overall survival, disease-free survival,...). However most of our findings can be easily extended to other disease types. Although one might argue that the strict conditions under which patients are treated within a clinical trial lead to less variability in outcome, one of the objectives of this paper is to show that this might also be advantageous. The availability of most important variables, measured in a standard way, makes the analysis more trustworthy as compared to an analysis based on registry data.

Once it has been shown that there is heterogeneity in outcome between institutions, a key issue in treatment outcome studies is to explain the heterogeneity using both "institution" and "patient" level factors. In most treatment outcome studies, these two steps

(showing heterogeneity in outcome and attempting to explain it) are performed using mainly descriptive statistics and sometimes fixed effect models. However such methods do not allow one to conclude that there exists more heterogeneity than what is to be expected only by chance nor to explain this heterogeneity. In these treatment outcome studies, trying to explain heterogeneity often consists in "looking at" heterogeneity in distribution of important prognostic factors across institutions. Such analyses are highly subjective and their interpretation therefore controversial. In this paper, we demonstrate how the frailty model can be used as an objective tool to demonstrate the presence of heterogeneity in time to event outcomes between institutions and to identify institution and patient level factors that explain this heterogeneity.

We first review the advantages of using clinical trial data over registry data to perform treatment outcome research (section 2) and the factors that may explain heterogeneity among institutions in the context of clinical trial data (section 3). Section 4 presents the classical techniques of treatment outcome research which are mainly based on descriptive methods while we show in section 5 how these methods can be supplemented by the frailty model. These two sections are illustrated by a case study investigating the heterogeneity in outcome found between institutions in an international breast cancer phase III trial (EORTC Trial 10854) comparing surgery followed by a short intensive course of peri-operative polychemotherapy versus surgery alone. This trial accrued a total of 2793 patients in 13 European and one South African institution. These results are discussed in the last section.

2 Use of clinical trial data to perform treatment outcome research

Treatment outcome research has mainly been based on cancer registries because they constitute a large amount of data that are representative of the whole population under study and are therefore not affected by selection biases as is the case for clinical trials. There are, however, also some major drawbacks. The major sources of information for most registries are medical records, pathology files and death certificates. However, the completeness of follow up and validity of the data (e.g. reliability of the diagnosis, classifications of lesions, staging information, ...) can be influenced by a number of factors which may differ from registry to registry [1]. Furthermore, time to event endpoints can hardly be considered when using data from registries. First, the start date from which the duration of the event is measured is often ill-defined and may vary from one registry to the other. If the date used as a starting point (e.g. date first seen in the hospital, date of diagnosis, date of first treatment, ...) in one of the databases is earlier in the course of the disease, then the time to event may seem to be longer for patients in this database. The choice of a starting date is linked to problems of lead-time bias (time to event appears to be longer because screening or early diagnosis accelerates the diagnosis while the date of event is not postponed). Second, the accuracy of the end date depends on the follow up methods. Some registries use active follow up procedures (contacting either the hospital or general practitioner responsible for the patient's care) to ascertain the long term outcome (e.g. progression or death), while others depend on more passive data collection [2].

Surprisingly, only few treatment outcome studies use data from randomized multicenter trials. Although one might argue that the strict conditions under which patients are treated within a clinical trial lead to less outcome variability, this can also be advantageous, especially if heterogeneity is detected, at the time of interpreting the results. Standardization of the treatment and a standardized measurement of important variables (pertaining to diagnosis, tumor staging, prognostic factors, treatment, follow up data, ...) makes the analysis more trustworthy as compared to an analysis based on registry data. Even if patients entered in a clinical trial are not a random sample of the population defined by the inclusion and exclusion criteria of the trial, the collection of important prognostic factors and eligibility criteria allow us to check whether differences in outcome are due to differences in the patient populations referred to the different institutions. Furthermore, the starting date for time to event endpoints is defined for all the patients as the date of randomization and all patients are followed in the same way, usually until death. Although the sample size of cancer clinical trials is generally much smaller than what is available through registries, we have shown in previous work that in most tumor types, the sample size achieved in large multicenter phase III trials is sufficient to perform treatment outcome research [3].

3 Explaining heterogeneity between institutions

In the presence of heterogeneity between institutions, the key point is to find factors that explain this heterogeneity. The location and the characteristics of an institution might influence the type of patients treated in this institution. Furthermore, each institution could decide to enter only a sub-population of the population described in the protocol

by the eligibility criteria, for example because they have a competitive study running. Two types of factors can be considered when trying to explain heterogeneity in outcome: factors that change only at the institution level and factors that change at the individual patient level.

Explaining heterogeneity by institution specific information

Institution level factors, i.e. factors that have the same value for all patients treated within the same institution but which vary between institutions, might explain heterogeneity in outcome. Typically such factors inherent to the institution are institution's location (geographical area, urban/rural location, ...), institution type (teaching institution, specialized cancer center, ...) and institution's size. However, once such factors are found to explain (part of) the heterogeneity between institutions, the interpretation of such findings remains challenging and requires further investigation. To avoid controversy at the time of interpreting the results, we have to keep in mind that differences in outcome between categories of institutions could simply be due to differences in patient populations referred to such institutions (e.g. worst cases are usually referred to university hospitals).

Explaining heterogeneity by patient specific information

Patient level factors are factors that do not have the same value for all patients treated within the same institution. Typical examples of such factors are the baseline characteristics of the patients treated. As mentioned above, differences in outcome between institutions or groups of institutions might be due to differences in the institution's patient population.

An individual patient level factor can only explain heterogeneity found between institutions if two requirements are fulfilled. First, the factor should have a significant impact on the outcome of the patient, i.e. should be of prognostic importance for the disease. Second, there has to be an imbalance in the distribution of the factor over the different institutions, so that some institutions have relatively more patients of good prognosis according to this factor than others.

A major question addressed by treatment outcome research based on non clinical trial data (e.g. registry data) is whether differences in outcome can be explained by differences in treatment policy between institutions. On the other hand, in randomized clinical trials, all patients are treated according to the same protocol and the treatment assignment is often stratified by institution so that the proportion of patients assigned to a particular treatment arm is virtually the same in all institutions. Therefore, even if the experimental treatment arm has a significant impact on the outcome of the patients as compared to the control treatment arm, the second requirement is not fulfilled and the treatment assignment can therefore never explain heterogeneity in outcome among institutions.

4 Statistical techniques most often used in treatment outcome research

Showing heterogeneity in outcome between institutions

The techniques most often used to evaluate whether there is heterogeneity in time to event outcomes between institutions are based on descriptive methods including tables of the

percentage of patients with an event per institution, Kaplan-Meier curves by institution, possibly with summary statistics such as the median time to event and the event free rate at some fixed time points. However such descriptive results are difficult to interpret and it is impossible to conclude whether there is more heterogeneity between institutions than what is expected only by chance.

