British Journal of Cancer (2003) 88, 1191 – 1198
© 2003 Cancer Research UK All rights reserved 0007 – 0920/03 $25.00
www.bjcancer.com

Reporting of prognostic markers: current problems and development of guidelines for evidence-based practice in the future

RD Riley*[,1], KR Abrams[1], AJ Sutton[1], PC Lambert[1], DR Jones[1], D Heney[2] and SA Burchill[3]

[1]Department of Epidemiology and Public Health, University of Leicester, 22-28 Princess Road West, Leicester, LE1 6TP, UK; [2]Department of Medical Education, University of Leicester, University Road, Leicester, UK; [3]Cancer Research UK Clinical Centre, St James's University Hospital, Beckett Street, Leeds, UK

Prognostic markers help to stratify patients for treatment by identifying patients with different risks of outcome (e.g. recurrence of disease), and are important tools in the management of cancer and many other diseases. Systematic review and meta-analytical approaches to identifying the most valuable prognostic markers are needed because (sometimes conflicting) evidence relating to markers is often published across a number of studies. To investigate the practicality of this approach, an empirical investigation of a systematic review of tumour markers for neuroblastoma was performed; 260 studies of prognostic markers were identified, which considered 130 different markers.

The reporting of these studies was often inadequate, in terms of both statistical analysis and presentation, and there was considerable heterogeneity for many important clinical/statistical factors. These problems restricted both the extraction of data and the meta-analysis of results from the primary studies, limiting feasibility of the evidence-based approach.

Guidelines for reporting the results of primary prognostic marker studies in cancer, and other diseases, are given in order to facilitate both the interpretation of individual studies and the undertaking of systematic reviews, meta-analysis and, ultimately, evidence-based practice. General availability of full individual patient data is a necessary step forward and would overcome the majority of problems encountered, including poorly reported summary statistics and variability in cutoff level, outcome assessed and adjustment factors used. It would also limit the problem of reporting bias, although publication bias will remain a concern until studies are prospectively registered. Such changes in practice would help important evidence-based reviews to be conducted in order to establish the most appropriate prognostic markers for clinical use, which should ultimately improve patient care.

British Journal of Cancer (2003) **88,** 1191 – 1198. doi:10.1038/sj.bjc.6600886 www.bjcancer.com

Keywords: prognosis; marker; survival analysis; meta-analysis; systematic review; guidelines

Prognostic markers (also called prognostic variables or factors) are relevant tools in the management of patients with cancer, and also many other medical conditions, because they help to stratify patients for treatment by identifying different risk groups in order to reduce morbidity and mortality. They include biological, clinical, genetic, histological and pathological features. For example, carcinoembryonic antigen (CEA) is a prognostic marker in colorectal cancer (Eche *et al*, 2001).

An *evidence-based* approach to identifying the most valuable prognostic markers for a given disease is clearly important because it is common for evidence relating to markers to be published across a number of studies, often with conflicting results (Altman, 2001, pp 228–247; Riley *et al*, 2003a). Furthermore, a frequent difficulty in assessing the clinical value of prognostic markers is the relatively small number of patients in primary research studies, sometimes a consequence of disease rarity and limited resources, such that each primary study has low statistical power for detecting any benefits of prognostic staging. The use of systematic review, and in particular meta-analysis, methodology may therefore be important and allow a useful assessment of the prognostic power of markers (Altman, 2001, pp 228–247). A *systematic review* is the preferred means of identifying and combining existing evidence (Egger *et al*, 2001). *Meta-analysis* is the statistical analysis of the review, which seeks to combine all the relevant results found from the literature identified in a quantitative way to produce results more precise than is possible from the individual studies (Sutton *et al*, 2000).

In this paper, we use a recently performed systematic review of prognostic tumour markers studied in neuroblastoma to demonstrate the problems encountered when using this approach, and highlight how they limit evidence-based practice. We then generalise the problems to other areas of oncology, and indeed other disease settings, and ultimately provide specific guidelines for reporting primary prognostic marker studies.

METHODS

Neuroblastoma is a neuroblastic tumour of the primordial neural crest and is the most common extracranial solid tumour of childhood. The study of prognostic markers for this disease forms an active research area within which a large body of evidence exists. This makes it an appropriate area for an empirical

*Correspondence: Dr RD Riley; E-mail: rdr3@leicester.ac.uk

Received 29 April 2002; revised 27 January 2003; accepted 29 January 2003

investigation, and as such the problems identified in this study are highly likely to generalise to other disease settings. A brief description of the systematic review strategy adopted is now given.

Search strategy

The three on-line bibliographic databases Medline, Embase and Cancerlit were chosen as a basis for identifying the relevant literature from 1966 to February 2000. Papers written in a non-English language were excluded. A full description of the search strategy and inclusion/exclusion criteria is provided in Riley *et al* (2003b). One investigator performed the assessment of the papers, with second and third investigators independently checking a sample of them. To be included in the review, a paper had to provide a quantitative result or give tabulated individual patient data evaluating the use of a tumour marker in neuroblastoma from a primary research study of humans. To be classified as relevant to *prognosis*, a paper had to present data, in the form of summary statistics or individual patient data, relating tumour marker levels at a measured point in time to the outcome of patients at the end of a specific follow-up period. Owing to the large number of potential markers, we focused on genetic/biological markers rather than histological markers.

