THE UNIVERSITY OF CALGARY

The Use and Effectiveness of Routine Baseline Bone Scans in Breast Cancer

by

Chester Joel Chmielowiec

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES

IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE

DEGREE OF MASTER OF SCIENCE

DEPARTMENT OF COMMUNITY HEALTH SCIENCES

CALGARY, ALBERTA

MAY, 1996

THE UNIVERSITY OF CALGARY

FACULTY OF GRADUATE STUDIES

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies for acceptance, a thesis entitled "The Use and Effectiveness of Routine Baseline Bone Scans in Breast Cancer" submitted by Chester Joel Chmielowiec in partial fulfilment of the requirements for the degree of Master of Science.

Santri Ramcharan

Supervisor, Dr.S.Ramcharan, Dept. of Community Health Sciences

Dr.R.Kloiber, Dept. of Radiology

Dr.P.M.A.Brasher, Dept. of Community Health Sciences

External Examiner, Dr.C.D.Black, University of Manitoba

June 19, 1996

Date

Abstract

The use of routine baseline bone scans in breast cancer has been a topic of controversy, especially in Stage II patients. Because of the current drive for efficiency, this study was undertaken to examine the use and effectiveness of such scans. Two time periods representing the beginning and the end of the 1980's were chosen and all breast cancer patients at a regional cancer centre were evaluated. Decreases in use of routine scans were expected in Stage I and Stage II patients but were not found. Survival analysis was performed for all Stage II patients to assess the effect of a baseline bone scan. No difference in five-year survival was found between those patients who were and were not scanned. Factors related to survival were identified through use of Cox's proportional hazard model. The most significant factors were age, menopausal status, estrogen receptor result and bone scan result.

Acknowledgements

Many thanks to Dr. Andrew Maksymiuk and the staff at the Saskatoon Cancer Centre for their assistance.

I would also like to thank the members of my supervisory committee - Drs. Ramcharan, Kloiber, and Brasher - for their suggestions, criticisms, and guidance.

Dedication

To my parents

Table of Contents

List of Tables

List of Figures

I. INTRODUCTION

In the current political and economic climate there is now a greater demand on the health care system to establish the effectiveness of its various procedures. For the most part, this has been initially directed to the therapeutic technologies because of the direct impact of therapy on patient outcome. Assessing the effect on patient outcome of diagnostic technologies has been ignored because of the difficulty of evaluating diagnostic technologies in this context. However, the effect of diagnostic tests on outcome can no longer be avoided due to the current impetus to increase the efficiency of the medical system. Many of the currently used diagnostic tests, including diagnostic imaging, have been inadequately assessed with regard to their effectiveness.

One of the diagnostic imaging technologies that is in use is nuclear medicine. Briefly, this technology uses unsealed radionuclides which are attached to various molecules and are introduced into the patient in a relatively non-invasive manner, usually through an intravenous injection. The radioactive tag allows one to trace the fate of the labelled molecules through the use of sophisticated detection systems. This allows visualization of the physiology of many organ systems. It is possible to make some interpretation of anatomical structure but this is not a strength of nuclear medicine. Although relatively new, nuclear medicine can be considered mature in some of its applications.

One of the most widely used nuclear medicine tests is the bone scan. An important indication for the use of this test is the detection of metastatic disease of the skeleton in those malignant diseases that are prone to spread to bone. A very common

example of such a malignant disease is breast cancer which is reported to metastasize frequently to bone(1). However, there has been a great deal of controversy about the routine use of bone scans during the initial work-up of breast cancer patients, especially for those patients with minimal disease. It is this controversy about the use and effectiveness of routine bone scanning in breast cancer patients that is the focus of this study, especially in those patients who are initially classified as Stage II.

II. BACKGROUND

1. Literature Review

a. Bone Scans in Breast Cancer

A review of the literature revealed no Canadian studies that examined the effectiveness of bone scans in breast cancer. Most of the research has been done in Europe and the United States but the number of published reports is fairly small. In the 1970's there were several studies on the use of bone scans in breast cancer that were mixed in terms of their recommendations for routine baseline bone scanning(2-5). During this time, the technology for bone scans was still developing and there tended to be a general enthusiasm for scanning everyone with breast cancer(6-8). However, as the test matured there began to be more discussion about the value of routine scanning. In 1981 there appeared a review of the use of bone scans in early breast cancer, i.e., clinical stages I and II, which suggested that there were too many misleading results in these patients. The author stated that in stage I only 4% of scans were positive and in stage II only 10% were positive. But what was disconcerting was the estimate that one half of the

studies were false positive. It was recommended that routine bone scans not be obtained in these patients(9). There were several other studies during the 1980's which came to similar conclusions(10-17). Yet there are at least two studies during this time frame that continued to recommend routine baseline scans in patients despite the problem of poor positive predictive values(18,19). The main reason cited was because of the propensity for the future development of bone metastases in breast cancer, which has been called scan conversion. It was argued that by having a baseline study it would be easier to detect early metastatic disease while at the same time avoiding false positives. Of interest, it appears that the studies recommending routine scans are from diagnostic imagers while the other papers are by other types of physicians, mainly oncologists.

In the 1990's this same difference of opinion continues to be reported. A group from England reported that although the number of positive scans was low in stages I/II, they found that individuals with a positive scan had a statistically reduced survival. This group recommended routine baseline studies in Stages II-IV as well as high risk Stage I patients (high risk in this study meant symptomatic bone pain or a raised alkaline phosphatase)(20). However, other authors rejected the use of routine baseline bone scans in stages I and II because of a very low yield for identifying bony metastases(21-25). This poor yield and the consequent unnecessary costs associated with the use of routine bone scans in these patients were reasons for not doing the test. This conclusion was reinforced by the finding that five-year survival in stages I and II was the same despite the result of the bone scan(21,26,27).

The consensus is that routine baseline bone scans are of little or no value for

asymptomatic stage I disease. The controversy is with regard to asymptomatic stage II patients. A main point of contention relates to the substantial number of false positives identified in this group. The marked sensitivity of the bone scan in detecting bone changes makes abnormal findings somewhat nonspecific. The low prevalence of bone metastases in these Stage II patients makes the positive predictive value of the bone scan very low. Several authors still recommend bone scans in these patients because of a high rate of scan conversion, i.e., a high rate of developing bony metastases within the first two years after initial diagnosis(20,28,29). Those individuals advocating these baseline studies seem to be diagnostic imagers whose reason for recommending these studies is to allow for the identification of early metastatic disease while avoiding potential false positives. Other authors, mainly oncologists, state that one should be more selective in choosing those patients who need a scan. They suggest that there are several indicators for those at a higher risk for developing metastases such as the presence of multiple axillary nodes or negative hormone receptor status of the primary tumour(23,25). It is important to note that the result of a baseline scan does not usually affect the initial treatment plan.

