PERGAMON

European Journal of Cancer 38 (2002) 1173–1180

European Journal of Cancer

www.ejconline.com

Review

Treatment outcome studies: pitfalls in current methods and practice

C. Legrand[a,*], R. Sylvester[a], L. Duchateau[b], P. Janssen[c], P. Therasse[a]

[a]*European Organization for Research and Treatment of Cancer, Av. E. Mounier 83, Box 11, B-1200 Brussels, Belgium*
[b]*Department of Physiology, Biochemistry and Biometrics, Faculty of Veterinary Medicine, Ghent University, Salisburylaan 133, B-9820 Merelbeke, Belgium*
[c]*Center for Statistics, Limburgs Universitair Centrum, Universitaire Campus, B-3590 Diepenbeek, Belgium*

Received 2 January 2002; accepted 11 January 2002

Abstract

The objective of a treatment outcome study is to investigate the heterogeneity in outcome between patients according to factors other than treatment, such as country, institution or physician. Results of treatment outcome studies have already been extensively presented in the medical literature. However, no clear methodology has emerged to perform treatment outcome studies and various methods have been used. This paper reviews the different types of questions addressed in treatment outcome studies, the different methodologies and the different endpoints used. Statistical techniques are mainly descriptive including tables, estimates of survival curves, but regression models have also been used. Most of the studies use registry data, while only a few use discharge data or data available from clinical trials.

Keywords: Treatment outcome; Registers; Endpoints; Mortality; Hospital; Clinical trial; Heterogeneity; Methodology

1. Introduction

While clinical trials evaluate which treatment is preferable for a disease, the objective of a treatment outcome study is to investigate the heterogeneity in outcome between patients according to factors other than treatment, such as country, institution or physician. By assessing which factors are associated with a better outcome, such studies can lead to improvements in the quality of patient care. However, these studies are also controversial as insurers, government and other third-party payers might use the results of such studies to determine the 'best' institutions in a competing healthcare market [1].

There is no unique way of conducting, analysing and drawing conclusions from a treatment outcome study. A broad set of questions can be investigated and a wide range of approaches can be followed in an attempt to answer these questions. The starting point of any treatment outcome study should be to clearly define the question of interest. The choice of the endpoints, the statistical analysis and the source of data will all influence the interpretation of the results and the conclusions that can be drawn.

The objective of this paper was to review the methodology of treatment outcome studies performed until now, to summarise the elements to be kept in mind when designing, conducting and interpreting such studies, and to provide a tool to guide the reader in understanding treatment outcome publications. Although a wide range of health problems has been covered by treatment outcome studies, this paper will concentrate on oncology.

2. Questions addressed in treatment outcome studies

A treatment outcome study investigates variations in patient outcome according to factors other than treatment. The factors most often considered can be classified into four broad categories: (i) the geographical area in which the patient lives or is treated; (ii) the institution in which the patient is treated; (iii) the type of physician by whom the patient is treated; and (iv) the participation or not of the patient in a clinical trial. Factors inherent to the patient (e.g. socio-economic status, race, marital status, insurance, etc.) will not be considered here.

* Corresponding author. Tel.: +32-2-774-1660; fax: +32-2-771-3810.
E-mail address: cle@eortc.be (C. Legrand).

0959-8049/02/$ - see front matter
PII: S0959-8049(02)00085-0

2.1. Geographical area

Although investigators are mainly interested by differences in outcome between countries, most of such studies are based on data from cancer registries covering different parts of different countries. A cancer registry might have been established in an area particularly well equipped with diagnostic and treatment facilities and might be representative of only a small proportion of the population [2]. Therefore, such studies should rather be interpreted as studies comparing outcome between defined areas of different countries. The most famous study of this type in Europe is probably the EUROCARE II study which compares patient outcome by tumour type between 45 cancer registries from 17 European countries [3].

Several studies compare outcome between different areas within the same country, such as for breast and prostate cancer in different areas of Finland [4]. The most frequently encountered examples are the studies investigating differences in outcome between states or cancer registry areas within the US [5,6].

Some studies focus rather on differences between types of geographical areas such as rural versus urban regions [7].

There are only a few studies investigating differences between continents or large regions of the world such as the study of La Vecchia and colleagues [8] comparing the mortality rate of childhood cancer between North America, Western Europe, Japan and Oceania. The rarity of such studies is probably due to the logistic difficulties encountered compared with the relatively low interest in the results.