Data extraction for meta-analysis

From each of the papers included, information was extracted on the tumour markers studied. Meta-analysis of those markers on which eight or more papers provided data was considered. The $\log_e$(hazard ratio) and its variance were the essential information required from each study, as they provide an important comparative estimate of the risk of death/disease recurrence between two groups of patients. Furthermore, there are several indirect estimation methods available when these statistics are not directly reported (Parmar *et al*, 1998), and the $\log_e$(hazard ratio) has an approximate normal distribution for large samples, making it particularly amenable to meta-analysis techniques. We make the assumption of proportional hazards throughout this paper.

It was common for a paper to report more than one prognostic result by relating one or more markers to overall survival and/or disease-free survival, and also by providing unadjusted and/or adjusted results (e.g. adjusted for age, stage of disease). Estimates of the $\log_e$(hazard ratio) and its variance comparing two groups defined by a single marker level were sought from *all* the overall survival and disease-free survival reports. An unadjusted estimate was preferred for each as prior knowledge indicated that adjusted results were likely to be highly inconsistent in the factors for which adjustment was made (Altman, 2001, pp 228–247). An adjusted estimate was sought in the absence of an unadjusted result. Although some markers take only binary values (e.g. chromosome 1p – deletion or no deletion), it was also usual for primary studies to dichotomise continuous variables using a cutoff level in order to categorise patients into high- and low-risk groups.

A five-step sequential process (Figure 1) using 10 different direct and indirect methods (Table 1), based on the approach of Parmar *et al* (1998), was used in an attempt to obtain the $\log_e$(hazard ratio) and its variance. Studies with samples smaller than 25 were not included in Steps 2–5 because they were not considered large enough to justify estimation methods. A more detailed version of Figure 1 and a more in-depth description of the extraction procedure are provided in Riley *et al* (2003b).

RESULTS

Literature search results

A total of 3415 papers were identified from the literature search. After assessment, 260 papers were classified as 'relevant' to prognosis, and these studied a total of 130 different tumour markers for risk stratification of patients (for references see Riley *et al*, 2003b).

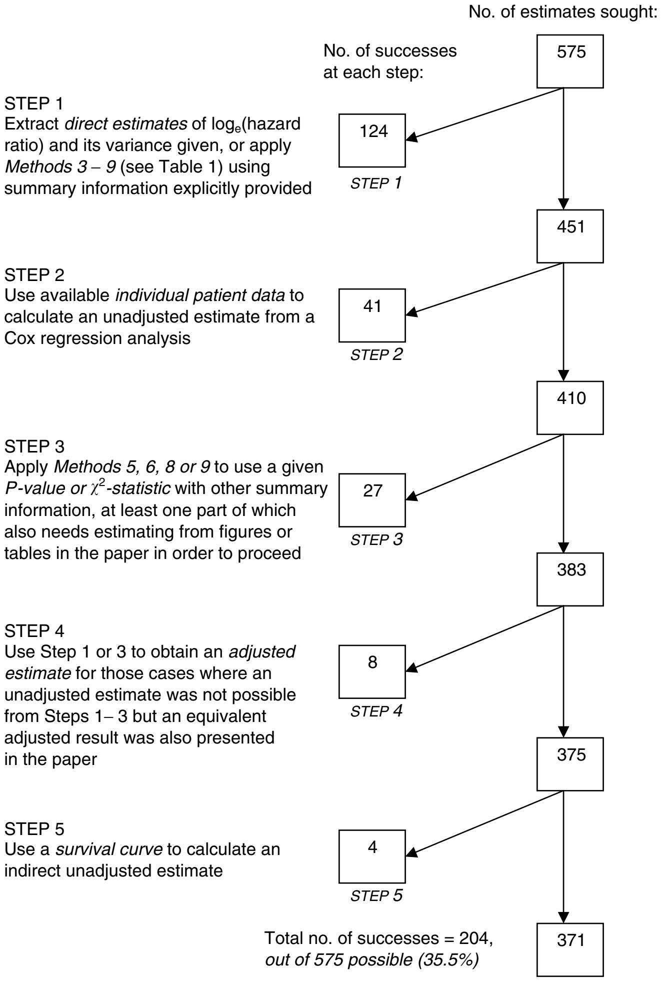
No. of estimates sought:
No. of successes at each step:
575
STEP 1
Extract *direct estimates* of log$_e$(hazard ratio) and its variance given, or apply *Methods 3 – 9* (see Table 1) using summary information explicitly provided
124
STEP 1
451
STEP 2
Use available *individual patient data* to calculate an unadjusted estimate from a Cox regression analysis
41
STEP 2
410
STEP 3
Apply *Methods 5, 6, 8 or 9* to use a given *P-value or χ^2-statistic* with other summary information, at least one part of which also needs estimating from figures or tables in the paper in order to proceed
27
STEP 3
383
STEP 4
Use Step 1 or 3 to obtain an *adjusted estimate* for those cases where an unadjusted estimate was not possible from Steps 1– 3 but an equivalent adjusted result was also presented in the paper
8
STEP 4
375
STEP 5
Use a *survival curve* to calculate an indirect unadjusted estimate
4
STEP 5
Total no. of successes = 204, *out of 575 possible (35.5%)*
371

Figure 1 Methods and results at each stage of the sequential process used to obtain a single direct or indirect estimate of the $\log_e$(hazard ratio) and its variance for each of the reports where one of the 13 tumour markers was related to overall or disease-free survival by summary statistics or individual patient data across the literature. Five steps were used, with unadjusted estimates sought primarily in each *unless* only an adjusted result was available or otherwise stated.

Data extraction of prognostic marker results

The 13 most commonly studied prognostic markers were each selected for an in-depth study to establish their individual value as a prognostic tool (Table 2). Expression of CD44 gene was studied in eight papers and the other 12 markers were studied in 10 or more prognosis papers (Table 2). This involved 211 (81.2%) of all the prognosis papers. Within these there were *575 reports* of prognostic power assessment where levels of any of these 13 tumour markers were related to overall survival or disease-free survival by summary statistics or IPD.

