INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI

A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor MI 48106-1346 USA
313/761-4700 800/521-0600

Université d'Ottawa · University of Ottawa

Non-Small Cell Lung Cancer: Assessment of methodologies to combine survival curves in meta-analysis

Craig Earle MD FRCPC

Thesis submitted to the School of Graduate Studies and Research in partial fulfillment of the requirements for the degree of Master of Science in Epidemiology

University of Ottawa

June 1998

Abstract

Purpose: To assess the accuracy of different methods of combining published survival curves into a summary curve that can be used in disease models or to make treatment comparisons in meta-analysis.

Methods: Five methods for combining survival curves were identified: 1) iterative generalized least-squares (IGLS, Dear 94), 2) meta-analysis (MA) of failure-time data with adjustment for covariates (MFD, Hunink 94), 3) non-linear regression (NLR, Shore 90), 4) log relative risk (LRR, Voest 89), and 5) weighted LRR (w-LRR). Published survival curves were scanned into a graphical package, coordinates along each curve were identified, and required data elements were extracted semi-automatically. Two separate analyses were done: 1) each method was used to combine the survival curves from eight single arm Phase II trials of chemotherapy in 918 patients with advanced non-small cell lung cancer (NSCLC); 2) each method was used to combine the survival curves from seven randomized trials in order to reproduce a MA of chemotherapy in advanced NSCLC. I compared the resulting summary curves using the Kolmogorov-Smirnoff goodness of fit test to a "gold standard" curve calculated from the corresponding individual patient data (IPD).

Results: Data extraction from the published survival curves was reliable. All methods were able to produce reasonably accurate summary survival curves, and no method was consistently more accurate than the others. Maximum discrepancies of the calculated summary curves from the IPD curves ranged from 1.8 – 4.1% for the first analysis, and up to 11% in the reproduction of the meta-analysis. None of these differences were

statistically significant. All methods were able to differentiate between the two treatment arms of the meta-analysis, which had 6 month survival rates of 33.5% and 51.8% in its two treatment arms. Only one method, IGLS, preserves randomization for theoretically justified comparisons of intervention arms. It also does not require provision of unusual detail in published reports. Characteristics of the component trials that affected the accuracy of the different techniques were 1) a high proportion of censored observations (MFD); 2) variability in the length of follow-up (IGLS, NLR, LRR, w-LRR); and 3) the heterogeneity of the treatment results (NLR, w-LRR).

Conclusions: All methods were able to accurately reproduce summary survival curves from the published literature and to detect treatment differences. Dear's method is preferred for comparing intervention groups because it preserves randomization.

Acknowledgements

This thesis could never have been completed without the help of several people.

I want to thank my supervisors, Dr. George Wells and Ba' Pham, for their continuous advice and support. I particularly want to give kudos to Ba' for his supreme patience in transmitting some of his programming expertise to a complete novice.

The Health Analysis and Modelling Group at Statistics Canada (B. Phyllis Will, Jean-Marie Berthelot, Christian Houle, and Bill Flanagan) were instrumental in providing the data necessary to carry out these analyses.

I also want to thank Drs. Jean Maroun and Bill Evans at the Ottawa Regional Cancer Centre for their moral support, as well as providing the necessary computer resources for this project.

Lastly, I want acknowledge the sacrifices made by my co-vivant, Tammy Binder. She was often without a spouse as a result this undertaking, yet always managed to give encouragement and invaluable feedback.

Table of Contents

Tables

Figures

1. Introduction

1.1 The Problem

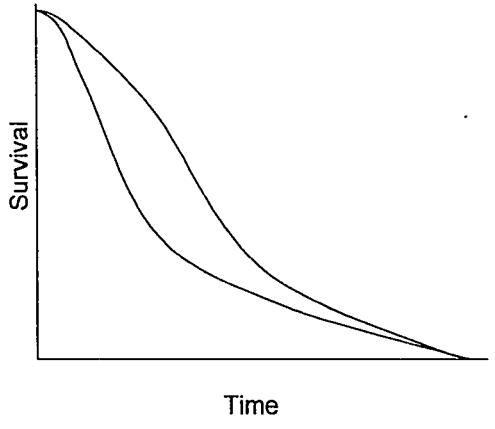
Clinical trials in oncology often involve survival analyses.[1] Meta-analyses of these trials are frequently undertaken to summarize their results and gain more precise estimates of treatment effects.[2] This thesis attempts to empirically assess some of the methodologies that have been employed to combine published survival figures into summary survival curves.

Common outcomes in cancer studies include the overall survival,[3] cause-specific survival,[4] and the time to relapse (or progression-free survival).[5] Meta-analyses of studies with these "time-to-event" end points usually synthesize the data by dichotomizing the results of a study into whether an event has occurred or not at a fixed point in time[6-9] such as the one-year survival.[7;9] Unfortunately, dichotomizing results discards potentially useful information about the timing of events and the resultant shape of the survival curve.[10] Moreover, it may lead to incorrect conclusions if the wrong time point is chosen.[11] For example, many cancers have only palliative treatments aimed at prolonging short-term survival and increasing quality of life.[12] An important effect could be missed if assessed too late (Figure 1-A). In addition, dichotomized outcomes provide less statistical power to compare treatments than time-to-event analyses.[11;13]

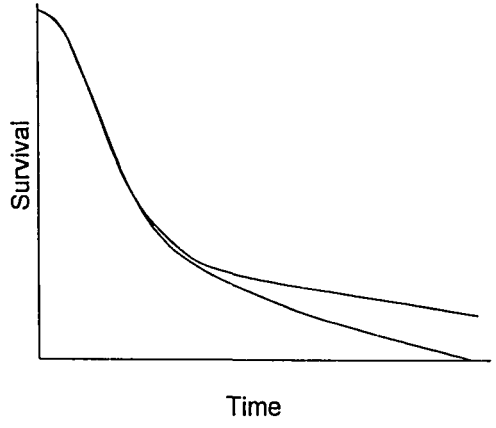
Accurate methods to combine survival curves are desirable. Research applications such as Markov modelling for decision analyses often require input of survival probabilities at many different points in time.[14;15] Also, economic analyses frequently rely

Figure 1. Information contained in the shape of a survival curve

A.



B.



Comparing the survival proportions between two curves at one specific point in time could lead to incorrect results if the time point compared is too late (A) or too early (B) to detect a treatment difference.

on calculation of the area between survival curves for estimates of effectiveness.[16-18] Moreover, it is sometimes necessary to use hazard ratios to estimate future survival patterns.[18; 19] Pooled survival data from several studies could provide representative estimates of survival for such analyses.[20; 21]

Clarke,[11; 22; 23] Stewart and Parmar[24] and others[16; 25-29] have advocated combining the individual patient data (IPD) from relevant studies as the best way to arrive at a summary survival curve. There are at least eight examples of meta-analyses that have been done separately with dichotomized results from published data and with IPD.[22; 24; 30-38] In five cases the effect size was overestimated in the analysis of published data.[24; 30-34; 38] The effect size was underestimated in the one,[22; 35] similar in another,[37] and not yet determined in the last.[36] These results suggest at least a tendency for meta-analyses of published data to be susceptible to Type 1 error.[27; 29]

Unfortunately, individual patient data are commonly not available. Meta-analyses of IPD are much more costly than analyses done with published data[37] and take more time.[36] As well, there may be an inability or reluctance on the part of the original investigators to provide data,[23; 25; 39] resulting in a possibly biased sample of studies being available.[37]

Without individual patient data, combining published survival curves from different studies presents several methodological challenges.[40; 41] Proportional survival generally must be extracted from figures in the publications, and published reports usually do not provide enough information about censored observations to accurately reconstruct life tables.[20; 42] As a result, these analyses are rarely done. Therefore, validated methods to combine survival curves using published data would be useful.[39]

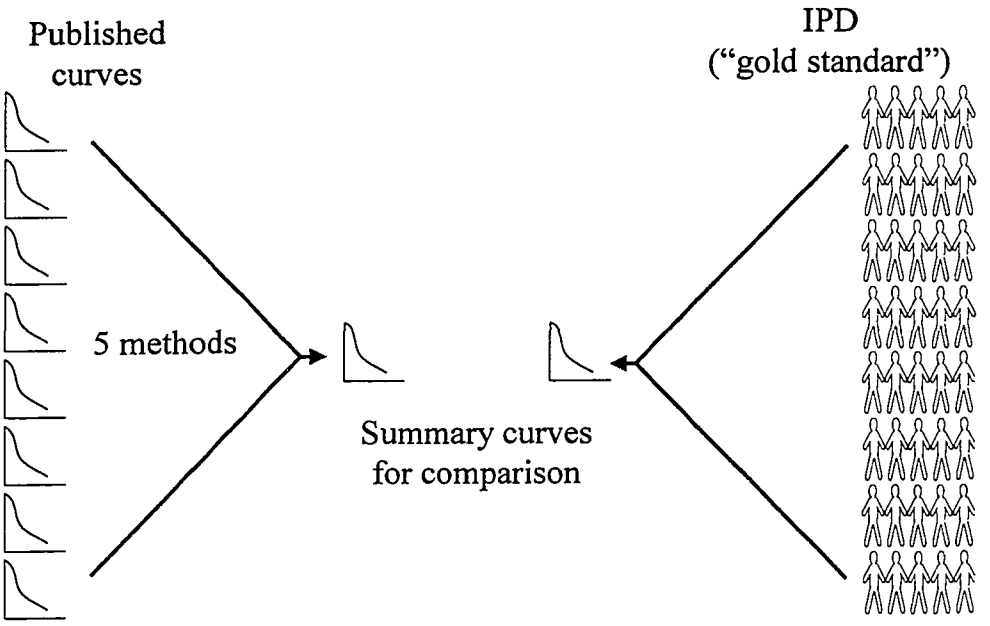
In this thesis I review the common methods used to generate survival curves, and methodologies that have been used to combine survival curves from published reports. I will demonstrate the accuracy of these methodologies empirically by using each to combine published survival curves from studies of patients treated with chemotherapy for advanced non-small cell lung cancer, and compare each method's summary curve to that derived from the corresponding IPD (Figure 2).

1.2 Example of lung cancer chemotherapy

Published randomized trials of chemotherapy versus best supportive care (BSC) in advanced non-small cell lung cancer (NSCLC) have produced a literature of conflicting small trials.[1; 43-48] Four meta-analyses have attempted to summarize the evidence.[7-9; 31] Three did not have access to individual patient data. Two of these studies (Souquet et al.[7] and Grilli et al.[9]) extracted the number at risk and the number of events from the survival curves at 3 month intervals, and then analyzed each of these time points as dichotomous outcomes. The third meta-analysis, by Marino and colleagues,[8] used the odds ratio of death at 6 months as their endpoint.

Each of these meta-analyses used different statistical methods and included different trials. All assumed that censoring did not occur, and all found statistically significant improvements in survival with chemotherapy. Marino et al.[8] had found a pooled odds ratio for death of 0.44 (95% confidence interval (CI): 0.32, 0.59) for patients taking chemotherapy compared to best supportive care, while Souquet et al. found odds

Figure 2. Overview of the thesis



I have used five methods to create summary survival curves from several published trials of chemotherapy in non-small cell lung cancer, and I have derived a summary curve from the corresponding individual patient data (IPD) to serve as the reference or "gold standard" curve for comparison.

ratios ranging from .05 at 3 months to .66 at 18 months. Unfortunately, the odds ratio is a poor predictor of the hazard ratio in diseases with high event rates like lung cancer. As a result, clinicians may interpret such results as being overly favourable.

Both the Souquet and Grilli studies found that significant benefit from chemotherapy occurred at the 6-month analyses, but did not persist beyond, illustrating the importance of the choice of the time point of comparison. The variability in results between meta-analyses and at different time points[18] has caused concern over the validity of such literature-based analyses.[27]

In 1995, the Non-Small Cell Lung Cancer Collaborative Group (NSCLCCG)[31] used individual patient data from 11 trials involving 1190 subjects with advanced lung cancer. They were able to construct survival curves, and found a pooled hazard ratio of death of 0.73 in favour of patients treated with eight cisplatin-containing chemotherapy regimens in seven trials. While there is still controversy over the clinical significance of this result, it has been accepted as proof of the biologic efficacy of chemotherapy.[49]

1.2.1 Stratification

A frequent source of difficulty for the reader of lung cancer chemotherapy studies is that patients with stage III (locally advanced) and stage IV (metastatic) lung cancer are usually analyzed together as "advanced disease". Although it may be appropriate to treat both of these patient groups with chemotherapy, their prognoses are quite different. The median survival for stage III patients is 10 months, while stage IV patients live a median of less than 6 months after diagnosis.[50] This difference is greater than that provided by

any treatment.[12] As a result, the relative proportions of patients with each stage in a study population can have a significant impact on the results.

Most meta-analytic techniques assume that the populations in the component studies are the same. However, disparities in the distribution of subgroups such as stage III and IV patients can affect the outcome of an analysis. A method to stratify analyses into subgroups of interest or adjust for the distributions of these prognostic features could improve the validity of such overviews.

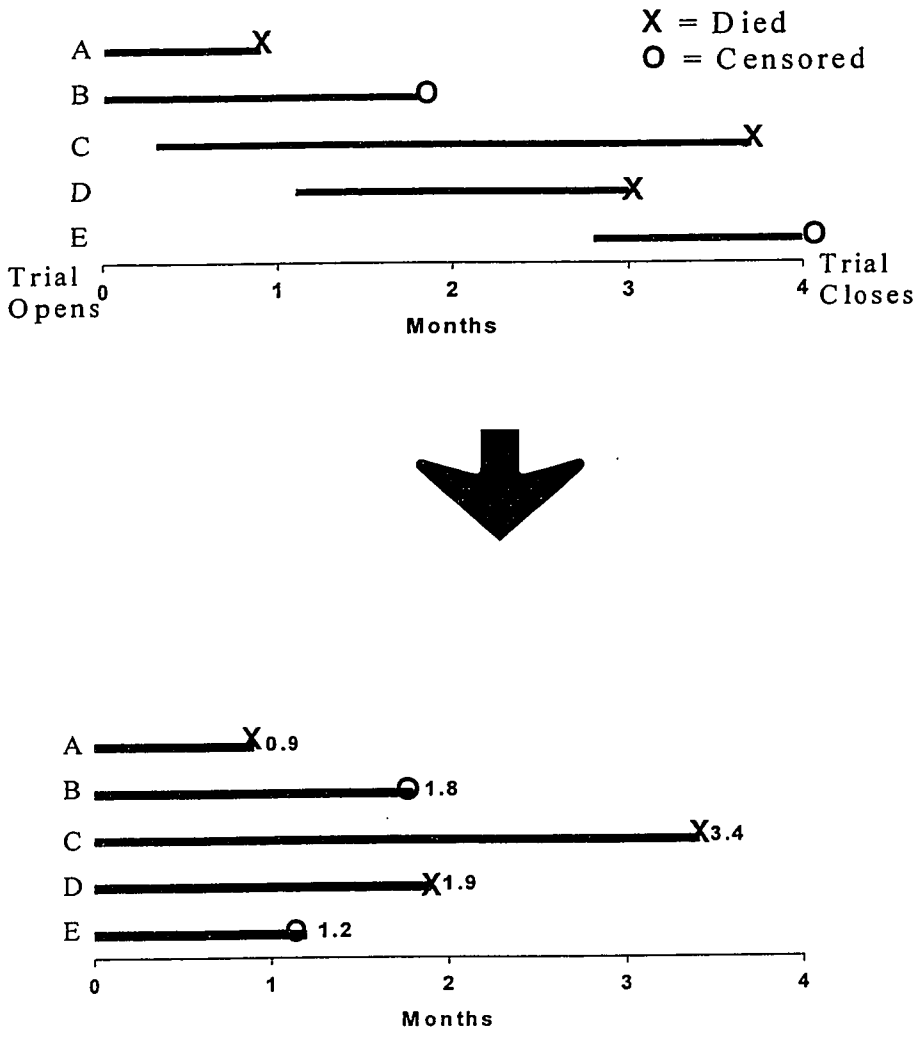
1.3 Background

Constructing a survival curve requires knowledge of two time points for each subject: the time of origin (e.g. the time of diagnosis, entry into a trial, or start of therapy), and the time of an event, (e.g. death, relapse, disease progression, or a combination of these). An inception cohort can then be constructed by artificially bringing each subject's starting time back to a common origin (Figure 3).

1.3.1 Censoring

Sometimes the exact time of origin, event, or both may be unknown. "Right censoring" occurs when a subject with a known time of origin who has not experienced an outcome event stops contributing data to the survival curve. All that is known is that the subject survived longer than the time for which he was observed. This can occur if the subject was lost to follow-up, the study ended and was analyzed before he could experience the outcome, or if he died from a cause not considered an outcome event (Figure 3). Less common is "left censoring" where the time of the event (e.g. death) is

Figure 3. Creating an inception cohort for survival analysis, and two examples of censoring



Subjects (A – E) that enter a trial at different times are brought back to a common origin for analysis. Subject B is right censored because he was lost to follow-up or died of a non-outcome event. Subject E is censored because no event occurred by the time the trial closed.

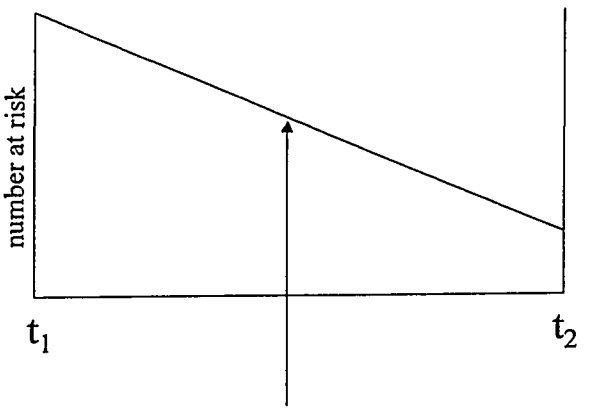
known, but the exact time of origin (e.g. onset of chronic lymphocytic leukemia) is not.[10] If neither the time of origin nor the time of event is known, the patient is "doubly censored." The only situation in which censoring does not occur is when the time of origin is known for all subjects, none are lost to follow-up or die of other causes, and all experience an outcome event before the study is analyzed.