2.2. Treating institution

Studies assessing differences in outcome between individual institutions [9,10] are rare, probably in view of the controversy that could arise when faced with the results. Most often differences in outcome are investigated according to the different characteristics of the institutions.

The institution's volume can be based on the number of patients discharged by the entire institution [11] or by one particular service, on the number of times a specific procedure has been performed [12–16] or on the number of patients included in a given treatment outcome study [9,17]. Volume categories are usually arbitrarily chosen to give a reasonable number of patients in each group: three is the most common number of groups, with institutions being classified as 'low', 'medium' or 'high' volume.

An institution can also be characterised according to its teaching status (university, non-university, ...) [9,18] ownership (non-profit, governmental, proprietary, ...) [1] or location (city, city suburban, rural, ...) [1]. A large part of the treatment outcome literature in the US compares the outcome achieved by Health Maintenance Organizations (HMOs) and Fee-for-Service settings [19]. However, such studies could also be considered as studies investigating differences in outcome according to factors inherent to the patient as sometimes the two types of health plans are present in the same institution [20].

Volume and other institution characteristics are sometimes considered jointly in order to compare, for example, outcome achieved by HMOs, small community hospitals, large community hospitals and teaching institutions [21].

2.3. Treating physician

Studies rarely look at differences between physicians themselves [9,22] but focus rather on differences in outcome according to the 'volume' or according to the 'specialisation' of the physician.

The 'physician's volume' can be based on the number of times they performed a specific procedure [9,23], on the number of patients they were responsible for [11] or on the number of patients included in the treatment outcome study [13]. In the latter situation, some physicians might be considered as 'low volume', while in fact they are 'high volume' when also considering the patients they treated outside the study (e.g. in other hospitals not participating in the study) [23]. Grouping is done arbitrarily to have sufficiently large groups for statistical analysis and/or a good balance of cases among groups.

The 'specialty' of the physician often has a different meaning in different studies. One of the problems in interpreting the results can be the lack of a clear definition of what the authors mean by 'specialist'. Besides the specialty, different 'types' of physician are sometimes defined according to the time since their certification as 'medical practitioner' [11] or 'specialist' [9].

Treatment outcome publications often do not specify upon which of the physicians taking part in the care of patients (admitting physician, operating surgeon, ...) the study is based. When discussing their results, Nguyen and colleagues [24] state that their conclusions are based on the assumption that the primary surgeon was also in charge of the patient's postoperative care and that survival is affected by the total management as provided by the primary surgeon.

Another problem when trying to compare results of different publications is that different physician specialties are often involved in different studies. For example, Kehoe and colleagues [25] compared the outcome of ovarian carcinoma operated upon by gynaecologists or by general surgeons, while Nguyen and colleagues [24] further subdivided the surgeon's specialty by looking at differences obtained by gynaecological oncologists, obstetrician gynaecologists and general surgeons.

2.4. Participation in a clinical trial

The strictly controlled conditions under which randomised clinical trials are performed can affect a physician's and a patient's compliance with the treatments under investigation, the quality of supportive care and possibly a patient's lifestyle [26]. This might result in differences in outcome between patients included or not in a clinical trial and is often linked to the problem of the generalisability of the results of a clinical trial [27–29]. Patients participating in clinical trials are a non-random sample selected in view of their baseline characteristics, known prognostic factors and disease status. Comparing survival between in- and out-trial patients when all patients who died before being entered in the trial are considered as out-trial patients increases the proportion of short survivors in this group and can lead to the impression of a survival advantage for the in-trial patients ('guarantee period bias') [27].

Moreover, institutions that participate in clinical trials may be quite different with respect to treatment facilities and the qualifications of investigators. Excluding these sources of variations and looking for a 'trial effect' by comparing a similar sample of in- and out-trial patients (in terms of baseline characteristics, known prognostic factors and disease status) treated in similar institutions by similar physicians is very difficult.

3. Endpoints

3.1. Overall survival

Overall survival is considered as a 'hard' endpoint in randomised clinical trials, but is more problematic in the context of treatment outcome studies.

The definition of the start date from which the duration of survival is measured is often different from one study or database to the other and should be clearly defined. Indeed, several dates might be considered as the starting point of the survival period [30,31]: date first seen for the condition by any medical practitioner, date first seen in the hospital, date of diagnosis, date of confirmation of diagnosis, date of first treatment, etc.

