available at www.sciencedirect.com

ScienceDirect

journal homepage: www.ejconline.com

Historical cross-trial comparisons for competing treatments in advanced breast cancer – An empirical analysis of bias

C.K. Lee [a,*,d], S.J. Lord [a,d], M.R. Stockler [b,d], A.S. Coates [c,d], V. Gebski [a,d], R.J. Simes [b,d]

[a] NHMRC Clinical Trials Centre, University of Sydney, NSW 1450, Australia
[b] Sydney Cancer Centre – Royal Prince Alfred and Concord Hospitals, Sydney, NSW, Australia
[c] International Breast Cancer Study Group, Bern, Switzerland

ARTICLE INFO

Article history:
Received 1 September 2009
Received in revised form 16 November 2009
Accepted 19 November 2009
Available online 23 December 2009

Keywords:
Randomised trials
Cross-trial comparisons
Historical comparisons
Treatment efficacy
Statistical models

ABSTRACT

Purpose: Randomised controlled trials (RCTs) provide optimal evidence to assess the benefits of new treatments. However, clinicians routinely rely on cross-trial comparisons to assess competing treatments when head-to-head randomised comparisons are unavailable. We investigate the validity of cross-trial comparisons using individual patient data (IPD) where patients received the same treatment protocol. We also examine the extent to which statistical adjustment for baseline characteristics can account for inter-trial differences in outcomes.

Patients and methods: We used pooled IPD of 378 women with advanced breast cancer assigned to oral cyclophosphamide, intravenous methotrexate and 5-fluorouracil (CMF) in the control arms of three first-line treatment RCTs (ANZ8101, ANZ8614 and ANZ0001) conducted between 1982 and 2001. The Kaplan–Meier method was used to compare progression-free survival (PFS) and overall survival (OS) across trials. Proportional hazard models were constructed to estimate the hazard rates across trials after adjustment for baseline characteristics.

Results: The distribution of baseline characteristics varied across trials. There was a statistically significant difference in survival among women treated with CMF in these trials (logrank $p = 0.009$). The median OS were 17.7, 10.3 and 10.1 months for 0001, 8101 and 8614, respectively. The hazard ratios for survival, adjusted for baseline characteristics differences, were 1.44 (8614) and 1.45 (8101) compared to 0001 ($p = 0.03$). PFS did not differ across trials (logrank $p = 0.38$).

Conclusions: Caution should be exercised when interpreting results from historical cross-trial comparisons even if the adjustment of baseline prognostic characteristics can be performed. Cross-trial comparisons have some role in hypothesis-generating, identifying and prioritising promising treatments for further investigation; however RCTs are still essential to guide sound clinical practice.

1. Introduction

Rapid advances in cancer medicine have produced an ever growing number of promising new treatment agents. Oncologists rely on randomised controlled trials (RCTs) for the best evidence of the relative efficacy of treatments to guide clinical practice.[1] However, direct head-to-head randomised comparisons of all potential treatment alternatives are rarely

* Corresponding author: Address: NHMRC Clinical Trials Centre, The University of Sydney, Locked Bag 77, Camperdown, NSW 1450, Australia. Tel.: +61 2 9562 5365; fax: +61 2 9565 1863.
E-mail address: chee.lee@ctc.usyd.edu.au (C.K. Lee).
[d] On behalf of Australian and New Zealand Breast Cancer Trials Group.
0959-8049/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ejca.2009.11.013

available. A recent meta-analysis investigating the efficacy of 22 chemotherapy and targeted therapies in advanced breast cancer identified RCTs reporting direct comparisons for only 45 (20%) of the possible 231 treatment comparisons.[2] Moreover, when performed, RCTs become outdated quickly with the introduction of newer, more promising agents or treatment combinations. Oncologists are often challenged to interpret the results of emerging research that does not include a comparison with current best standard practice. Thus, we frequently have to rely on *post hoc* comparisons using data from separate trials to compare the efficacy and safety of competing treatment strategies.

Cross-trial comparisons can be defined as *contemporaneous* when outcomes are compared between trials conducted during the same time period, or *historical* when trials compared are conducted at different time periods. Both types of cross-trial comparisons are problematic and susceptible to the same biases as those associated with non-randomised studies.[1,3] Any differences in the prognostic characteristics of the trial populations, supportive care, treatment compliance, drop-outs and definition and measurement of end-points will reduce the validity of these comparisons and may lead to erroneous conclusions. Despite these significant short-comings, cross-trial comparisons are common. They are necessary to interpret results from single-arm phase II trials and to prioritise these treatments for further investigation in phase III studies. Empirical studies investigating the validity of this evidence provide valuable examples for education and ongoing debate on the appropriate use and interpretation of cross-trial comparisons.[4]

Using individual patient data (IPD) from three first-line advanced breast cancer trials, we investigate the validity of historical cross-trial comparisons by comparing best objective tumour response, progression-free and overall survival outcomes for patients assigned to the same combination chemotherapy across the different trials. We also examine the extent to which statistical adjustment accounts for differences in baseline prognostic characteristics across the three trials.