The exact time of censoring may be unknown as well. This occurs when assessments for the outcome event or data collection only happen at periodic intervals. If a subject is lost to follow-up, it is only known that they were lost at some point during the last interval. For simplicity, the most common assumption made to deal with this is that censoring occurs with equal probability over time. Therefore, the average time of censoring will be the midpoint of the interval (Figure 4). This assumption is applied when calculating actuarial life tables, described later.

There are several possible approaches to censored observations. Analysis could be restricted to subjects for which there are complete data. This has the advantage of simplicity, but because of the substantial loss of sample size it is rarely done. Another approach is to make extreme assumptions about the fate of censored subjects, such as assuming that all censored subjects experienced an event at the time of censoring, or that none of them did. While this may give an indication of the analysis' sensitivity to censoring, it would introduce its own bias into the calculation of the curve.

Most often, a "likelihood-based approach" is used. It relies on an assumption of "non-informative"[10] or "independent"[51] censoring; that is, the reason for censoring is not related to the subject's probability of experiencing an event. This implies that censored subjects have a subsequent survival experience that is similar to that of the subjects with

Figure 4. Censoring in actuarial analysis



Subjects lost to follow-up between two time intervals, t_1 and t_2, are assumed to be censored at a constant rate. On average, they would be censored at the midpoint of the time interval (arrow) and be at risk of an event for only half of the interval. Therefore, actuarial calculations subtract half of the censored subjects from the number at risk at the start of an interval to arrive at the effective sample size for the interval.

longer follow-up. This is likely a reasonable assumption for subjects censored at the end of a study, as there is usually no reason to believe that the time of entry into the study (which determines the length of follow-up and is usually random) is related to their likelihood of an event. However, if there had been a secular trend in the occurrence of the outcome during the time the study was open, this might not be true.[52]

The validity of the non-informative censoring assumption is much less clear for subjects who are lost to follow-up. If a subject is too ill to come for follow-up because of an impending event, their censoring may be "informative". The subject likely has a higher probability of an event than the remaining subjects. The resultant calculation would overestimate survival after this subject is censored. On the other hand, subjects could feel so well that they decide not to return. In this case, the survival function will be underestimated. In reality, studies can contain a mixture of informative censoring with overestimation of survival, informative censoring with underestimation of survival, and non-informative censoring.[10]

1.3.2 Estimating the survival function

The two most frequently used[53] methods to estimate survival curves are:

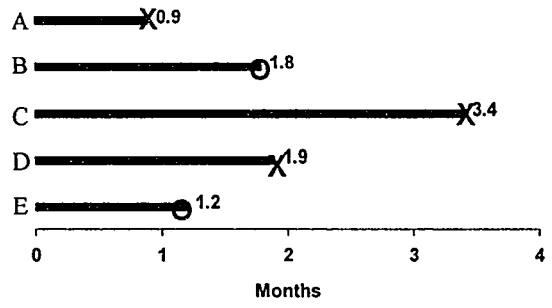
1. the actuarial (life table, Cutter-Ederer) method, and
2. the Kaplan-Meier (product-limit) estimate.

The actuarial method (Figure 5) is computationally simpler than the Kaplan-Meier method. Data are grouped into intervals and calculations are made based on the number of subjects who experienced an outcome event or were censored during the interval. The intervals do not need to be equal.[54] Before widespread access to microcomputers, the

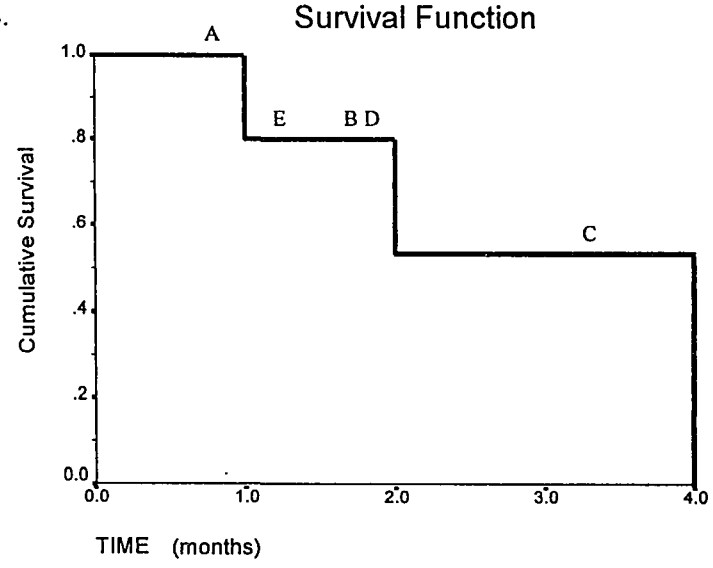
Figure 5. Calculating an actuarial survival curve

Interval (months)	# alive at start ($r_i = r_{i-1} - c_{i-1} - d_{i-1}$)	Censored (c_i)	# at risk during interval ($n_i = r_i - \frac{1}{2}c_i$)	# deaths during interval (d_i)	Prob. surviving interval ($P_i = 1 - d_i/n_i$)	Survival subsequent to the interval ($S_i = S_{i-1} * P_i$)
-	-	-	-	-	1.0	1.0
0 - 1	5	0	5	1	.8	.8
1 - 2	4	2	3	1	.66	.53
2 - 3	1	0	1	0	1.0	.53
3 - 4	1	0	1	1	0	0

A.



B.



Knowledge of the number of censored observations and outcome events in each interval (A) allows calculation of an actuarial survival curve (B).

actuarial method was used to reduce the number of calculations required to plot a large data set. Today it is mostly used when data are naturally grouped into intervals, such as in the case of census data.[55] As described previously, because the actual time of censoring within the interval is not known, it is assumed that they would be observed for half of the interval on average. Therefore, the number of patients at risk is the number that were at risk at the start of the interval minus half of the number censored during the interval.

If the time intervals of an actuarial analysis are made smaller and smaller toward a single unit of time, it becomes a Kaplan-Meier plot.[54] Here, the interval of interest is the time between two distinct events. Figure 6 shows the calculation of a Kaplan-Meier estimate. The proportional survival is recalculated each time an event occurs. Therefore, usually only one event occurs in each time interval, the intervals are of varying length, and the events occur at the beginning of the interval, while censoring only affects the start of the next interval.

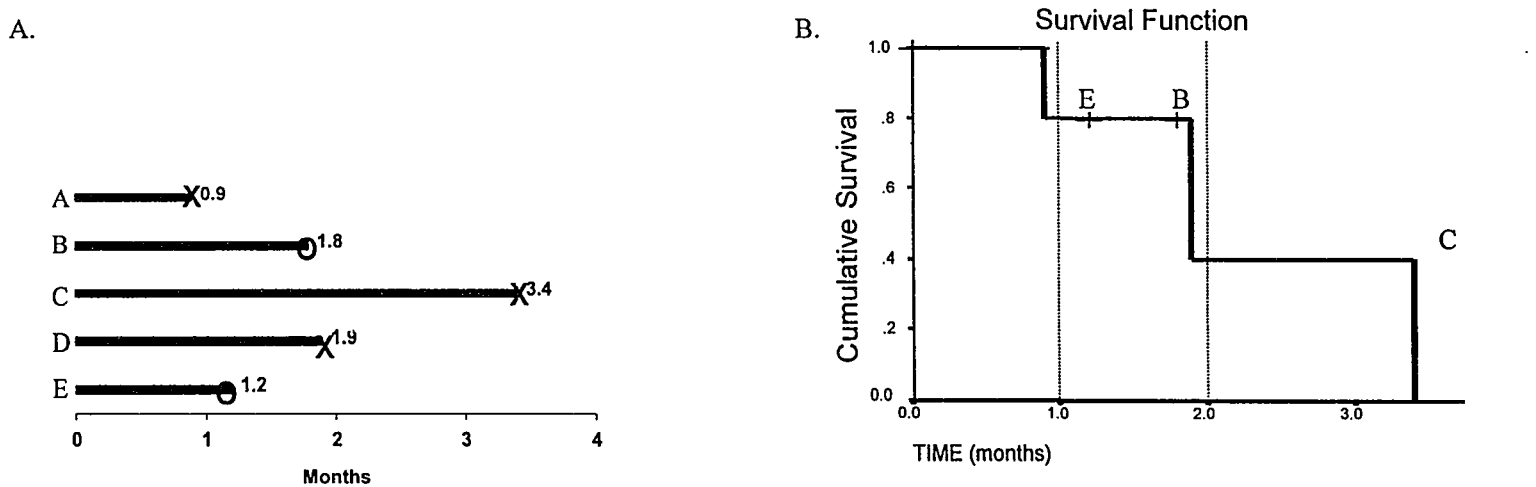
The principle underlying calculation of the survival function $S(t)$ with either the Kaplan-Meier or actuarial method is that the probability of surviving two time intervals, $P_{0\text{-}2}$, is the product of the probability of surviving the first interval $P_{0\text{-}1}$ and the probability of surviving the second $P_{1\text{-}2}$:

$$P_{0\text{-}2} = (P_{0\text{-}1})(P_{1\text{-}2})$$

As a result, the probability of surviving longer than a given time is the product of the conditional probabilities of the preceding time intervals.[52] The Kaplan-Meier method is generally preferred because it makes use of more of the observed information about survival than the actuarial method. However, both rely on the assumption of non-informative censoring. Also, because censored subjects are only included in the total

Figure 6. Calculating a Kaplan Meier survival curve

Interval between Deaths	# alive immediately before end of interval ($n_i = n_{i-1} - d_{i-1} - c_i$)	Censored (c_i)	# Deaths (d_i)	Prob. surviving interval ($P_i = 1 - d_i/n_i$)	Survival subsequent to the interval ($S_i = S_{i-1} * P_i$)
0	-	-	-	1.0	1.0
0 - 0.9	5	0	1	0.8	0.8
0.9 - 1.9	2	2	1	0.5	0.4
1.9 - 3.4	1	0	1	0	0



Knowledge of the exact timing of censored observations and outcome events (A) allows calculation of a Kaplan-Meier survival curve (B).

number of patients at risk up until their time of censoring, the tail of a survival curve estimated by either method may be based on observations from only a few subjects.

1.3.3 Relationships between functions

The distribution of survival times is characterized by three inter-related functions:

1. the survivorship function $S(t)$ is the proportional survival up to time t, or the probability that an individual will survive longer than time t. The graph of this function is the commonly known survival curve;
2. the probability density function, $F(t) = 1 - S(t)$, is the unconditional failure rate, the probability of having an event before time t; and
3. the hazard function:

$$h(t) = \frac{\text{Number dying in an interval starting at } t}{\text{Number at risk during that interval}}$$

This is known as the conditional failure rate because it describes the instantaneous probability of experiencing an event in a small time interval Δt given that the subject has survived to time t. As the time intervals become infinitely small the hazard function becomes a continuous function, and $h(t) = -\ln(S(t))$, the instantaneous probability of failure.

As can be seen, if one of these functions is known, the other two can be derived. The methods examined in this thesis either calculate the summary survivorship function directly,[56-58] or estimate the combined hazard function and use it to derive the survival curve.[41]

Combining survival curves introduces a special case of censoring. When studies are of different lengths, any remaining subjects in a shorter trial are censored at the time that trial ends. The termination of a trial with a large sample size or an extreme result can cause important changes in the remainder of the summary curve. I will refer to this type of censoring as "study termination" censoring.

1.4 Relevance

As the example of chemotherapy in lung cancer illustrates, reliable methods to produce a summary plot of published survival curves would be very useful for both research applications and clinical decisions. Furthermore, a method to allow assessment of stratified results could be very informative. Several authors have proposed methods of constructing summary actuarial survival curves.[41; 56-58] If such procedures are accurate, more precise survival estimates could be available for disease modelling, and systematic reviews could make more complete comparisons of treatment groups. Empirical assessment of these methods has never been attempted before; this thesis will be the first such undertaking.

2. Methods

2.1 Primary objective

To assess the accuracy of five methods for creating summary survival curves from published studies. I assessed the methods in two ways:

1) I combined published survival data from eight treatment groups in four Phase II studies of patients with advanced NSCLC treated with chemotherapy and compared the resultant survival curves with one derived from the corresponding individual patient data;
2) I applied each method to the studies in the NSCLCCG meta-analysis in order to replicate the treatment and control survival curves. I then compared these with the summary curve derived from IPD in the meta-analysis.

2.2 Secondary objectives

i. To examine a method to approximate the survival curves for patient subgroups by stratifying data between Stage III and Stage IV NSCLC patients.
ii. To assess the effects of censoring assumptions on the summary survival curves.
iii. To assess the effect of the heterogeneity of survival between trials on each method.
iv. To assess each method's ability to detect treatment differences.

2.3 Literature Search

I searched the English literature in the MEDLINE database from January 1966 to September 1998 to identify current methodologies to combine survival curves. Search

terms used were: life tables, survival analysis, actuarial analysis, proportional hazards model, and meta-analysis. I reviewed abstracts and titles identified by the searches, as well as those found in personal files and cited in relevant papers and reviews. I also contacted the authors of the methods and other experts in the field for references. Relevant articles were retrieved. This search revealed the four distinct methods described below as well as a fifth method that was a refinement of one of the techniques. Of note, it would appear that each method was developed for and used in only one specific study.

2.4 Data Sources

The data for the first part of this project (objective 2.1.1) came from four published studies of chemotherapy in advanced NSCLC for which I had access to individual patient data. Two were single arm Phase II studies and two were three arm randomized Phase II studies (Table 1). This yielded a total of eight treatment arms and 918 patients. As different treatments have only modest if any effect on survival in lung cancer patients,[59] these were considered as eight single arm studies. The IPD comparator curve was an actuarial survival curve created from the pooled IPD. The use of these data was confined to the objectives outlined above. The studies will not be identified to avoid implying any conclusions about the effectiveness of certain treatments.

Data for replicating the NSCLCCG meta-analysis (objective 2.1.2) were extracted from the published reports of the primary studies[1; 43-48](Table 2). The IPD comparator curve was extracted from the published report of the NSCLCCG meta-analysis of IPD.[31]

I also created eight simulated trials with a total of 2500 patients. Survival times followed an exponential decay function[60] with a constant, randomly generated hazard.[61]

Table 1. Characteristics of the eight single-arm Phase II studies

Trial	sample size	number censored	Stage III/IV	% III/IV	Follow-up (weeks)
1	27	2	3/24	11/89	212
2	161	36	57/104	35/65	76
3	25	0	0/25	0/100	232
4	48	0	0/48	0/100	184
5	45	0	0/45	0/100	248
6	206	19	91/115	44/56	156
7	206	29	85/121	41/59	140
8	200	17	75/125	38/63	132

Table 2. Characteristics of the randomized trials in the NSCLCCG meta-analysis

Study	Treatment	sample size	number censored	Stage III/IV	% III/IV	Follow-up (weeks)
Cartei [1]	CMP	52	0	0/52	0/100	132
	BSC	50	0	0/52	0/100	88
Cellerino [47]	CEP/VP-MTX	62	4	25/37	40/60	200
	BSC	61	1	26/35	43/57	176
Ganz [48]	VBL-P	22	3	0/22	0/100	100
	BSC	26	1	0/26	0/100	80
Kaasa [49]	VP-P	44	0	0/44	0/100	64
	BSC	44	3	0/44	0/100	220
Rapp [50]	CAP	43	0	6/37	14/86	56
	VDS-P	44	0	8/36	18/82	52
	BSC	50	0	5/45	10/90	44
Quiox [51]	VDS-P	24	3	0/24	0/100	96
	BSC	22	0	0/22	0/100	32
Woods [52]	VDS-P	97	0	25/72	26/74	228
	BSC	91	0	39/52	43/57	228

NSCLCCG = Non-Small Cell Lung Cancer Collaborative Group
chemo = chemotherapy
BSC = best supportive care
C = cyclophosphamide
M = mitomycin-C
P = cisplatin
E = epirubicin
VP = etoposide
MTX = methotrexate
VBL = vinblastine
A = adriamycin
VDS = vindesine

Right-censored observations were distributed randomly, assuming a constant accrual rate.[62]

2.5 Extracting data from published reports

The data elements required to carry out the analyses were:

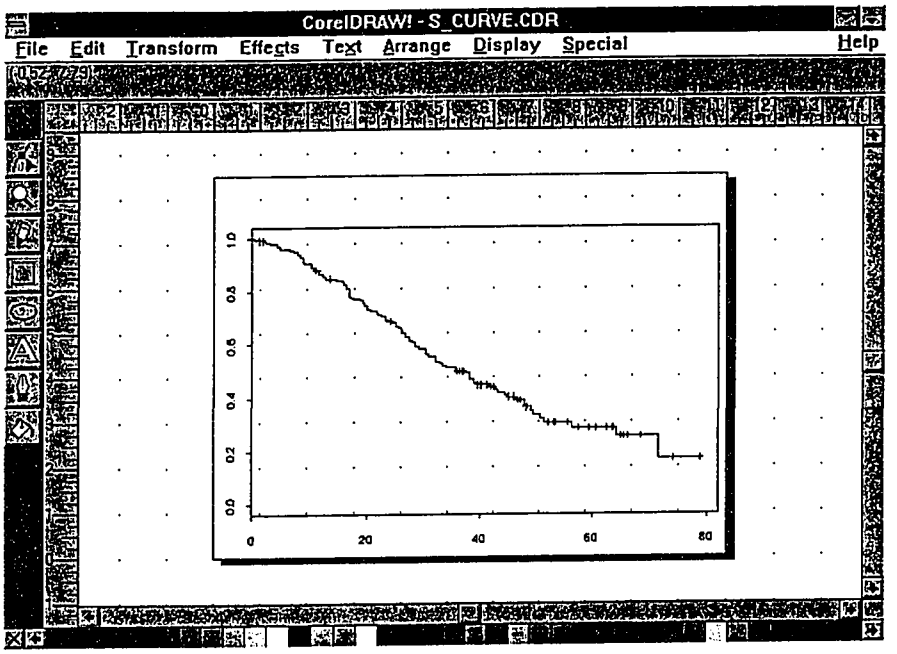
- the number of patients at risk in each time interval
- the number of events in each time interval
- the proportional survival in each time interval

However, this information is usually not in the text of published reports, and obtaining accurate data from published survival figures can be difficult because of the small size or distortions in reproduction of the survival graph. In order to extract data from the survival curves, I scanned them as Windows metafiles and imported them into CorelDRAW! 3.0. In CorelDRAW! the mouse pointer moves around a co-ordinate system. By enlarging the survival curves and re-tracing them with the mouse pointer, exact measurements of the survival proportion $S(t)$ could be made at any point in time (Figure 7). The axes were checked to ensure that there was no significant distortion of the image. All data were extracted on two separate occasions by the same reviewer several weeks apart to ensure that there were no errors, and minor differences were resolved by reviewing those points on the curve a third time to decide which value was most appropriate.