The Cox proportional hazards model with institution included as a fixed effect or a log-rank test between institutions is sometimes used. Note that a fixed effect for institution assumes that the levels of these factors (namely the institutions considered) are by themselves of interest and have been intentionally "fixed" by the study design. Andersen et al. [4] showed that a fixed effect test for institutions rejects the hypothesis of no institution effect too often when the null hypothesis is true unless the number of subjects in each institution is very large. This type of test should therefore not be used.

Explaining heterogeneity in outcome between institutions

When investigating whether institution level factors could explain potential heterogeneity in outcome, similar techniques are used grouping institutions according to the factor considered rather than looking at individual institutions.

At the patient level, the prognostic importance of factors is usually studied by means of the logrank test or the univariate Cox model. Once important prognostic factors are identified, descriptive tables are most often used to study heterogeneity in the distribution of these factors over institutions. Conclusions drawn from these analyses too often assume that imbalances in an important prognostic factor is enough to explain heterogeneity in

outcome.

A better, more objective method to describe the heterogeneity in distribution of factors over institutions is based on the generalized mixed model [5] with a random institution effect.

The generalized linear mixed model

The heterogeneity in distribution of baseline characteristics over institutions can be modeled using the generalized linear model.

Consider the binary baseline characteristic X with two levels say x_0 and x_1. The observed number of patients with baseline characteristics x_0 in a particular institution i (with corresponding observed probability p_i) can be assumed to follow a binomial distribution with parameters n_i, the number of patients in the institution and π_i the probability of having baseline characteristics x_0 in institution i. To study the heterogeneity in the distribution of X over institutions we will consider the heterogeneity of π_i over institutions fitting a generalized linear mixed model with a logistic link function and a random institution effect c_i:

$$\eta_i = \ln \frac{\pi_i}{1-\pi_i} = \beta_0 + c_i \qquad \text{with} \qquad n_i p_i \sim B(n_i, \pi_i). \tag{1}$$

The random institution effect c_i is assumed to come from a normal distribution $c_i \sim N(0, \sigma_c^2)$. Therefore, the estimate of the variance σ_c^2 describes the heterogeneity among institutions. The overall probability over the different institutions is estimated by

$$\hat{\pi}_0 = \frac{\exp(\hat{\beta}_0)}{1+\exp(\hat{\beta}_0)}. \tag{2}$$

The predicted probability for the i^{th} institution is estimated by:

$$\hat{\pi}_i = \frac{\exp(\hat{\eta}_i)}{1+\exp(\hat{\eta}_i)}. \tag{3}$$

Note that $\eta_i \sim N(\beta_0, \sigma_c^2)$ with β_0 and σ_c^2 the population parameters. To get a feel for the heterogeneity between institutions we consider the interval spanned by the 5^{th} and the 95^{th} quantiles of the distribution of η_i, i.e. $[\beta_0 - 1.65\sigma_c, \beta_0 + 1.65\sigma_c]$. By back transformation we obtain the following inter-quantile interval around π_i :

$$\left[\frac{\exp(\beta_0 - 1.65\sigma_c)}{1+\exp(\beta_0 - 1.65\sigma_c)}, \frac{\exp(\beta_0 + 1.65\sigma_c)}{1+\exp(\beta_0 + 1.65\sigma_c)}\right] \tag{4}$$

with 90% of the institutions having probabilities within these limits. Inserting the estimated values for β_0 and σ_c^2 will thus provide us with an interpretation of the variance component σ_c^2.

This model is fitted with the SAS procedure PROC NLMIXED [6] which integrates the random effect out of the likelihood by using its density function. Approximate standard errors of the estimates are based on the information matrix of the likelihood function and predictions of the random effect are based on empirical Bayes estimation.

A case study: EORTC Trial 10854

EORTC trial 10854 is a multicenter randomized phase III trial comparing one course of peri-operative chemotherapy (experimental arm) with no further treatment (control arm) following potentially curative surgery of carcinoma of the breast [7,8]. Between May 1986 and March 1991, a total of 2793 T_{1-2-3} N_{0-1-2} M_0 breast cancer patients were accrued by 14 institutions in 8 countries (7 European countries and South Africa). Note that 2 institutions recruited more than half of the patients, 4 institutions recruited an

intermediate number of patients and 8 institutions recruited less than 100 patients each (Table 1). After a median follow up of 10.7 years, no significant differences between the two treatment arms were found for overall survival (HR: 0.906; 95%CI: 0.786-1.043, p-value: 0.1701) while a difference in disease free survival (DFS) was significant in favor of the experimental arm (HR: 0.855; 95%CI: 0.755-0.968, p-value: 0.0133). In this study, we primarily consider heterogeneity in disease free survival among institutions; secondary endpoints are overall survival and time to loco-regional recurrence.

In all tables and figures, institutions are ordered in the same way according to 5 a-priori defined geographical areas: institutions 1 to 3 are located in The Netherlands, institutions 4 and 5 in Poland, institutions 6 to 9 are in France, while institutions 10 to 13 are located in South Europe and institution 14 in South Africa.

Heterogeneity in outcome between institutions

There seems to be some variation in outcome among institutions (Table 1). The 5 year disease free survival varies between 57.8% (institution 14; 95% CI: 51.0-64.5%) and 81.5% (institution 9; 95% CI: 79.0-84.1%). Kaplan-Meier disease free survival curves for each institution are shown in Figure 1. However, the large number of curves plotted on the same figure makes the interpretation difficult. Furthermore we have to keep in mind that the width of the confidence interval associated with each curve varies with the number of events in each institution. A classical logrank test leads to a significant difference between the institutions ($p < 0.001$). Therefore substantial heterogeneity in outcome among institutions seems to be present.

Explaining heterogeneity by institution specific information

There seems to be a relationship between the geographical area and the disease free survival in each institution (Table 1). Pooling data by geographical area, South Africa and South Europe have the lowest 5 year disease free survival (57.8%, 95%CI: 51.0-64.5 and 66.8%, 95%CI: 58.9-73.8 respectively) with institution 5-year disease free survival rates between 57.8% and 70.2%. At the other extreme France seems to have the highest 5 year survival (79.2%, 95%CI: 77.4-81.0) with 5-year disease free survival rates between 75.2% and 81.5%. The situation in Poland and in The Netherlands is more difficult to interpret.

Explaining heterogeneity by patient specific information

Univariate analyses identified the type of surgery (mastectomy versus breast conservative surgery), axillary nodal status (negative versus positive), tumor size (not palpable or less than 2 cm versus 2 cm or more or advanced disease) and presence/absence of other concomitant diseases as having a prognostic impact on the disease free survival of the patients (Table 2). Note that according to the design of this study, randomization occured after surgey, therefore type of surgery should indeed be considered as a baseline characteristic

As expected, the treatment allocation over institutions is perfectly well balanced due to the stratification applied and can thus not explain heterogeneity (Table 1 - Figure 2a).