2. Patients and methods

We use data from three RCTs of first-line treatments for patients with advanced breast cancer conducted by Australia and New Zealand Breast Cancer Trials Group (ANZ8101, ANZ8614 and ANZ0001). Trial participants were recruited from teaching hospitals across Australia and New Zealand. ANZ8101, activated in June 1982, was a two-by-two factorial RCT comparing the efficacy of doxorubicin and cyclophosphamide versus cyclophosphamide, methotrexate and 5-fluorouracil (CMF), administered continuously versus intermittently.[5] ANZ8614, activated in January 1988, was a two-arm RCT comparing the efficacy of mitoxantrone versus CMF plus prednisone.[6] ANZ0001, activated in June 2001, was a three-arm RCT comparing the efficacy of intermittent capecitabine versus continuous capecitabine versus CMF plus prednisone.[7]

2.1. Patients

The eligibility criteria for patients enrolled in the three trials were similar. Patients had histologically confirmed breast carcinoma with measurable or evaluable recurrent or metastatic disease, adequate bone marrow, hepatic and renal function, and were available for follow-up. Patients were excluded if they had received cytotoxic chemotherapy for recurrent or metastatic disease or extensive radiotherapy, or had a history of other cancer, diabetes mellitus or cardiac failure. Patients assigned to the intermittent CMF arm of ANZ8101 were excluded from the present analysis because this treatment arm was not considered comparable to the CMF regimen used in the continuous CMF arm of ANZ8101[5] or the two other trials. All patients provided a written informed consent for participation in the trials.

2.2. Treatments

For the present analysis, patients in each of the three trials were treated with CMF which was administered in 28-d cycles with oral cyclophosphamide, 100 mg/m^2 daily for 14 d, and intravenous methotrexate 40 mg/m^2 and intravenous 5-fluorouracil 600 mg/m^2 on days 1 and 8. Oral prednisone 40 mg/m^2 for first 14 d was routinely administered in patients from ANZ8101 and ANZ8614, and according to the oncologist's discretion in ANZ0001.

All patients continued the initial chemotherapy regimen until disease progression, intolerance, or unacceptable toxicity. Therapy beyond initial treatment failure was at the discretion of the treating oncologist.

2.3. Outcomes

The outcomes were best objective tumour response (OTR), progression-free survival (PFS) and overall survival (OS). OTR rate was measured as the proportion of patients with evaluable disease who achieved a complete response (CR) or partial response (PR), where World Health Organisation (WHO) criteria[8] were used for ANZ8101 and ANZ8614 and Response Evaluation Criteria in Solid Tumours (RECIST)[9] were used for ANZ0001. PFS was measured from randomisation to the date of first documented disease progression. OS was measured from randomisation to the date of death or last known date of follow-up.

2.4. Analysis

Discrete data were compared using Fisher's exact test. PFS and OS were estimated using the Kaplan–Meier method, and differences between trials were compared using the log-rank test.[10] Cox proportional-hazard models were used to estimate differences in PFS and OS between trials.[11] Multivariable analyses with backward stepwise selection of variables were performed for both PFS and OS to calculate the adjusted hazard ratios (AHRs) to account for any differences in baseline characteristics. All the analyses were two sided with no adjustment for multiple comparisons.

3. Results

3.1. Patient characteristics

A total of 378 patients, with a median follow-up of 4.8 years, were included in this pooled analysis. All patients received

Table 1 – Baseline characteristics of women receiving CMF chemotherapy.

Trial Characteristic	ANZ8101 $n = 75$ (%)	ANZ8614 $n = 194$ (%)	ANZ0001 $n = 109$ (%)	p
Age (years)				
⩽50	27	33	21	0.02
51–60	37	25	23	
>60	36	43	56	
Menstrual status				
Premenopausal	17	22	11	0.02
Postmenopausal	68	68	84	
Unknown	15	10	5	
Extent of disease				
Local or regional disease only	8	9	21	0.003
Distant disease only	49	47	52	
Local or regional and distant disease	43	44	27	
Disease-free interval				
⩽2 years	56	50	28	<0.001
>2 years	44	50	72	
Time from recurrence to randomisation				
⩽1 year	73	55	44	<0.001
>1 year	27	45	56	
Performance status				
0	34	30	37	0.008
1	26	42	48	
2	28	21	11	
⩾3	12	7	5	
Previous adjuvant chemotherapy	12	18	37	<0.001
Previous endocrine therapy	64	81	79	0.008
Oestrogen-receptor status				
Negative	21	27	23	<0.001
Positive	24	31	59	
Unknown	55	41	18	
Progesterone-receptor status				
Negative	24	26	32	0.001
Positive	17	27	40	
Unknown	59	47	28	
Metastatic sites[a]				
Bone	64	68	70	0.7
Liver	32	38	47	0.1
Lymph node	37	36	33	0.8
Lung	32	32	34	0.9
Pleura	27	19	22	0.4
Brain	4	3	3	0.9