With regard to patients who are symptomatic or staged higher than I/II, there is no real controversy about the use of a bone scan. In this case, the scan is used for confirmation of metastatic disease rather than as a screening or staging tool.

b. Technology Assessment

Technology assessment is becoming a very important part of health care as society strives to increase the efficiency of the health care system. Part of the need for assessment comes as a result of the dramatic expansion of technology over the past few decades. This expansion plus the increasing costs of health care lead to the necessity for the proper evaluation of technology, especially as some estimate that medical technologies may contribute as much as 50% to these increased costs(30). Other reasons for increasing costs include population growth, population ageing, inflation and increased litigation. Most of these factors are difficult, if not impossible, to influence. However, it should be possible to identify those patients who need to receive health care and to determine the level of services provided to these patients and in this way "control" costs. Underlying this is the need for sound scientific evidence on safety, effectiveness and cost-benefit of medical technology. This will assist in the making of good policy decisions(30).

With specific regard to diagnostic technologies, it is necessary to work within a well-defined framework with a set of guidelines. This has been often structured as a model of hierarchy which ranges from technologic capability to patient outcome, with the number of intervening levels in the model varying between authors(31-33). These intermediate levels usually relate to matters of diagnostic accuracy, possible uses of the technology and impacts on patients and providers. While this hierarchical model is a common one, there are also other models in existence such as the accuracy/receiver operating characteristic curve analysis model, the technologic impact model and the

change model(34).

Guyatt et al have proposed criteria for clinical evaluation of diagnostic technologies(32,33). These criteria form a hierarchy of progressively more rigorous evaluation. The first step is in fact pre-clinical and is the laboratory assessment of the capability of the technology in physical terms. Following this phase, there comes a level where the range of possible uses needs to be assessed. This exploratory stage where the technology is applied to a wide range of patients and conditions allows for the development of hypotheses which can be tested at a later time. It is after this exploratory phase that the level of formal testing is entered. This has been called the diagnostic accuracy level by Guyatt(32,33).

Before proceeding further it is necessary to define the terms efficacy and effectiveness as they relate to this context. Efficacy refers to the performance of a test under ideal circumstances. Effectiveness refers to a test's performance under more usual clinical conditions.

At the level of diagnostic accuracy, more rigorous methodology must be used. Comparison to a "gold standard" is the ideal although long term follow-up may be a suitable substitute for such a standard. If neither is possible, then comparison to other measures may suffice in that this may allow for construct validity. Thorough evaluation should produce output such as receiver operating characteristic (ROC) curves, sensitivity, specificity, positive/negative predictive values and likelihood ratios. Other things which may be specified include correlation coefficients and chance corrected agreement (kappa) values. It is important that effectiveness be tested by including patients with varying

degrees of pathology, including easily confused diseases, in order to simulate usual clinical practice. This level is often the last one where appropriate, high quality evaluation is performed(31).

Beyond the level of diagnostic accuracy, most evidence is usually fairly weak and often relies on consensus or opinion. The first level beyond diagnostic accuracy is that level where the impact on health care providers is assessed. This is separate from the impact on therapy or patient outcome. It appears to relate to the effect of the technology on anxiety or confidence of the provider. It may also be related to the concept of "defensive medicine." Evaluation tools are poorly developed at this level and need further investigation.

The level after provider impact is that of therapeutic impact. In order for the technology to change the patient's outcome, it is necessary that it provide information which will influence decisions on therapy. The expectation is that this will be a positive impact where patients are benefitted although it is possible a negative impact where patients are harmed could occur. This question of patient impact leads to the final level where patient outcome is evaluated.

It becomes obvious that there is a need for high quality, rigorous research in technology assessment. Many of these levels can be assessed simultaneously with the use of a well designed randomized controlled trial (RCT). Certain requirements are necessary to run a good trial. There should be a written protocol with a thorough review of the literature, including an attempt to find unpublished information. Efforts need to be made to minimize or measure observer error and bias. This is especially true for

diagnostic imaging trials where blinding of the observer is often difficult. If possible, all measures or observations should be done in duplicate. Sound statistical analysis is important and should incorporate sensitivity analysis. Results should include confidence intervals. In trials with no significant results, there should be a discussion of sample size and power. Enough information should be found in the report such that someone else could replicate the study(31,32,35-37).

Although the RCT is held to be the best method for obtaining evidence, there are other types of study design. Case series or case studies are handicapped by lack of a control group and so may have problems with internal validity related to placebo effect, spontaneous regression, and bias due to a lack of blinding. Cohort and case control studies may be biased because of the presence of confounding. Strategies which may be used to minimize confounding include stratification, restriction, matching or random allocation. The RCT design controls confounding but has other problems. One of the most troublesome is the lack of generalizability. This comes about because the study groups are highly selected due to time and financial constraints. As these trials often have small numbers, there may be insufficient power to detect clinically significant differences. Within the trial it may be difficult to determine at which level the technology causes an impact on outcome. Some other problems which can affect a RCT are contamination and co-intervention(32,38).

On occasion a stringent test of benefit may not be needed. This can occur when there is a dramatic benefit for the patient through application of the technology. If the technology provides the same information with greater accuracy, fewer side effects and at

a lower cost then a RCT may not be necessary. When a controlled trial demonstrates that the application of a diagnostic technology leads to the institution of therapy of known benefit then this may also obviate the need for a RCT(32,38).

Because of the expensive nature of most diagnostic imaging technologies, it is usually necessary to perform an economic evaluation(33). This can often be incorporated into the assessment of patient outcome. Several points need to be addressed in an economic investigation. To begin with, one needs to determine from which perspective to do the analysis, i.e., from a micro or service perspective versus a macro or system perspective. This leads to the actual viewpoint taken which is usually from that of society as a whole. If this is the case, then a wide range of costs and benefits need to be considered. Other viewpoints that can be taken are that of the patient and family, third party payer or institution. Since economics is essentially about choice, it is necessary to allow for a relevant range of alternatives in any analysis. The alternative of "doing nothing" must always be kept in mind. Many forms of economic evaluation exist but they can be summarized into five types, viz., cost analysis, cost-minimization analysis, cost-effectiveness analysis, cost-utility analysis, and cost-benefit analysis. All of these forms consider costs but it is in the assessment of benefit where they differ. Cost-minimization refers to cases where effectiveness of alternatives is not known but can be shown to be equivalent. When the alternatives are not equivalent then it is necessary to do a cost-effectiveness study where the outcome can be equated to a common quantifiable objective such as years of life gained. When the quality of outcome is considered, then one is doing a cost-utility study. If benefits can be assessed in monetary terms then a

cost-benefit analysis can be done(33,39).

It becomes obvious that it is crucial to determine the relevant range of costs and benefits as well as appropriate estimates of these quantities. Other considerations in economic analysis are the concepts of average versus marginal costs, discounting because of the differential timing of costs and benefits, and allowance for uncertainty in all estimates - at least the important ones (33).