If the date used as a starting point for calculating survival in one of the databases is earlier in the course of the disease, then survival may seem to be longer for this patient population. When using date of diagnosis as the starting point, the survival in a given patient population might appear longer than in the other because screening or early diagnosis accelerates the diagnosis while the date of death is in fact not postponed (lead-time bias) [19,31,32].

It seems clear that the accuracy of survival estimates depends on the registration and the follow-up methods. In the specific case of long-term lung cancer survivors, Sant and colleagues [30], based on a sample from several EUROCARE II registries, assessed the impact of different follow up procedures and concluded that inaccuracies in death records influenced survival patterns to only a limited extent. However, this conclusion might not be generalisable to other disease types and especially not to short survivors.

The overall survival rate is defined as the proportion of patients alive after a specified number of years of follow-up with respect to the patients alive at the starting point of the observation period (e.g. date of diagnosis). This proportion is thus specific to a particular length of follow-up after the start of the observation period.

3.2. Relative survival

As it is not always possible to conclude from the available data whether a death is due or not to cancer, most of the studies consider overall survival irrespective of the attributed cause of death. However, when comparing different populations, the observed survival can be different for reasons not necessarily related to the cancer under study and therefore several authors correct for mortality due to other causes by calculating the relative survival rate. Hakulinen [33] and Hakama [34] has defined the relative survival rate as the ratio of the survival rate observed in a patient group under consideration to the survival rate expected in a group of people similar to the patient group with respect to all possible factors (e.g. age and sex) affecting survival, except the disease under study. A simpler approach consists of calculating the ratio of the observed survival rate in the patient group to the overall survival rate derived from the general population [2,6,30,35]. The ratio of observed survival rate to the expected survival rate of the general population of the same gender and age is sometimes referred to as standardised relative survival or age-standardised relative survival [3,32,36].

3.3. Mortality rate

The mortality rate is defined as the ratio of the number of deaths due to any cause in a particular population at risk for a specified time period to the size of this population. The mortality rate is affected by several factors other than the disease under consideration [1,14,37] and particularly by the morbidity of the population being cared for. For example, well-run hospitals that care for many acutely ill patients may spuriously be classified as providing poor-quality care on the basis of mortality rates. Mortality rates ignore the problem of patients lost to follow-up and moreover, for lethal diseases, the length of survival is often considered more important than mortality itself.

In-hospital mortality rate (as well as re-admission rate) should be considered as a 'weak' endpoint, as it is affected by hospital policy concerning length of stay and discharge of patients. To overcome the influence of hospital policy concerning length of stay, 30- or 90-day mortality should be preferred to in-hospital mortality [15].

The disease-specific mortality rate [8,9,11] is defined as the ratio of the number of deaths due to the disease in a particular population to the size of this population. However, this leads to serious interpretation problems as higher rates can be due to an elevated incidence of the disease or poorer overall survival [38]. Moreover, populations having more efficient systems for recording deaths from the disease under study will appear to have a higher disease-specific mortality rate.

3.4. Postoperative outcome

Postoperative complications considered independently of the long-term outcome may be a 'misleading endpoint' for the comparison of outcome between surgeons. For example, radical surgery could worsen the patient's postoperative status, but improve their quality and duration of life, while a more conservative approach might minimise the risk of postoperative complications, but eventually compromise long-term survival [39]. In addition, it would be wrong to assume that postoperative mortality relates only to the (technical) ability of the surgeon as it could also be due to factors related to the preoperative preparation of the patients, their anaesthetic management or a variety of aspects of postoperative care [40].

As local recurrence clearly represents a failure of surgical technique, differences in local recurrence rate or time to local recurrence after potentially curative surgery may be of interest when considering differences in outcome between surgeons [9,22].

4. Statistical methods

The statistical methods used in treatment outcome studies are mostly descriptive presenting data in tables or graphs which poses a number of problems when interpreting the results.

4.1. Adjustments

Treatment outcome studies are obviously not based on randomised evidence. Non-randomised studies are known to be highly susceptible to bias and the populations compared in treatment outcome studies might therefore be very different at baseline (e.g. different dietary habits, different socio-economic levels and different lifestyles). Differences in outcome might thus simply reflect differences in patient population characteristics at baseline. It is thus necessary to adjust for patient characteristics that are strongly associated with the outcome and that could, without proper attention, confound the results [9,41].