CMF – cyclophosphamide, methotrexate and 5-fluorouracil chemotherapy.
a More than one site could have been involved, so percentages sum to more than 100%.

at least one cycle of CMF chemotherapy. The baseline prognostic characteristics varied widely across the three trials. For example, patients in ANZ0001 were older, had a longer disease-free interval and tended to receive previous adjuvant chemotherapy as compared to patients in ANZ8101 and ANZ8614. Table 1 summarises the baseline characteristics of patients across the three trials.

3.2. *Objective tumour response*

The OTR rates were 52%, 38% and 20% for ANZ8101, ANZ8614 and ANZ0001, respectively (Table 2). Of the 69 (92%) patients assessable for response in ANZ8101, 5 (7%) had a CR, 31 (45%) had a PR, 16 (23%) had stable disease (SD) and 17 (25%) had progressive disease (PD). In ANZ8614, 178 (92%) patients were assessable for response and 9 (5%) had a CR, 61 (34%) had a PR, 78 (44%) had SD and 30 (17%) had PD. In ANZ0001, 95 (87%) patients were assessable for response and 1 (1%) had a CR, 18 (19%) had a PR, 45 (47%) had SD and 31 (33%) had PD.

3.3. *Progression-free survival*

The median time to first disease progression was 5.5 months in ANZ8101, 5.6 months in ANZ8614 and 7.1 months in

Table 2 – Comparisons of outcomes across three trials.

Trials	OTR[a]	Median PFS (months)	Median OS (months)	Hazard ratio[b] (PFS)	Hazard ratio[b] (OS)
ANZ0001	20%	7.1	17.7	1.00	1.00
ANZ8614	38%	5.6	10.1	1.12	1.45
ANZ8101	52%	5.5	10.3	0.92	1.47
	$p < 0.0001$			$p = 0.38$	$p = 0.009$

[a] Combined complete response (CR) and partial response (PR) rates. The p-value is based on Fisher's test.
[b] Hazard ratio is unadjusted for baseline prognostic characteristics. The p-value is based on logrank test.

ANZ0001 (logrank $p = 0.38$, Table 2 and Fig. 1A). The unadjusted hazard ratio (HR) for PFS was 1.12 (95% CI, 0.87–1.43) for ANZ8614 and 0.92 (95% CI, 0.67–1.27) for ANZ8101 when compared with ANZ0001. Trial was not statistically significantly associated with PFS with or without adjustment for baseline characteristics.

3.4. *Overall survival*

Median overall survival times were 10.3 months in ANZ8101, 10.1 months in ANZ8614 and 17.7 months in ANZ0001 (logrank $p = 0.009$, Table 2 and Fig. 1B). The unadjusted HRs were 1.45 (95% CI, 1.13–1.88) for ANZ8614 and 1.47 (95% CI, 1.08–2.01) for ANZ8101 when compared with ANZ0001.

Table 3 summarises the unadjusted HR for overall survival for all baseline characteristics that were clinically important in patients with advanced breast cancer. Age, performance status, liver and brain metastasis, total tissue sites of metastasis, oestrogen receptor status, serum haemoglobin, neutrophil and alkaline phosphatase were baseline characteristics that remained statistically significantly associated with OS in multivariate analysis (Table 4).

3.5. *Multivariable analysis*

Trial was statistically significantly associated with OS in the multivariable model. The AHRs were 1.44 (95% CI, 1.08–1.93) for ANZ8614 and 1.45 (95% CI, 1.00–2.09) for ANZ8101 when compared with ANZ0001 ($p = 0.03$) (Table 4 and Fig. 2).

3.6. *Survival after disease progression*

Median survival after disease progression was 5.5 months in ANZ8101, 5.4 months in ANZ8614 and 10.1 months in ANZ0001 (logrank $p = 0.10$, Fig. 1C). The unadjusted HRs were 1.29 (95% CI, 0.99–1.68) for ANZ8614 and 1.36 (95% CI, 0.98–1.88) for ANZ8101 when compared with ANZ0001 ($p = 0.10$). The AHRs were 1.38 (95% CI, 1.03–1.85) for ANZ8614 and 1.39 (95% CI, 0.95–2.03) for ANZ8101 when compared with 0001 ($p = 0.07$).