It is important to reinforce the idea that assessment of diagnostic technologies needs to be rigorous and as complete as possible. This will allow for rational decision making and should lead to a more efficient and effective provision of diagnostic service.

2. Definitions and Major Concepts

a. Breast Cancer

Breast cancer is the second leading cause of cancer deaths in women with an estimated lifetime risk of approximately 10 percent. In addition, the incidence of this disease has been increasing in North America although mortality has been stable or decreasing(1,40,41). Screening may have contributed in part to the increase in incidence(41). There are many risk factors involved and this continues to be an area of research. Strong risk factors include older age, country of birth (especially North America and northern Europe), a personal history of previous breast cancer as well as a very strong family history. Moderate risks include upper socioeconomic class, obesity, postmenopausal status, certain mammographic patterns, a first degree relative with breast cancer, a personal history of ovarian or endometrial cancer, radiation to the chest, age at

first pregnancy greater than thirty, and being nulliparous. The weakest risk factors are a moderate alcohol intake, menarche before age twelve and menopause after age fifty-five(1,41-43).

It is difficult to estimate the true natural history of breast cancer but if left untreated the median survival time seems to be around 2.7 years with an 18% five-year survival and a 4% ten-year survival. It appears that there are at least two main subgroups of patients. One subgroup comprises about 60% of patients and has a less aggressive form of disease with about a 2.5% mortality rate per year while the other subgroup has more aggressive disease and a mortality rate of 25% per year. Depending on the definition of cure there is either no cure (statistical evidence for increased mortality for up to 40 years) or some cure (personal definition of living symptom-free from breast cancer and dying from other causes) in the order of 21% to 26%(1).

Prognostic factors in breast cancer have been evaluated. Those that are generally accepted include axillary lymph node status, size of the primary tumour, tumour histology including grade, as well as estrogen/progesterone receptor status(44). Other factors which may have prognostic significance are patient age at diagnosis and initial stage of disease. Some potential prognostic factors may be genetic markers, DNA content and markers of proliferation. It should be pointed out that many of these factors are not independent. A poor prognosis is seen in large primary tumours with positive axillary lymph nodes (especially if more than four) and histology which is relatively poorly differentiated. There is a suggestion that women who are less than thirty-five years of age may have a poorer prognosis(45).

A recent review of breast cancer in the United States demonstrated that over two-thirds of the cases of breast cancer are of the infiltrating duct type. The 5-year survival for this histology is 79%. The next most common type of tumour was lobular which represented 6% of cases with an 84% 5-year survival. Medullary carcinoma was third at 3% of cases and had an 82% 5-year survival(46). In another review the authors gave the 5-year survival based on the initial stage. For Stage I patients the 5-year survival was 84% while it was 71% for Stage II patients. In Stage III, the 5-year survival was 48% and in Stage IV the 5-year survival was only 18%(45).

Clinical staging of breast cancer is based on the TNM system with the 1987 manual from the International Union Against Cancer having the following criteria(47).

Primary tumour

TX	primary tumour cannot be assessed
T0	no evidence of primary tumour
Tis	carcinoma in situ
T1	tumour 2 cm or less in greatest dimension
	a. 0.5 cm or less
	b. between 0.5 cm and 1 cm
	c. between 1 cm and 2 cm
T2	tumour between 2 cm and 5 cm
T3	tumour greater than 5 cm

T4 tumour any size with direct extension to chest wall or skin (chest wall includes ribs, intercostal muscles, and serratus anterior muscles)

a. extension to chest wall

b. edema, ulceration of skin, or ipsilateral skin nodules in breast

c. both a and b

d. inflammatory carcinoma

Nodes (N)

NX regional nodes cannot be assessed

N0 no regional node metastases

N1 metastases to moveable ipsilateral axillary node(s)

N2 metastases to ipsilateral axillary nodes fixed to one another or to other structures

N3 metastases to ipsilateral internal mammary node(s)

Distant Metastases (M)

M0 no evidence distant metastases

M1 distant metastases (includes metastases to ipsilateral supraclavicular nodes)

The clinical stages can then be assigned based on this system with the staging based on the American Joint Committee on Cancer: Manual for Staging Breast Carcinoma 3rd edition 1988(48).

Stage 0	TisN0M0
Stage I	T1N0M0
Stage IIA	T0N1M0 or T1N1M0 or T2N0M0
Stage IIB	T2N1M0 or T3N0M0
Stage IIIA	T0N2M0 or T1N2M0 or T2N2M0 or T3N1M0 or T3N2M0
Stage IIIB	T4 any N M0 or any T N3 M0
Stage IV	any T any N M1

b. Bone Scans

The patterns of abnormality that are seen in malignant involvement of bone on a bone scan are many and varied. The most common pattern for metastases is multiple lesions that are randomly distributed throughout the axial and proximal appendicular skeleton and involving the marrow space. It should be noted that while this pattern is not pathognomonic for metastases, it is nearly 100% specific in the appropriate clinical context. Some other common causes of multiple abnormalities on a bone scan are degenerative bone and joint disease (especially in the elderly) as well as multiple fractures. Rarely, multiple septic emboli can mimic a metastatic pattern- this has been

described in tuberculosis. There are patterns of local or regional bone abnormalities that are seen in breast cancer and these often involve the sternum and nearby ribs in an asymmetric fashion. Lastly, mention should be made that although the patterns described above consist of lesions with increased uptake of the radiotracer, there can be photon-deficient abnormalities that are often much more difficult to detect(49).

c. Statistical concepts

A technique for estimating survival functions for censored data on relatively small samples is the product-limit method which was developed by Kaplan and Meier. This method can be thought of as a special case of a life table estimate where each interval contains only one observation. The survivorship function is estimated each time a death occurs. This results in a step type of function as survival is considered constant between deaths. The probability of surviving k intervals is the product of the proportion of patients surviving all of the preceding intervals as well as the kth one, i.e., $S(k)=P_1 * P_2 * \ldots * P_k$, where P_k denotes the conditional probability of surviving the kth interval given that one has survived all of the previous intervals. Of note is that these estimates are maximum likelihood estimates. Censored individuals are handled by assuming they are alive up to the point of censoring and after this point they are removed from the denominator for the estimates of the proportion of patients surviving at any later time interval. Median survival can be estimated from the output by identifying the position where the survival function is equal to 0.50. Methods exist to estimate the variance of these estimates(50).

It is often important to compare survival distributions. Many nonparametric tests are available to perform this task. One of the most commonly used tests is the logrank test. This test is based on a set of scores assigned to the various observations. These scores are a function of the logarithm of the survival function where large uncensored observations have small scores and censored observations have negative scores. The scores of the two groups when taken together sum to zero. The logrank test is based on the sum of the scores of one group. This sum plus an estimate of its variance leads to a test statistic which has an asymptotically normal distribution. It can be assessed against the standard normal function with regard to significance(51).