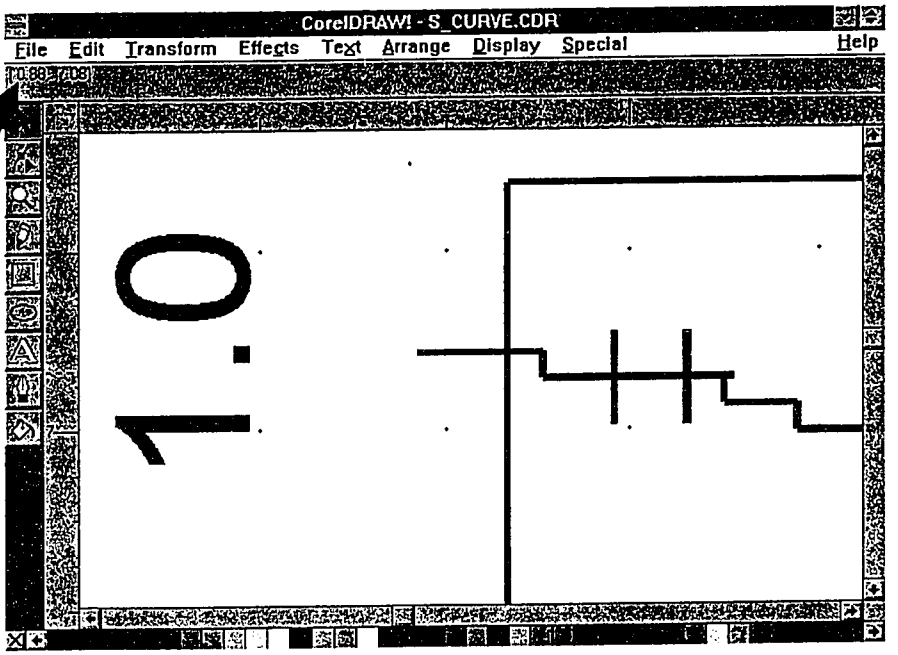
I obtained the initial sample size, stage distribution, and number of censored observations from the text and tables of each study. Censored patients were those who were described as being either lost to follow up or still alive at the end of the trial. I calculated the probability of surviving each one-month interval P_i using the formula:

Figure 7. Scanning survival curves for data extraction

A.



B.



Scanning survival curves as graphics (A) allows enlargement (B) and reading of the coordinates of each step in the curve (arrow).

$P_i = S_i/S_{i-1}$ (Figure 8). The standard error of the survival proportion at each time point was calculated using the Greenwood formula for the variance: a sum of terms $d/(n(n-d))$, where d is the number of deaths during a given time interval and n is the number of individuals at risk during that interval.[51] Where the curves allowed reading of censored observations (c_i), these were also recorded in the appropriate time interval. From these data, and knowing the initial sample size, the number at risk ($n_i = r_i - 1/2c_i$, where r_i is the number of patients alive at the beginning of the interval, and censored observations are assumed to contribute to half of the interval), and the number of failures ($d_i = n_i(1 - P_i)$) in each interval could be calculated. This completed the actuarial life table. Where the censored data was not available, I considered different assumptions as described in section 2.12.

2.6 Method 1: "Hunink" - Meta-analysis of failure-time data with adjustment for covariates

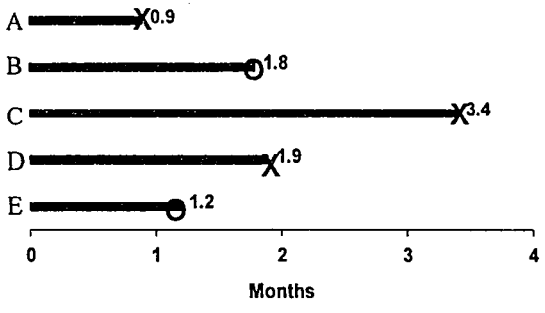
Hunink and Wong[41] present a method for combining survival data from various cohort studies. As a result, it does not preserve randomization. It uses an actuarial life-table approach in order to simplify the data to be extracted from published reports. The steps involved are:

1. Extract and summarize the data from each study into life tables
2. Pool the number of patients at risk and the number of events in each time interval.
3. Calculate the hazard-rate in each interval.
4. Convert the hazard function into an actuarial survival function.

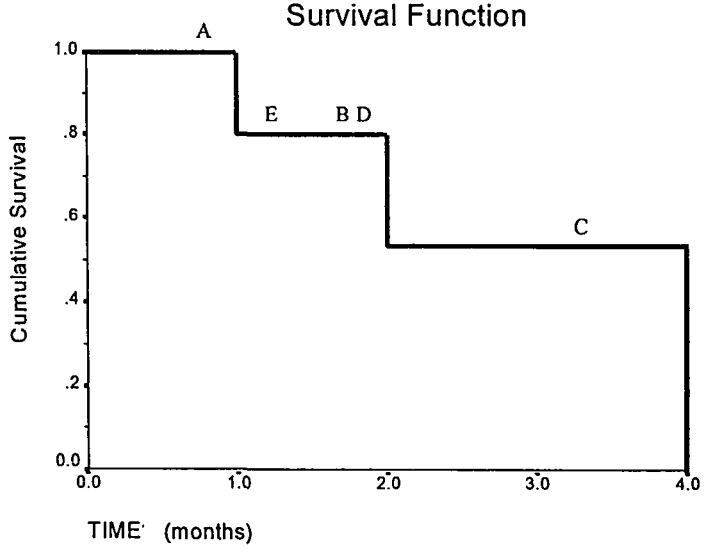
Figure 8. Calculation of the number of events from the proportional survival measured on a graph

Interval	Survival subsequent to the interval (S_i, measured from graph)	Prob. surviving interval ($P_i = S_i/S_{i-1}$)	Censored (c_i, read off of graph)	# alive at start ($r_i = r_{i-1} - c_{i-1} - d_{i-1}$)	# at risk during interval ($n_i = r_i - ½c_i$)	# deaths during interval ($d_i = (1 - P_i)*n_i$)
0	1.0	1.0	0	5	5	0
0 - 1	.8	.8	0	5	5	1
1 - 2	.53	.66	2	4	3	1
2 - 3	.53	1.0	0	1	1	0
3 - 4	0	0	0	1	1	1

A.



B.



(based on the actuarial survival curve generated in Figure 5)

Because it requires generation of a life table, exact information about the timing of censored observations is important for the accuracy of this method.

2.6.1 Estimating the effective sample size in each stratum for unstratified trials

Hunink's method was developed to address the issue that variation across studies is largely due to heterogeneity of the case-mix of the component studies. Combining survival data from studies with different distributions of important prognostic covariates can lead to results that are difficult to interpret. Hunink and Wong suggest that data be extracted separately for important subgroups where possible. One subgroup is chosen as the reference stratum, and its survival function calculated as described above. The hazard-rate ratios of each of the other subgroups are calculated relative to the reference stratum. This allows adjustment for the distribution of covariates before pooling results.

However, studies usually do not present separate survival data for each stratum of interest. If a proportional hazard can be assumed (Appendix 2), it can be used to estimate the stratum-specific survival functions. This requires:

1. knowledge of the initial distribution of subjects in the each stratum in each study;
2. an estimate of the hazard ratio(s) between strata; and
3. an assumption that censoring occurs with equal probability across strata within a study, as publications rarely provide stratum-specific censoring information.

The initial sample size in each stratum for the first interval is determined by the initial distribution of patients between the strata (Figure 9). Censored observations

Figure 9. Hunink's method to stratify survival data

Interval	# at start $r_i = r_{i-1} - c_{i-1} - d_{i-1}$	Censored c_i	Deaths d_i	# at risk $n_i = r_i - 1/2c_i$
0 – 1	100	2	3	99
1 – 2	95	4	9	93
2 - 3	82	...	...	...

⇓

Interval	Stage III				Stage IV			
	r_i	c_i	d_i	n_i	r_i	c_i	d_i	n_i
0 – 1	50	1	1	49.5	50	1	2	49.5
1 – 2	48	2	3	47	47	2	6	46
2 - 3	43	...	...	...	40	...	...	...

This life table for a hypothetical trial of 100 patients is separated into life tables for Stage III and IV strata using Hunink's method. The initial stage distribution was 50 Stage III and 50 Stage IV patients. For this illustration, assume the estimate of the hazard ratio for Stage IV versus Stage III is 2. Censored observations (c_i) are distributed between the two strata according to the effective sample size, and deaths (d_i) are distributed proportional to the hazard ratio.

during the interval, if known, are distributed among strata proportionally to their effective sample sizes in that interval, and are assumed to occur at the midpoint of the interval. The events in the interval are allocated between the two strata according to the hazard ratio (hazard ratios can be obtained from a study reporting a Cox proportional hazards model (see Appendix 2) or using the Mantel-Haenszel method;[40] otherwise, they have to be estimated). In this way, the effective sample size in each stratum for the second interval can be calculated by subtracting the number of events and half of the number of censored observations allocated to that stratum in the first interval from the effective sample size of the first interval. This process can be repeated for each subsequent interval, and stratified life tables constructed.

For my study, the strata of interest were Stage III vs. Stage IV NSCLC patients. I chose stage IV as the reference stratum because it had the largest sample size. Since I had access to the corresponding IPD, I used a Cox proportional hazard analysis to calculate the actual hazard ratio for stage IV versus Stage III patients in my data set. Such analysis would not be possible in the usual application of this method. Because the procedure relies on an assumption of proportional hazards, I tested this assumption graphically by plotting the logarithm of the cumulative hazard functions of the IPD for each stage over time. Proportional hazards means that the functions for Stages III and IV have a constant ratio ($e^{b(x1-x2)}$, where $x1$ and $x2$ are the covariates for Stage III and IV) over time. Hence the two lines should be parallel. I also tested the assumption formally by plotting the beta coefficient for the hazard ratio against time.[63] If the hazards are proportional, this should be a horizontal straight line. Therefore, the null hypothesis is that the slope is 0. Non-proportional hazards occur if the ratio of the baseline hazard functions varies with

time, e.g. if their survival curves cross. An example of the latter might be seen if patients in a treatment group experience cumulative serious toxicity with chemotherapy. As time goes on, the hazard of death in the chemotherapy treated group may increase. If the assumption of proportional hazards is violated over any time period, the method cannot be applied over that interval.

I used this method to distribute data from the eight single arm Phase II studies between stage III and IV subgroups and compared the resultant stage-specific curves to actuarial stage-specific curves calculated with IPD. I used the method to similarly distribute the data of the NSCLCCG meta-analysis by stage to see whether the results of that study might have been different if only one stage group had been considered.

2.7 Method 2: "Dear" - Iterative generalized least squares for meta-analysis of survival data at multiple times

Dear[56] presents a method that allows survival proportions reported at multiple times during a trial to be analyzed together. It only requires extraction of the survival proportions in each time interval and calculation of their standard errors, $(pq/n)^{1/2}$, where p is the proportion surviving, $q = 1-p$, and n is the effective sample size. A multiple linear regression model is made with the survival proportion as the dependent variable, and both between and within study covariates as independent variables: each separate time interval, study, treatment group, and an interaction term for time by treatment. In this way, intervention effects can be evaluated while preserving randomization. This feature is very important because it allows theoretically justified treatment comparisons by only measuring the effect of an intervention within a randomized trial. The other methods

combine patients in each arm from all the trials, thereby creating two cohorts. This leads to patients being compared to others that they had not been randomized against.

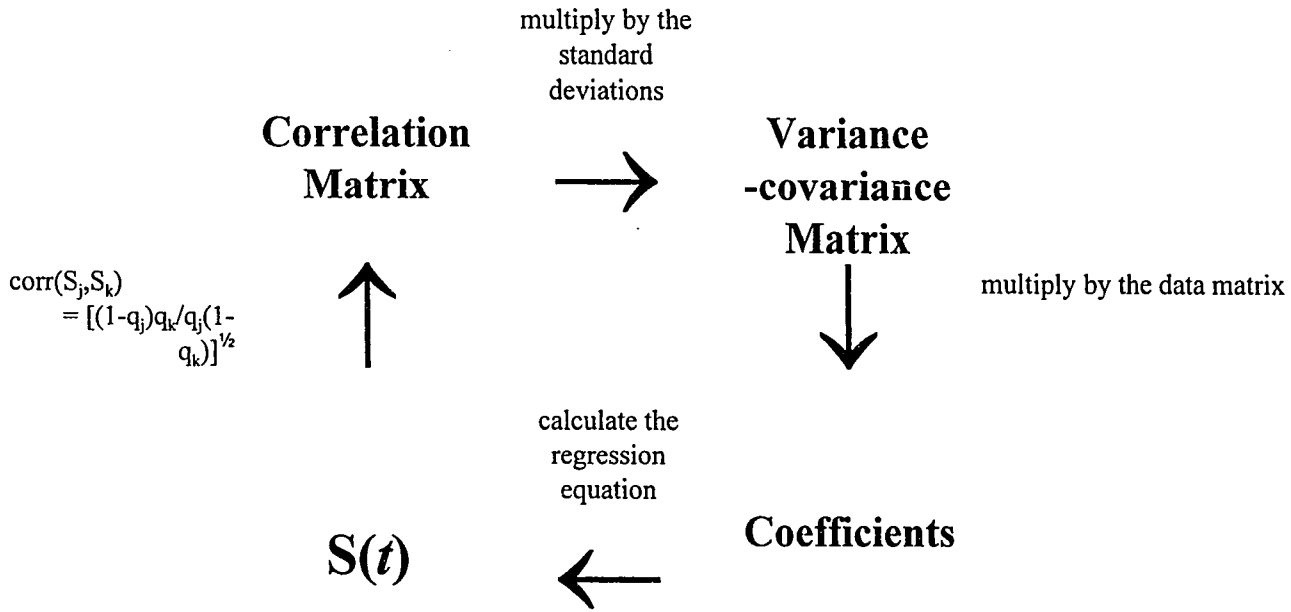
Dear's model can also incorporate both controlled and single arm studies, if desirable, to improve the precision of survival estimates by increasing the sample size. Linear regression yields coefficients for each covariate, allowing pooled survival proportions for each intervention subgroup during each time interval to be calculated from the resulting linear equation.

An iterative generalized least squares method is used to derive parameter estimates for the linear regression model (see Appendix 3). However, in such an analysis each observation is not independent, but consists of serial observations on the same cohort. The survival at time t is related to the survival at time t-1. This violates the assumption of independent observations required for the regression analysis. Therefore, Dear provides a formula to calculate a correlation structure between successive survival proportions that is derived iteratively from their fitted values and incorporated into the model via a variance-covariance matrix (Figure 10). The variance-covariance matrix is multiplied by the data matrix, and the resulting linear regression problem solved by least squares.

2.8 Method 3: "Nelson" - Iteratively reweighted least square analysis

This method[57] assumes that the survival curve takes the form of an exponential decay function for nonlinear regression (Appendix 3), with the contribution of each study at each time point being the covariates. It estimates the survival function by fitting a

Figure 10. Estimation for the Dear method



S_j = the proportion of subjects surviving until time j.
q_j = the probability of surviving up to time j
Time k is the next time point.

mathematical model to ln(S) at each time interval using iteratively reweighted least squares. It is simpler than Dear's method in that it does not consider either between or within study covariates, and does not include a correlation structure. However, each treatment group is analyzed as a separate cohort, so randomization is not preserved.

2.9 Method 4: "Voest" - Log(relative risk)

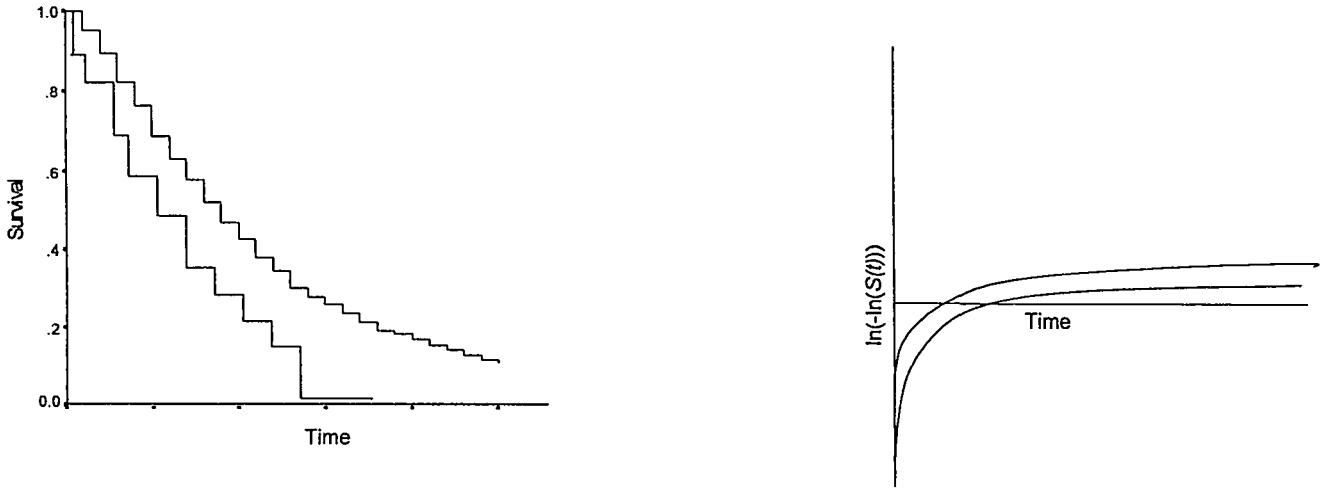
The only data elements required for this method[58] are the proportional survivals in each time interval for each study. To perform this analysis, the ln(–ln($S(t)$)) of the survival functions are calculated and plotted. This transforms the survival stepfunctions into essentially parallel continuous curves (Figure 11). Voest called this summary measure the "log relative risk" (LRR), and used it to describe survival curves in the context of a meta-analysis of prognostic factors in advanced ovarian cancer. Using these curves, an average curve is computed. An inverse-transformation yields a corresponding summary survival curve. This method does not preserve randomization.