4. Discussion

This study provides empirical evidence of the limited validity of historical cross-trial comparisons. This pooled dataset provides the unique opportunity to compare outcomes for patients treated with the same combination chemotherapy in three different ANZBCTG trials. These trials were conducted in similar settings but over a time span of 20 years. Patients enrolled in the most recent trial, ANZ0001 (enrolment year 2001), had a statistically significantly longer median OS of 17.7 months when compared with patients enrolled in the two earlier trials (median OS 10.3 months for ANZ8614 (enrolment year 1982) and 10.1 months for ANZ8101 (enrolment year 1988). Paradoxically, ANZ8101 had the best tumour response rates of 52% when compared with ANZ8614 (38%) and ANZ0001 (20%). Furthermore, there was no inter-trial difference in PFS.

We demonstrate that statistical adjustment is able to account for some but not all inter-trial differences in OS (unadjusted trial p-value = 0.009, adjusted trial p-value = 0.03). As compared to ANZ0001, the AHRs for ANZ8614 and ANZ8101 were 1.43 and 1.45, respectively. Assuming no true difference in CMF efficacy over time, the 43–45% excess in adjusted hazard observed in ANZ8614 and ANZ8101 is likely to be due to bias from differences in patient selection and/or other factors such as changes in treatment, practice or measurement over time as well as some chance effect. The bias is likely to be from multiple sources apart from differences in baseline prognostic characteristics.

Our findings provide a timely reminder that historical cross-trial and other non-randomised comparisons should always be interpreted with caution. This and other studies[3,4] have consistently highlighted the significance of this problem. In addition to prognostic factor bias, non-randomised studies are susceptible to other common sources of bias: selection bias due to differences in physician- or self-selection into the study; diagnostic and staging bias due to differences in the methods and criteria for identifying eligible patients for the study; performance bias due to differences in supportive care and other auxiliary patient management; detection bias due to differences in methods and criteria of patient evaluation; and, attrition bias due to differences in patient adherence and follow-up.[1,3,12] One or more of these sources of bias can be present in any non-randomised study where their magnitude and direction cannot be easily ascertained or minimised.

The three trials shared important similarities that are not common to all historical cross-trial comparisons. In particular, they were conducted by the same cooperative trial group and followed similar protocols. Patients were prospectively recruited using similar eligibility criteria and outcomes were assessed using similar definitions for PFS and OS. Enrolment took place in hospital settings located within the same geographical region with similar treatment approaches for advanced breast cancer. Thus, this study might be less vulnerable to bias due to differences in patient selection,