Non-adjusted studies must be interpreted with care and more weight should be attached to studies that attempt to adjust for confounding factors [9,29]. As a large set of variables may describe a patient's condition, the aim of risk adjustment is to find a parsimonious representation of those patient characteristics that have a strong relationship to the endpoint and which could confound the results. When only descriptive tables are used, the problem of adjustment is difficult to deal with as adjustment can only be done by looking at different subgroups (race, age, ...). The use of regression models (e.g. logistic regression, Cox proportional hazards model, ...) allows for adjustment or stratification for confounding factors.

4.2. Regression models

In the analysis of treatment outcome data, postoperative mortality, (in-hospital) mortality and surgical complications are considered as binary outcomes and are often investigated using logistic regression with adjusted odds ratios [9]; also the Poisson regression model is sometimes used [13]. For time to event endpoints, e.g. survival or time to progression of the disease, one often finds Kaplan–Meier curves, logrank tests, and the Cox proportional hazards model. Although the Cox model has the advantage of using the actual survival time and allowing for varying lengths of patient follow-up [21], logistic regression using the proportion of survivors/deaths at a fixed time-point is sometimes preferred [7]. However, when considering mortality rates, the way in which patients who are lost to follow-up before the fixed time-point are dealt with is often not specified. Considering them as dead at the time they were lost to follow-up, like Matthews and colleagues [40], or excluding them from the analysis would both introduce a source of bias.

Factors of interest (e.g. centres, countries, ...) are nearly always introduced as fixed effects in the regression model. It assumes that the levels of these factors are by themselves of interest and have been intentionally 'fixed' by the study design. The way the factor is introduced in the model and the choice of the reference level are both of importance when interpreting the estimated coefficients. In some studies, the outcome for each category is compared with the outcome of all the other categories combined [9,22].

Heterogeneity in outcome according to a factor of interest might also be investigated using random effects. In this case, the levels of the factors taken into account in the study are considered as a random selection of all

the possible levels and it is the distribution, and especially the variance, of these levels that is of interest. For time to event endpoints, proportional hazard models with random effects (frailty models) have recently received considerable attention in the statistical literature [42,43].

4.3. Type I error, power and sample size

Most treatment outcome studies take the form of retrospective studies and are essentially data-driven which leads to an increased risk of type I errors (false-positive conclusions) [17]. Few studies mention the problem of multiple comparisons and, although rarely mentioned, a type I error (alpha) of 0.05 is usually used.

Sample sizes vary widely among treatment outcome publications and power considerations are almost never discussed. In some circumstances, one can have a small number of cases in some categories leading to unreliable estimates. This may be problematic in studies investigating differences in outcome between institutions or physicians when the number of patients treated by each institution/physician is too small to perform a meaningful analysis. It is therefore recommended to pay close attention to the confidence intervals [3] which should always accompany estimated values. However, the number of patients studied in treatment outcome studies might be extremely large and in this case the statistical significance of observed differences has to be distinguished from clinical or public health significance. Small non-meaningful differences might be statistically significant given the large number of cases examined [8,44]. For example, differences in outcome between large geographical areas found in studies considering thousands of patients, although statistically significant, might be irrelevant from the standpoint of the practising physicians [5].

When considering time to event analyses, the power depends on the number of events and not the number of patients [45]. Although median follow-up at the time of the analysis is quite often specified, the number of events observed in each category is often not reported.

5. Source of data

5.1. Cancer registries

Most treatment outcome studies are based on data from cancer registries. An advantage of population-based cancer registries is that as they aim to record all the cancer cases in a defined area and they are not affected by selection biases as are seen with hospital or trial populations [30,46,47]. Cancer registries also have the advantage of being a neutral and independent organisation outside the purchaser provider framework. This might be a key factor in obtaining the agreement of all providers to participate in a treatment outcome study comparing outcomes between institutions or between physicians, for example [48].

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 can be influenced by a number of factors which may differ from registry to registry [2].

When using cancer registry data, unexpected results should bring the attention of the investigator to a possible problem in the quality and validity of diagnosis and vital status assessment. For example, an unexpectedly high survival observed for cancers known to have a very poor prognosis might indicate an inaccurate diagnosis, or deaths not known to the registry or both. An unexpectedly high death rate the first month after diagnosis might indicate that the date of first diagnosis may have been missed and a date of complications used instead [3].