Finally, it is often useful to try and identify those factors which are important to survival. Usually there are multiple factors to assess. This entails methods of multiple regression. A problem is created by the presence of censored survival data. Many of the proposed regression models for survival distributions involve an assumption of a proportional hazard function, i.e., that the hazard of the study group is proportional to that of some underlying baseline hazard function. One such model was proposed by Cox. This is a nonparametric model which is useful for censored and uncensored data. In this model, an attempt is made to assess the relation between the distribution of survival time and multiple explanatory variables (continuous or categorical). If survival times are continuously distributed, a patient's (i) hazard at time t can be represented as $h_i(t)=h_0(t) * \exp(B_1x_{1i} + B_2x_{2i} + \ldots + B_px_{pi})$ where $h_0(t)$ is the underlying hazard when all x's are ignored. If one divides by the baseline function and takes the logarithm of this function, one is left with a linear combination of the variables which provides an estimate

of the relative hazard of the ith individual. The resultant equation is a standard multiple regression equation and can be solved using maximum likelihood estimation techniques. This allows one to identify those variables which are related significantly to survival. The coefficients attached to each variable can be converted into a hazard ratio which provides an estimate of the relative hazard attributed to that variable when all other variables are held constant. In contrast to parametric models of survival, the underlying survival function does not need to be specified in the Cox model. This is a very brief overview of Cox's proportional hazard model(52).

III. STUDY PURPOSE

One purpose of this study was to evaluate the use of a routine baseline bone scan in the initial diagnosis of breast cancer for all stages. Because of the conflicting recommendations found in the literature for Stage II patients, the other major purpose was to focus on this group with regard to the effectiveness of a routine baseline bone scan. Effectiveness for the purposes of this study was defined in terms of patient outcome, in this case five-year mortality.

Data were obtained from a single centre, the Saskatoon Cancer Centre. Two time periods were selected, viz., 1981/2 and 1988/9. These were chosen in order to insure that there would be five years of follow-up. Each period involved two years so that there would be an adequate number of patients to analyse.

The initial part of the study was to determine if there had been a change in the use of routine bone scans in breast cancer between the two time periods. From the literature

review, it was hypothesized that there might be a decrease in the use from 1981/2 to 1988/9 for stage I/II patients. The remainder of the study focussed on the stage II patients where the use of routine baseline bone scans is controversial.

It was hypothesized that in the earlier time period that the use of the baseline bone scan was very routine such that its use alone would not predict a difference in survival. At the same time, the actual result of the scan for this earlier time period group would predict a decreased survival for those patients with an abnormal study. For the later time period, it was hypothesized that the baseline study may not have been as routine, possibly as a result of the literature. Therefore, patients who were sent for a baseline scan may in fact have been a select group with a poorer prognosis. If this was true then survival analysis should reveal a decreased survival for those patients who were scanned as opposed to those who were not. Finally, if survival of those patients who were scanned in this later time period was analysed by the result of the scan, it can be suggested that there might be no difference in survival, possibly as a result of having an already poorer prognosis merely by being selected for the scan.

While the above hypotheses suggest that there may be an influence on survival of the baseline bone scan, it seemed prudent to corroborate this by assessing the scan's significance through the use of a statistical model. One of the most commonly used models for assessing factors related to survival is Cox's proportional hazards model. The data will be investigated with this model.

IV. METHODS AND RESULTS

1. Methods

The basic methodology used in this project was that of a chart review. All patients who were registered with the Saskatoon Cancer Centre during the years 1981/2 and 1988/9 with a diagnosis of breast cancer were identified. All available charts were reviewed with the information entered directly into a computer using Epi Info (53) (see Appendix 1). Patients were excluded if they were male, if the current diagnosis was not the first diagnosis of breast cancer and if the histology was not a carcinoma.

Because of the retrospective nature of the study, the number of variables that were examined was kept to a relatively small number. This was done to avoid the problem of missing data. The use of a baseline bone scan and its result were included as this was the focus of the study. Some variables were selected because of a prior association with the diagnosis or prognosis in breast cancer. These included a baseline alkaline phosphatase measurement, the patient's age at diagnosis, the initial stage, the histology of the tumour, the histologic grade, the presence or absence of unusual bone pain, the menopausal status at diagnosis, as well as the result of an estrogen receptor assay. In order to assess survival, the initial date of diagnosis was recorded as was the last date seen. The patient's status on this last date was noted. Other variables that were recorded were the chart number and whether or not the patient was on a well-defined protocol. While TNM information was available, a decision was made not to use these variables separately but rather to use only the initial stage which indirectly includes this information. This was done to try and keep the number of variables to a reasonable number.

Most of the data were categorical except for the patient's age. The use of a baseline bone scan was classified as yes or no while the result was noted as normal, abnormal or equivocal with regard to the presence of skeletal metastases. Alkaline phosphatase was coded as normal, abnormal or unavailable and the local laboratory's definition of normal was used. Similarly, estrogen receptor results were recorded as positive, negative or unavailable with equivocal results classified as negative. Histology and histologic grade were recorded from the pathologist's report. Menopausal status was classified into premenopausal, postmenopausal or uncertain. For the purposes of this study, perimenopausal women were placed into the postmenopausal category. Bone pain was noted as either present or absent. Protocol status was dichotomised into two categories which categorized patients as being on a well-defined treatment protocol or not. The initial stage was based on a combination of clinical and pathological assessment as it seemed none of the physicians would stage the patient without the results of an axillary node dissection. In order for a measurement to be considered baseline, it had to have been done within six weeks of the initial diagnosis. This date of initial diagnosis was taken to be the date a specimen was received by pathology for the first positive report. Finally, on the last date available or seen, it was recorded whether or not the patient was alive or dead.

Those patients who were diagnosed in 1981/2 were considered to be in group 1 while the patients from 1988/9 were considered to be in group 2. Although not recorded, the number of referring physicians was noted.

After reviewing all the charts, the data were then edited and checked for errors.

This allowed the data to be read into Stata (54). With this computer program it was possible to generate descriptive statistics of the data along with appropriate graphics. For the purposes of this analysis, Stages III and IV were combined. Survival analysis was very simply done in Stata using a Kaplan Meier technique. This resulted in the output of survival curves. Differences in survival were tested through the use of a logrank test. Proportional hazards modelling was also a simple task in Stata. Initially a large main effects model was created in order to identify significant variables. When the important variables for this data set were identified, further testing was performed which included looking at two way interaction terms as well as higher order terms to look for non-linearity. The level of significance was alpha = 0.05.

This study received ethical approval from the respective ethics committees of the University of Saskatchewan and the University of Calgary.