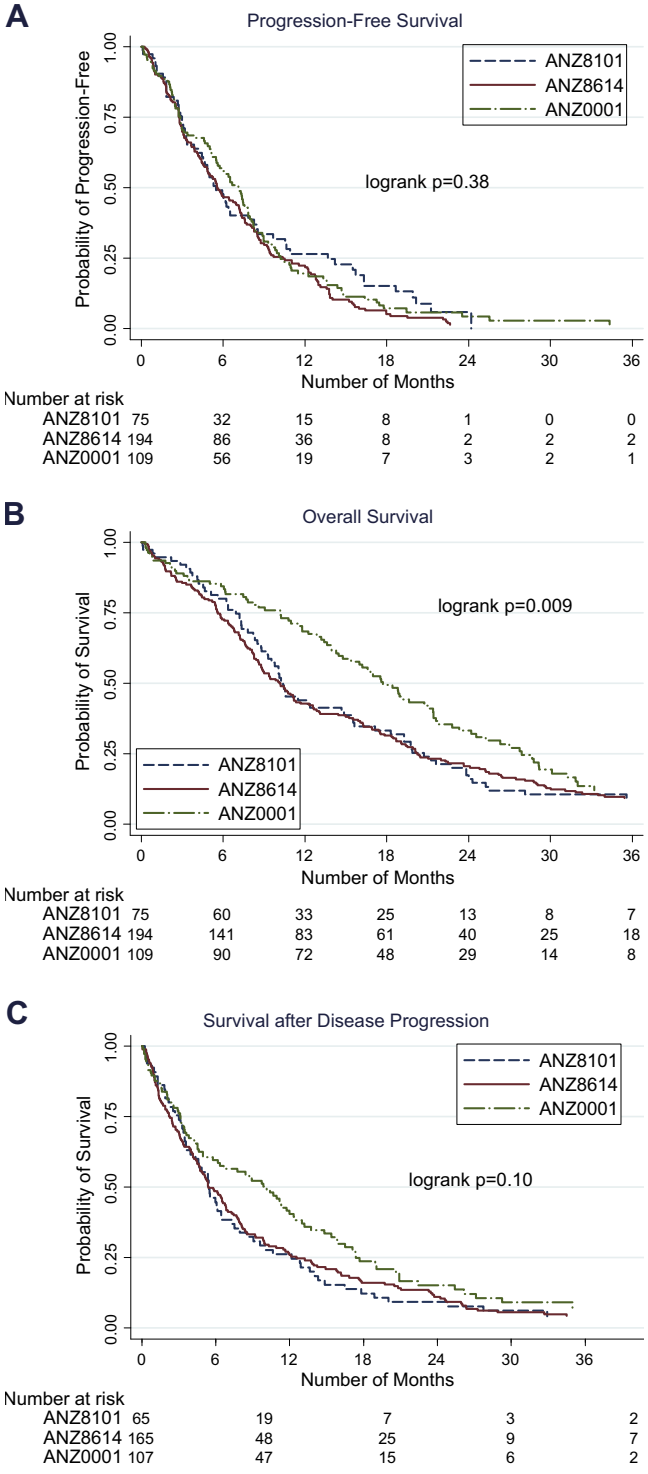

Fig. 1 – Kaplan–Meier estimation of PFS, OS and survival after disease progression.

evaluation and follow-up than historical comparisons conducted across RCTs conducted by different groups of oncologists in different settings and different countries. Furthermore, the availability of IPD allowed rigorous statistical adjustment for baseline patient characteristics to minimise prognostic factor bias. The statistical adjustment accounts for established prognostic factors such as age, performance status, disease free interval, hormone receptor status and prior adjuvant treatments.[13–16] Other prognostic factors, such as baseline neutrophil and alkaline phosphatase, were also identified and included in the statistical model.[17,18]

There are number of possible explanations for some of the differences in the inter-trial response rates and survival outcomes observed in this study. Firstly, the evaluation of tumour responses was based on the WHO criteria[8] for ANZ8101 and ANZ8614 and the more rigorous RECIST criteria[9] for ANZ0001. Secondly, advances in imaging technology, such as the 64-slice computed tomography scanners, may have led to both a 'stage migration' effect with the inclusion of patients with an earlier diagnosis of advanced disease, and the earlier detection of disease progression for patients in ANZ0001who were enrolled after 2001 as compared with patients enrolled in earlier trials. Earlier diagnosis may have contributed to the improved survival observed in ANZ0001 due to lead time bias without producing a difference in progression-free survival due to the corresponding effect of earlier detection of progression. Thirdly, advances in breast cancer therapeutics have led to the availability of less toxic and more effective salvage treatments upon disease progression that were not available to patients enrolled in earlier trials. These treatments include capecitabine, taxanes and other newer chemotherapeutic and biologic agents. For example, a substantial number of patients from ANZ0001, at disease progression, crossed over to the experimental arm to receive salvage treatment with capecitabine, a treatment with OS advantage over CMF within the trial.[7] In addition, the improvement of supportive care management such as the availability of 5-hydroxytryptamine3 antagonists as effective anti-emetic agents[19] for patients enrolled in ANZ0001 would likely have improved tolerance and compliance to chemotherapy. Finally, some important baseline prognostic characteristics known today were either inconsistently measured or not determined in these trials and hence not adequately accounted for in the statistical model. For example, oestrogen receptor status was unknown in 55% of patients from 8101 as compared with 18% of patients from ANZ0001. HER-2/neu recognised today as an important prognostic factor was not determined in patients enrolled in these trials.[20,21]

Despite attempts to adjust for most of the known baseline prognostic characteristics using a standard statistical approach,[11] the multivariable model presented (Table 4) only explains 21% of the variation in the survival of these patients with advanced breast cancer. Others have reported that prognostic models, including those that incorporate modern genetic classifiers, only explain 7–50% of the variation in survival.[22] Our current knowledge of baseline characteristics with prognostic significance for survival remains limited and impairs our ability to compare survival outcomes between different patient groups even when IPD with all known prognostic factors are available.

This study does not demonstrate any statistically significant difference in the PFS across trials. This result supports the findings in other studies[23,24] that treatment effects on PFS may be poor surrogates for treatment effects on OS in advanced breast cancer. After disease progression, there was a trend, but not statistically significant, of survival difference across trial (unadjusted trial p-value = 0.10, adjusted trial

Table 3 – Univariable Cox regression analysis of baseline characteristics for OS.