5.1.1. Staging and diagnosis

Reliability of the diagnostic data and the proportion of cases confirmed by histological examination vary from one registry to the other. Misclassification of certain types of lesions due to a lack of histological confirmation might inflate the outcome for these sites as such tumours are likely to be associated with a poor prognosis [31]. Some types of cancer have a common and well-defined premalignant stage with a better outcome. If a registry does not differentiate these cases from invasive carcinoma, a too optimistic estimation of outcome will be obtained for the invasive carcinoma patients. In addition, the staging information routinely available to registries may be incomplete and insufficiently standardised, making more detailed comparisons not always possible [46]. The more accurately patients are staged, the more likely they are to be classified as inoperable or metastatic and therefore, stage by stage, outcome will improve (Will Rogers or stage-migration phenomenon) [30]. In some situations, comparisons without classification by stage may be less subject to bias as, in some registries, metastatic tumours may have been mistakenly classified as localised more often than in others [4].

5.1.2. Death certificate only

Some registries use active follow-up procedures (contacting either the hospital or general practitioner responsible for the patient's care) to ascertain death, while others depend on passive measures [31]. The proportion of lost cases and of cases identified by death certificate only (DCO) can be used as an indicator of the completeness of cancer registration [30]. A low proportion of DCO cases would indicate that the registry actively seeks clinical information on cases and might

therefore be considered as an indicator of high quality. DCO cases are typically excluded from analyses [2,4,5,32,49,50], as no data are available for these patients. More specifically, their survival times cannot be defined. Registries with a high proportion of DCO will therefore exclude from analysis relatively more patients who, by definition, are all fatal cases and are often patients with a very poor prognosis who received only palliative care or no therapy and remained at home for the terminal phase of their illness. Varying proportions of DCOs between registries should be considered as a warning for a possible bias limiting to some extent the comparability of outcomes between registries as it might indicate that different methods are used among registries to collect cases from autopsies and to manage DCO cases [2,31,46,50,51]. The EUROCARE study pointed out large variations in the DCO among European cancer registries [30], however Nordic countries are known to have a very low percentages of DCO [4,50].

5.1.3. SEER

Most treatment outcome studies performed in the US use data from the Surveillance, Epidemiology and End Results programme (SEER) [5,6,14,16,18,52,53] maintained since 1973 by the US NCI (National Cancer Institute). The SEER programme uses a standardised data collection procedure among a system of 11 population-based tumour registries covering approximately 14% of the US. As limitations of the SEER programme, one has to keep in mind that (i) SEER areas tend to be somewhat more rural and have a higher socio-economic status than the rest of the US; (ii) only treatment delivered or planned within 4 months after the initiation of treatment is recorded and (iii) no information is available about comorbidity. Linking the SEER database to the Medicare database from the US government's health insurance programme provides details of the surgical procedures performed, information on comorbidities, and follow-up data on survival. The Medicare database encompasses 97% of individuals aged 65 years or older [14,54].

5.2. Hospital discharge records

Hospital discharge data could be affected by hospital policies regarding the length of stay. No follow-up information is available and therefore one must use in-hospital mortality as endpoint. In addition, one cannot effectively identify individual patients and link them to their cancer diagnosis [14] and case mix adjustments are limited by the availability and quality of data on disease severity in the discharge database.

5.3. Clinical trials

Surprisingly, only a few treatment outcome studies use data from randomised multicentre trials. It is clear that the strict conditions under which patients are treated within a clinical trial protocol leaves less space to variability in outcome and one might argue that in such a context everything is done to remove all variability. However, the collection and standardisation of all-important data (diagnosis, tumour staging, prognostic factor, treatment, follow-up data, ...) makes the comparisons and the interpretation of differences easier. Adjustment for baseline characteristics remains necessary as randomisation balances baseline characteristics only between the treatment arms. However, this adjustment is often possible as most clinical trial data collect information on all known prognostic factors of importance for the disease under study.

Sample sizes when using clinical trial data are generally much smaller than what is available through registry data. However, in many tumour types, the sample size achieved in large international phase III trials is adequate to perform treatment outcome research [43]. The major drawback of such an approach is of course the large selection-bias that affects clinical trial data. Institutions and physicians participating in clinical trials are often selected based on their 'quality' and patients randomised in clinical trials are a non-random sample of the patient population making the generalisation of the results to the entire population difficult. Therefore, absence of a difference in outcome within a clinical trial does not necessarily mean that no difference exists when looking at the entire population or that the results are homogeneous across different patient populations.