2.10 Method 5: "Weighted Voest"

After examining the Voest method, it seemed reasonable to apply the standard meta-analytic technique of using the inverse variance[64] of the survival function to weight each of the transformed survival functions, resulting in a "weighted average" LRR curve which can then be transformed back into a survival function.[20; 26] Like the Voest method, this analysis is attractive because of its relative computational simplicity. However, neither method preserves randomization for making valid treatment comparisons.

Figure 11. ln(-ln(S(*t*))) transformation of a survival stepfunction

A. B.



The difference between two survival stepfunctions (A) depends on the time at which the comparison is made. A ln(-ln) transformation (B) yields continuous, parallel curves from which an average curve can be determined.

2.11 Software

As previously described, CorelDRAW! 3.0 was used to extract the survival function from each trial. I used an Excel spreadsheet (version 7.0) to calculate the life tables from the extracted data, generate simulated survival data, and calculate the survival function from the beta coefficients resulting from the Dear model. I developed procedures in the S-Plus™ statistical program, version 3.3, to perform each of the required manipulations for the Hunink, Voest, and weighted Voest methods. Where possible, the procedures were tested by replicating the examples provided by the authors. I obtained the SAS implementation of the Dear method directly from its author. The nonlinear regression procedure (NLIN) in SAS (version 6.12) was used to develop the Nelson procedure.

2.12 Assessing the reliability of data extraction

The accuracy of manual data extraction from the published graphs was assessed by comparing the resultant proportional survivals with the actual proportional survivals of curves calculated from the IPD. To do this, I took the IPD of the eight single arm Phase II survival curves and extracted the numeric Kaplan-Meier survival estimate at each 4 week time period in each study. These corresponded to the same values read from the graph. I then repeated the analyses with each method using this "computer-extracted" data.

To assess intra-observer reliability, I extracted all data from the eight Phase II studies on two separate occasions several weeks apart. To examine inter-observer variability, a non-expert observer was trained to independently reproduce the data extraction. Intraclass correlations using a two way mixed effects model for intra-observer

reliability, and a two way random effects model for inter-observer reliability were calculated.[65] Additionally, the curves generated from each observer's data were compared.

2.13 Testing assumptions about censoring

The number of subjects lost to follow-up or still alive at the time of analysis is usually described in the text of a paper. Survival curves sometimes indicate censored observations as tick marks on the curve, or present a running tally of the number of patients at risk underneath the x-axis. However, in most cases I did not know exactly when censored patients were censored. The Hunink procedure requires calculation of the number of events in each time interval, which depends on knowledge of the number of censored patients in each interval. As a result, only an estimate of the number of events was possible. The Dear method only incorporates censored observations indirectly in the standard error calculation, and the weighted Voest in the calculation of the variance. The other methods did not require knowledge of censoring for their calculations.

Censoring could not be determined by calculating the change in proportional survival on a line by line basis. Assumptions of random or continuous distributions of censoring throughout the intervals put too many censored observations at the tail of the curve, resulting in time points where there were no patients left at risk while the proportional survival continued to change.

Therefore, I attempted to empirically determine an approach to censoring for the Hunink method (and the Dear method for comparison) by exploring alternative sets of assumptions:[18]

1. Uncensored analysis only: An assumption that subjects who were censored did not contribute to the analysis. The effective sample size became the initial sample size minus the number of observations known to have been censored. this would decrease the influence of trials with high levels of censoring.
2. "median censoring" analysis: An assumption that all censored observations occurred at the point of median follow-up. This assumes that censoring followed a skewed distribution similar to that of the survival function. However, like the random distribution assumption, this often placed too many censored observations late in the trial, resulting in the number at risk becoming 0 while events were still occurring. Because mean survival times tend to be longer than median, an assumption of normally distributed censoring would have worsened this problem.
3. "full sample analysis": An assumption that no subjects were censored during a trial. From a practical point of view, this essentially ignored the issue of censoring. It has been used by other authors with good results.[9; 18]

I derived summary survival functions from the eight single arm Phase II survival curves under each of these sets of assumptions using the Hunink and Dear methods, and compared them with the reference IPD curve. I also examined the effect of having complete knowledge of censoring by extracting the correct distribution of censoring from the IPD and incorporating it into the life tables derived from each survival curve. Finally, I repeated these analyses with simulated trials of larger sample sizes and higher degrees of censoring to further test each set of assumptions under more extreme conditions.

2.14 Assessing the resultant survival curves

I compared the curves resulting from each methodology with that from individual patient data by looking at the maximum discrepancy between the curves. This was done with a two-sample Kolmogorov-Smirnoff goodness of fit test (Appendix 4). The null hypothesis was that each method produced survival curves that were not different from that derived from individual patient data.

The Kolmogorov-Smirnoff test assumes that the data are from mutually independent random samples. However, this was not the case in my comparisons. The same subjects were contributing information for the IPD curve as well as each method's summary curve. Because of this, I could not formally test the hypothesis that the two survival curves were the same. Non-independent samples provide less information than independent ones and may, therefore, lead to false positive results. As a result, I do not provide p-values for these comparisons, but only indicate where the maximum discrepancy is more than would be expected by chance for distributions with these sample sizes if the observations were independent.

2.15 Heterogeneity and jack-knife type analyses

Hunink and Wong recommend comparing the survival curve of each study with the pooled survival curve using the log rank test to assess heterogeneity between studies.[41] They also recommend determining the contribution of each study using a "jack-knife" type method. I did this by repeating the analyses systematically excluding a different study each time, to see if any one trial was inordinately influencing the results.

2.16 Detecting Treatment Differences

The reproduction of the NSCLCCG meta-analysis allowed me to test the null hypothesis that there was no difference between the chemotherapy and best supportive care arms. I made life tables from the summary curves from each method, calculated the hazard ratios, and did log rank tests to compare the treatment arms (Appendix 5). I made comparisons of the six-month and one-year survival proportions, as well as the survival at the median survival times for each arm. I also made comparisons between treatment arms with the stage-specific data derived from Hunink's method.

3. Results

3.1 Reliability of data extraction

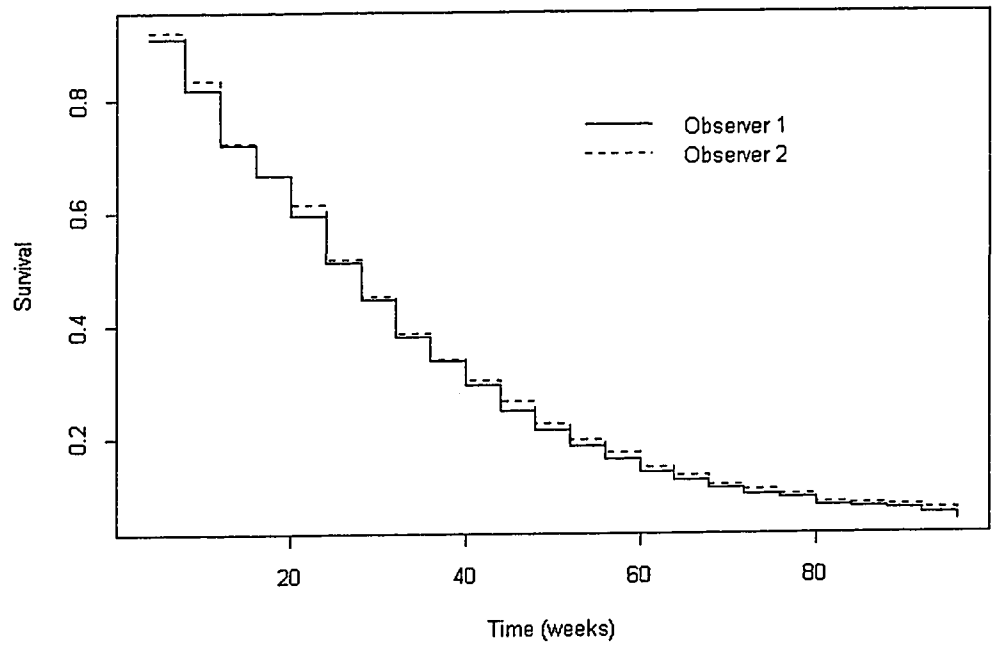
There was no obvious distortion in the reproductions of any of the published figures after scanning and enlargement. Occasionally broken lines required interpolation between the dashes, or a reading became somewhat subjective when lines crossed, overlapped, or when a time point occurred on the shoulder of a step. There was no difference between curves calculated with the proportional survival data I extracted and those made with data calculated directly from the IPD. The intra-class correlation for both intra- and inter-observer reliability assessments was 0.99 (see Appendix 6), and the survival curves produced from the two observer's data were similar (Figure 12).

3.2 Censoring

Only one of the eight single arm Phase II survival curves indicated censored observations on the survival curve. Similarly, only one of the seven trials in the NSCLCCG meta-analysis displayed the time of censoring on the survival figure. None indicated the number of patients still being followed along the bottom of the survival graph.

Table 3 shows the results of my empiric examination of the effect of different censoring assumptions on the results of the Hunink and Dear procedures, applied to the single arm Phase II studies and limited to a follow-up time when all trials were contributing. Whether analyses were carried out under assumptions of "uncensored

Figure 12. The effect of inter-observer variability on the summary survival curve



Curves from data extracted independently by two observers (using Hunink's method to reproduce the chemotherapy arm of the NSCLCCG meta-analysis).

Table 3. The effect of censoring at the start, time of median survival, or at the end of each trial on the results of the Hunink and Dear procedures

Method	Maximum discrepancy (compared with IPD)
Hunink	
"uncensored only"	0.046
"median censoring"	0.023
"full sample analysis"	0.021
Dear	
"uncensored only"	0.026
"median censoring"	0.026
"full sample analysis"	0.018

Note: no discrepancy was more than would be expected by chance if independent samples were being compared. The critical value was a maximum deviation of 0.063 for the "median censoring" and "full sample" analyses, and 0.069 for the "uncensored only" analysis.

IPD = individual patient data

"uncensored only" = removes censored patients from the sample at the start of the trial

"median censoring" = removes censored patients from the effective sample at the time of median survival

"full sample analysis" = ignores censoring during a trial

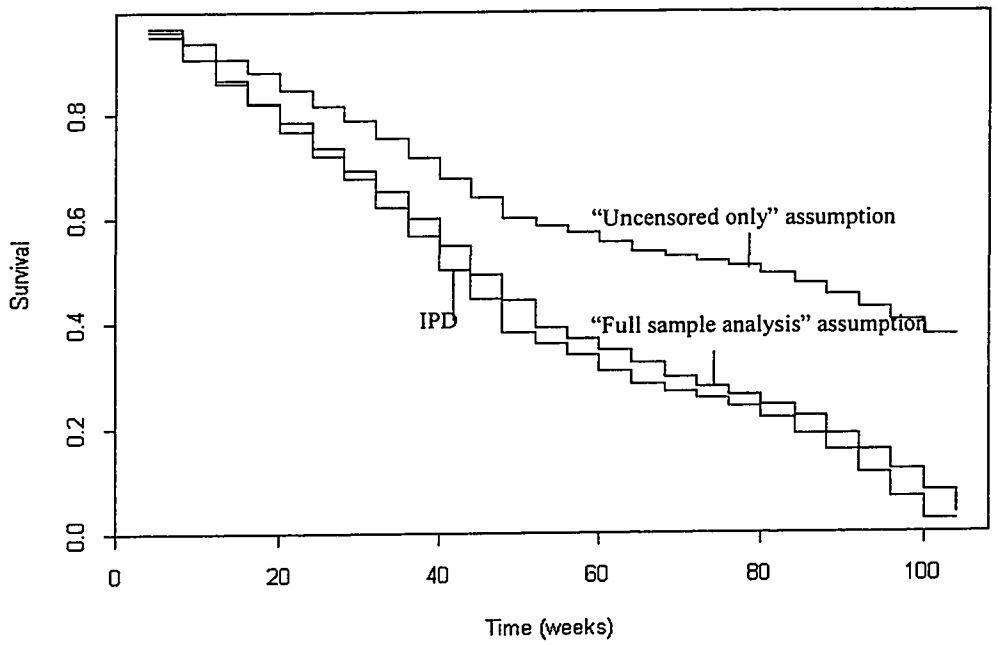
analysis only", "median censoring", or "full sample analysis" had little effect on the results. However, there appeared to be a trend towards a smaller maximum deviation between the curves with "full sample analysis", when all censoring was assumed to occur at the end of the trials. When the exact times of censoring were incorporated into the Hunink analysis, the maximum discrepancy decreased from 0.021 to 0.006. This had minimal effect on the Dear analysis.

To create an extreme example I simulated eight trials (2500 patients in total) with high rates of censoring (range 8.6% to 53%). I only examined the effects of censoring patients at the start ("uncensored analysis only") and at the end of the trial ("full sample analysis") because "median censoring" often resulted in no patients left alive beyond that median time point. Again, the Dear method was affected little by the different assumptions. As Figure 13 illustrates, the Hunink method deviated significantly from the IPD curve when censored observations were removed at the beginning. As a result, censoring was ignored in all subsequent analyses where detailed censoring information was not provided on the survival curve.

3.3 Accuracy

Figure 14 shows the summary survival curves generated with each method compared to the IPD curve. All were quite accurate in the first part of the curves, but became much less accurate in their tails. The shortest trial ended at 76 weeks follow-up, and that appears to be the time that most curves start to noticeably deviate from the IPD curve. In fact, the Dear, Nelson, Voest, and weighted Voest procedures all allowed the

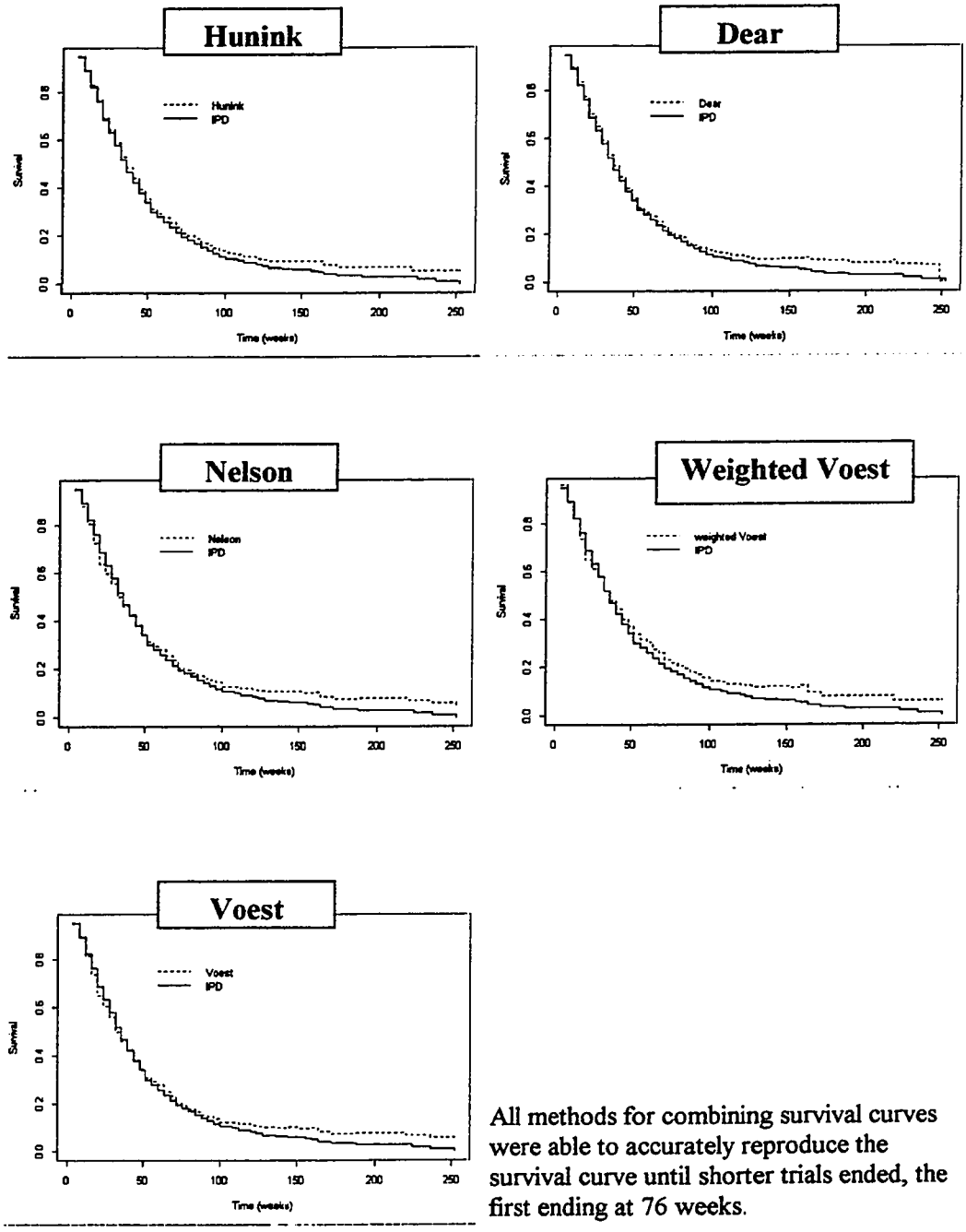
Figure 13. The effect of different censoring assumptions on simulated trials with high levels of censoring



Data from eight simulated trials (n = 2500) with 8.6% to 53 % censored observations were combined into an actuarial survival curve of individual patient data (IPD). Data extracted from the survival curves of each trial were combined using Hunink's method under two sets of assumptions:

- "Uncensored only," which removes censored patients from the sample size at the start of the trial
- "Full sample analysis," which ignores censoring during a trial

Figure 14. Comparison of each method with IPD-derived summary survival curves from eight single-arm survival curves



All methods for combining survival curves were able to accurately reproduce the survival curve until shorter trials ended, the first ending at 76 weeks.

Table 4. The maximum difference between survival curves generated with each method and IPD from eight single arm Phase II curves limited to 76 weeks

Method	Maximum discrepancy (compared with IPD)
Hunink	0.021
Dear	0.018
Nelson	0.047
Weighted Voest	0.021
Voest	0.037

Note: no discrepancy was more than would be expected by chance if independent samples were being compared. The critical value for these analyses was 0.063.

IPD = individual patient data

survival to actually rise after a trial ended, although these were generally not the points of maximum discrepancy. Table 4 shows the maximum deviation of each method's curve from the IPD curve with the follow-up restricted to the first 76 weeks. None was more than would be expected by chance alone.