On the other hand, type of surgery has the largest imbalance in distribution over in-

stitutions. The observed percentage of patients who undergo mastectomy varies between 27.1% (institution 9) to 97.5% (institution 4). The generalized linear model estimates the average probability of undergoing mastectomy to be 72.8% with a wide inter-quantiles interval (Table 3 - Figure 2b) which confirms this apparent heterogeneity in distribution. Therefore type of surgery fulfils the two criteria as it is a strong prognostic factor for disease free survival and it shows a large variability in distribution over institutions. We would therefore be tempted to argue that the imbalance in the type of surgery over institutions explains the heterogeneity found in disease free survival between institutions. However, the classical methods do not allow us to conclude that type of surgery actually explains this heterogeneity and we need additional statistical tools to draw any conclusions. One of the main problems is that the observed association is based on means over institutions, both for the outcome and the prognostic factor, whereas the association should be demonstrated on an individual patient level.

The other important prognostic factors, namely axillary nodal status, tumor size and presence of other concomitant diseases also have quite a large variability in distribution over institutions and might therefore also be seen as a possible explanation for the heterogeneity in outcome between institutions.

5 Treatment outcome research using the frailty model

Heterogeneity between institutions in time to event endpoints can be modeled by the frailty model [9,10,11]. Furthermore, both institution level and patient level factors can be introduced in the model to assess their effect on reducing (and thus explaining) the

between institution heterogeneity.

In the shared frailty model a random effect is introduced for each institution so that patients from one institution share the same random effect. The random effect describes the unobserved influences common to all patients of that particular institution. The variance of this random effect is a measure of the heterogeneity in outcome between institutions.

As a further exploratory tool we can also plot the value of the predicted random institution effect obtained from a model with and without fixed effects for the factors of interest.

The frailty model

The frailty model is an extension of the Cox model allowing for a random institution effect. The hazard rate for subject j in institution i is given by

$$\lambda_{ij}(t) = \lambda_0(t) \exp(w_i + \beta' Z_{ij}) \qquad j = 1, \ldots, n_i; i = 1, \ldots, G \tag{5}$$

where λ_0 is the baseline hazard rate, Z_{ij} is the incidence vector of the covariates for patient j in institution i and β is the vector of unknown regression coefficients and w_i is the random effect for institution i. The w_i's are assumed to be independent and identically distributed.

An equivalent formulation of the frailty model is

$$\lambda_{ij}(t) = \lambda_0(t) u_i \exp(\beta' Z_{ij}) \qquad j = 1, \ldots, n_i; i = 1, \ldots, G \tag{6}$$

where $u_i = \exp(w_i)$ is now called the frailty term for institution i. Institutions with a frailty term larger than 1 will have a higher hazard rate and therefore patients in that institution will experience, on average, the event at an earlier time than patients in an institution with a frailty term smaller than 1. In the further discussion the w_i's will mainly be used because they can be seen as random effect terms in a linear model for the log-hazard function, which makes a graphical interpretation easier.

We will assume that the frailties come from a one-parameter gamma density with mean 1 and variance θ, i.e.$f_u(u) = \frac{1}{\Gamma(\frac{1}{\theta})}\theta^{-\frac{1}{\theta}}u^{\frac{1}{\theta}-1}\exp(\frac{-u}{\theta})$. The variance θ of the frailty term represents the heterogeneity among institutions and is thus called the heterogeneity parameter. If θ equals 0 there is no heterogeneity between institutions.

The two most commonly encountered approaches for fitting semiparametric frailty models are based on the EM algorithm [9] and on the penalized partial likelihood [11,12]. For the one-parameter gamma frailty density, the two approaches give the same solution [11,3] and the penalized partial likelihood implemented in Splus was used to fit the models.

Interpretation of the heterogeneity parameter in a gamma frailty model

To get a better interpretation of a particular value of the heterogeneity parameter θ, we look at the impact of such a value on the spread of the median time to event from institution to institution by considering the density function of the median time to event over institutions. We show in the appendix that, if we assume a constant baseline hazard

λ_0, this density function $f_{T_{0.5,z}}(t)$ is given by

$$f_{T_{0.5,z}}(t) = \left(\frac{\ln(2)}{\theta\lambda_0 \exp(\beta z)}\right)^{\frac{1}{\theta}} \frac{1}{\Gamma(\frac{1}{\theta})} \left(\frac{1}{t}\right)^{1+\frac{1}{\theta}} \exp\left(-\frac{\ln(2)}{\theta\lambda_0 t \exp(\beta z)}\right) \tag{7}$$

In trials with a long median time to event where in most institutions the median has not been reached, such as the early breast cancer clinical trial considered in this manuscript, the median time to event is a rather meaningless summary statistic. Although not commonly used in the medical context, the 25% quantile instead of the median time to event is more relevant in this setting. While the median disease free survival represents the time point at which half the patients have experienced the event, the 25% quantile can be interpreted as the time at which one fourth of the patients have experienced the event. Formula (7) can be easily adapted to derive the density function of the 25% quantile time to event:

$$f_{T_{0.5,z}}(t) = \left(\frac{\ln(4/3)}{\theta\lambda_0 \exp(\beta z)}\right)^{\frac{1}{\theta}} \frac{1}{\Gamma(\frac{1}{\theta})} \left(\frac{1}{t}\right)^{1+\frac{1}{\theta}} \exp\left(-\frac{\ln(4/3)}{\theta\lambda_0 t \exp(\beta z)}\right) \tag{8}$$

By considering this density function and its tails, for example the 5% (q_5) and 95% (q_{95}) quantiles of this density function, we get an immediate idea of the heterogeneity in outcome that can be expected between institutions. Indeed we can deduce that with a particular value of θ, and under the assumption made about the constant baseline hazard and the hazards ratio and assuming a gamma distribution of the random effect, 90% of the centers would have their median or 25% quantiles of the disease free survival distribution (depending on the formula used) between q_5 and q_{95}.

Physicians are more used, in such a setting, to consider fixed time point disease free

survival estimates (e.g. 5 years DFS) rather than the 25% quantile of the disease free survival distribution. Similarly the density function $f_{S_{5,z}}(s)$ of the 5 year DFS over centers can be derived assuming a proportional hazards model:

$$f_{S_{5,z}}(s) = \frac{s^{\frac{1}{5\lambda_0\theta\exp(\beta z)}-1}}{(5\lambda_0\theta\exp(\beta z))^{\frac{1}{\theta}}\Gamma\left(\frac{1}{\theta}\right)}\left(\ln\frac{1}{s}\right)^{\frac{1}{\theta}-1}. \tag{9}$$

A case study: EORTC Trial 10854

Heterogeneity in outcome between institutions

The estimated heterogeneity parameter θ for disease free survival with treatment as a fixed effect equals 0.0665 (Table 4) and the predicted random institution effects are plotted in Figure 3. To ensure that such a value of θ contradicts the null hypothesis of no heterogeneity between institutions, we ran large scale simulations using the same simulations parameters and the same distribution of patients among centers as the values observed from the data [3]. From these simulations we observed that in case the true value of θ is 0 (no heterogeneity), 95% of the estimated values would be below 0.006. Therefore a value above the 0.006 contradicts the hypothesis of no heterogeneity between institutions.