Only 204 (35.5%) estimates of both the $\log_e$(hazard ratio) and its variance could be calculated from Steps 1–5 using Methods 1–10 (Figure 1, Table 1). In particular, the $\log_e$(hazard ratio) and its variance were both *directly* provided on only three occasions in the 575 reports (0.005%) (Table 1, Method 1), and all were from a single paper (Berthold *et al*, 1997). Fortunately, individual patient data were frequently presented within this literature, and from this

Table 1 Description of the methods used to obtain estimates of the $\log_e$(hazard ratio) and its variance

Method	Summary statistics or data required	Step 1	Step 2	Step 3	Step 4	Step 5	Total
1	HR or $\log_e$(HR) and V	3	—	—	0	—	3
2	Individual patient data	—	41	—	—	—	41
3	$\log_e$(HR) and CI	0	—	—	1	—	1
4	HR and CI	30	—	—	3	—	33
5	$\log_e$(HR) and *P*-value	2	—	0	2	—	4
6	HR and *P*-value	10	—	2	2	—	14
7	HR, group numbers and total events	2	—	—	0	—	2
8	χ^2-statistic, group numbers and total events	10	—	4	0	—	14
9	*P*-value, group numbers and total events	67	—	21	0	—	88
10	Survival curve	—	—	—	—	4	4
	Total	124	41	27	8	4	204

The summary statistics required for each method are shown together with the number of times each method was successfully used in Steps 1–5 of the extraction process. Methods 3–10 were used in order of preference shown. HR=hazard ratio; V=variance of the $\log_e$(HR); CI=confidence interval.

Table 2 Names of the 13 markers grouped by tumour marker class, with the number of prognosis papers identified for each, the number of reports when it was related to either overall or disease-free survival by summary statistics or individual patient data, and the number of successful estimates made of the $\log_e$(hazard ratio) and variance; evidence of heterogeneity is shown for outcome, cutoff levels, age, stage and adjustment factors

Marker class	Marker name	Papers[a]	OS and DFS reports[a]	Total successful estimates (ψ)[b]	OS/DFS successes[c]	(i) Different cutoff groups[d]	(ii) Different stage groups[d]	(iii) Different age groups[d]	U/A[e]	Different sets of adjustment factors[f]
DNA or chromosome abnormalities	*MYC-N*	151	194	94	48/46	9	9	4	77/17	16
	DNA index	44	62	19	11/8	8	3	3	18/1	1
	Chromosome 1p	40	49	20	11/9	1	5	2	18/2	2
Urinary catecholamines	VMA	36	40	4	3/1	4	3	2	4/0	—
	HVA	26	29	2	2/0	2	2	1	2/0	—
	VMA:HVA	20	28	5	2/3	3	4	2	5/0	—
	Dopamine	10	11	2	1/1	2	2	1	2/0	—
Biological markers	CD44	8	8	3	0/3	1	1	2	3/0	-
	TrkA	16	21	11	4/7	7	1	1	9/2	2
	NSE	28	39	9	4/5	6	3	1	8/1	1
	LDH	26	30	12	5/7	5	4	1	8/4	4
	Ferritin	33	41	7	3/4	5	4	2	6/1	1
	MDR	16	30	16	9/7	8	3	3	13/3	2

VMA=vanillylmandelic acid; HVA=homovanillic acid; NSE=neuron-specific enolase; MDR=multi-drug resistance protein; LDH=lactate dehydrogenase. [a]Number of papers reporting prognostic marker results or IPD, with the overall no. of OS and DFS reports within them. There were 211 papers overall which reported overall survival (OS) and/or disease-free survival (DFS) results or individual patient data (IPD) for one or more of the markers. [b]Number of OS and DFS reports for which successful estimates were extracted. [c]Number of total successful estimates (ψ) by OS and DFS. [d]The total successful estimates (ψ) could also be grouped by those (i) using the same cutoff level (cutoff includes a group for when it was 'unknown' (for an example see Table 3)); (ii) relating to patients with the same stages of disease (stages of disease groups were 'unknown' and combinations of stages 1, 2, 3, 4, 4s); and (iii) relating to patients with the same age range (age groups were 'all ages', '< 1 year', '> 1 year' and 'unknown'). Columns 7–9 show the number of different subgroups in each case. [e]Number of total successful estimates (ψ) that were unadjusted (U) and adjusted (A) (unadjusted estimates were preferred where possible). [f]Number of successful adjusted estimates (A) that related to different sets of adjustment factors.

a further 41 direct estimates were made (Step 2, Method 2). The remaining 160 successful estimates were obtained using Methods 3–10 (Table 1), the most frequently required of which used a *P*-value/χ^2-statistic in combination with group numbers and total number of events, that is, deaths/recurrences of disease (102 times) (Methods 8 and 9).

Problems limiting meta-analysis

Poor reporting of primary studies Primary studies of prognostic tumour markers are clearly essential and we observed many important results across the literature that have implications for clinical practice. However, the general standard of reporting primary studies was inadequate, and it was disappointing that we only managed to obtain 35.5% of the estimates required despite the intensive, time-consuming extraction procedure (Figure 1). This hindered the use and interpretation of meta-analysis because we could not incorporate the majority of results reported in the literature and consequently introduced a strong potential for bias. Among the 371 reports that did not enable estimates to be made, there were five common reporting problems, most of which can be simply addressed (Figure 2). Encouragingly, there was some evidence that the reporting of prognostic markers has improved over the last 10 years because all the papers that did provide a hazard ratio or $\log_e$(hazard ratio) were published after 1990. However, these papers still only represented approximately 17% of the total literature identified over this period (i.e. only 26 out of 157 papers published after 1990 reported a hazard ratio).