2. Results

There were a total of 409 patients registered by the Cancer Centre for the years 1981/2 (group 1). Of these, 36 patients were excluded. This left 373 eligible patients. Unfortunately, not all of the charts were available and only 361 were reviewed. However, this was still approximately 97% of the eligible charts. For group 2 (1988/9 patients), there were a total of 515 patients registered. Forty-seven patients were excluded which left a total of 468 eligible patients. The number of charts actually reviewed was 457 which was almost 98% of those eligible. The main reason for exclusion was that the diagnosis represented a second episode of breast cancer. A few

patients were excluded because they did not have the appropriate histology. With regard to the unavailable charts of eligible patients, the major reason for not seeing them was that the records department were unable to locate them within the time frame available. Although the exact number of referring physicians was not recorded, there were certainly more than twelve which suggests that this data are a fair representation of the usual practice in the population.

The first part of the analysis consisted of generating various descriptive statistics. Tabulations of the variables were performed with regard to group and stage. For the sake of brevity most of these results will be demonstrated through the use of tables.

Tables 1 and 2 describe the number of patients in each group by stage and age category where age is represented by five categories. For each of these categories the number of deaths as well as the number of censored observations is recorded.

The mean age for group 1 was 60.6 years with a standard deviation (SD) of 14.11 years. For group 2, the numbers in the corresponding categories were 62.4 years with a SD of 13.53 years. A t test of the mean age for the two groups was performed which yielded a t value of -1.94 which has a p value of 0.052.

Age was also examined by stage and group which also demonstrated no significant differences in age for any stage (Table 3). Specifically, there was no difference in age for stage II patients.

TABLE 1 Stage, Age, Deaths, Censored Observations: Group 1

Stage	Age	# patients	# deaths	# censored*
I	<35	7	1	1
	36-50	33	5	0
	51-60	42	6	0
	61-74	50	7	0
	75+	26	3	0
II	<35	5	0	0
	36-50	39	11	1
	51-60	37	14	1
	61-74	52	14	0
	75+	31	17	0
III/IV	<35	0	0	0
	36-50	8	5	0
	51-60	13	10	0
	61-74	12	7	0
	75+	6	5	0

* with respect to survival times less than 5 years

Table 2 Stage, Age, Deaths, Censored Observations: Group 2

Stage	Age	# patients	#deaths	#censored*
I	<35	2	0	1
	36-50	34	2	3
	51-60	41	4	3
	61-74	69	6	5
	75+	35	14	3
II	<35	9	3	0
	36-50	46	10	1
	51-60	38	7	2
	61-74	95	27	3
	75+	36	15	4
III/IV	<35	1	1	0
	36-50	7	5	0
	51-60	5	4	0
	61-74	22	16	0
	75+	17	15	0

* with respect to survival times less than 5 years

TABLE 3 Mean Age By Stage and Group

Stage	Group	# Obs	Mean Age (SD)	t value	p value
I	1	158	60.37 (14.25)	-1.50	0.14
	2	181	62.61 (13.90)		
II	1	164	60.58 (14.35)	-0.50	0.62
	2	224	61.30 (13.61)		
III/IV	1	31	61.28 (12.73)	-2.02	0.05
	2	31	66.88 (13.36)		

(#Obs=number of patients)

TABLE 4 Baseline Bone Scans By Stage and Group

Stage	Group	Bone Scan Done		chi2	p value
		No (%)	Yes (%)		
I	1	126 (79.7)	32 (20.3)	0.54	0.46
	2	150 (82.9)	31 (17.1)		
II	1	59 (36.0)	105 (64.0)	0.86	0.35
	2	91 (40.6)	133 (59.4)		
III/IV	1	9 (23.1)	30 (76.9)	0.47	0.49
	2	9 (17.3)	43 (82.7)		

From Table 4 one can see that there is no significant difference in the use of a baseline bone scan between groups for any stage. Although the statistical tests are not shown, there are significant differences among the stages for the use of scans.

TABLE 5 Bone Scan Results By Stage and Group

Stage	Group	Bone Scan Result			chi2	p value
		N (%)	Eq (%)	Ab (%)		
I	1	28 (88)	4 (13)	0 (0)	1.06	0.59
	2	26 (84)	4 (13)	1 (3)		
II	1	89 (85)	14 (13)	2 (2)	0.30	0.86
	2	111 (84)	18 (14)	4 (3)		
III/IV	1	18 (60)	6 (20)	6 (20)	4.01	0.14
	2	20 (46)	5 (12)	18 (42)		

(N=normal, Eq=equivocal, Ab=abnormal)

In Table 5 one can see that the distribution of scan results in any stage is similar between the groups. From this table it is obvious that most of the studies are normal in stages I-II and that in these stages most of the other studies are equivocal. It is only in Stage III/IV that there is a substantial number of abnormal studies.

TABLE 6 Alkaline Phosphatase Results By Stage and Group

Stage	Group	Alk Phos Result			chi2	p value
		N (%)	Ab (%)	Un (%)		
I	1	55 (35)	16 (10)	87 (55)	12.46	0.00
	2	79 (44)	3 (2)	99 (55)		
II	1	89 (54)	34 (21)	41 (25)	8.46	0.02
	2	133 (59)	23 (10)	68 (30)		
III/IV	1	20 (51)	16 (41)	3 (8)	3.45	0.18
	2	28 (54)	14 (27)	10 (19)		

(Alk Phos=alkaline phosphatase, Un=unavailable)

TABLE 7 Estrogen Receptor Results By Stage and Group

Stage	Group	Estrogen Receptor			chi2	p value
		+ (%)	- (%)	Un (%)		
I	1	62 (39)	44 (28)	52 (33)	7.70	0.02
	2	96 (53)	32 (18)	53 (29)		
II	1	83 (51)	46 (28)	35 (21)	1.32	0.52
	2	126 (56)	58 (26)	40 (18)		
III/IV	1	13 (33)	11 (28)	15 (38)	8.48	0.01
	2	15 (29)	4 (8)	33 (63)		

(+=positive, -=negative)

TABLE 8 Menopausal Status By Stage and Group

Stage	Group	Menopausal Status			chi2	p value
		Pre (%)	Post(%)	Un (%)		
I	1	36 (23)	111 (70)	11 (7)	3.27	0.20
	2	30 (17)	131 (72)	20 (11)		
II	1	39 (24)	120 (73)	5 (3)	0.76	0.68
	2	45 (20)	172 (77)	7 (3)		
III/IV	1	8 (20)	30 (77)	1 (3)	1.44	0.49
	2	6 (11)	44 (85)	2 (4)		

(Pre=premenopausal, Post=postmenopausal)

As can be seen from the tables 6-8, there are some statistically significant differences for alkaline phosphatase and estrogen receptor results between the groups. With regard to the alkaline phosphatase results, there may have been a change in the lab test between the two time periods which may account for some of the difference. Because of the large numbers of unavailable results, the meaning of these differences may be open to criticism. It should be kept in mind that unavailable in this context means that the test was not done within six weeks of initial diagnosis.