To get a better understanding of what a value of 0.0665 for the estimated heterogeneity parameter means, we plotted the density function of the 25% quantile of the disease free survival time over institutions (Figure 4a) and the density function of the 5 year disease free survival over institutions (Figure 4b) in the control arm and in the experimental arm, assuming a constant yearly baseline hazard of 0.034, as observed in our data. With these assumptions, 90% of the institutions would have a 25% quantile of disease free survival time between 5.8 and 13.6 years in the experimental arm and between 4.9 and 11.6 years

in the control arm. This corresponds to 90% of the institutions having a 5 year disease free survival between 78.0% and 90.0% in the experimental arm and between 74.8% and 88.4% in the control arm.

We can therefore conclude that substantial heterogeneity in disease free survival exists in this trial. Considering the secondary endpoints, we observed even larger heterogeneity for overall survival but a smaller heterogeneity for time to first loco-regional recurrence (LRR) (Table 4). As surgery was the primary treatment of these patients, the fact that time to LRR exhibits the lowest heterogeneity among institutions may look surprising. However, time to LRR was censored at the time of distant metastases and we can suspect a "dilution" effect due to problem of under-reporting of LRR. In the remainder we will concentrate on disease free survival.

Explaining heterogeneity by institution specific information

We previously observed a relationship between the geographical area and the disease free survival in each institution and the plot of the predicted random institution effects for disease free survival seems to confirm this relationship (Figure 3). All the institutions from Southern Europe (institutions 10 to 13) and South Africa (institution 14) show a positive random effect for institution, corresponding to a higher event risk. On the other hand, all the institutions located in France (institutions 6 to 9) show a negative random effect for institution, corresponding to a lower event risk. For the Netherlands (institutions 1 to 3), one institution has a positive random institution effect and seems therefore to behave differently than the two other ones. In Poland (institutions 4, 5), one institution has a

positive random institution effect, an opposite result from the other Polish institution. When adding a fixed effect for geographical area to the frailty model, the estimate of the heterogeneity parameter decreases dramatically to 0.0078 (Table 4) and the geographical area has a significant effect ($p < 0.001$). This model confirms that patients treated in France and to a lesser extent in Poland have a better outcome while patients treated in Southern Europe and South Africa seem to have a higher risk of event. Figure 5 shows the predicted random institution effects for the model with and without fixed effects for geographical area with predictions in the model with geographical area getting closer to 0 for all the institutions. This shows that geographical area largely explains the heterogeneity in disease free survival among institutions.

Explaining heterogeneity by patient specific information

As expected, fitting a gamma frailty model on disease free survival with a random effect for institution, with or without treatment as a fixed effect, leads to estimates of θ of respectively 0.0665 and 0.0657, very similar to each other. Figure 3 clearly shows that treatment does not reduce the heterogeneity in outcome among institutions.

Our previous results suggested that type of surgery might explain heterogeneity in outcome as this factor has a strong prognostic impact and shows a large variability in distribution over institutions. When the predicted probability of mastectomy is compared with the value of the random institution effect shown in Figure 3, it seems that institutions with a positive value of the random effect and therefore a higher risk of event also have a higher proportion of patients treated by modified radical mastectomy while institutions

with a negative value of the random effect show a lower proportion of patients treated with mastectomy. This seems especially true for patients treated in France and South Europe but does not really hold for the other institutions.

However, when introducing type of surgery as fixed effect in our frailty model we observe that, although this factor has strong prognostic impact, the value of the estimated heterogeneity parameters does not decrease (0.0695). Figure 5 confirms that in fact this factor does not help to explain heterogeneity among institutions in this trial.

Considering axillary nodal status, tumor size and presence of other concomitant diseases, the use of the frailty model confirms that none of these important prognostic factors explain heterogeneity in outcome between institution in our data.

6 Discussion

As clinical trial protocols are written with the objective of suppressing as much variability as possible, investigating heterogeneity in outcome between institutions within large phase III cancer clinical trials has rarely been done. However, in case heterogeneity is found between institutions, availability of all important patient information and standardization of the treatment, the data collection and the follow up of the patients are major advantages when trying to explain the observed heterogeneity.

Using frailty models, we find substantial heterogeneity in disease free survival between institutions within a large early breast cancer clinical trial (EORTC 10854). We also show

that geographical area largely explains this heterogeneity in outcome. Institutions from South Africa and Southern Europe seems to have worse outcomes than the average while institutions from France and the Netherlands, with the exception of one institution, seem to have a better outcome. Contrary to what the descriptive approach suggests, it seems that none of the baseline patient characteristics considered can explain the heterogeneity. Interpreting the geographical area effect is a further challenge. Further steps consist of confirming these results and trying to explain this heterogeneity using data from other clinical trials in the same disease type, and eventually in other disease types.

In the context of breast cancer, we could not find any studies in the literature based on clinical trial data which investigated differences in outcome between institutions or geographical area. On the other hand, several cancer registry-based treatment outcome studies show differences in outcome of breast cancer according to geographical area in different parts of the world (within Nordic countries [13, 14], within the UK [15], and within the US [16,17, 18, 19], ...). The Eurocare II project [Quinn 1998], also based on cancer registry data, shows wide differences in relative survival among 17 European countries: survival was above the European average in Iceland, Finland, Sweden, Switzerland, France and Italy; about average in Denmark, The Netherlands, Germany and Spain; below average in Scotland, England and Slovenia; and well below average in Slovakia, Poland and Estonia. In most of these studies differences in timing of the diagnosis (and therefore in stage distribution) or in the treatment procedure are suggested as a possible explanations. By using clinical trial data, these two factors are controlled in the type of analysis we performed.

Based on these results, we advocate the use of the frailty model with an institution random effect as an efficient tool to assess heterogeneity in outcome and the impact of institution and patient level factors on this heterogeneity. However, we have to keep in mind that such treatment outcome research remains retrospective and mainly data-driven. Conclusions from such results should be drawn with caution and any link to quality of care without any additional information is dangerous. Comparison of our results with results obtained using other clinical trial data, eventually with registry based information available from the literature and with the opinion of medical experts, is the best way to bring these important conclusions into practice in order to improve the quality of cancer care.

Acknowledgement

Paul Janssen acknowledges partial support by the Ministry of the Flemish Community (Project BIL00/28, International Scientific and Technological Cooperation) and by the Interuniversity Attraction Pole Programme research network P5/24 of the Belgian Government (Belgian Science Policy).

References

[1] Legrand C, Sylvester R, Duchateau L, Janssen P, Therasse P. Treatment outcome studies: pitfalls in current methods and practice. Eur J Cancer 2002; 38: 1173-1180.

[2] Prior P, Woodman CB, Collins S. International difference in survival from colon-cancer: more effective care versus less complete registration. Br J Surg 1998; 85: 101-101.

[3] Duchateau L, Janssen P, Lindsey P et al. The shared frailty modem and the power for heterogeneity tests in multicenter trials. Comput Stat Data An 2002; 40: 603-620.

[4] Andersen PK, Klein JP, Zhang MJ. Testing for centre effects in multi-centre survival studies: a Monte carlo comparison of fixed and random effects tests. Stat Med 1999; 18: 1489-1500.