Key problems in the reporting of prognostic marker studies

We studied 13 tumour markers in 211 papers, and identified 575 *reports* (involving summary statistics or IPD) that assessed their prognostic value. On trying to extract the log(hazard ratio) and variance from these reports, we found five main problems:

(1) *No appropriate statistical analysis performed or reported.*
In 133 (23.1%) reports a paper reported prognostic data (e.g. the number of patients who had an event in each group) but no results from a Cox regression analysis or log-rank/Wilcoxon test, often because no such analyses had been performed. Hence, Methods 1–10 could not be used.
It is clearly important that where one of the purposes of the study is to assess the prognostic value of markers, appropriate statistical analyses should be performed to calculate a comparative group estimate, for example, hazard ratio, with some measure of precision, for example, confidence interval.

(2) *Hazard ratio not calculated or not reported*
In only 57 of the 575 reports (9.9%) were direct estimates of the hazard ratio or $\log_e$(hazard ratio) provided. For these, a variance (or a standard error) was given three times and a confidence interval 34 times. In the other 20, at most a *P*-value was given.
In 222 reports a Cox regression analysis or log-rank/Wilcoxon test had been performed and results given, but without a hazard ratio being stated. Instead either a *P*-value/χ^2-statistic from the analysis (n=210) or only a survival curve (n=12) was presented.
This illustrates the tendency of authors/journals to base the importance of a result on a P-value rather than a comparative group estimate with some measure of precision.

(3) *Inexact P-values provided*
Overall, 273 of the 575 reports presented a *P*-value from a Cox regression analysis or log-rank/Wilcoxon test. In 126 of these the *P*-value was stated as '$P<X$' or '*P*=significant', and in 13 reports the *P*-value was stated as '$P>X$' or '*P*=not significant'.
This again shows inappropriate emphasis placed on the P-value for a statistically significant result.

(4) *Group numbers and group events not given*
There were 210 reports where a *P*-value/χ^2-statistic from a Cox regression analysis or log-rank/Wilcoxon test was presented but without a hazard ratio or $\log_e$(hazard ratio). From the 194 of these that had a sample size >25, only 104 indirect estimates were obtained because the group numbers and/or group events were not reported and could not be estimated from figures or tables.
The number of patients and events in groups defined by marker levels are often smaller than the overall numbers because of missing or incomplete patient data. Hence, it is important to report numbers for the groups themselves.

(5) *Marker studies too small*
In Steps 2–5, estimation methods were only considered appropriate if the sample size was greater than 25, as it was in only 196 of 318 reports otherwise suitable for these steps.

When necessary and possible, research groups need to collaborate to achieve larger sample sizes and thus increase statistical power.

Figure 2 Description of the key reporting problems that prevented estimation of the $\log_e$(hazard ratio) and its variance in 371 (64.5%) of the reports

Heterogeneity of clinical and statistical factors The synthesis of our estimates was also restricted by the large variability in both clinical and statistical factors. For each estimate of the $\log_e$(hazard ratio) and its variance obtained, the cutoff level used to dichotomise the continuous markers, stage of disease, age of patients and outcome (overall or disease-free survival) were recorded, and also whether the estimate was unadjusted or adjusted and, if so, what adjustment factors were used. There was great diversity in these features (Table 2). For example, for the marker *MYC-N* there were 94 estimates of the $\log_e$(hazard ratio) and variance obtained but these involved nine different cutoff points, nine different stage groups, four different age groups, 77 unadjusted/17 adjusted estimates and two different outcomes (Table 3). Furthermore, of the 17 estimates that were adjusted for other prognostic markers or clinical features (using a Cox regression model) only two were adjusted for exactly the same set of factors, and these were from the same article (Maris *et al*, 2000).

Table 3 Heterogeneity in the 94 estimates of the $\log_e$(hazard ratio) and its variance obtained for marker *MYC-N*

	n
Outcome	
DFS	46
OS	48
Result type	
Unadjusted	77
Adjusted	17
Stage groups	
All	68
1	2
3	2
4	4
1, 2, 3	3
1, 2, 3, 4	5
2, 3, 4, 4S	2
3, 4	3
Unknown	5
Cutoff point	
1 copy	23
2 copies	1
3 copies	17
4 copies	5
5 copies	2
10 copies	18
Mean gene expression	2
Positive *vs* negative protein (or staining *vs* no staining)	9
Unknown	17
Age groups	
All	78
<1 year	2
>1 year	5
Unknown	9

OS=overall survival; DFS=disease-free survival.

This inconsistent and variable reporting was reflected equally in the estimates obtained for the other 12 markers (Table 2). The type of treatment of patients and the method of measuring the markers were not recorded, but both would have added further heterogeneity to that observed.

Publication bias and reporting bias The common problem of publication bias, and other reporting biases, may still affect our data extraction; some results that do not generate formal statistically significant or clinically valuable findings may not have been published, because of a reluctance of journals to report or of researchers to present negative findings. Such problems severely limit the conclusions that can be drawn from meta-analyses because not all the available evidence can be included, and therefore the pooled results are likely to be biased. We investigated the estimates obtained for *MYC-N* and indeed there did appear to be evidence of publication bias, with a number of studies with smaller hazard ratios considered to be missing (Riley *et al*, 2003). This problem is likely to be closely related to the problem of small sample sizes in some primary studies (Figure 2, key problem 5).

Should we proceed with meta-analysis?