	Hazard ratio	95% confidence interval	*p*-value
Age (per unit decade)	1.09	0.99–1.20	0.07
PS 0	1.00		0.0002
PS 1	1.38	1.07–1.77	
PS 2	1.74	1.29–2.34	
PS 3+	2.12	1.38–3.25	
Disease-free interval (>2 years)	1.01	0.98–1.04	0.4
ER–	1.00		0.002
ER+	0.67	0.52–0.88	
ER unknown	0.99	0.76–1.30	
PR–	1.00		0.003
PR+	0.83	0.63–1.10	
PR unknown	1.28	1.00–1.65	
Liver metastasis	1.27	1.02–1.57	0.03
Lung metastasis	1.39	1.11–1.75	0.004
Pleural metastasis	1.47	1.14–1.89	0.003
Brain metastasis	2.35	1.28–4.31	0.006
Total tissue sites of metastasis[a]	1.19	1.09–1.31	0.0003
Previous adjuvant chemotherapy	0.83	0.64–1.07	0.1
Previous endocrine treatment	1.15	0.90–1.48	0.3
Haemoglobin (g/dL)	0.91	0.85–0.97	0.004
Neutrophil > ULN	2.01	1.47–2.73	<0.001
Platelet > ULN	1.34	1.04–1.73	0.03
Total bilirubin > ULN	1.73	1.14–2.61	0.009
Alkaline phosphatase > ULN	1.34	1.08–1.66	0.007

PS, performance status; ER, oestrogen receptor; PR, progesterone receptor; ULN, upper limit normal.
[a] Total sites of metastasis.

Table 4 – Multivariable Cox regression analysis for OS.

	Hazard ratio	95% confidence interval	p-value
ANZ0001	1.00	–	0.03
ANZ8101	1.45	1.00–2.11	
ANZ8614	1.43	1.07–1.90	
Age (per decade unit)	1.19	1.07–1.33	0.002
PS 0	1.00	–	0.01
PS 1	1.29	0.99–1.69	
PS 2	1.45	1.04–2.01	
PS 3+	1.52	0.95–2.45	
Brain metastasis	2.57	1.37–4.83	0.003
Liver metastasis	1.47	1.14–1.89	0.003
Total tissue sites of metastasis[a]	1.13	1.02–1.26	0.02
Haemoglobin (g/dL)	0.91	0.85–0.98	0.01
Neutrophil > ULN	1.72	1.22–2.44	0.002
Alkaline phosphatase > ULN	1.39	1.07–1.82	0.02
ER–	1.00	–	0.004
ER+	0.62	0.46–0.82	
ER unknown	0.80	0.60–1.07	

PS, performance status; ER, oestrogen receptor; ULN, upper limit normal.
[a] Total sites of metastasis exclude liver and brain metastases.

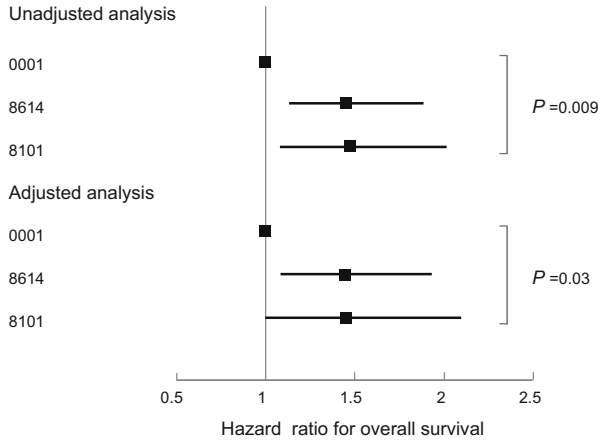

Fig. 2 – Unadjusted and adjusted hazard ratios of trials for overall survival.

p-value = 0.07). This suggests that the management of disease after disease progression, differences in the assessment of disease progression and other factors might have influenced the subsequent survival of these patients.

In this study, we compare trial outcomes over a 20-year period. Although, it can be argued that *contemporaneous* cross-trial comparisons may be more valid than *historical* cross-trial comparisons due to changes in care over time, we suggest that similar caution should be applied for all cross-trial comparisons regardless of their temporal similarity. The critical issue is the timing of pivotal changes in diagnosis, staging, treatment and supportive care, rather than the length of time between trials per se. In some cases, these changes may introduce bias between trials performed within short periods of time of each other or within similar time periods at different locations. This issue has been illustrated by Gennari et al.[16] in their analyses of pooled data from six consecutive trials from the same cooperative group conducted over a 20-year period. The investigators demonstrated that the temporal trend did not influence OS in patients with metastatic breast cancer and concluded that the availability of effective treatment with taxane chemotherapy was responsible for the observed improvements in survival over time.