[5] Dobson AJ. An introduction to generalized linear models. 1st ed. London: Chapman & Hall; 1990.

[6] Clahsen PC, van de Velde CJH, Julien JP et al. Improved Local control and Disease-free survival after perioperative chemotherapy for early-stage breast cancer: a European Organization for Research and Treatment of Cancer Breast Cancer Cooperative Group study. J Clin Oncol 1996; 14: 745-753.

[7] van der Hage JA, van de Velde CJH, Julien JP. Improved survival after one course of perioperative chemotherapy in early breast cancer patients: long term results from the European Organization for Research and Treatment of Cancer (EORTC) Trial 10854. Eur J Cancer 2001; 37: 2184-2193.

[8] Klein JP. Semiparametric estimation of random effects using the Cox model based on the EM algorithm. Biometrics 1992; 48: 795-806.

[9] Klein JP, Moeschberger ML. Survival analysis. Techniques for censored and truncated

data. New York: Springer Verlag; 1997.

[10] Therneau TM, Grambsch PM. Modeling survival data. Extending the Cox model. New York: Springer Verlag; 2000.

[11] McGilchrist CA. REML estimation for survival models with frailty. Biometrics 1993; 49: 221-225.

[12] SAS/STAT User's Guide, Version 8. SAS Institute Inc., Cary, NC 1999.

[13] Engeland A, Haldorsen T, Dickman PW at al. Relative survival of cancer patients, a comparison between Denmark and the other Nordic countries. Acta Oncol 1998; 37: 49-59.

[14] Kliukiene J, Andersen A. Survival of breast cancer patients in Lithuania and Norway, 1988-1992. Eur J Cancer 1998; 34: 372-7.

[15] Purushotham AD, Pain SJ, Miles D, Harnett A. Variations in treatment and survival in breast cancer. Lancet Oncol 2001; 2: 719-25.

[16] Farrow DC, Samet JM, Hunt WC. Regional variation in survival following the diagnosis of cancer. J Clin Epidemiol 1996; 49: 843-7.

[17] Goodwin JS, Freeman JL, Freeman D, Nattinger AB. Geographic variations in breast cancer mortality: do higher rates imply elevated incidence or poorer survival? AM J Public Health 1998; 88: 458-60.

[18] Goodwin JS, Freeman JL, Mahnken JD, Freeman DH, Nattinger AB. Geographic variations in breast cancer survival among older women: implications for quality of breast cancer care. J Gerontol A Biol Sci Med Sci 2002; 57: M401-6.

[19] Maggard MA, Thompson JE, Ko CY. Why do breast cancer mortality rates vary across states ? Am Surg 2003; 69: 59-62.

[20] Quinn MJ, Martinez-Garcia C, Berrino F. Variations in survival from breast cancer in Europe by age and country, 1978-1989. EUROCARE Working Group. Eur J Cancer 1998; 34: 2204-11.

Tables

Table 1. Distribution of patients per institutions

Geographical area	Institution	Total accrual (N=2793)	Randomization exp./standard	5 yrs DFS (%) (95% CI)
The Netherlands	Institution 1	53	49.1%/50.9%	71.7 (59.6-83.8)
	Institution 2	25	48.0%/52.0%	80.0 (64.3-95.7)
	Institution 3	184	50.0%/50.0%	65.7 (58.8-72.6)
Poland	Institution 4	40	47.5%/52.5%	69.4 (55.0-83.9)
	Institution 5	78	51.3%/48.7%	76.9 (67.6-86.3)
France	Institution 6	311	49.5%/50.5%	80.7 (76.3-82.8)
	Institution 7	622	49.8%/50.2%	75.2 (71.8-78.6)
	Institution 8	185	49.2%/50.8%	78.9 (73.0-84.8)
	Institution 9	902	50.4%/49.6%	81.5 (79.0-84.1)
South Europe	Institution 10	54	51.9%/48.1%	60.6 (47.4-73.7)
	Institution 11	60	51.7%/48.3%	68.3 (56.6-80.1)
	Institution 12	25	56.0%/44.0%	68.6 (49.0-82.2)
	Institution 13	48	50.0%/50.0%	70.2 (56.6-83.9)
South Africa	Institution 14	206	49.5%/50.5%	57.8 (51.0-64.5)

Table 2. Univariate analysis of baseline prognostic characteristics

Variable	Events/Patients (%)	5 yrs DFS (95%CI)	HR* (95%CI)	P-value
Type of surgery				<0.0001
Mastectomy	505/1231 (41%)	69.8 (67.2-72.4)	1	
Breast conserving therapy	486/1542 (32%)	80.2 (78.2-82.2)	0.69 (0.61-0.78)	
Axillary nodal status				<0.0001
Negative	417/1467 (28%)	82.9 (81.0-84.9)	1	
Positive	573/1303 (44%)	67.4 (64.8-69.9)	1.80 (1.58-2.04)	
Tumor size				<0.0001
Not palpable/less than 2cm	218/823 (27%)	84.5 (82.0-87.0)	1	
$\geq$ 2cm/advanced disease	759/1915 (40%)	72.0 (70.0-74.0)	1.69 (1.46-1.97)	
Other concomittant disease				0.0064
No	887/2542 (35%)	76.2 (74.5-77.8)	1	
Yes	101/232 (44%)	70.5 (64.6-76.4)	1.33 (1.08-1.64)	

*Cox PH model including factor of interest as fixed effect - Patients with missing values for the factor of interest are excluded.

Table 3. Distribution of baseline characteristics over institutions - Generalized linear model

Baseline characteristics	N	$\hat{\beta}_0$	(se)	$\hat{\sigma}_0^2$	(se)	Average probability (inter-quantile range)
Type of surgery (mastectomy)	2773	0.9847	(0.4281)	2.3299	(1.072)	72.8% (11.9%-98.2%)
Axillary nodal status (positive status)	2770	0.0504	(0.1186)	0.1427	(0.081)	51.3% (33.4%-69.0%)
Tumor size ($\geq$ 2cm/advanced disease)	2738	0.9981	(0.1167)	1.1268	(0.069)	69.6% (57.5%-84.5%)
Other concomittant disease (yes)	2774	-2.3458	(0.2731)	0.7896	(0.380)	8.7% (1.7%-35.3%)

Table 4. Results frailty models

Endpoint	Model	$\hat{\theta}$	$\hat{\beta}$	(se)	HR	P
DFS	Random institution effect	0.0665				
	Fixed treatment effect					
	Experimental vs. standard		-0.158	(0.063)	0.855	0.013
OS	Random institution effect	0.1090				
	Fixed treatment effect					
	Experimental vs. standard		-0.100	(0.072)	0.900	0.160
LRR	Random institution effect	0.0586				
	Fixed treatment effect					
	Experimental vs. standard		-0.445	(0.122)	0.640	0.0003
DFS	Random institution effect	0.0078				
	Fixed treatment effect					
	Experimental vs. standard		-0.157	(0.063)	0.855	0.013
	Geographical area*					
	The Netherlands		0.025	(0.102)	1.025	0.810
	Poland		-0.183	(0.142)	0.833	0.200
	France		-0.342	(0.073)	0.710	< 0.0001
	South Europe		0.157	(0.109)	1.170	0.150
	South Africa		0.344	(-)	1.410	-