The poor reporting, potential for publication bias and, in particular, the large heterogeneity across studies meant it was practically impossible to perform reliable meta-analyses that would determine the clinical importance of each marker studied. Even the analysis of subgroups of estimates was not considered realistic because it was virtually impossible to obtain subgroups that reflected patients with similar features. For example, for marker *MYC-N* there were 48 overall survival estimates obtained, of which 41 were unadjusted, and 30 related to 'all' stages and 'all' ages. Furthermore, only eight of these 30 estimates related to the most commonly used cutoff level of '1 copy number', and there is then the additional problem of heterogeneity for treatment used and method of measuring the marker, not to mention the potential impact of publication/reporting bias. The subgroup numbers were even smaller for the other, less-studied markers; for example, lactate dehydrogenase had only two unadjusted overall survival estimates relating to the most common cutoff level ($1500\,\mathrm{U\,l^{-1}}$) and patients of 'all' ages and 'all' stages.

The only possible benefit of meta-analysis using the estimates that we extracted is to highlight the results of previous studies and help prioritise which markers should be studied in the future. We take such an approach elsewhere (Riley *et al*, 2003b), but for the purposes of this feasibility study it is clear that no firm clinical policy decisions can be made from our evidence-based review.

DISCUSSION

Appraisal of the systematic review and data extraction

During the systematic review, we evaluated 3415 papers overall and identified 260 with results from studies assessing the prognostic power of tumour markers. This will have identified the majority of the English-language literature, but inevitably some papers will have been excluded unintentionally. However, it seems plausible that the reporting in such papers, and equally non-English papers, would be equally poor and heterogeneous.

We used the indirect methods suggested by Parmar *et al* (1998) to increase the number of occasions an estimate of the $\log_e$(hazard ratio) and its variance could be obtained. However, the estimates they provide are only approximate and simply make the best possible use of the results presented. Questions still exist about how best to combine indirect estimates with direct estimates. For this reason, we did not use other indirect methods. For example, given further assumptions, we could have used estimates of the proportion surviving to 2, 3, 5 or 10 years to obtain estimates of the $\log_e$(hazard ratio) and its variance (Vale *et al*, 2002). However, the papers were equally inadequate at presenting these survival statistics. For example, in the 26 prognosis papers for the serum marker lactate dehydrogenase, only 12 gave actuarial estimates of the proportion surviving, and only six of these also gave a confidence interval or standard error. They were also heterogeneous – five estimates were for overall survival, six were for disease-free survival and one was unspecified; estimates were made at 2, 3, 4 or 5 years. Further, very few reported numbers at risk explicitly, as required for reliable estimation.

Generalisations to other prognostic markers

Although these reporting problems were observed for tumour markers within the neuroblastoma literature, they have also limited reviews in other paediatric cancers (Riley *et al*, 2003a), and it seems plausible that the reporting will be equally poor for prognostic markers in other areas of oncology, and indeed other disease areas. Altman (2001, pp 228 – 247) discusses the potential problems involved in systematic reviews of prognostic markers, in particular that of poor and heterogeneous reporting of primary studies. Cutoff points are frequently used to dichotomise continuous markers and define groups, while different outcomes, adjustment factors and groups of patients are common features across prognostic studies. Inadequate reporting and presentation of survival data has been shown to be a concern in the cancer literature (Altman *et al*, 1995).

Reliable and clinically useful meta-analyses of observational and nonrandomised studies, such as the majority of prognostic marker studies, are generally difficult to perform (Fleiss and Gross, 1991). Other recent systematic reviews of prognostic markers have encountered similar problems to the ones we identified. Parker *et al* (2001) performed a systematic review in prostate cancer to establish whether age is a prognostic marker, but the incomplete and heterogeneous nature of the reports prohibited any quantitative overview. Similarly, a systematic review of prognostic laboratory variables in patients with unresected colorectal liver metastases was limited by the heterogeneity and poor quality of individual studies (Friedburg *et al*, 2001). Zandbergen *et al* (2001) performed a systematic review of biochemical markers of brain damage for identifying poor outcome in anoxic-ischaemic coma, but conclusions were limited by small sample sizes and different cutoffs and/or laboratory techniques.

Meta-analyses of prognostic markers have been facilitated when individual patient data were available (Look *et al*, 2002), in particular to determine a consistent cutoff level (Sakamoto *et al*, 1996). For those investigators currently interested in performing a quantitative review of prognostic markers, we recommend that they consider asking authors for individual patient data and/or the extra information they require, such as the $\log_e$(hazard ratio) and its variance, as this approach is likely to be the most productive.

Towards guidelines for improved reporting of prognostic markers

It is clearly important that the quality of primary studies, and the reporting of their results improve if clear conclusions and policy recommendations are to be formed about prognostic markers. Altman and Lyman (1998) have proposed important guidelines for both conducting and evaluating prognostic marker studies, including the need for prospective registration of studies. Alongside these, we have developed simple guidelines on how to report results to facilitate both interpretation of individual studies and the undertaking of systematic reviews, meta-analysis and, ultimately, evidence-based practice (Figure 3). Collaboration of research groups is required to promote such practice and achieve both the consistency and standards required. Ideally, *both* summary data and individual patient data should be reported according to our guidelines. It is important that *time to event* is incorporated within prognostic marker analyses, and thus the hazard ratio is preferred to other measures of relative risk such as the odds ratio, which relates to a fixed time-point and ignores censoring. However, in addition authors may wish to present the more familiar actuarial % survival at *n* years preferably with a confidence interval and the number of patients at risk at that time in each group.