6. Discussion

The analysis and interpretation of a treatment outcome study is not an easy task and can be misleading. Publications of such studies have often led to debate.

Most of the treatment outcome studies published in the medical literature are data-driven and only a few are performed using a specified protocol. In most studies, considerations on statistical power are lacking and statistical significance is rarely discussed in terms of clinical relevance.

Randomised treatment outcome studies are not feasible and comparisons are based on retrospective non-randomised groups of patients. Therefore, the populations compared can be very different at baseline and emphasis must be placed on adjustment for important prognostic factors. Treatment outcome studies are usually done using data from cancer registry databases, leading to difficulties in the interpretation of the results. The difficulty of interpreting the results increases with the incompleteness of the data on which the study is based. The quality of the data can also influence the results. Differences in the registries themselves can be confounding factors and one should take these

into account when drawing conclusions from such comparisons.

Treatment outcome studies based on randomised clinical trials are rare, although the interpretation of the results is made easier by the fact that the patients are all treated according to the same protocol, controlled for the entry criteria and the treatment, and the data collection is standardised. However, the generalisation of results from such studies to real life is probably less obvious than for population-based studies due to the non-random selection of patients who are entered in clinical trials.

Whatever approach is used, and even when one can assume that methodological differences account for only a fraction of the differences found, the explanation of the results is never straightforward. The differences in outcome might be due to variability of access to, or effectiveness of, therapy, or variability in the utilisation of early detection programmes, or perhaps more generally to differences in the level of health provision or healthcare funding [30].

Studies investigating differences in outcome between institutions or between physicians themselves are rare. Such studies should be conducted with care as their results may be controversial and not easily accepted by the medical world. To minimise controversy, one could recommend that such studies should be anonymous and performed with the objective of improving the standard level of care rather than promoting the institutions/physicians with better outcome.

Until now, too few treatment outcome studies have been based on a prespecified protocol, stating in advance the hypothesis to be tested and the significance level to be used for each comparison. Although data-driven retrospective studies are valid tools for developing hypotheses, such hypotheses should always be further tested in a prospective fashion.

As they aim at improving the quality of patient care, well designed treatment outcome studies are of importance and should be encouraged. A new project investigating heterogeneity in outcome using data from large international multicentre clinical trials is actually under way at the European Organization for Research and Treatment of Cancer (EORTC) and research on the methodology of such studies is being carried out in order to specifically cover the case of survival data.