TABLE 9 Histology By Group

Group	Histology (%)					
	infilt duct	ca/adeno	lobular	medullary	tubular	other
1	56	32	3	2	1	6
2	68	19	5	1	1	6

(infilt duct=infiltrating duct, ca/adeno=carcinoma/adenocarcinoma)

Since the number of subjects in most of the histologic groups were small, statistical testing was not done. However, the data suggest that the two groups are similar with infiltrating duct being the dominant histological type (Table 9). With regard to protocol and histologic grade, there is not much information in these categories. There were only 22 patients identified as being on protocol and classified as follows: 3/339 in stage I, 17/388 in stage II, 2/62 in stage III and 0/29 in stage IV. As for histologic grading, only 29 reports included this information. Finally, unusual bone pain was recorded for only 18 patients in total. Except for one Stage I patient and three Stage III patients, the remainder were all Stage IV.

The foregoing description summarizes the findings for the entire 818 patients, and also provides the information for determining if there had been a change in the use of routine baseline bone scans for Stage I/II breast cancer patients between the two selected time periods.

However, it is the stage II individuals that are a primary focus of this study. An aim of this study is to assess the effectiveness of routine baseline bone scans in these patients by comparing 5-year mortality in those who were, and those who were not, scanned. This comparison was carried out separately in the two groups who were identified by the time of their initial diagnosis. Because there was some suggestion anecdotally that very young and very old patients were treated differently with regard to scanning, age was coded into three categories which consisted of ages 45 or less, ages 46 to 74, and ages 75 or more and cross tabulated against bone scan. Except for the comparison of the youngest age category versus the mid-age category in group one, there were statistically significant differences in the use of baseline bone scans. When the groups were combined there remained a significant difference in usage between the oldest age category and the mid-age one (Table 10). Because of this difference, the analysis of survival was done twice for each group. One analysis was for ages 46 to 74 while the second analysis was for all ages. The 46-74 age range was selected because it was felt that these individuals were representative of the "typical" patient. In addition, the proportion of patients scanned in this age range was similar for the two groups.

TABLE 10 Bone Scan By Age Category

Group	Age	Bone Scan Done (%)		chi2*	p value
		No	Yes		
1	min/45	26	74	0.12	0.73
	46-74	29	71		
	75/max	68	32	15.1	0.00
2	min/45	53	47	4.83	0.03
	46-74	33	67		
	75/max	58	42	7.69	0.01
1+2	min/45	42	58	2.27	0.13
	46-74	32	68		
	75/max	63	37	21.71	0.00

* for all tests, the comparison group is always the 46-74 age group. All comparisons are between 2 groups, e.g., min/45 vs. 46-74.

The results of the survival analysis are summarized by means of graphs and tables. Based on the criterion of whether or not a baseline bone scan was done, the survival curves for five years of follow-up do not show any statistical difference for either group 1 or group 2, especially for the 46-74 ages (see figures 1-6).

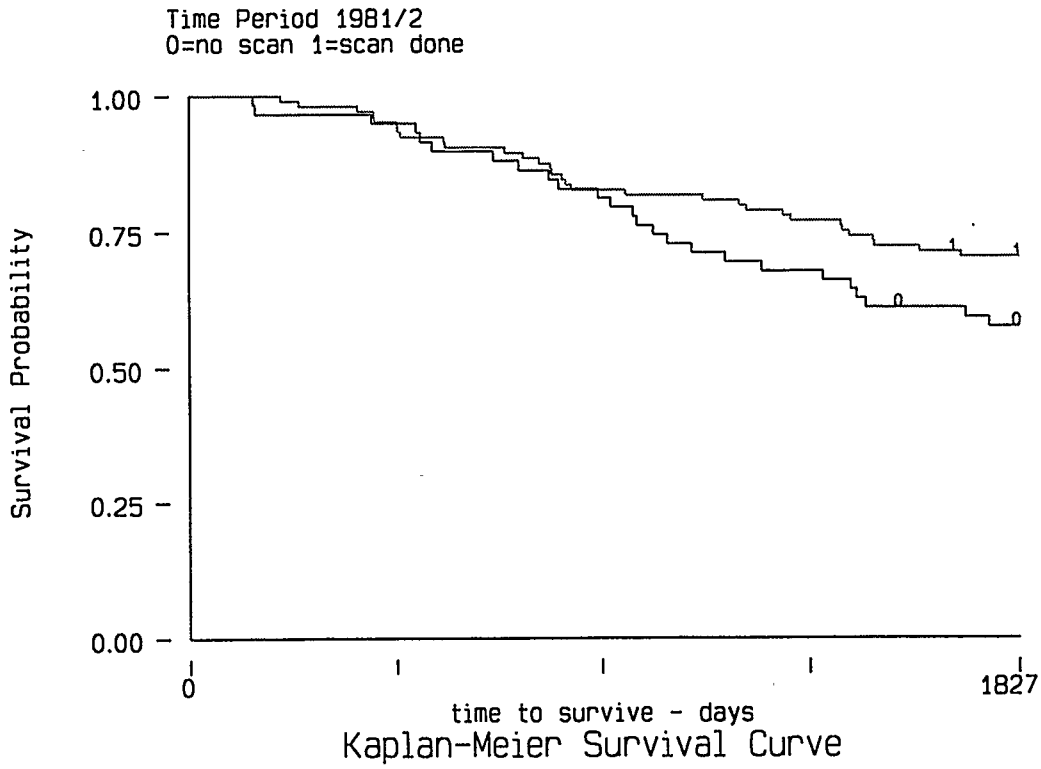

Figure 1 Survival curves, group 1, ages 46-74 based on performance of baseline bone scan. Logrank test result: chi2=1.31 p=0.25