Benefits of individual patient data Although improved reporting of summary statistics is very important, the availability of individual patient data is the most viable way forward in order to produce valid and clinically useful evidence-based reviews of prognostic markers. Subject to any restrictions imposed by data protection laws and guidelines, presentation or availability of full individual patient data using our guidelines would overcome

Guidelines for reporting prognostic marker studies

Objective: To improve reporting of prognostic marker results and facilitate access to individual patient data for evidence-based reviews

Results of *all* the marker analyses should be presented—both significant and nonsignificant results—and we recommend the following:

Essential to present:

(1) The *hazard ratio* and its *confidence interval*, or the $\log_e$(hazard ratio) and its variance. Markers that have a continuous function should be modelled as a continuous variable using appropriate methods. If there is a justifiable reason for using a cutoff level for a continuous marker, it should be specified *at the start* of the study and clearly reported.

(2) The *number of patients* and *number of events* in total. For binary markers (and continuous markers if a cutoff level is used) also report the numbers within each group.

(3) Both *unadjusted* and *adjusted* results for each marker. For adjusted results, clearly state what variables have been adjusted for. Ideally, a consistency in the set of adjustment factors used across studies should be sought through *collaborative groups* working towards prospectively planned pooled analyses. Otherwise, (i) always present results adjusted for age and stage of disease and (ii) consider using the same set of adjustment factor as in important earlier studies.

(4) *Individual patient data* in the paper or on the Internet, or make available with details clearly indicated within the paper. Data on markers that were not analysed should be included. Subject to any restrictions imposed by data protection laws and guidelines, include:

- Exact initial marker level and how marker was measured.
- Time of disease recurrence (if appropriate).
- Follow-up time.
- Final disease status.
- Levels of other existing prognostic markers of recognised and accepted importance for current clinical practice.
- Patient subgroup information, for example, age, stage of disease, type of treatment received.
- Details of inclusion/exclusion criteria would also be beneficial.

Highly desirable to present:

(5) *Exact P-values*. Reporting of results as 'significant' or 'not significant' is insufficient. Very small *P*-values can be given as $P<X$ (e.g. $P<0.0001$), but in this case the exact χ^2-statistic is also needed.

(6) *Survival curves* showing the difference in survival over time between the groups, with clear *step* and *censoring points*; also the initial numbers in each group, and the number of events and remaining numbers at various time-points during follow-up are needed.

(7) *% survival at n years* with a confidence interval using Kaplan–Meier or other methods that allow for censoring, together with the number of patients at risk at that time in each group.

Figure 3 Guidelines on how to report primary prognostic marker studies in order to improve current reporting standards and allow clinically useful evidence-based reviews to be made

variability in cutoff level, type of estimate (unadjusted or adjusted), outcome assessed (overall or disease-free survival) and adjustment factors; the study of markers in subgroups of patients (e.g. different ages, treatments) would also be easier. It would also eliminate the problem of extracting estimates when inexact *P*-values are presented, and would remove the need for arbitrary extraction decisions when an individual study presents a marker's results for a range of cutoff values. Furthermore, if levels of all the prognostic markers measured (even those producing nonsignificant results) are provided, then the problem of reporting bias would be reduced. However, publication bias might still be a concern if some studies are not published and do not make IPD available; prospective registration of studies is therefore also important to counteract this.

Individual patient data would also enable direct estimates of the hazard ratio, and other statistics of interest, when data were available but not used, analysed or presented properly in the primary study. A total of 41 (20%) of the 204 estimates that we obtained in the neuroblastoma review were direct estimates calculated from individual patient data that would not have otherwise been possible. It is clearly important to include predominately direct estimates in any quantitative synthesis. In fact, the potential for substantial differences in meta-analysis of survival data when using results provided within the literature instead of individual patient data has recently been shown in the head and neck cancer literature (Duchateau *et al*, 2001). Individual patient data would also allow model assumptions, for example proportional hazards, to be checked as necessary, and enable the baseline survival function to be estimated.

Presentation or availability of individual patient data would permit more appropriate meta-analyses (Stewart and Parmar, 1993), and would further facilitate the identification of different publications whose results relate to the same or overlapping set of patients. It would also allow an evaluation of combinations of markers, which may produce more specific and accurate prognostic assessments. If it is not appropriate or feasible to provide individual patient data within a paper itself, then there is the opportunity to publish on the Internet (Hutchon, 2001). Of course, even making individual patient data available on the web is not without its problems, with the nonpermanency of individual web-pages, and so perhaps a central repository to collate and manage individual patient data is needed within each disease area. The United Kingdom Children's Cancer Study Group have already initiated this type of approach within paediatric oncology (Mott *et al*, 1997). Authors may also wish to state in their paper that the IPD is available upon request (with contact details indicated) for those requiring it for evidence-based reviews.

We acknowledge that there are additional issues that arise when conducting individual patient data reviews (Stewart and Clarke, 1995), especially cost and time, but these have to be weighed against the substantial problems we encountered. Of course, even when prioritising the IPD approach, the meta-analyst will in practice end up with a mixture of estimates obtained from IPD and estimates obtained from summary statistics; hence, meta-analysis methods that take these different sources into account are needed.