*Sum contrast

Figures

Figure 1. Disease free survival by institution

Figure 2a. Distribution of allocated treatment arm over institutions. Predicted probability of control arm

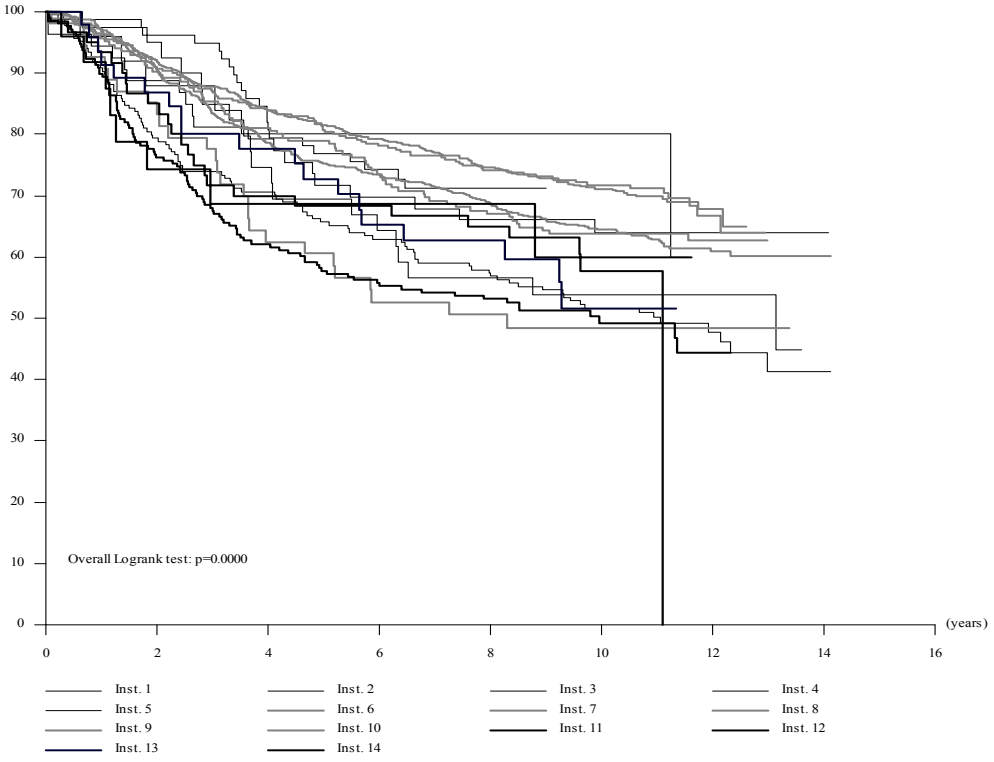
Figure 2b. Distribution of type of surgery over institutions. Predicted probability of mastectomy

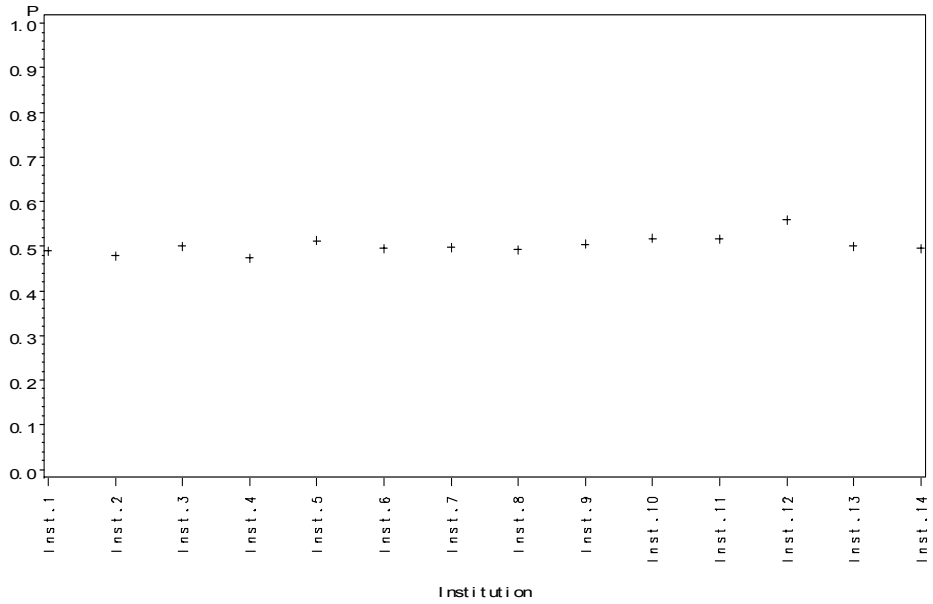
Figure 3. Predicted random effects w_i in a model with ($\circ$) / without ($\times$) treatment included as fixed effect

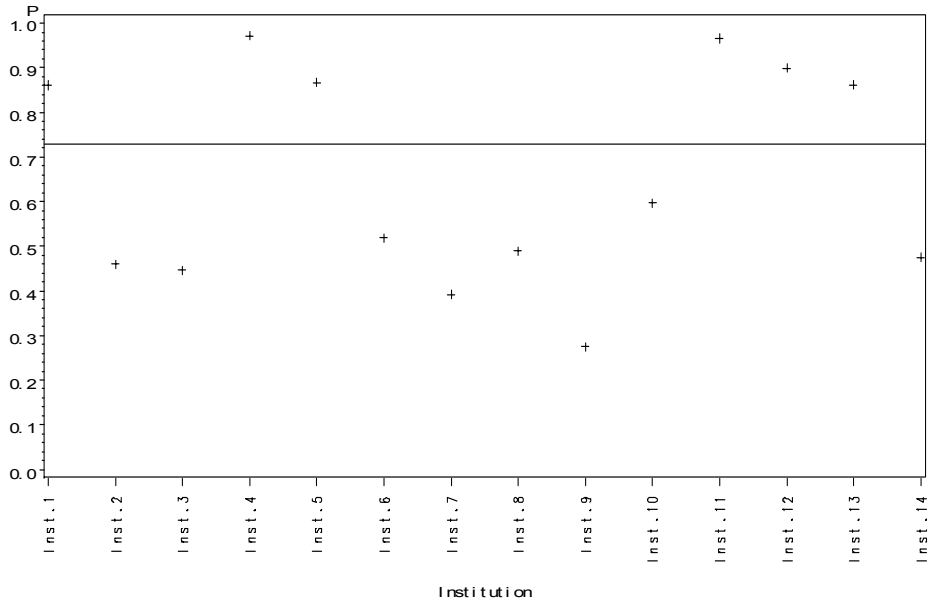
Figure 4a. Density function of 25 % quantile time to event (DFS) over institutions