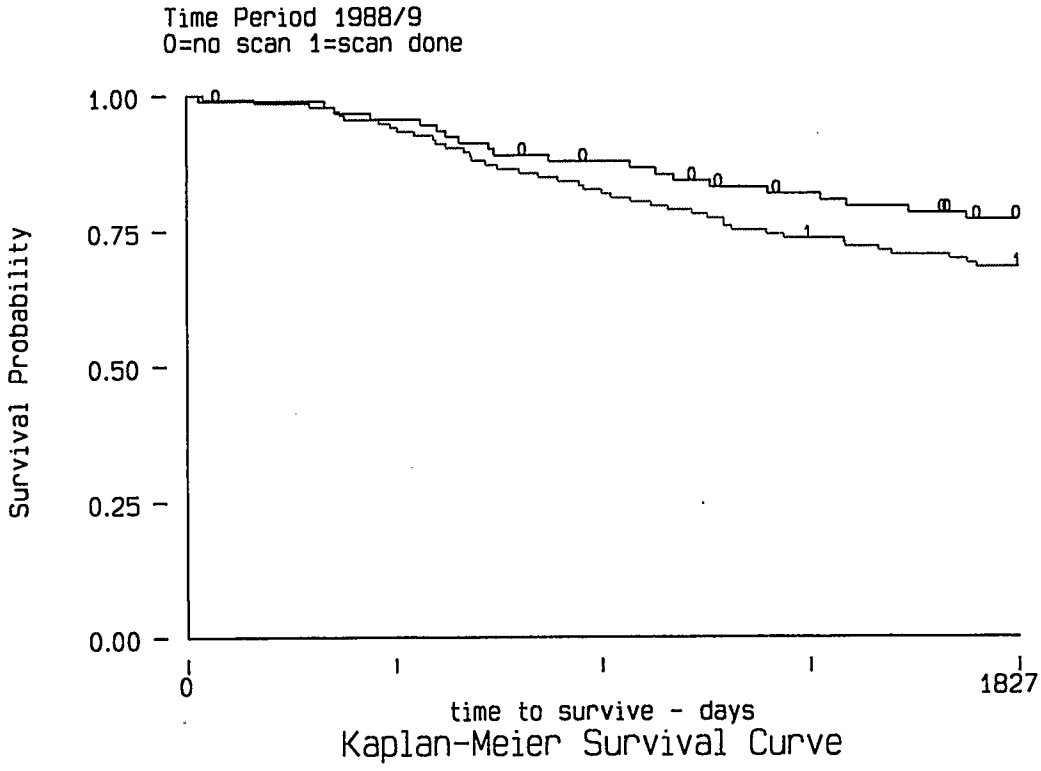

Figure 2 Survival curves, group 2, ages 46-74 based on performance of baseline bone scan. Logrank test result: chi2=0.22 p=0.63

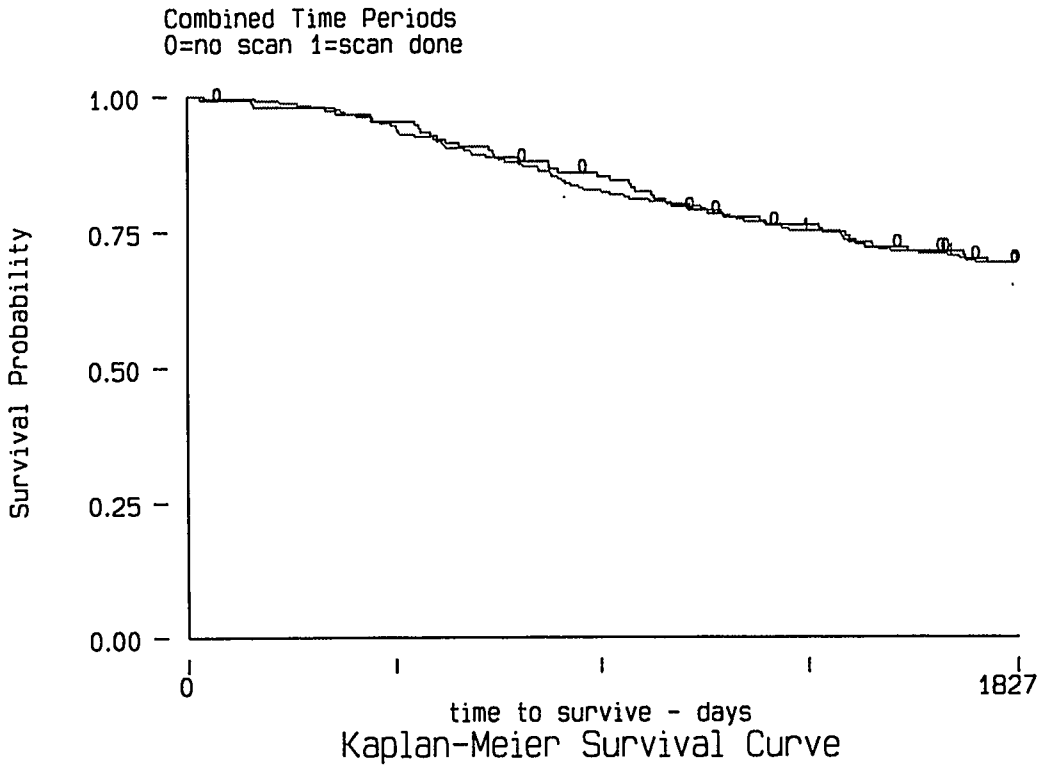

Figure 3 Survival curves, both groups, ages 46-74 based on performance of baseline bone scan. Logrank test result: chi2=0.13 p=0.72

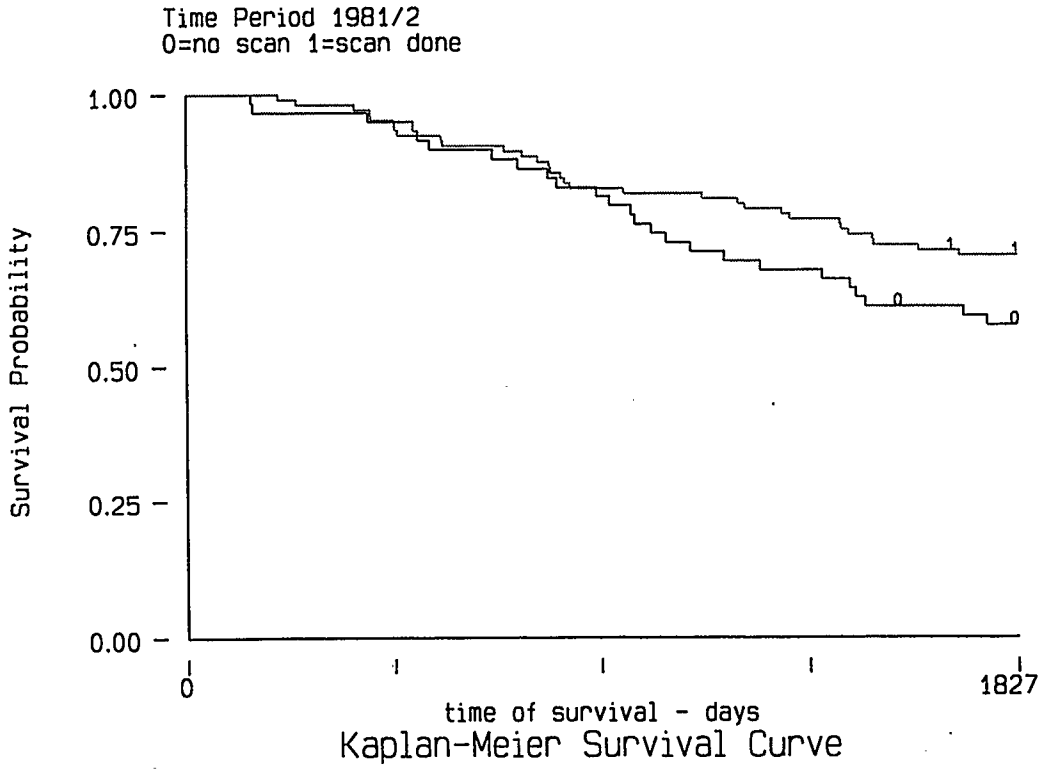

Figure 4 Survival curves, group 1, all ages based on performance of baseline bone scan. Logrank test result: chi2=2.59 p=0.11