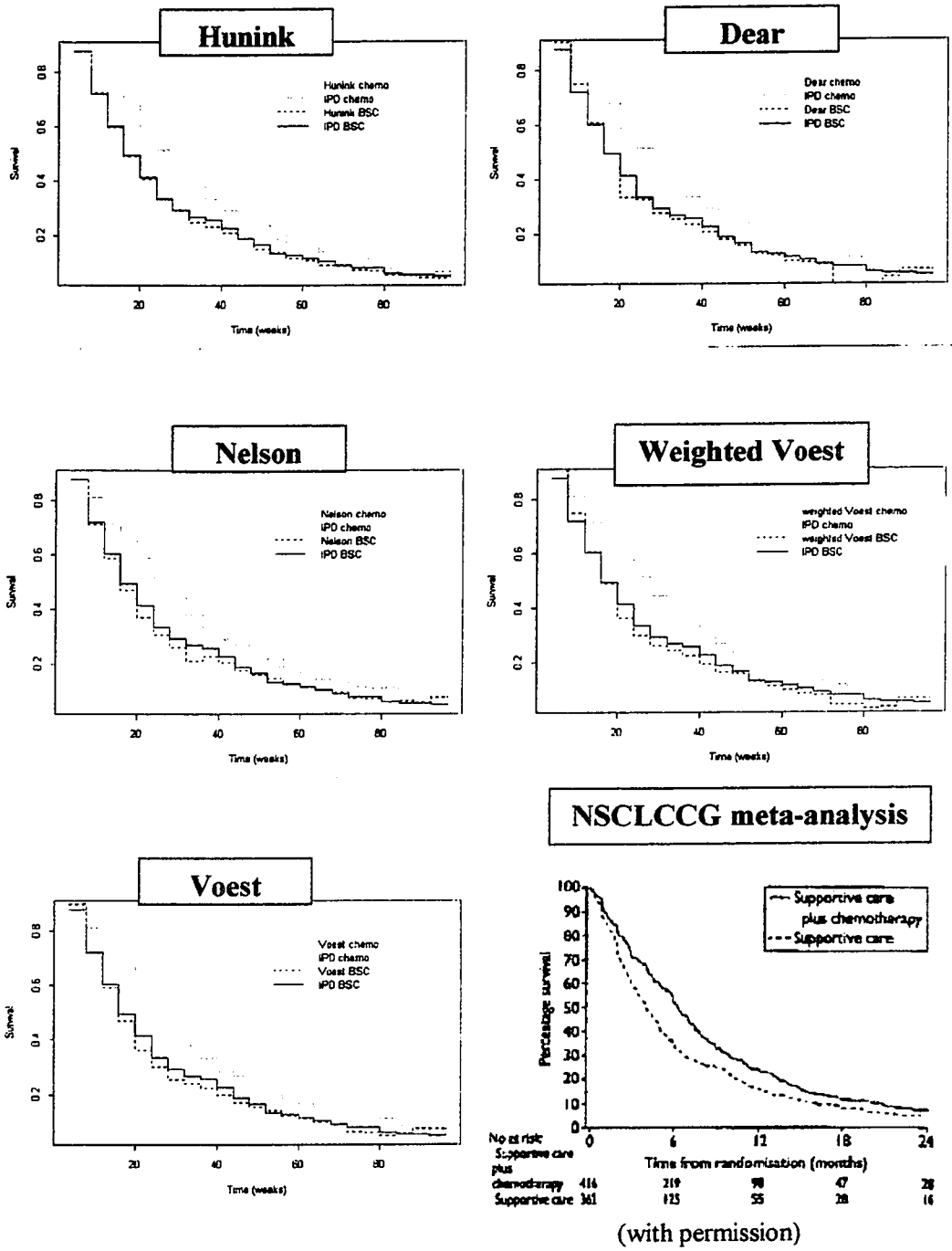
Figure 15 shows the replication of the NSCLCCG meta-analysis. Again, all methods were reasonably accurate. The Hunink procedure had the smallest deviation for both arms. The other procedures once more calculated instances of increasing survival after shorter trials ended. As opposed to the overestimation of the curves seen with the eight single arm Phase II studies, here the methods tended to underestimate survival.

Figure 16 compares the accuracy of each method when used to combine the eight single arm Phase II studies, to replicate the NSCLCCG meta-analysis, and to combine a series of computer-simulated trials. Although the Kolmogorov-Smirnoff test was occasionally able to detect a curve that deviated more than expected from the IPD curve, no consistently superior method could be identified. All were quite accurate, with the greatest maximum discrepancy (11%) in the NSCLCCG meta-analysis.

3.3.1 Hunink's method to split strata

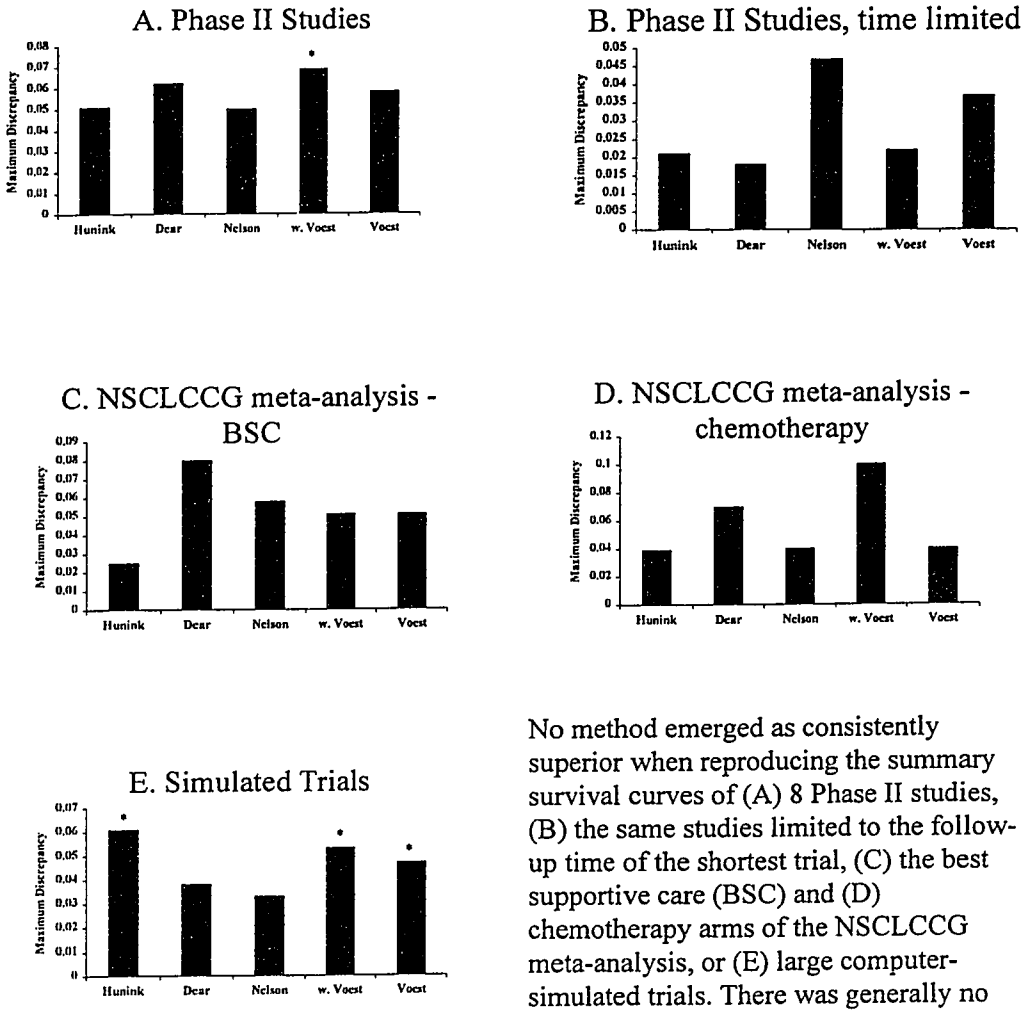
I initially tested the assumption of proportional hazards between stage III and stage IV patients in the eight single arm Phase II studies graphically (Figure 17). The lines were mostly parallel. However, they started to diverge after about 100 weeks, suggesting that the proportional hazards assumption did not hold over time. This was confirmed on statistical testing of the proportional hazards assumption.[63] The hazard ratio for Stage IV versus Stage III was 1.35 (95% CI: 1.16, 1.57) as calculated from IPD.

Figure 15. Reproduction of the NSCLCCG meta-analysis with each method



(with permission)

Figure 16. Comparison of each method's ability to accurately reproduce summary survival curves

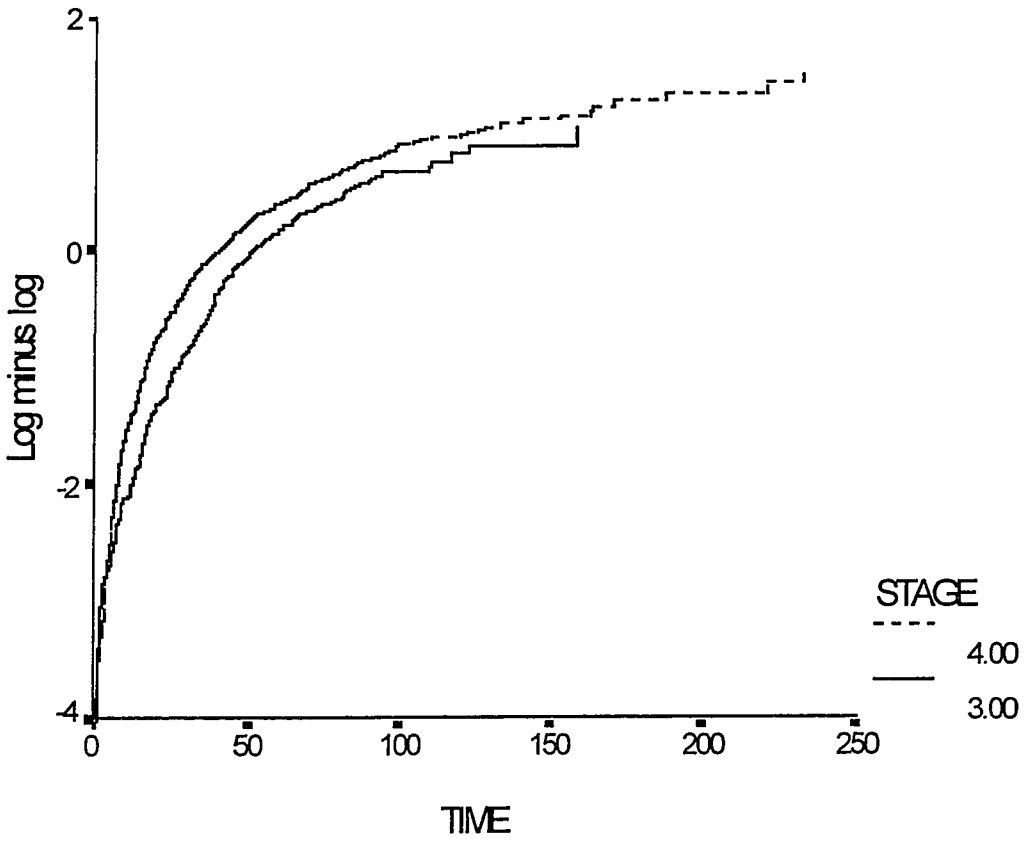


No method emerged as consistently superior when reproducing the summary survival curves of (A) 8 Phase II studies, (B) the same studies limited to the follow-up time of the shortest trial, (C) the best supportive care (BSC) and (D) chemotherapy arms of the NSCLCCG meta-analysis, or (E) large computer-simulated trials. There was generally no statistical difference between the curves produced by each method and those derived from individual patient data.

* Maximum discrepancy more than would be expected by chance if independent samples were being compared.

NSCLCCG = Non-Small Cell Lung Cancer Collaborative Group

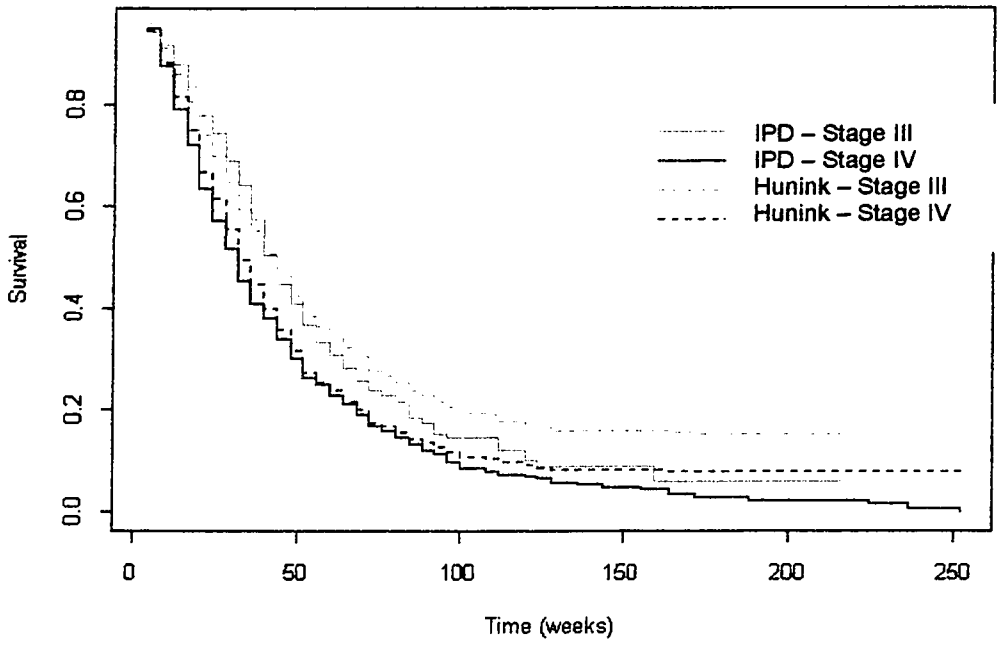
Figure 17. ***ln(-ln(S(t)))*** **Plot of Stage III versus Stage IV patients**



Stage III - - -
Stage IV ——

The hazard functions appear to be parallel over the first 100 weeks, indicating that the assumption of proportional hazards is correct over this time interval. However, they converge at the end, suggesting that the proportional hazards assumption is later violated.

Figure 18. Hunink's method to stratify survival data between stage III and IV in eight single-arm trials



The method appears to produce accurate curves initially, but becomes inaccurate later when the proportional hazards assumption is violated.

Using Hunink's method to stratify results into stages III and IV, I obtained curves that were similar to those generated by the IPD only over the first 100 weeks (Figure 18).

When I applied Hunink's method to the NSCLCCG meta-analysis I found that it predicted six month survivals of 41.3% (95% CI: 39.0, 43.8) for the BSC arm and 59.4% (95% CI: 58.1, 60.8) in the chemotherapy arm if only Stage III patients were considered (Table 5). Stage IV patients had 6 month survivals of 30.3% (95% CI: 27.3, 33.6) and 49.5% (95% CI: 47.6, 51.5) in the BSC and chemotherapy arms, respectively. The six-month survival proportions were significantly different between treatments and between stage groups. As expected, because the method relies on an assumption of proportional hazards to split the two groups, the hazard ratio for chemotherapy versus BSC was the same for both stages at 0.75. However, this was different from that seen when both stages were analyzed together (0.83). Even limiting the analysis to the first half of follow-up, the stratified hazard ratio was less than the combined.

3.4 Detecting treatment differences: comparison of chemotherapy vs. best supportive care by replicating the NSCLCCG meta-analysis

All methods detected a statistically significant difference between the chemotherapy and best supportive care arms by log-rank testing ($p < 0.001$). The hazard ratio between the two arms ranged from 0.74 (Dear) to 0.83 (Hunink) (Table 6). These results compare with 0.73 found in the NSCLCCG meta-analysis, with $p < 0.0001$ on log rank testing.

Table 6 also shows comparisons of six-month and one year survival by each of the methods. The proportional survivals were similar with all methods. All were able to

Table 5. Percent survival at six months and one year in the NSCLCCG meta-analysis with the data split between Stages III and IV by Hunink's method

	6 month survival (95% CI)		1 year survival (95% CI)		
Stage	BSC	chemo	BSC	chemo	hazard ratio
III	41.3 (39.0, 43.8)	59.4 (58.1, 60.8)	17.9 (16.0, 20.0)	26.8 (24.7, 29.15)	.75
IV	30.3 (27.3, 33.6)	49.5 (47.6, 51.5)	9.8 (7.7, 12.4)	16.9 (14.4, 19.9)	.75
combined	33.6 (30.9, 36.5)	51.1 (49.3, 53.0)	13.9 (12.1, 16.1)	18.4 (15.9, 21.3)	0.83

Table 6. Comparison of six month and one year survival rates in the NSCLCCG meta-analysis replicated with each method

	Six month survival		One year survival		
Method	BSC % (95% CI)	chemo % (95% CI)	BSC % (95% CI)	chemo % (95% CI)	hazard ratio
Hunink	33.6 (30.9, 36.5)	51.1 (49.3, 53.0)	13.9 (12.1, 16.1)	18.4 (15.9, 21.3)	0.83
Dear	32.5 (31.4, 33.7)	51.5 (49.7, 53.4)	12.9 (9.9, 16.7)	16.2 (12.8, 20.3)	0.74
Nelson	30.8 (28.1, 33.8)	50.4 (48.6, 52.3)	14.7 (12.8, 17.0)	19.0 (16.7, 21.7)	0.78
Weighted Voest	29.9 (27.2, 33.2)	51.4 (49.6, 53.4)	13.0 (10.2, 16.6)	14.7 (12.7, 16.8)	0.78
Voest	30.1 (27.4, 33.2)	50.8 (48.9, 52.7)	14.2 (12.9, 16.7)	18.0 (15.5, 21.0)	0.78

chemo = chemotherapy
BSC = best supportive care

The six month survival for the NSCLCCG meta-analysis was 33.5% for BSC and 51.8% for chemotherapy. The one year survival was 13.4% for BSC and 22.0% for chemotherapy. The hazard ratio in the NSCLCCG meta-analysis was 0.73

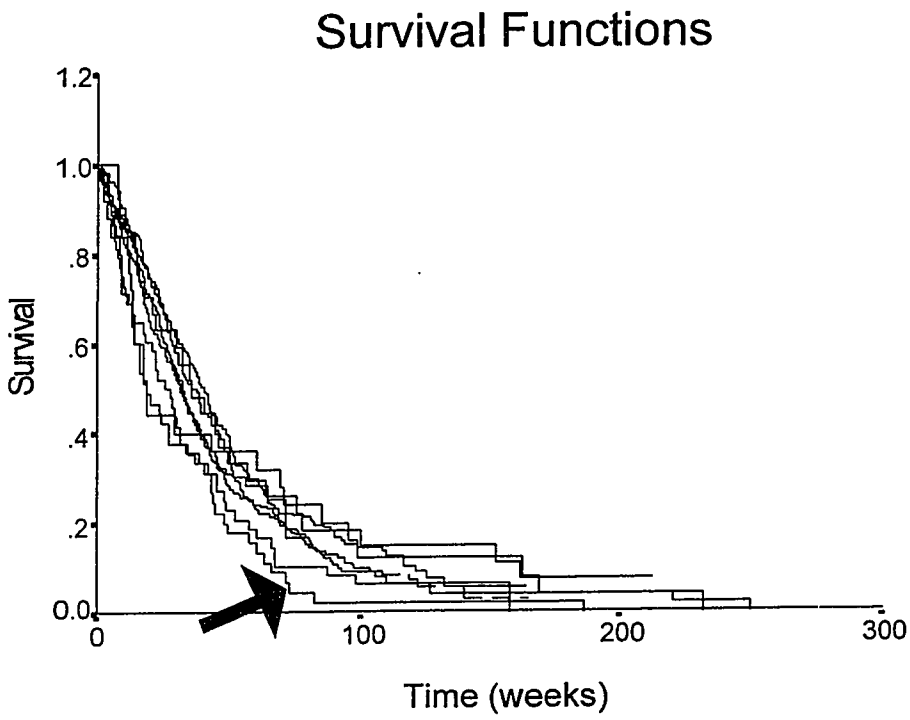
distinguish between the two treatments at six months, but none could at one year. This is similar to the findings of the meta-analyses done from the published literature. The NSCLCCG meta-analysis from IPD did not report differences at six months and one year. All methods arrived at median survivals of 28 weeks for the chemotherapy arm and 16 weeks for the BSC arm, a difference of 12 weeks. The NSCLCCG meta-analysis had found a median survival improvement of 6 weeks with chemotherapy.