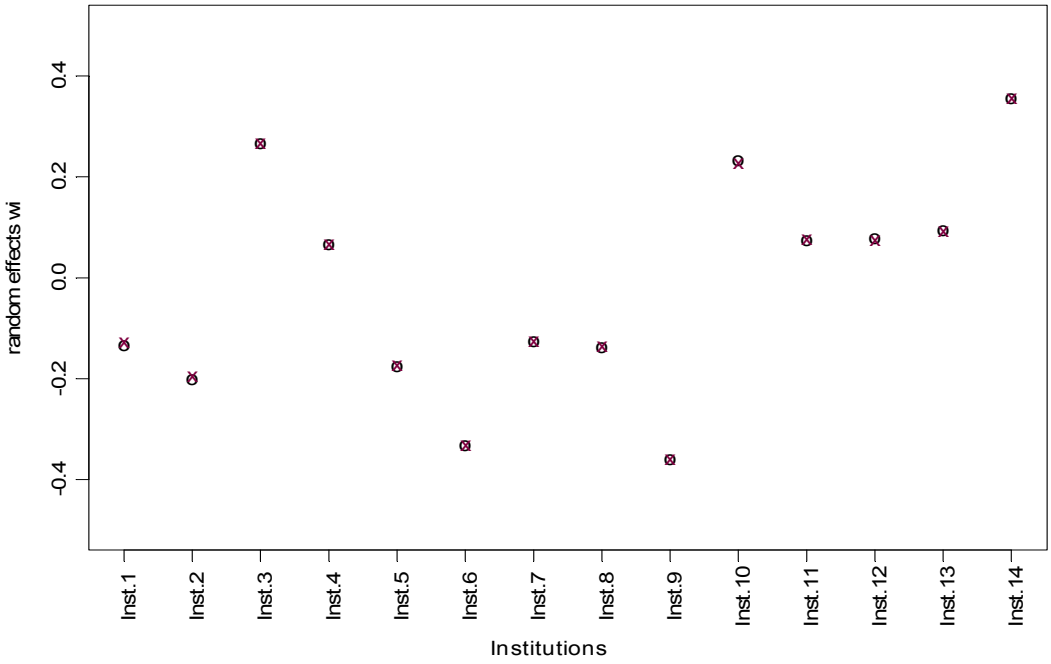
Figure 4b. Density function of 5 year DFS over institutions

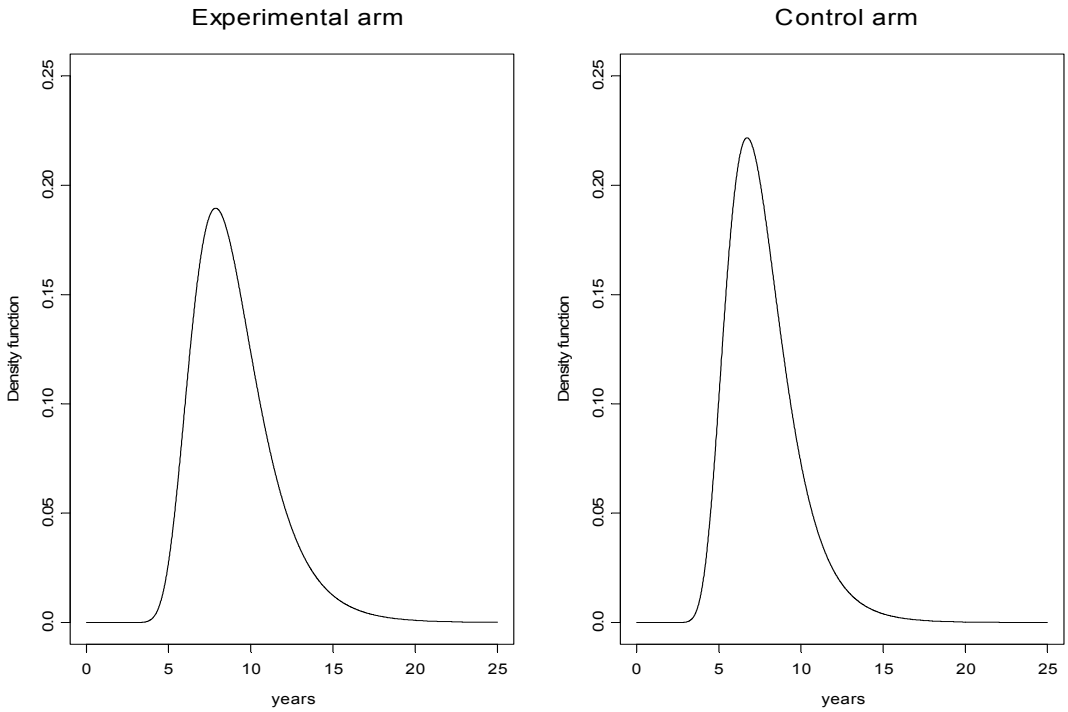
Figure 5. Predicted random effect in a model with treatment as fixed effect ($\circ$) , with treatment and geographical area as fixed effects ($\times$) , and with treatment and type of surgery as fixed effects ($+$).

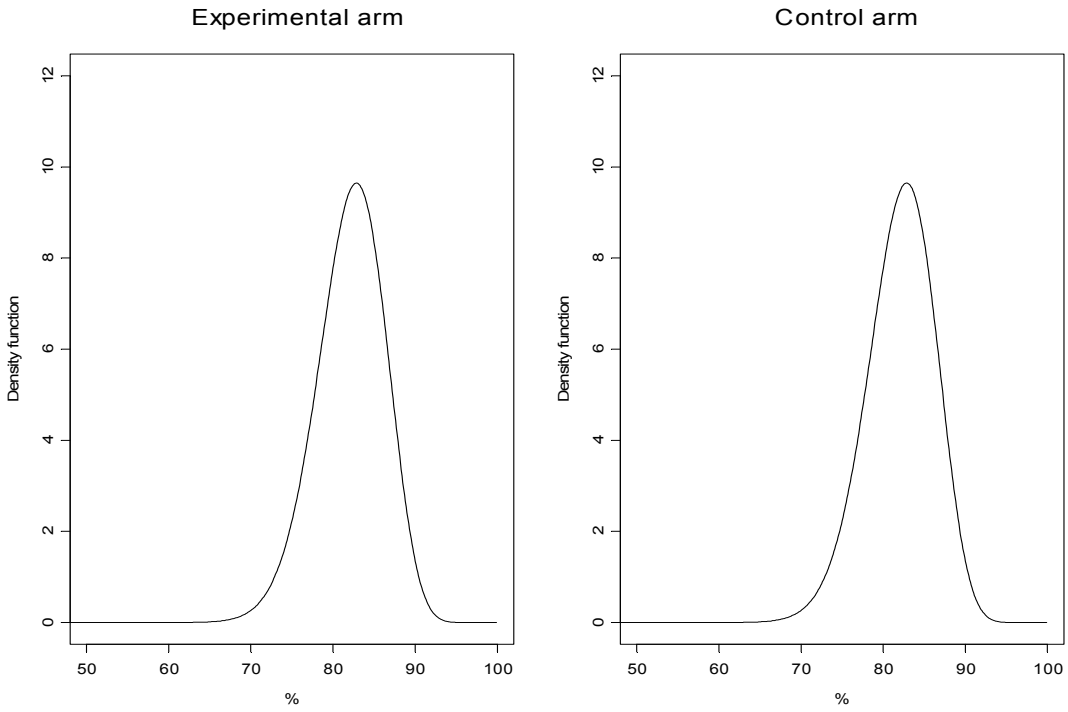

100
90
80
70
60
50
40
30
20
10
0
Overall Logrank test: p=0.0000
0
2
4
6
8
10
12
14
16
(years)
Inst. 1
Inst. 2
Inst. 3
Inst. 4
Inst. 5
Inst. 6
Inst. 7
Inst. 8
Inst. 9
Inst. 10
Inst. 11
Inst. 12
Inst. 13
Inst. 14