References

1. Brennan TA, Herbert LE, Laird NM, *et al*. Hospital characteristics associated with adverse events and substandard care. *JAMA* 1991, **265**, 3265–3269.
2. Quinn MJ, Martinez-Garcia C, Berrino F. Variations in survival from breast cancer in Europe by age and country, 1978–1989. *Eur J Cancer* 1992, **34**, 2204–2211.
3. Berrino F, Gatta G, Chessa E, Valente F, Capocaccia R. Introduction: the Eurocare II study. *Eur J Cancer* 1998, **34**, 2139–2153.
4. Karjalainen S. Geographical variation in cancer patient survival in Finland: chance, confounding, or effect of treatment? *J Epidemiol Community Health* 1990, **44**, 210–214.
5. Farrow DC, Samet JM, Hunt WC. Regional variation in survival following the diagnosis of cancer. *J Clin Epidemiol* 1996, **49**, 843–847.
6. Weinstock MA, Reyners JF. The changing survival of patients with mycosis fungoides: a population-based assessment of trends in the United States. *Cancer* 1999, **85**, 208–212.
7. Launoy G, Le Coutour X, Gignoux M, Pottier D, Dugleux G. Influence of rural environment on diagnosis, treatment and prognosis of colorectal cancer. *J Epidemiol Community Health* 1992, **46**, 365–367.
8. La Vecchia C, Levi F, Lucchini F, Lagiou P, Trichopoulos D, Negri E. Trends in childhood cancer mortality as indicators of the quality of medical care in the developed world. *Cancer* 1998, **83**, 2223–2227.
9. Holm T, Johansson H, Cedermark B, Ekelund G, Rutqvist LE. Influence of hospital- and surgeon-related factors on outcome after treatment of rectal cancer with or without preoperative radiotherapy. *Br J Surg* 1997, **84**, 657–663.
10. Harding MJ, Paul J, Gillis CR, Kaye SB. Management of malignant teratoma: does referral to a specialist unit matter. *Lancet* 1993, **341**, 999–1002.
11. Kee F, Wilson RH, Harper C, *et al*. Influence of hospital and clinician workload on survival from colorectal cancer: cohort study. *Br Med J* 1999, **318**, 1381–1386.
12. Glasgow RE, Mulvihill SJ. Hospital volume influences outcome in patients undergoing pancreatic resection for cancer. *West J Med* 1996, **165**, 294–300.
13. Harmon JW, Tang DG, Gordon TA, *et al*. Hospital volume can serve as a surrogate for surgeon volume for achieving excellent outcomes in colorectal resection. *Ann Surg* 1999, **230**, 404–411.
14. Begg CB, Cramer LD, Hoskins WJ, Brennan MF. Impact of hospital volume on operative mortality for major cancer surgery. *JAMA* 1998, **280**, 1747–1751.
15. Ellison LM, Heaney JA, Birkmeyer JD. The effect of hospital volume on mortality and resource use after radical prostatectomy. *J Urol* 2000, **163**, 867–869.
16. Bach PB, Cramer LD, Schrag D, Downey RJ, Gelfand SE, Begg CB. The influence of hospital volume on survival after resection for lung cancer. *N Engl J Med* 2001, **345**, 181–188.
17. Collette L, Sylvester RJ, Stenning SP, *et al*. Impact of the treating institution on survival of patients with "poor prognosis" metastatic non-seminoma. European Organisation for Research and Treatment of Cancer Genito-Urinary Cancer Collaborative Group and the Medical Research Council Testicular Cancer Working Party. *J Natl Cancer Inst* 1999, **91**, 839–846.
18. Wolfe CD, Tilling K, Raju KS. Management and survival of ovarian cancer patients in South East England. *Eur J Cancer* 1997, **33**, 1835–1840.
19. Potosky AL, Merrill RM, Riley GF, *et al*. Breast cancer survival and treatment in health maintenance organization and fee-for-service settings. *J Natl Cancer Inst* 1997, **89**, 1683–1691.
20. Vernon SW, Hughes JI, Heckel VM, Jackson GL. Quality of care for colorectal cancer in fee-for-service and health maintenance organization practice. *Cancer* 1992, **69**, 2418–2425.
21. Lee-Fedstein A, Anton-Culver H, Feldstein PJ. Treatment differences and other prognostic factors related to breast cancer survival. *JAMA* 1994, **271**, 1163–1168.
22. McArdle CS, Hole D. Impact of variability among surgeons on postoperative morbidity and mortality and ultimate survival. *Br Med J* 1991, **302**, 1501–1505.
23. Munoz E, Mulloy K, Goldstein J, Tenenbaum N, Wise L. Costs, quality and the volume of surgical oncology procedures. *Arch Surg* 1990, **125**, 360–363.