Cutoff levels The use of different cutoffs makes synthesis of results particularly difficult. Of added concern is the possibility that the choice of cutoff level in a report may be specifically chosen to optimise the difference between the groups and produce a result with the maximum statistical or clinical significance possible (Altman *et al*, 1994; Altman and Lyman, 1998). If there is good clinical reason to use a cutoff level, then it should be specified at the start of a study and clearly reported within the results (Figure 3). However, Altman (2001) suggests that continuous markers should not be dichotomised because, among other reasons, this approach discards potentially important quantitative information and considerably reduces the power to detect a real association between the marker and outcome. Hence, we encourage researchers to analyse and report results (e.g. hazard ratio) of continuous markers on their original continuous scale. Importantly, availability of individual patient data including *exact* marker levels would allow data to be reanalysed where cutoff levels were not consistent, and also where continuous marker results were desired but results using a cutoff level were given (or *vice versa*) (Figure 3). Indeed, the most appropriate analysis of continuous prognostic markers may require nonlinear modelling techniques, as highlighted by Sauerbrei *et al* (1999); consultancy with statisticians or others experienced with such techniques is recommended in this situation.

Adjustment factors It is clear that once important prognostic markers have been identified, they need to be evaluated against, and also used in combination with, other known clinically useful prognostic factors, such as clinical characteristics (e.g. age, stage of disease) or indeed other marker levels. Prognostic marker results that are adjusted for other known prognostic factors will have the greatest implications for clinical practice, and subsequently meta-analyses of adjusted results are the necessity. However, if authors are inconsistent in the sets of adjustment factors they use, it becomes very difficult and impractical to pool results across studies and make a proper evaluation of markers over and above other factors. For the 17 adjusted *MYC-N* estimates, there were 16 different sets of adjustment factors, each containing one or more of age, stage of disease, Shimada index, lactate dehydrogenase and eight other prognostic markers. Individual study estimates of risk (e.g. hazard ratio) can be influenced by which adjustment factors are used (Rushton and Jones, 1992), and so there may be an additional reporting bias concern if researchers specifically only report those adjusted estimates with the most statistically significant result.

We recommend that research groups collaborate and identify the most commonly used prognostic tools in current practice, so that adjusted results of new prognostic marker studies can use consistent sets of adjustment factors. These identified prognostic tools should also be presented within the available individual patient data alongside the new markers being studied (Figure 2). This would allow adjusted results to be calculated independently across studies, using consistent sets of adjustment factors. Prognostic indexes could also be calculated across studies and evaluated in a meta-analysis if desired. Lambert *et al* (2002) have shown that individual patient data are generally required when investigating patient characteristics as effect modifiers in a meta-analysis, and, for prognostic markers, our study shows that the most valid and clinically useful meta-analysis results will only be obtained from an individual patient data approach.

CONCLUSION

Prognostic markers are important tools in the management of patients with cancer and many other diseases, and as such primary studies of prognostic markers are essential. However, the design and evaluation of such studies can be greatly improved (Altman and Lyman, 1998). Furthermore, we have shown that a change in how prognostic marker studies are reported is needed to provide more effective and meaningful results, and also allow important *evidence-based* reviews to be conducted. To facilitate such improved reporting, we have attempted to compile guidelines regarding how summary statistics and individual patient data should be presented. In particular, the availability of full individual patient data, including all markers considered, is the most viable way forward to produce valid and clinically useful evidence-based reviews and meta-analyses. Individual patient data would limit the large problems of poor and heterogeneous reporting that we observed, and also reduce the potential impact of reporting bias. Prospective registration of studies alongside the availability of IPD would also help restrict the potential for publication bias. These guidelines all point to researchers working together towards planned pooled analyses, currently a particularly important concept for epidemiological research (Blettner *et al*, 1999).

Research groups within each disease area should be encouraged to collaborate and facilitate these changes in practice; for example, by defining a clear set of important adjustment factors and by initiating central repositories to collate and manage individual patient data. This move towards a more evidence-based approach to the study and reporting of prognostic markers will help properly establish the most appropriate individual, and potential combinations of markers to be used in clinical practice, and should thereby improve patient care.

Clinical

ACKNOWLEDGEMENTS

We wish to thank Doug Altman for his helpful comments and advice during the drafting of this paper. This work was funded by the NHS Health Technology Assessment (HTA) programme [Project no. 97/15/03].

REFERENCES

Altman DG, Lausen B, Sauerbrei W, Schumacher M (1994) Dangers of using 'optimal' cutpoints in the evaluation of prognostic factors. *J Natl Cancer Inst* **86:** 829 – 835

Altman DG, Lyman GH (1998) Methodological challenges in the evaluation of prognostic factors in breast cancer. *Breast Cancer Res Treat* **52:** 289 – 303

Altman DG (2001) Systematic reviews of studies of prognostic variables. In *Systematic Reviews in Health Care: Meta-analysis in Context,* Egger M, Davey Smith G, Altman DG (eds) pp 228 – 247. London: BMJ Publishing Group

Altman DG, De Stavola BL, Love SB, Stepniewska KA (1995) Review of survival analyses published in cancer journals. *Br J Cancer* **72:** 511 – 518

Berthold F, Sahin K, Hero B, Christiansen H, Gehring M, Harms D, Horz S, Lampert F, Schwab M, Terpe J (1997) The current contribution of molecular factors to risk estimation in neuroblastoma patients. *Eur J Cancer* **33**(12)**:** 2092 – 2097

Blettner M, Sauerbrei W, Schlehofer B, Scheuchenpflug T, Friedenreich C (1999) Traditional reviews, meta-analyses and pooled analyses in epidemiology. *Int J Epidemiol* **28**(1)**:** 1 – 9

Duchateau L, Pignon JP, Bijnens L, Bertin S, Bourhis J, Sylvester R (2001) Individual patient-versus literature-based meta-analysis of survival data: time to event and event rate at a particular time can make a difference, an example based on head and neck cancer. *Controlled Clin Trials* **22**(5)**:** 538 – 547