3.5 Heterogeneity and jack-knife type analysis

In Figure 19, the curve indicated by the pointer was found to describe a significantly inferior survival compared to the rest. This corresponds to trial 3 in Table 1. It was a small trial with only Stage IV patients. When I did the jack-knife type analysis (Figure 20), removal of this trial had an important effect on the analysis. The Nelson and Voest procedures had had the greatest maximum discrepancy, but when Trial 3 was removed they became much more accurate and yielded results very similar to those seen with the other methods.

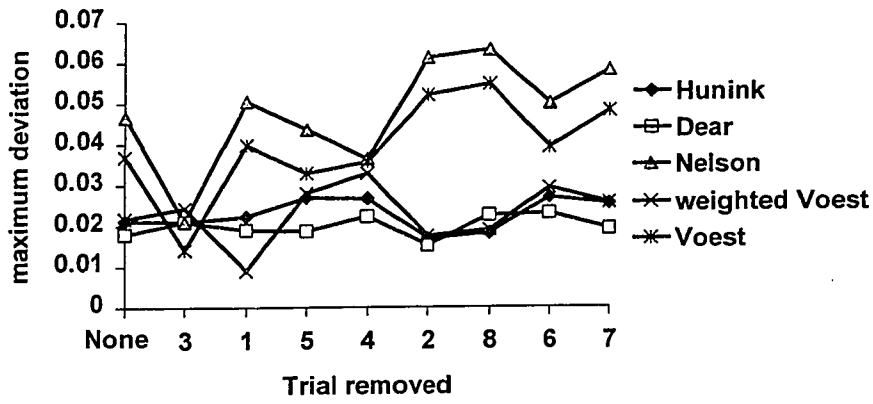
Similarly, the NSCLCCG[31] had assessed heterogeneity in their meta-analysis of IPD and found that a small trial (n = 46) by Quiox et al.[47] had a significantly inferior BSC arm compared to the other trials. In fact, all patients in the BSC arm of this trial were dead before 32 weeks. Therefore, I tested the effect of removing this trial from the analysis. As can be seen in Table 7, removing the Quiox trial did cause the maximum discrepancy between the BSC curves to decrease, though less dramatically than in the previous analysis.

Figure 19. Testing for heterogeneity in the eight single-arm Phase II curves



The curve with the pointer is heterogeneous compared to the others.

Figure 20. Jack-knife type analysis of the eight single-arm Phase II curves



Trials were removed in order of increasing size. When Trial 3, the smallest and most heterogeneous one, was removed, all methods were able to produce curves with near equal accuracy.

Table 7. Maximum discrepancy between BSC survival curves generated with each method compared to IPD when replicating the NSCLCCG meta-analysis, with and without the trial by Quiox et al.

Method	All Trials	Without Quiox [51]
Hunink	0.025	0.017
Dear	0.081	0.062
Nelson	0.058	0.029
Weighted Voest	0.051	0.047
Voest	0.051	0.033

Note: no discrepancy was more than would be expected by chance if independent samples were being compared. The critical value for these analyses was a maximum deviation of 0.104

The Quiox trial was removed because it was heterogeneous compared to the other trials.

BSC = best supportive care

IPD = individual patient data

NSCLCCG = Non-Small Cell Lung Cancer Collaborative Group

4. Discussion

4.1 Reliability

This study has shown the feasibility of extracting reliable, accurate data from published survival curves. As others have found,[9] scanning the curves and enlarging them facilitated this. The innovation of using the co-ordinate system of CorelDRAW! to measure survival proportions made data extraction quite objective, as indicated by the small inter- and intra-observer variability.

4.2 Accuracy of the methodologies

In these analyses, all methods created reasonably accurate summary survival curves, especially when limited to a time frame in which all studies were contributing data. Furthermore, all methods allow incorporation of both controlled and single arm or retrospective data if desired to obtain an accurate description of the survival experience of a group of patients (e.g. for disease modelling).

No method was consistently more accurate than the others. The Kolmogorov-Smirnoff test was occasionally able to detect a curve that differed from the IPD curve. However, because of censoring and because the assumption of independent samples was violated, these comparisons were actually based on less information than the test assumed, so there was the potential for finding differences when none existed. Furthermore, the multiple comparisons made increases the risk of Type I errors. However, there did appear to be situations in which some methods performed better than others depending on the characteristics of the trials being combined.

4.2.1 Censoring

Because it requires extrapolation of the number of events in a time interval, the Hunink procedure is especially susceptible to problems when a high proportion of subjects are censored. Incorporating true censoring times into the data improved the accuracy of the method. Unfortunately, this information is rarely available in publications.[20] Even when the time of censoring is indicated directly on a survival curve, there is generally no way of knowing whether more than one patient was censored at any one time.

In the absence of detailed information about censoring, the analysis was found to be most accurate if censoring was ignored. Fenn et al.[18] have made similar observations. They explain that when censoring is ignored, bias is only created in the time intervals in which censoring took place but is assumed not to have occurred. Making artificial assumptions about censoring may create bias both in the interval in which the censored observation was omitted, as well as the interval to which a censored observation is incorrectly attributed.[26] Furthermore, censoring often increases towards the end of a survival curve.[18] As a result, the effect of any incorrect assumption about censoring attributed to the early part of a curve is felt throughout the distribution, while ignoring the issue until the end more closely approximates the actual time of most censoring.

Other authors have observed the accuracy of survival estimates to decrease towards the end of a trial[26] and to be worsened with increasing rates of censoring.[42] This can lead to either an overestimation or underestimation of survival. For example, the methods all overestimated the curve when combining the eight single arm Phase II

studies, but tended to underestimate it in the reproduction of the NSCLCCG meta-analysis. This indicates that the data determines the direction of error. When censoring is ignored, trials that have a high degree of censoring gain increasingly inappropriate weight as time goes on.[24]

Other authors have also noted inaccuracy in estimating the effect size and direction of an intervention when using published data in meta-analysis.[66 18; 27] However, in situations where the average proportion of censored observations is not large, censoring may not pose a serious problem.[9] Hunink and Wong[41] found their results on patency rates after femoropopliteal angioplasty were relatively insensitive to alternative assumptions about censoring probabilities. Dear[56] reports a simulation study with 10^5 trials and variable rates of censoring. He found survival estimates were within 1% of the actual estimate with average censoring of 6%, and suggests that for most purposes, censoring can be ignored with his method.

4.2.2 Study-termination censoring

Not all trials have the same duration of follow-up. Combined curves that extend beyond the duration of the shortest trial might introduce bias, because the shorter trials no longer contribute data to the tail of the curve. The tail of the curve will be more heavily influenced by the longer trials, regardless of their precision (e.g. sample size) or heterogeneity. Furthermore, if these longer trials have different distributions of covariates, they may not describe the survival experiences that would be seen if the shorter trials had had longer follow-up. Therefore, all methods, including IPD, may be biased in their estimation of the tail of the survival curve. Abel and Edler have recognized

this problem as being particularly important when the treatment effect is time dependent and if the number of subjects at risk changes markedly.[67] Because of these factors, analyses should be restricted to a time frame in which all studies are contributing data. The other option would be to exclude shorter trials. However, this could also result in unacceptable bias if the length of a study was not independent of the survival experience of its subjects.

Hunink's procedure appears to handle study-termination censoring better than the other methods because it recalculates the number of subjects at risk in every time interval. The other methods rely primarily on combining the proportional survivals. As a result, if a trial ends when its proportional survival is less than the summary survival calculated by one of these methods, the summary survival estimate can actually go up in the next interval.

The impact of censoring and varying lengths of follow-up were illustrated in the replication of the NSCLCCG meta-analysis. The Hunink procedure performed well because there were very few censored patients in the constituent trials. Most of the other procedures performed worse in the BSC arm because several of the constituent studies had BSC arms in which of all of the patients died quite quickly.

4.2.3 Heterogeneity

Heterogeneity of the constituent trials also emerged as a potentially important factor in the accuracy of some of the methods. Meta-analyses in oncology commonly combine trials with different chemotherapy combinations in similar patients.[31;34;35] As a result, clinical heterogeneity is likely.[16]

The Hunink, Dear, and weighted Voest methods all assume a fixed effects model, in which the inference is conditional on the specific studies included. Therefore, it is only concerned with within-study variation. This differs from a random effects model, which would assume that the studies being combined were a random sample of all potential studies, necessitating consideration of the heterogeneity between studies as well. When studies are homogeneous, both models weight them according to sample size and yield essentially identical results. However, if studies are heterogeneous the results may be different since the random-effects model takes the between-study variation into account while the fixed-effects model does not.

None of the methods in this study used a random effects model. The Nelson and Voest procedures, because they do not incorporate a measure of the precision of the component trials, are sensitive to both within-study and between-study variation. Therefore, they are especially prone to being unduly affected by a small trial with an extreme result. This explains the increased accuracy of these two methods when the heterogeneous trials in both the analysis of the single arm Phase II studies and the reproduction of the NSCLCCG meta-analysis were removed.

4.3 Stratifying survival curves

Hunink's method to stratify survival data yielded its most accurate results when it was limited to a time interval over which the proportional hazards assumption held. The hazard plot indicated that the assumption of proportional hazards between the two strata was violated after longer follow-up. This may be because patients with stage IV NSCLC will all eventually die of lung cancer, while some patients with stage III will be cured.

With time, the group with some cures will have its cumulative hazard plateau while the other group continues to experience events.[68] Unfortunately, the hazard plot required access to IPD, so the assumption of proportional hazards could not usually be verified in a real application of the method.

Even when limited to the time interval in which the proportional hazard assumption appeared to hold, the hazard ratio between treatment arms for the stratified analyses of the NSCLCCG meta-analysis were still different from the combined results. This supports the notion that stage III and IV patients should be reported separately. However, it may also be caused by the assumption that censoring occurs with equal probability across strata being incorrect. Stage III patients are generally healthier than Stage IV patients. As outlined previously, this may lead to either increased or decreased censoring. Furthermore, this method is still susceptible to the same problems with unknown censoring times and study termination described above, and these effects might be magnified by the smaller sample size in each subgroup.

The other important impediment to using Hunink's method is its reliance on knowing the stratum-specific hazard-rate ratios. I was able to calculate the hazard ratio directly from the IPD. However, these data may not be readily available and would often have to be estimated.[40] Any biases would be perpetuated throughout the analysis.

4.4 Hypothesis testing

If the goal of combining survival curves is to compare treatments, patients within a treatment group of a given study should only be directly compared with patients of the other treatment groups in that study.[69] Although not significantly more accurate than the

other methods, the Dear procedure does have the theoretical advantage of being the only one that preserves randomization in this way. The other methods essentially create two cohorts of patients. If there are inequalities in the distribution of known or unknown prognostic features, these could be perpetuated and bias the result.

Despite these concerns, all methods were able to produce summary curves that could differentiate the two arms of the NSCLCCG meta-analysis by log-rank testing with similar hazard ratios. They were also able to provide estimates to compare survival at different points in time, although with less power. The reproduction of the NSCLCCG meta-analysis could distinguish chemotherapy from BSC at six months but not at one year, illustrating the importance of selecting the correct time point for such analyses. All four of the previously reported meta-analyses of chemotherapy in NSCLC had found a similar result.[7-9;31] Because of the lack of precision inherent in grouping events into four-week blocks, comparison of median survivals between treatment arms was not really possible. The median survivals of 28 and 16 weeks in the chemotherapy and BSC arms, respectively, could have been due to a difference in median survival ranging from 8 (24 and 16 weeks, respectively) to 16 weeks (28 and 12 weeks, respectively). However, this is only a technical limitation. If smaller time intervals had been chosen, such as one-week periods, comparison of median survival may have been rational.

4.5 Other sources of error

A potential source of error in the replication of the NSCLCCG meta-analysis came from having a less than ideal "IPD curve" with which to compare each method's summary curve. Even though the meta-analysis used IPD, I did not have access to it. The

reference curve was extracted from the published summary curve in the meta-analysis. However, the analysis of the accuracy and reliability of data extraction suggests that this is a small source of error.

Another potential error in the replication of the NSCLCCG meta-analysis is that the meta-analysis actually contained *updated* IPD. Therefore, the reference curve did not contain exactly the same data as the curves I was trying to combine. However, if censoring was truly non-informative in each of the trials, the survival experience should not change with longer follow-up. The fact that each method was still able to generate curves that were statistically similar to the extracted curve from the meta-analysis suggests that this assumption was likely true.

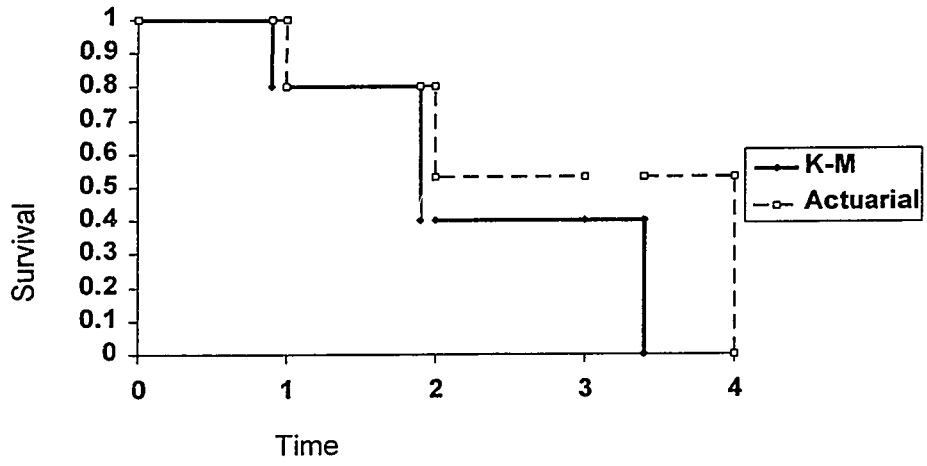
Error could occur if different methods had been used to generate the original published survival curves. The eight single arm Phase II curves were all Kaplan-Meier estimates. However, the NSCLCCG meta-analysis contained two trials in which the results were presented as actuarial survival curves. While this actually simplified data extraction, Kaplan-Meier and actuarial analyses of the same data can provide slightly different survival estimates at the same point in time (Figure 21). Furthermore, it is unclear whether these methods would be applicable to a published curve adjusted with a Cox proportional hazard model. Though the effect of these differences is likely to be small, they may contribute to the overall inaccuracy of combined results.

4.6 Summary and Recommendations

Overall, I prefer the Dear method. It was consistently as accurate as the other methods and preserves randomization, allowing theoretically justified comparisons of

Figure 21. Comparison of actuarial and Kaplan-Meier curves

(based on the data in Figures 5 and 6)



intervention groups. Moreover, it does not require extraordinary detail in the presentation of survival data in published reports (an important feature for journal editors). Because it, like the others, is prone to inaccuracy in the tail of the curve after shorter trials terminate, it should only be applied to the time period in which all component studies are contributing data. Its main disadvantage is that it is computationally the most intense method.

All methods were able to produce reasonably accurate survival curves, with the following caveats:

- If there is a high degree of censoring and you do not have detailed information about when censoring occurred, do not use the Hunink method.
- If it is important to estimate the survival beyond the follow-up of the shortest trial, the Hunink method may be more accurate than the others.
- If heterogeneous trials are to be included, do not use the Nelson or Voest methods.
- Hunink's method to stratify survival data must be used with caution. Error can come from inaccurate estimates of the hazard ratio between subgroups, and incorrect assumptions of proportional hazards and equally distributed censoring.

4.7 Future directions

Computer simulated experiments could further define the roles of each of these methods. This study primarily looked at combining the survival curves of advanced NSCLC patients. It is possible that different survival characteristics, such as a disease

with fewer deaths over a longer follow-up, could have different effects on the performance of some of the methods.

Also, a Kaplan-Meier-type method of combining survival curves would be useful because it would be able to preserve more information from each trial. For example, it would allow comparison of median survival times, which are often of interest to oncologists. Because of the accuracy and reproducibility of the method of enlarging the curves and extracting the data on computer, it would be possible to decrease the unit of time further, e.g. from weeks to days. In fact, each step-point on most Kaplan-Meier curves could be extracted relatively easily. I have developed a procedure in S-Plus that can fill in data for the other time intervals. Combining these data with any of the five methods examined here would produce a summary curve approximating a Kaplan-Meier curve. Computing power would be the only limiting factor in such an analysis.