24. Nguyen HN, Averette HE, Hoskins W, Penalver M, Sevin BU, Steren A. National survey of ovarian carcinoma. Part V: the impact of physician's specialty on patients' survival. *Cancer* 1993, **72**, 3663–3670.
25. Kehoe S, Powell J, Wilson S, Woodman C. The influence of operating surgeon's specialisation on patient survival in ovarian carcinoma. *Br J Cancer* 1994, **70**, 1014–1017.
26. Marubini E, Mariani L, Salvadori B, *et al*. Results of a breast-cancer surgery trial compared with observational data from routine practice. *Lancet* 1996, **347**, 1000–1003.
27. Davis S, Wright PW, Schulman SF, *et al*. Participants in prospective, randomized clinical trials for resected non-small cell lung cancer have improved survival compared with non-participants in such trials. *Cancer* 1985, **56**, 1710–1718.
28. Stiller CA. Centralised treatment, entry to trials and survival. *Br J Cancer* 1994, **70**, 352–362.
29. Ward LC, Fielding JW, Dunn JA, Kelly KA. The selection of cases for randomised trials; a registry survey of concurrent trial and non-trial patients. The British Stomach Cancer Group. *Br J Cancer* 1992, **66**, 943–950.
30. Sant M, Capocaccia R, Verdecchia A, *et al*. Comparisons of colon-cancer survival among European countries: the Eurocare study. *Int J Cancer* 1995, **63**, 43–48.
31. 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**, 101–104.
32. Kliukiene J, Andersen A. Survival of breast cancer patients in Lithuania and Norway, 1988–1992. *Eur J Cancer* 1998, **34**, 372–377.
33. Hakulinen T. Cancer survival corrected for heterogeneity in patient with withdrawal. *Biometrics* 1982, **38**, 933–942.
34. Hakama M, Kajalainen S, Hakulinen T. Outcome-based equity in the treatment of colon cancer patients in Finland. *Int J Technol Assess Health Care* 1989, **5**, 619–630.
35. Engeland A, Haldorsen T, Dickman PW, *et al*. Relative survival of cancer patients. A comparison between Denmark and other Nordic Countries. *Acta Oncol* 1998, **37**, 49–59.
36. Hakulinen T, Tenkanen L, Abeywickrama K, Paivarinta L. Testing equality of relative survival patterns based on aggregated data. *Biometrics* 1987, **43**, 313–325.
37. Rosenthal GE, Shah A, Way LE, Harper DL. Variations in standardized hospital mortality rates for six common medical diagnosis. Implications for profiling hospital quality. *Med Care* 1998, **36**, 955–964.
38. 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**, 458–460.
39. McArdle C. ABC of colorectal cancer: primary treatment, does the surgeon matter? *Br Med J* 2000, **321**, 1121–1123.
40. Matthews HR, Powell DJ, McConckey CC. Effect of surgical experience on the results of resection for oesophageal carcinoma. *Br J Surg* 1986, **73**, 621–623.
41. Wu AW. The measure and mis-measure of hospital quality: appropriate risk adjustment methods in comparing hospitals. *Ann Intern Med* 1995, **122**, 149–150.
42. Vaida F, Xu R. Proportional hazards model with random effects. *Stat Med* 2000, **19**, 3309–3324.
43. Duchateau L, Janssen P, Lindsey P, Legrand C, Nguti R, Sylvester R. The shared frailty model and the power for heterogeneity tests in multicenter trials. *Comp Stats Data Anal* (in press).
44. Mettlin CJ, Menck HR, Winchester DP, Murphy GP. A comparison of breast, colorectal, lung and prostate cancers reported to the National Cancer Data Base and the Surveillance, Epidemiology, and End Results Program. *Cancer* 1997, **79**, 2052–2061.
45. Parmar MKB, Machin D. *Survival Analysis: A Practical Approach*. Chichester, John Wiley and Sons, 1995.
46. Gatta G, Sant M, Coebergh JW, Hakulinen T. Substantial variation in therapy for colorectal cancer across Europe: Eurocare analysis of cancer registry data for 1987. *Eur J Cancer* 1996, **32A**, 831–835.
47. Nilsson B, Gustavson-Kadaka E, Hakulinen T, *et al*. Cancer survival in Estonian migrants to Sweden. *J Epidemiol Community Health* 1997, **51**, 418–423.
48. Ma B, Bell J, Campbell S, Basnett I, Pollock A, Taylor I. Breast cancer management: is volume related to quality? *Br J Cancer* 1997, **75**, 1652–1659.
49. Janssen-Heijnen ML, Gatta G, Forman D, Capocaccia R, Coebergh JW. Variations in survival of patients with lung cancer in Europe, 1985–1989. EUROCARE Working Group. *Eur J Cancer* 1998, **34**, 2191–2196.
50. Vercelli M, Quaglia A, Casella C, Parodi S, Capocaccia R, Martinez-Garcia C. Relative survival in elderly cancer patients in Europe. EUROCARE Working Group. *Eur J Cancer* 1998, **34**, 2264–2270.
51. Verdecchia A, De Angelis R, Capocaccia R, *et al*. The cure for colon cancer; results from the Eurocare study. *Int J Cancer* 1998, **77**, 322–329.
52. Harlan L, Brawley O, Pommenrenke F, Wali P, Kramer B. Geographic, age, and racial variation in the treatment of local/regional carcinoma of the prostate. *J Clin Oncol* 1995, **13**, 93–100.
53. Zippin C, Lum D, Hankey BF. Completeness of hospital cancer case reporting from the SEER Program of the National Cancer Institute. *Cancer* 1995, **76**, 2343–2350.
54. Nattinger AB, Gottlieb MS, Veum J, Yahnke D, Goodwin JS. Geographic variation in the use of breast-conserving treatment for breast cancer. *N Engl J Med* 1992, **326**, 1102–1107.