Eche N, Pichon MF, Quillien V, Gory-Delabaere G, Riedinger JM, Basuyau JP, Daver A, Buecher B, Conroy T, Dieu L, Bidart JM, Deneux L (2001) Standards, options and recommendations for tumor markers in colorectal cancer. *Bull Cancer* **88**(12)**:** 1177 – 1206 [French]

Egger M, Davey Smith G, Altman DG (eds) (2001) *Systematic Reviews in Health Care: Meta-analysis in Context*. London: BMJ Publishing Group

Fleiss JL, Gross AJ (1991) Meta-analysis in epidemiology, with special reference to studies of the association between exposure to environmental tobacco smoke and lung cancer: a critique. *J Clin Epidemiol* **44:** 127 – 139

Friedberg B, Watine J, Miedouge M (2001) Unresected colorectal liver metastases: prognostic value of laboratory variables. *Gastroenterol Clin Biol* **25**(11)**:** 962 – 966

Hutchon DJR (2001) Publishing raw data and real time statistical analysis on e-journals. *Br Med J* **322**(7285)**:** 530

Lambert PC, Sutton AJ, Abrams KR, Jones DR (2002) A comparison of summary patient-level covariates in meta-regression with individual patient data meta-analysis. *J Clin Epidemiol* **55**(1)**:** 86 – 94

Look MP, van Putten WL, Duffy MJ, Harbeck N, Christensen IJ, Thomssen C, Kates R, Spyratos F, Ferno M, Eppenberger-Castori S, Sweep CG, Ulm K, Peyrat JP, Martin PM, Magdelenat H, Brunner N, Duggan C, Lisboa BW, Bendahl PO, Quillien V, Daver A, Ricolleau G, Meijer-van Gelder ME, Manders P, Fiets WE, Blankenstein MA, Broet P, Romain S, Daxenbichler G, Windbichler G, Cufer T, Borstnar S, Kueng W, Beex LV, Klijn JG, O'Higgins N, Eppenberger U, Janicke F, Schmitt M, Foekens JA (2002) Pooled analysis of prognostic impact of urokinase-type plasminogen activator and its inhibitor PAI-1 in 8377 breast cancer patients. *J Natl Cancer Inst* **94**(2)**:** 116 – 128

Maris JM, Weiss MJ, Guo C, Gerbing RB, Stram DO, White PS, Hogarty MD, Sulman EP, Thompson PM, Lukens JN, Matthay KK, Seeger RC, Brodeur GM (2000) Loss of heterozygosity at 1p36 independently predicts for disease progression but not decreased overall survival probability in neuroblastoma patients: a Children's Cancer Group study. *J Clin Oncol* **18**(9)**:** 1888 – 1899

Mott MG, Mann JR, Stiller CA (1997) The United Kingdom Children's Cancer Study Group – the first 20 years of growth and development. *Eur J Cancer* **33**(9)**:** 1448 – 1452

Parker CC, Gospodarowicz M, Warde P (2001) Does age influence the behaviour of localized prostate cancer? *BJU Int* **87:** 629 – 637

Parmar MKB, Torri V, Stewart L (1998) Extracting summary statistics to perform meta-analyses of the published literature for survival endpoints. *Stat Med* **17:** 2815 – 2834

Riley RD, Burchill SA, Abrams KR, Heney D, Sutton AJ, Jones DR, Lambert PC, Young B, Wailoo AJ, Lewis IJ (2003a) A systematic review of molecular and biological markers in tumours of the Ewing's sarcoma family. *Eur J Cancer* **39:** 19 – 30

Riley RD, Burchill SA, Abrams KR, Heney D, Lambert PC, Jones DR, Sutton AJ, Young B, Wailoo AJ, Lewis IJ (2003b) Systematic review and evaluation of the use of tumour markers in paediatric oncology: Ewing's sarcoma and neuroblastoma. *NHS Health Technology Assessment* (No. 97/15/03) 2003. Vol. 7, No. 5

Rushton L, Jones DR (1992) Oral contraceptive use and breast cancer risk: a meta-analysis of variations with age at diagnosis, parity and total duration of oral contraceptive use. *Br J Obstet Gynaecol* **99:** 239 – 246

Sakamoto J, Teramukai S, Koike A, Saji S, Ohashi Y, Nakazato H (1996) Prognostic value of preoperative immunosuppressive acidic protein in patients with gastric carcinoma. Findings from three independent clinical trials. Tumor Marker Committee for the Study Group of Immunochemotherapy with PSK for Gastric Cancer. *Cancer* **77**(11)**:** 2206 – 2212.

Sauerbrei W, Royston P, Bojar H, Schmoor C, Schumacher M (1999) Modelling the effects of standard prognostic factors in node-positive breast cancer. German Breast Cancer Study Group (GBSG). *Br J Cancer* **79**(11 – 12)**:** 1752 – 1760

Stewart LA, Clarke MJ (1995) Practical methodology of meta-analyses (overviews) using individual patient data. Cochrane Working Group. *Stat Med* **14:** 2057 – 2079

Stewart LA, Parmar MKB (1993) Meta-analysis of the literature or of individual patient data: is there a difference? *The Lancet* **341:** 418 – 422

Sutton AJ, Abrams KR, Jones DR, Sheldon TA, Song F (2000) *Methods for Meta-analysis in Medical Research*. London: John Wiley

Vale CL, Tierney JF, Stewart LA (2002) Effects of adjusting for censoring on meta-analyses of time-to-event outcomes. *Int J Epidemiol* **31**(1)**:** 107 – 111

Zandbergen EGJ, de Haan RJ, Hijdra A (2001) Systematic review of prediction of poor outcome in anoxic-ischaemic coma with biochemical markers of brain damage. *Intens Care Med* **27:** 1661 – 1667