References

1. Cartei G, Cartei F, Cantone A, Causarano D, Genco G, Tobaldin A. Cisplatin-cyclophosphamide-mitomycin combination chemotherapy with supportive care versus supportive care alone for treatment of metastatic non-small cell lung cancer. J Natl Cancer Inst 1993; 85:794-800.

2. Altman DG. Statistics in medical journals: developments in the 1980s. Statistics in Medicine 1991; 10:1897-913.

3. Saunders M, Dische S, Barrett A, Harvey A, Gibson D, Parmar M. Continuous hyperfractionated accelerated radiotherapy (CHART) versus conventional radiotherapy in non-small-cell lung cancer: a randomised multicentre trial. Lancet 1997; 350:161-5.

4. Omenn G, Goodman GE, Thornquist MD, et al. Effects of a combination of beta carotene and vitamin A on lung cancer and cardiovascular disease. N Eng J Med 1996; 334:1150-5.

5. Rosell R, Gomez-Codina J, Camps C, et al. A randomized trial comparing preoperative chemotherapy plus surgery with surgery alone in patients with non-small-cell lung cancer. N Eng J Med 1994; 330:153-8.

6. Berlin JA, Laird NM, Sacks HS, Chalmers T. A comparison of statistical methods for combining event rates from clinical trials. Statistics in Medicine 1989; 8:141-51.

7. Souquet PJ, Chauvin F, Boissel JP, et al. Polychemotherapy in advanced non small cell lung cancer: a meta-analysis. Lancet 1993; 342:19-21.

8. Marino P, Pampallona S, Preatoni A, Cantoni A, Invernizzi F. Chemotherapy vs. Supportive Care in Advanced Non-Small-Cell Lung Cancer: Results of a Meta-analysis of the Literature. Chest 1994; 106:861-5.

9. Grilli R, Oxman AD, Julian JA. Chemotherapy for Advanced Non-Small-Cell Lung Cancer: How Much Benefit Is Enough? J Clin Oncol 1993; 11:1866-72.

10. Leung K-M, Elashoff RM, Abdelmonem AA. Censoring issues in survival analysis. Annu Rev Public Health 1997; 18:83-104.

11. Clarke MJ, Stewart LA. Systematic reviews of randomized controlled trials: the need for complete data. J Evaluation in Clin Practice 1995; 1:119-26.

12. Thatcher N, Ranson M, Lee M, Niven R, Anderson H. Chemotherapy in non-small cell lung cancer. Annals of Oncology 1995; 6:S83-S95

13. Clarke MJ, Stewart LA. Obtaining data from randomised controlled trials: how much do we need for reliable and informative meta-analysis. BMJ 1994; 309:1007-10.

14. Beck JR, Pauker SG, Gottlieb JE, Klein K, Kassirer JP. A convenient approximation of life expectancy (The "DEALE"). Am J Med 1992; 73:889-97.

15. Schrag D, Kuntz KM, Garber JE, Weeks JC. Decision analysis - effects of prophylactic mastectomy and oophorectomy on life expectancy among women with BRCA1 or BRCA2 mutations. N Engl J Med 1997; 336:1465-71.

16. Messori A. Current controversies in the application of meta-analysis (with special reference to oncological treatments). Pharm World Sci 1997; 19:152-8.

17. Sheldon TA. Problems of using modelling in the economic evaluation of health care. Health Economics 1996; 5:1-11.

18. Fenn P, McGuire A, Phillips V, Backhouse M, Jones D. The analysis of censored treatment cost data in economic evaluation. Medical Care 1995; 33:851-63.

19. Quigley HA, Vitale S. Models of open-angle glaucoma prevalence and incidence in the United States. Invest Ophthalmol Vis Sci 1997; 38:83-91.

20. Whitehead A, Whitehead J. A general parametric approach to the meta-analysis of randomized clinical trials. Statistics in Medicine 1991; 10:1665-77.

21. Goodwin PJ, Boyd NF. Mammographic parenchymal pattern and breast cancer risk: a critical appraisal of the evidence. Am J Epidemiology 1988; 127:1097-108.

22. Clarke MJ. Why the early Breast Cancer Trialists' Collaborative Group (EBCTCG) individual patient data meta-analysis was needed. Control Clin Trials 1995; 16:67S-8S.

23. Clarke MJ, Stewart LA. Meta-analyses using individual patient data. J Evaluation in Clin Practice 1997; 3:207-12.

24. Stewart LA, Parmar MKB. Meta-analysis of the literature or of individual patient data: is there a difference? Lancet 1993; 341:418-22.

25. Stewart LA, Clarke MJ. Practical methodology of meta-analysis (overviews) using updated individual patient data. Statistics in Medicine 1995; 14:2057-79.

26. Srinivasan C, Zhou M. A note on pooling Kaplan-Meier estimators. Biometrics 1993; 49:861-4.

27. Pignon JP. Randomized trials of radiotherapy alone versus combined chemotherapyand radiotherapy in stages IIIa and IIIb nonsmall cell lung cancer: a meta-analysis. Cancer 1996; 77:2413-4.

28. Pignon JP, Arriagada R. Meta-analyses of randomized clinical trials: how to improve their quality? Lung Cancer 1994; 10 Suppl. 1:S135-S141

29. Oxman AD, Clarke MJ, Stewart LA. From science to practice: Meta-analyses using individual patient data are needed. JAMA 1995; 274:845-6.

30. Marino P, Preatoni A, Cantoni A. Randomized trials of radiotherapy alone versus combined chemotherapy and radiotherapy in stages IIIa and IIIb nonsmall cell lung cancer: a meta-analysis. Cancer 1995; 76:593-601.

31. Non-small Cell Lung Cancer Collaborative Group. Chemotherapy in non-small cell lung cancer: a meta-analysis using updated data on individual patients from 52 randomised clinical trials. BMJ 1995; 311:899-909.

32. Pignon JP, Arteaga C. Role of thoracic radiotherapy in limited-stage small cell lung cancer: quantitative review based on literature versus meta-analysis based on individual data. J Clin Oncol 1992; 10:1819-20.

33. Tierney JF, Mosseri V, Stewart LA, Souhami RL, Parmar MKB. Adjuvant chemotherapy for soft-tissue sarcoma: review and meta-analysis of the published results of randomised clinical trials. Br J Cancer 1995; 72:469-75.

34. Sarcoma Meta-analysis Collaboration. Adjuvant chemotherapy for localised resectable soft-tissue sarcoma of adults: meta-analysis of individual data. Lancet 1997; 350:1647-54.

35. Early Breast cancer Trialists' Collaborative Group. Systemic treatment of early breast cancer by hormonal, cytotoxic or immune therapy: 133 randomised trials involving 31 000 recurrences and 24 000 deaths among 75 000 women. Lancet 1992; 339:1-15, 71-85.

36. Pignon JP, Bourhis J. Meta-analysis of chemotherapy in head and neck cancer: individual patient data vs. literature data. Br J Cancer 1995; 72:1062-3.

37. Steinberg KK, Smith SJ, Stroup DF, et al. Comparison of effect estimates from a meta-analysis of summary data from published studies and from a meta-analysis using individual patient data for ovarian cancer studies. Am J Epidemiol 1997; 145:917-25.

38. Jeng GT, Scott JR, Burmeister LF. A comparison of meta-analytic results using literature vs. individual patient data. Paternal cell immunization for recurrent miscarriage. JAMA 1995; 274:830-6.

39. Marino P. Author Reply. Cancer 1996; 77:2414

40. Hasselblad V, Mosteller F, Littenberg B, et al. A survey of current problems in meta-analysis. Medical Care 1995; 33:202-20.

41. Hunink MGM, Wong JB. Meta-analysis of failure-time data with adjustment for covariates. Med Decis Making 1994; 14:59-70.

42. Altman DG, De Stavola BL, Love SB, Stepniewska KA. Review of survival analyses published in cancer journals. Br J Cancer 1995; 72:511-8.

43. Cellerino R, Tummarello D, Guidi F, Isidori P, Raspugli M, Giscottini B. A randomized trial of alternating chemotherapy versus best supportive care in advanced non-small cell lung cancer. J Clin Oncol 1991; 9:1453-61.

44. Ganz PA, Figlin RA, Haskell CM, LaSoto M, Siau J. Supportive care versus supportive care and combination chemotherapy in metastatic non-small cell lung cancer. Cancer 1989; 63:1271-8.

45. Kaasa S, Lund E, Thorud E, Hatevoll R, Host H. Symptomatic treatment versus combination chemotherapy for patients with extensive non-small cell lung cancer. Cancer 1991; 67:2443-7.

46. Rapp E, Pater JL, Willan A, et al. Chemotherapy can prolong survival in patients with advanced non-small-cell lung cancer - Report of a Canadian multicenter randomized trial. J Clin Oncol 1988; 6:633-41.

47. Quoix E, Dietemann A, Charbonneau J, Boutin C, Meurice JC, Orlando JP. La chimiotherapie comportant du cisplatin est-elle utile dans le cancer bronchique non microcellulaire au stade IV? Resultats d'une etude randomisee. Bull Cancer 1991; 78:341-6.

48. Woods RL, Williams CJ, Levi J, Page J, Bell D, Byrne M. A randomised trial of cisplatin and vindesine versus supportive care only in advanced non-small cell lung cancer. Br J Cancer 1990; 61:608-11.

49. Masters GA, Vokes EE. Should non-small cell carcinoma of the lung be treated with chemotherapy? Pro: Chemotherapy is for non-small cell lung Cancer. Am J Respir Crit Care Med 1995; 151:1285-7.

50. DeVita VTJr, Hellman S, Rosenberg SA. Cancer: Principles and Practice of Oncology. Philadelphia: J.B. Lippincott Co., 1993:

51. Kalbfleisch JD, Prentice RL. The Statistical Analysis of Failure Time Data. Toronto: John Wiley and Sons, 1980:

52. Coldman A, Elwood M. Examining survival data. Can Med Assoc J 1979; 121:1065-71.

53. Lee ET, Go OT. Survival analysis in public health research. Annu Rev Public Health 1997; 18:105-34.

54. Armitage P, Berry G. Statistical Methods in Medical Research. Oxford: Blackwell Scientific Publications, 1994:469-92.

55. Mould RF. The life table method using grouped data for survival rate calculations. Current Oncology 1997; 2:158-61.

56. Dear KBG. Iterative generalized least squares for meta-analysis of survival data at multiple times. Biometrics 1994; 50:989-1002.

57. Shore T, Nelson N, Weinerman B. A meta-analysis of stages I and II Hodgkin's disease. Cancer 1990; 65:1155-60.

58. Voest EE, Van Houwelingen JC, Neijt JP. A meta-analysis of prognostic factors in advanced ovarian cancer with median survival and overall survival (measured with the log(relative risk)) as main objectives. Eur J Cancer Clin Oncol 1989; 25:711-20.

59. Douglas IS, White SR. Con: Therapeutic Empiricism - The Case against Chemotherapy in Non-small Cell Lung Cancer. Am J Respir Crit Care Med 1995; 151:1288-91.

60. Mike V, Stanley KE. Statistics in Medical Research Methods and Issues with Applications in Cancer Research. New York: John Wiley & Sons, 1982:

61. Yateman NA, Skene AM. The use of simulation in the design of two cardiovascular survival studies. Statistics in Medicine 1993; 12:1365-72.

62. Goldman AI, Hillman DW. Exemplary data: sample size and power in the design of event-time clinical trials. Controlled Clin Trials 1992; 13:256-71.

63. Grambsch P, Therneau T. Proportional hazards tests and diagnostics based on weighted residuals. Biometrika 1994; 81:515-26.

64. Reimold SC, Chalmers T, Berlin JA, Antman EM. Assessment of the efficacy and safety of antiarrhythmic therapy for chronic atrial fibrillation: observations on the role of trial design and implications of drug-related mortality. Am Heart J 1992; 124:924-32.

65. Shrout PE, Fleiss JL. Intraclass correlations: Uses in Assessing rater reliability. Psychological Bull 1979; 86:420-8.

66. Williamson P, Marson A, Hutton J, Chadwick D. Individual patient versus aggregate data meta-analysis for time-to-event outcomes: empirical evidence from epilepsy. Systematic Reviews: Beyond the basics 1998; (Abstract)

67. Abel UR, Edler L. A pitfall in the meta-analysis of hazard ratios. Control Clin Trials 1988; 9:149-51.

68. Buyse M, Piedbois P. On the relationship between response to treatment and survival time. Statistics in Medicine 1996; 15:2797-812.

69. Buyse M, Parmar MKB. Letter to the Editor. Eur J Cancer Clin Oncol 1988; 25:1901

70. Weisberg S. Applied Linear Regression. New York: John Wiley & Sons, 1985:

71. Seber GAF, Wild CJ. Nonlinear Regression. New York: John Wiley & Sons, 1989:

72. Hamilton LC. Regression with Graphics: A Second Course in Applied Statistics. Belmont, California: Duxbury Press, 1992:

73. Conover WJ. Practical Nonparametric Statistics. New York: John Wiley & Sons Inc., 1971:

74. Dudewicz EJ, Mishra SN. Modern Mathematical Statistics. John Wiley & Sons, 1988:

75. Saeki S, Ogata H, Okubo T, Takahashi K, Hoshuyama T. Return to work after stroke: a follow-up study. Stroke 1995; 26:399-401.

76. Fleiss JL. Reliability of Measurement. In: Barnett V, Bradley RA, Hunter JS, et al., eds. The Design and Analysis of Chemical Experiments. John Wiley & Sons, 1986: 1-31.

Appendix 1 - Glossary of terms, abbreviations, and a summary of the methods

BSC: best supportive care

Dear method: a multiple linear regression model is fit with survival proportion as the dependent variable, and each time interval, study, treatment, and interaction terms as independent variables. It is the only method examined that preserves randomization for treatment comparisons.

"full sample analysis": ignores censoring during a trial.

Hunink method: data is extracted from component studies and organized into a summary life table. The hazard rate is calculated in each interval and converted into a survival function.

Individual Patient Data: knowledge of the time to event or censoring for each subject in an analysis.

Informative censoring: when the reason for censoring is related to the subject's probability of experiencing an outcome event.

IPD: See Individual Patient Data.

"median censoring": removes censored patients from the effective sample size at the time of median survival.

Nelson method: A non-linear regression model fits an exponential decay function to the data from the component studies.

Non-informative censoring: when the reason for censoring is unrelated to the subject's probability of experiencing an outcome event.

Phase II study: a single arm study in which patients are given an intervention and observed for an indication of efficacy.

Proportional survival: The proportion of patients alive at a certain time; the same as $S(t)$.

Randomized Phase II study: a multiple arm Phase II study in which patients are randomly assigned to two or more investigational interventions, without a standard control arm.

Study termination censoring: censoring when patients with the longest follow-up in a study have not experienced an event.

"uncensored only": removes censored patients from the sample size at the start of the trial.

Voest method: a $\ln(-\ln(S(t)))$ transformation turns the survival stepfunctions into parallel curves. An average curve is calculated and transformed back into a summary survival stepfunction.

Weighted Voest method: the same as Voest, except the trials are weighted by their inverse variance before an average curve is calculated.

Appendix 2 - Proportional Hazards

A proportional hazards assumption is that there is an underlying "true" shape of the survival curve, and that the relative hazard for the variable of interest (e.g. the prognostic or treatment group) is constant throughout.[51] The Cox proportional hazards model is a regression model that uses all observations to derive the base shape of the underlying survival curve in a manner similar to the Kaplan-Meier estimate. It then uses regression analysis to calculate the hazard ratio (the instantaneous relative risk) for each variable, and adjusts the curve up or down accordingly (Figure 22). It can be used:

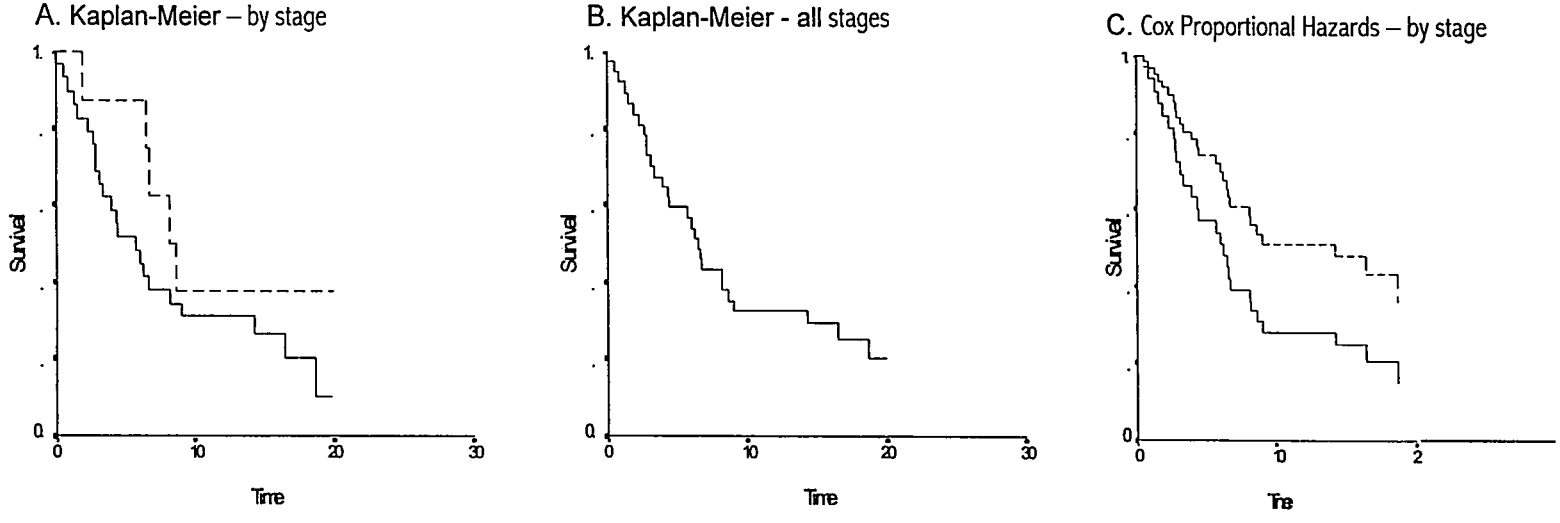
1. to identify prognostic variables (e.g. age, sex, previous treatment) that significantly affect survival,
2. to compare two treatments by estimating treatment effect as one of the variables in the model, and
3. to adjust survival curves based on the distribution of prognostic features, thereby "controlling for" those variables to arrive at an "adjusted hazard ratio". This is useful when comparing two groups that may not be entirely balanced for these features.[53]

In this way, the hazard rate h is a function of both time t and the prognostic covariates x. The basic structure of the model can be given by:

$$h_i(t,x_i) = h_0(t)e^{b_1x_1 + b_2x_2 + b_3x_3 + \cdots}$$

where h_i is the hazard ratio for a particular covariate x_i at time t. Therefore, the hazard is the product of a baseline hazard $h_0(t)$ and an exponential linear regression function of the covariates.