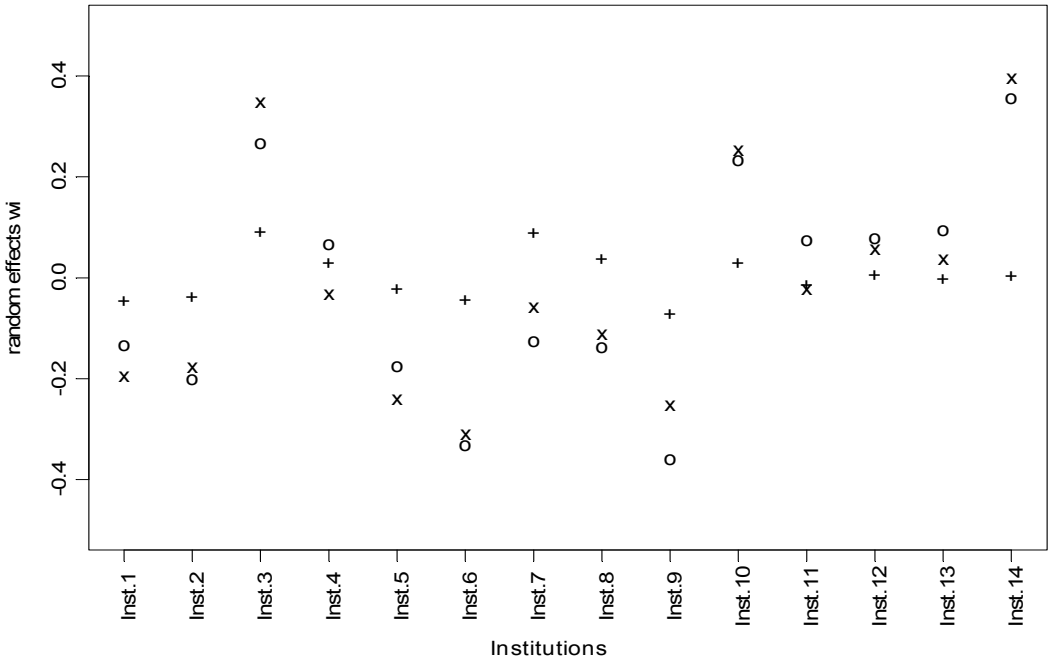

P
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Inst.1
Inst.2
Inst.3
Inst.4
Inst.5
Inst.6
Inst.7
Inst.8
Inst.9
Inst.10
Inst.11
Inst.12
Inst.13
Inst.14
Institution


P
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Inst.1
Inst.2
Inst.3
Inst.4
Inst.5
Inst.6
Inst.7
Inst.8
Inst.9
Inst.10
Inst.11
Inst.12
Inst.13
Inst.14
Institution


random effects wi
0.4
0.2
0.0
-0.2
-0.4
Inst.1
Inst.2
Inst.3
Inst.4
Inst.5
Inst.6
Inst.7
Inst.8
Inst.9
Inst.10
Inst.11
Inst.12
Inst.13
Inst.14
Institutions


Experimental arm
Control arm
Density function
0.0
0.05
0.10
0.15
0.20
0.25
0
5
10
15
20
25
years


Experimental arm
Control arm
Density function
0
2
4
6
8
10
12
50
60
70
80
90
100
%


random effects wi
0.4
0.2
0.0
-0.2
-0.4
Inst.1
Inst.2
Inst.3
Inst.4
Inst.5
Inst.6
Inst.7
Inst.8
Inst.9
Inst.10
Inst.11
Inst.12
Inst.13
Inst.14
Institutions

Appendix

In the case of a gamma distributed frailty and constant baseline hazard, the density of the median time to event is given by

$$f_{T_{0.5,z}}(t) = \left(\frac{\ln(2)}{\theta\lambda_0 \exp(\beta z)}\right)^{\frac{1}{\theta}} \frac{1}{\Gamma(\frac{1}{\theta})}\Big(\frac{1}{t}\Big)^{1+\frac{1}{\theta}} \exp\left(-\frac{\ln(2)}{\theta\lambda_0 t \exp(\beta z)}\right) \tag{10}$$

Indeed, for a constant baseline hazard the conditional survival curve is given by

$$S(t \mid u, z) = \exp(-\lambda_0 u \exp(\beta z)t) \tag{11}$$

and $T_{0.5,z}$, the median time to event, satisfies $S(T_{0.5,z} \mid u, z) = 0.5$ or (dropping the z dependence)

$$T_{0.5,z} = h(u) = \frac{\ln 2}{\lambda_0 u \exp(\beta z)}. \tag{12}$$

Since $T_{0.5,z}$ is a monotone transformation of u, we have for $t \geq 0$,

$$f_{T_{0.5,z}}(t) = f_u\Big(h^{-1}(t)\Big)\left|\frac{d}{dt}h^{-1}(t)\right| \tag{13}$$

with $f_u(.)$ the gamma density, now (10) is immediate from (13).

In the same way, we obtain the density for $T_{0.25,z}$, the 25% quantile,

$$f_{T_{0.5,z}}(t) = \left(\frac{\ln(4/3)}{\theta\lambda_0 \exp(\beta z)}\right)^{\frac{1}{\theta}} \frac{1}{\Gamma(\frac{1}{\theta})}\Big(\frac{1}{t}\Big)^{1+\frac{1}{\theta}} \exp\left(-\frac{\ln(4/3)}{\theta\lambda_0 t \exp(\beta z)}\right) \tag{14}$$

Similarly we have from (11) that $S_{5,z}$, the 5 year event free survival time, satisfies (dropping the z dependence) $S_{5,z} = \exp(-5\lambda_0 u \exp(\beta z))$. Using thefact that $S_{5,z}$ can be seen as a monotone transformation of the random variable u, we find that

$$f_{S_{5,z}}(s) = \frac{s^{\frac{1}{5\lambda_0\theta\exp(\beta z)}-1}}{(5\lambda_0\theta\exp(\beta z))^{\frac{1}{\theta}}\Gamma\Big(\frac{1}{\theta}\Big)}\Big(\ln\frac{1}{s}\Big)^{\frac{1}{\theta}-1}. \tag{15}$$