This study does not investigate the validity of using indirect comparisons of the relative efficacy of two competing treatments when each treatment has been investigated in separate trials that use a common control arm. For example, a comparison of the HRs, for treatment A versus treatment C from one trial, with treatment B versus treatment C of another trial, is made to draw conclusions about the relative efficacy of treatment A versus treatment B. Comparisons of relative risks help to control for inherent differences in the selection of patients and provision of supportive care between RCTs. Thus, indirect comparisons of HR are preferred over cross-trial comparisons of absolute risk measures, such

as median survival times. Even so, these analyses still have important limitations, may still be prone to some bias, and the results are not always consistent with the results of direct head-to-head comparisons.[12,25]

Despite the limitations discussed, cross-trial comparisons remain a common practice. They are frequently used in conference presentations and some published literature for various purposes. Firstly, they are used, explicitly or implicitly, to interpret uncontrolled phase II trial results for feasibility, safety and efficacy; and to prioritise test agents for more definitive investigation of efficacy in RCTs. Given the consistent findings that comparisons with historical controls tend to report larger benefits in favour of the new or later treatment,[26,27] cross-trial comparisons that produce highly favourable results present an ongoing challenge to clinicians and their patients who are tempted to adopt new treatments early before evidence is available from RCTs.

Secondly, cross-trial comparisons are often used to discuss and compare the findings of new phase III RCTs in the context of the existing evidence about the study treatments. Here they have a role for hypothesis generation to raise questions about treatment efficacy when head-to-head randomised comparisons are unavailable. For example, Boccardo and colleagues[28] compared the results of their study with other 'similar' phase III trials, to highlight the consistency of their findings of superiority in efficacy from switching to aromatase inhibitor after 2 or more years of tamoxifen for adjuvant treatment of early stage breast cancer. However, a subsequent head-to-head randomised comparison of sequential versus monotherapy with aromatase inhibitor demonstrates no difference in treatment efficacy,[29] indicating that the consistency of results from cross-trial comparisons do not necessary provide a stronger evidence of efficacy. In another example, Ozols et al.[30] performed cross-trial analysis to question the benefit of intraperitoneal chemotherapy for stage III ovarian cancer when compared with the standard practice of intravenous carboplatin/paclitaxel. In this situation, a finding of minimal difference between treatments does not exclude the possibility of a true clinical benefit. In other situations, Marksman has demonstrated that cross-trial comparisons have suggested positive benefits but direct head-to-head comparisons have concluded no difference or even harm.[4] Together, these examples illustrate that the interpretation of results from cross-trial comparisons should always be treated with caution. The magnitude and the direction of bias are unpredictable and confirmatory evidence from direct head-to-head comparison is always desirable.

Given it will never be possible to conduct RCTs to test all possible treatment comparisons, others have suggested that a large relative treatment effect in the magnitude of ten or more times in a non-randomised comparison may be sufficiently large to overcome the combined effect of plausible confounders and conclude treatment efficacy.[31,32] Unfortunately, cancer treatments rarely provide such benefits. Direct randomised head-to-head comparison remains essential for evidence of treatment efficacy in these situations.

In conclusion, historical cross-trial and other non-randomised comparisons of treatment efficacy can be misleading even when IPD are available to adjust for differences in patients' baseline prognostic characteristics. Well-conducted head-to-head randomised comparisons should remain the gold standard to provide sound evidence to guide clinical practice.

Conflict of interest statement

None declared.

REFERENCES

1. Altman DG, Bland JM. Statistics notes: Treatment allocation in controlled trials: why randomise? *Br Med J* 1999;**318**(7192):1209.
2. Mauri D, Polyzos NP, Salanti G, Pavlidis N, Ioannidis JPA. Multiple-treatments meta-analysis of chemotherapy and targeted therapies in advanced breast cancer. *J Natl Cancer Inst* 2008;**100**(24):1780–91.
3. Zelen M. The role of statistics in the design and evaluation of trials in cancer medicine. In: Veronesi U, Bonadonna G, editors. *Clinical trials in cancer medicine*. Orlando: Academic Press; 1985. p. 561–8.
4. Markman M. The dangers of "cross-trial" and "cross-retrospective experience" comparisons. *Cancer* 2007;**109**:1929–32.
5. Coates A, Gebski V, Bishop J, et al. Improving the quality of life during chemotherapy for advanced breast cancer. A comparison of intermittent and continuous treatment strategies. *New Engl J Med* 1987;**317**(24):1490–5.
6. Simes R, Gebski V, Coates A, et al. Quality of life with single agent mitozantrone or combination chemotherapy for advanced breast cancer, a randomised trial. *Proc Am Soc Clin Oncol* 1994;**13**:73.
7. Stockler M, Sourjina T, Harvey V, et al. A randomized trial of capecitabine given intermittently versus continuously versus classical CMF as first line chemotherapy for women with advanced breast cancer unsuited to more intensive treatment. *Breast Cancer Res Treat* 2007;**100**:S278.
8. Miller A, Hoogstraten B, Staquet M, Winkler A. Reporting results of cancer treatment. *Cancer* 1981;**47**(1):207–14.
9. Therasse P, Arbuck SG, Eisenhauer EA, et al. New guidelines to evaluate the response to treatment in solid tumors. *J Natl Cancer Inst* 2000;**92**(3):205–16.
10. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *J Am Stat Assoc* 1958;**53**(282):457–81.
11. Cox DR. Regression models and life-tables. *J Roy Stat Soc* 1972;**34**(2):187–220.
12. Deeks J, Dinnes J, D'Amico R, Sowden A. Evaluating nonrandomised intervention studies. *Health Technology Assessment* 2003;**7**. 27.
13. Chang J, Clark G, Allred D, et al. Survival of patients with metastatic breast carcinoma: importance of prognostic markers of the primary tumor. *Cancer* 2003;**97**:545–53.
14. Ryberg M, Nielsen D, Osterlind K, Skovsgaard T, Dombernowsky P. Prognostic factors and long-term survival in 585 patients with metastatic breast cancer treated with epirubicin-based chemotherapy. *Ann Oncol* 2001;**12**(1):81–7.
15. Largillier R, Ferrero J-M, Doyen J, et al. Prognostic factors in 1038 women with metastatic breast cancer. *Ann Oncol* 2008:mdn424.
16. Gennari A, Conte P, Rosso R, Orlandini C, Bruzzi P. Survival of metastatic breast carcinoma patients over a 20-year period. *Cancer* 2005;**104**(8):1742–50.
17. Hortobagyi G, Smith T, Legha S, et al. Multivariate analysis of prognostic factors in metastatic breast cancer. *J Clin Oncol* 1983;**1**:776–86.