Figure 22. Cox Proportional Hazards Model



The Cox Proportional Hazards model takes the Kaplan-Meier curves for two groups (A) and combine them to determine the shape of the underlying curve (B). It calculates the hazard ratio between the two groups, and shifts the combined curve up and down to create group-specific survival curves (C)

The hazard ratio for individuals with two different covariates, given that all others are equal, is:

$$e^{b(x_1 - x_2)}.$$

Thus, the hazard ratio does not depend on the baseline hazard and is constant over time. The two hazard functions can never cross.

Appendix 3 - Least squares estimation

Least squares estimation was used to estimate the final survival curves for the Dear and Nelson methods. In linear regression, least squares estimation finds the straight line that would minimize the residual sum of squares (Figure 23),[70] described by the equation:

$$Y = b_0 + b_1X$$

where b_0 is the intercept and b_1 is the change in Y for a unit change in X.

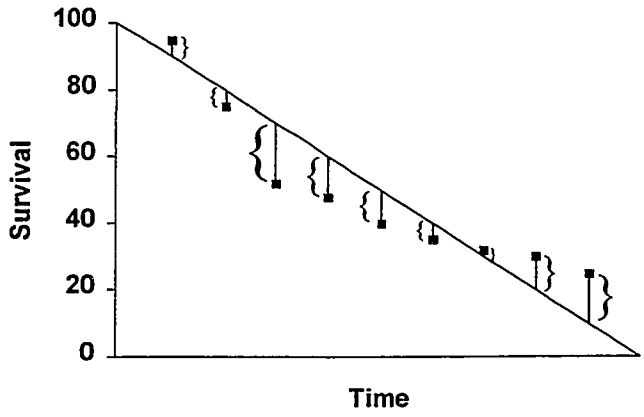
Non-linear regression is similar to linear regression, except that the underlying model is not considered to be a straight line (Figure 24). They are usually associated with a logistic function. For example, an exponential decay function (as was used in the Nelson method) would be written by:

$$Y = b_0 + exp(-b_1X)$$

Finding the least squares estimates generally requires using one of several iterative algorithms available in statistical computer programs, such as the Newton-Raphson iterative algorithm used in the Hunink method.[41;71] They begin with a set of initial parameter estimates, then try to find better estimates that reduce the residual sum of squares.[72] The initial estimates may be arbitrary, the result of a simple linear regression, or more complex estimates provided by the user.[70]

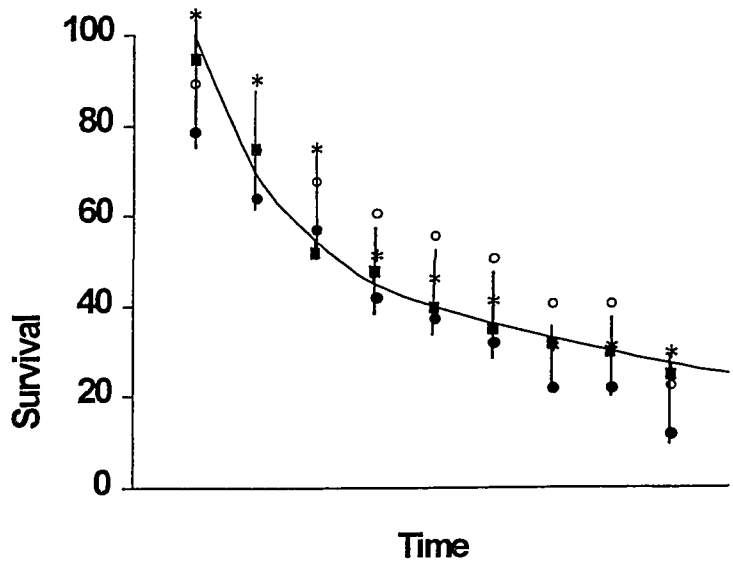
The analyses described above assume that error variances are unknown, equal (homoscedasticity), and uncorrelated with each other.[72] However, when information is known about the variance of data, it can be incorporated into the analysis in the form of "generalized" least squares analyses. The model can be linear or non-linear. The variance

Figure 23. Least squares estimation in linear regression



The regression will fit a straight line that minimizes the sum of the squares of the distances between it and each of the observed points.

Figure 24. Non-linear regression



As in Figure 23, the regression model will fit a line to minimize the sum of the squares of the distance between it and each of the observed points, in this case proportional survival from different trials at each point in time. However, in non-linear regression, the line is not necessarily straight. For example, in the Nelson method it follows an exponential decay function.

of the survival function can be calculated, for example by the Greenwood formula (section 2.5). By considering the variance in the form of a variance-covariance matrix, this method recognizes that residuals corresponding to errors with a larger variance are given less weight in the estimation procedure. The diagonal entries of the variance-covariance matrix are the variances, while the off-diagonal entries are the covariances between the elements. However, calculating the covariances can be difficult. "Weighted least squares" can be used when the errors have different variances but are uncorrelated. In this case, the covariance matrix reduces to a diagonal matrix of the variances, which determine the weights, with the off-diagonal covariances being 0. This greatly simplifies the calculations.

If there is a large sample size and the estimates depend on a small number of parameters, "iteratively reweighted least squares" can be used, as in the Dear method. Arbitrary initial coefficients are used to estimate the survival function (Figure 10). From this, a new correlation matrix is determined. This can yield a new variance-covariance matrix as described above. The new coefficients are estimated, and the procedure repeated until the estimates of the coefficients converge, i.e., they do not change more than a preset amount with each iteration. If the sample size is too small, however, the precision of each estimate will be poor. Since the weighting is based on precision, the analysis will not converge quickly.

Appendix 4 - The Kolmogorov-Smirnoff test

I chose the Kolmogorov-Smirnoff test to assess the ability of each method to produce a summary curve similar to the curve derived from IPD. There are other tests for goodness of fit, such as the chi-square test, however, this is designed for nominal data and has less power as a result.[73] Furthermore, the Kolmogorov-Smirnoff test does not assume an underlying distribution, as do tests such as the Lilliefors test.[73]

Let $S1_m(t)$ be the survival function for one population of sample size m, and $S2_n(t)$ be the survival function for another independent population of sample size n. The discrepancy $D_{m,n}$ between the two functions at time t is the absolute value of the vertical distance between them (Figure 25):

$$D_{m,n} = |\, S1(t) - S2(t)|$$

To test the null hypothesis that the two survival functions are the same,

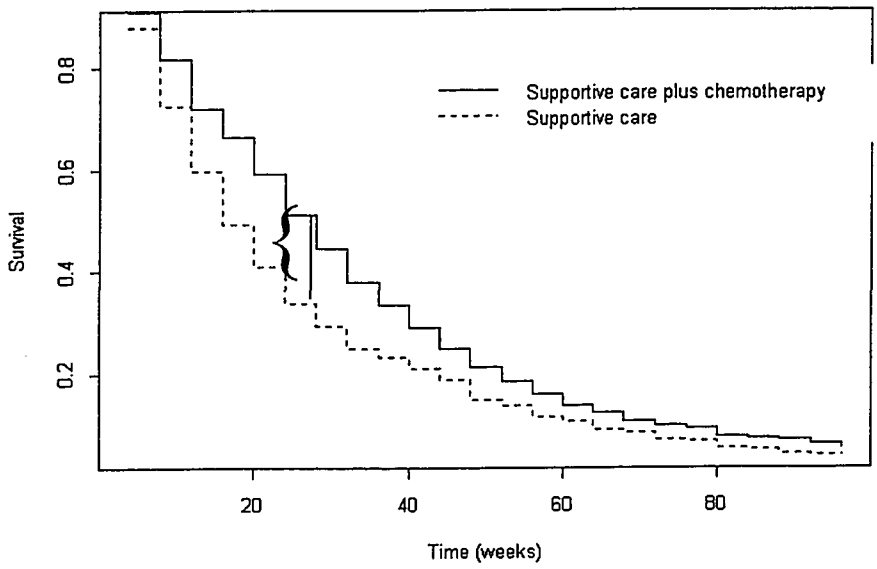
$$H_0: S1(t) = S2(t), \qquad \text{versus } H_1: S1(t) \neq S2(t)$$

I expect the largest distance between the two graphs, $D_{m,n}$, to be "small".[73] The test rejects the null hypothesis and accepts the alternate hypothesis that the distributions are not equal at level α if $D_{m,n} > d_{m,n,\alpha}$ according to the following critical points:[74]

α	.01	.05	.10
$d_{m,n,\alpha}$	$1.63\sqrt{(m+n)/mn}$	$1.36\sqrt{(m+n)/mn}$	$1.22\sqrt{(m+n)/mn}$

The confidence interval for the function is given by $S(t) \pm d_{m,n,\alpha}$

Figure 25. The Kolmogorov-Smirnoff Test



The Kolmogorov-Smirnoff Goodness of Fit Test examines the maximum vertical distance between two curves and determines whether it is more than would be expected by chance alone if the distributions were equal.

The Cramer-von Mises Goodness of fit test is similar to the Kolmogorov-Smirnoff test. It considers several vertical distances between the two curves, not just the largest. This is intuitively appealing because it makes more complete use of the data. The Kolmogorov-Smirnoff test only measures the maximum discrepancy between two curves, so it is sensitive to marked differences in survival that may occur at only one point in time.[61] Despite this, the Cramer-von Mises test has generally been found to be no better than the Kolmogorov-Smirnoff test when they have been compared empirically. Furthermore, it is much more difficult to compute and does not improve power over the Kolmogorov-Smirnoff test. As a result, it is less commonly used.[73]

Appendix 5 - Hypothesis Testing

Statistical comparison of two survival curves, *S1(t)* and *S2(t)* is usually done by testing the null hypothesis that $S1(t) = S2(t)$. The simplest analysis is to do a t-test comparing the survival rates at a time point of interest (e.g. the one-year survival). However, the shape of the curve also contains important information that could be missed if the wrong time point is chosen. Because survival curves tend to be skewed,[60] parametric tests are usually not appropriate. However, there are some parametric distributions known to approximate survival times. These include the exponential, Weibull, Gompertz, and log-logistic. The latter three differ from the exponential distribution in that the hazard can change with time.[54] Nevertheless, non-parametric statistics, such as the median survival rather than the mean, are usually used for describing survival data.

The log rank and Wilcoxan tests are the most commonly used tests to analyze survival curves.[53] Both are non-parametric tests that make a combined set of observations from the survival functions being compared and ranks them. The null hypothesis for the log rank test is that the risk of death is the same in the two groups (A and B) being compared. Therefore, the number of deaths, d, at any particular time will be expected ($E(d)$) to be distributed between the two groups in proportion to the numbers at risk (n):

$$E(d_A) = d_{AB}(n_A/n_{AB})$$

If O is the sum of the observed number of events d_A over all time points, and E is the sum of the expected number of events $E(d_A)$ over all time points, then any difference between O and E is evidence against the null hypothesis.[54]

The logrank statistic (Figure 26), a chi square statistic with 1 degree of freedom (one less than the number of groups being compared) can be approximated by:

$$X^2 = (O_A - E_A)^2/E_A + (O_B - E_B)^2/E_B$$

The log-rank test is appropriate for survival distributions whose hazard functions are proportional over time, e.g., for curves that do not cross. If this is not the case, other tests such as the Wilcoxan should be used. The Wilcoxan test also relies on ranking observations, but puts more emphasis on events in the first part of the curve. As a result, these tests can occasionally give different results when comparing the same data.[75]

For comparison of the proportional survival at time points of interest, each method allows calculation of the standard error of the survival proportion at each time using the Greenwood formula for the variance described in section 2.5. With this, 95% confidence intervals (CI) can be calculated around the proportional survival[54] $S(t)$:

$$95\%\ \text{CI} = S(t)EXP(\pm 1.96(\text{standard error})/-S(t)lnS(t))$$

Figure 26. Calculation of the log rank statistic from summary survival proportions

Interval	# deaths during interval in one group (o_i)	# deaths during interval in both groups (d_i)	Number of patients alive in one group immediately prior to the end of the interval (m_i)	Number of patients alive in both groups immediately prior to the end of the interval (n_i)	$e_i = d_i * m_i / n_i$	$v_i = \frac{e_i(n_i - m_i)(n_i - d_i)}{n_i(n_i - 1)}$
0	0	0	m	n	0	0
0 - 1	o_1	d_1	m_1	n_1	e_1	v_1
1 - 2	o_2	d_2	m_2	n_2	e_2	v_2
2 - 3	o_3	d_3	m_3	n_3	e_3	v_3
...	...	...	...	...	...	...

$O = o_i$
$E = e_i$
$V = v_i$
Log-rank statistic = $(O-E)^2/V$, which follows a chi-square distribution with one degree of freedom

Appendix 6 - Intraclass correlation

To examine inter-observer reliability, k observers independently make observations on the same N time points. Consider the two observers in my inter-observer reliability assessment to be a sample from a larger population of potential observers. Each made measurements at 306 distinct time points when assessing the component trials of the NSCLCCG meta-analysis. An analysis of variance (ANOVA) was carried out, with the dependent variable being the survival proportion measured, and the independent variables being the observer and the time point (Table 8).

Because the observers were a sample of possible observers, a random effects model is appropriate for analysis. The statistical model that describes the observed measurement X is:

$$X = T + r + e$$

where T is the true value of the observation, r is the effect of the rater (observer) on an observation, and e is an error term.[76] T, r, and e are assumed to be mutually independent. The variance (VAR) of a single measurement is:

$$\mathrm{VAR}_X = \mathrm{VAR}_T + \mathrm{VAR}_r + \mathrm{VAR}_e$$

and the intraclass correlation coefficient is:

$$\mathrm{R} = \frac{\mathrm{VAR}_T}{\mathrm{VAR}_T + \mathrm{VAR}_r + \mathrm{VAR}_e}$$

From Table 8 we can calculate that an estimator of the intraclass correlation coefficient is given by:

$$\mathrm{R} = \frac{\mathrm{N(PMS - EMS)}}{\mathrm{(N)PMS + (k)RMS + (Nk - N - k)EMS}}$$

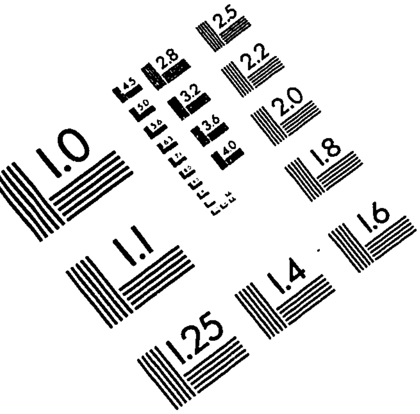
$$= 0.99$$

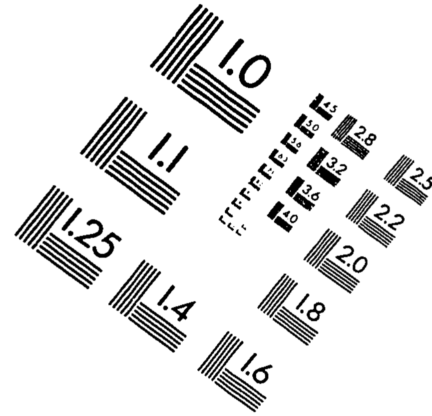
Table 8. Analysis of variance for assessing inter-observer reliability

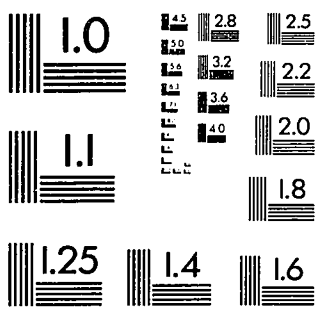
Source of variation	df formula	df	SS	MS name	MS	Expected (MS)
Time point	$N-1$	305	39.36899125	PMS	0.12907866	$VAR_e + k*VAR_T$
Observer	$k-1$	1	0.00111417	RMS	0.00111417	$VAR_e + N*VAR_r$
Error	$(N-1)(k-1)$	305	0.04180788	EMS	0.00013708	VAR_e
Total	Nk - 1	611	39.41191330			

df = degrees of freedom
SS = sum of squares
MS = mean square
N = number of time point observations
k = number of observers
VAR = variance
T = true value of the observation
r = the observer's effect on an observation
e = error

IMAGE EVALUATION TEST TARGET (QA-3)

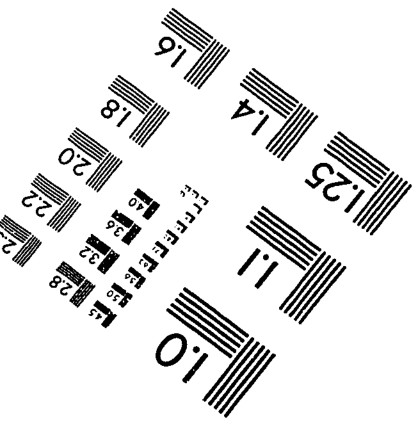


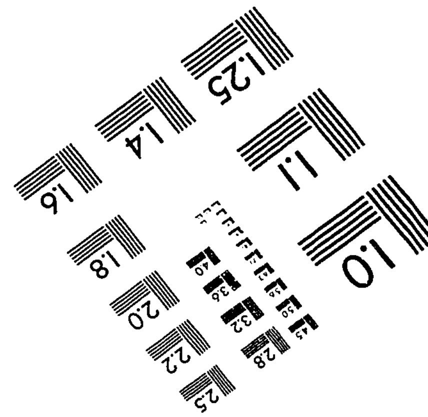




150mm

6″





APPLIED IMAGE, Inc
1653 East Main Street
Rochester, NY 14609 USA
Phone: 716/482-0300
Fax: 716/288-5989