18. Yamamoto N, Watanabe T, Katsumata N, et al. Construction and validation of a practical prognostic index for patients with metastatic breast cancer. *J Clin Oncol* 1998;**16**(7):2401–8.
19. Jantunen IT, Kataja VV, Muhonen TT. An overview of randomised studies comparing 5-HT3 receptor antagonists to conventional anti-emetics in the prophylaxis of acute chemotherapy-induced vomiting. *Eur J Cancer* 1997;**33**(1):66–74.
20. Seshadri R, Firgaira F, Horsfall D, et al. Clinical significance of HER-2/neu oncogene amplification in primary breast cancer. The South Australian Breast Cancer Study Group. *J Clin Oncol* 1993;**11**(10):1936–42.
21. Slamon D, Godolphin W, Jones L, et al. Studies of the HER-2/neu proto-oncogene in human breast and ovarian cancer. *Science* 1989;**244**(4905):707–12.
22. Dunkler D, Michiels S, Schemper M. Gene expression profiling: does it add predictive accuracy to clinical characteristics in cancer prognosis? *Eur J Cancer* 2007;**43**(4):745–51.
23. Burzykowski T, Buyse M, Piccart-Gebhart MJ, et al. Evaluation of tumor response, disease control, progression-free survival, and time to progression as potential surrogate end points in metastatic breast cancer. *J Clin Oncol* 2008;**26**(12):1987–92.
24. Miksad RA, Zietemann V, Gothe R, et al. Progression-free survival as a surrogate endpoint in advanced breast cancer. *Int J Tech Assess Health Care* 2008;**24**(04):371–83.
25. Song F, Altman D, Glenny A, Glenny J. Validity of indirect comparison for estimating efficacy of competing interventions: empirical evidence from published meta-analyses. *Br Med J* 2003;**326**(7387):472–7.
26. Sacks H, Chalmers T, Smith H. Randomized versus historical controls for clinical trials. *Am J Med* 1981;**72**:233–40.
27. Kunz R, Oxman A. The unpredictability paradox: review of empirical comparisons of randomised and non-randomised clinical trials. *Br Med J* 1998;**317**:1185–90.
28. Boccardo F, Rubagotti A, Puntoni M, et al. Switching to anastrozole versus continued tamoxifen treatment of early breast cancer: preliminary results of the italian tamoxifen anastrozole trial. *J Clin Oncol* 2005;**23**(22):5138–47.
29. The BIG 1-98 Collaborative Group. Letrozole therapy alone or in sequence with tamoxifen in women with breast cancer. *New Engl J Med* 2009;**361**(8):766–76.
30. Ozols R, Bookman M, du Bois A, et al. Intraperitoneal cisplatin therapy in ovarian cancer: comparison with standard intravenous carboplatin and paclitaxel. *Gynecol Oncol* 2006;**103**(1):1–6.
31. Glasziou P, Chalmers I, Rawlins M, McCulloch P. When are randomised trials unnecessary? Picking signal from noise. *Br Med J* 2007;**334**(7589):349–51.
32. Howick J, Glasziou P, Aronson JK. The evolution of evidence hierarchies: what can Bradford Hill's 'guidelines for causation' contribute? *J Roy Soc Med* 2009;**102**(5):186–94.