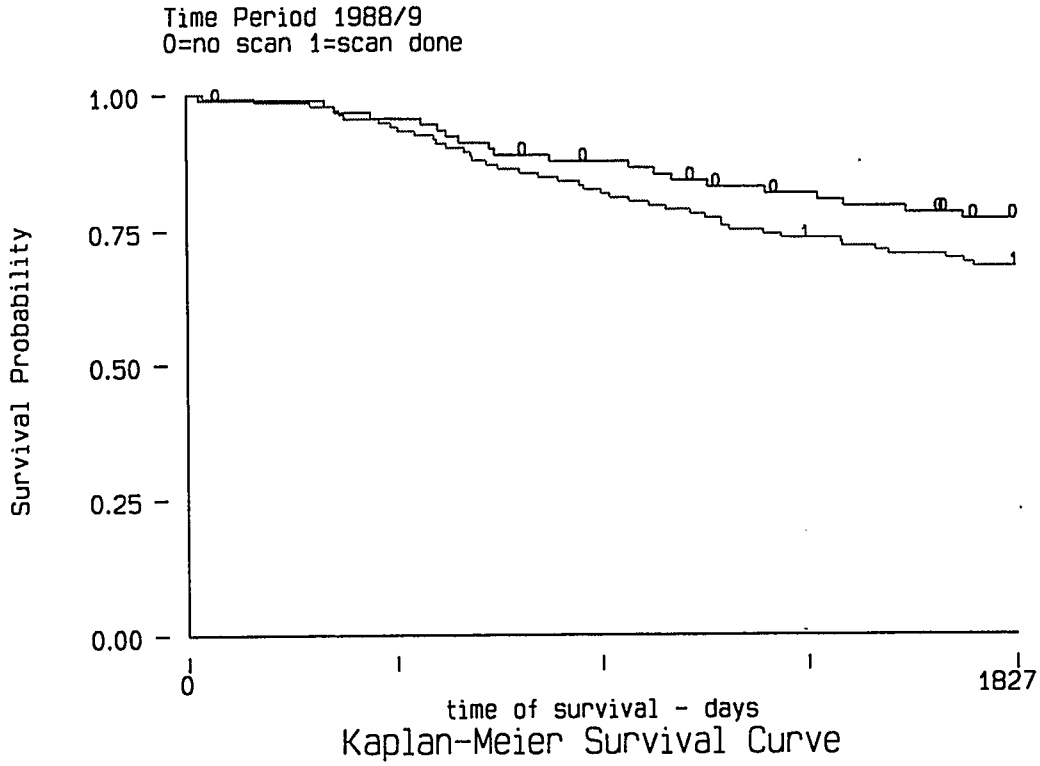

Figure 5 Survival curves, group 2, all ages based on performance of baseline bone scan. Logrank test result: chi2=1.98 p=0.16

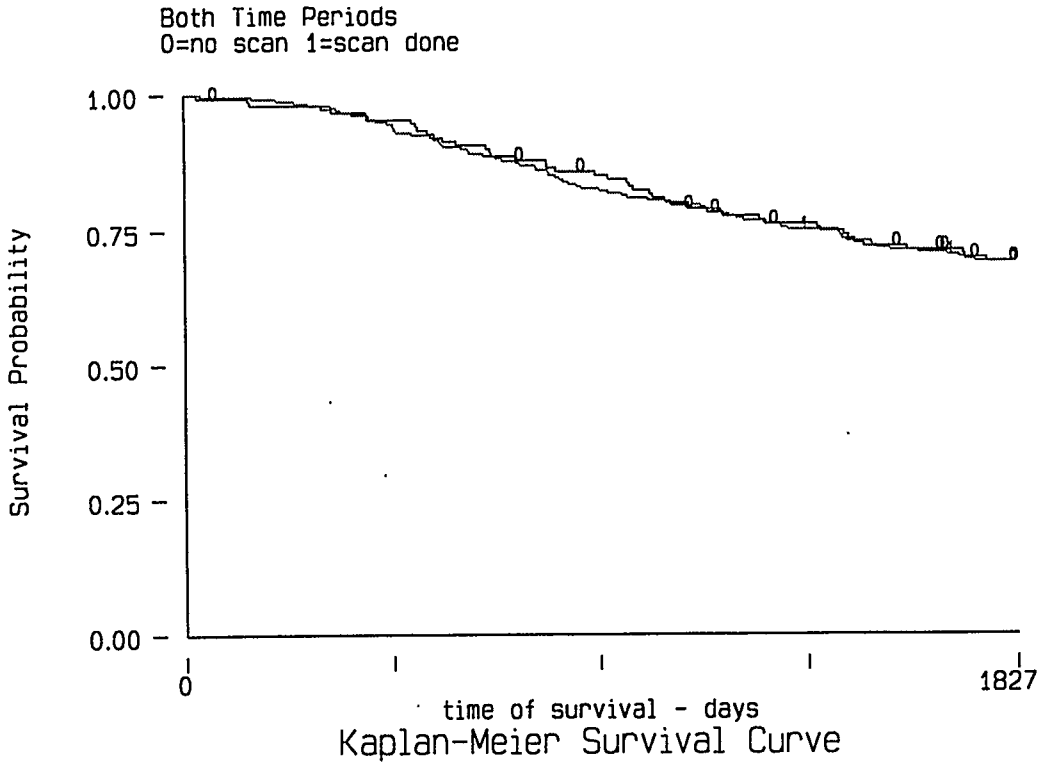

Figure 6 Survival curves, both groups, all ages based on performance of baseline bone scan. Logrank test result: chi2=0.00 p=0.97

When the survival curves are examined, one can see that there is a different trend between the groups. For group 1, those individuals who were scanned appear to have a greater survival than those who were not scanned while for group 2 this pattern is reversed. Neither of these differences is statistically significant. Because there is no significant difference in survival based on the use of a baseline bone scan, it seemed reasonable to perform the remaining analysis on the combined groups. This argument is strengthened if one examines the previous descriptive statistics which fail to demonstrate any differences between the two groups in Stage II, except for the alkaline phosphatase results. Even though the mean age is similar, it would be useful to look at the cumulative age distribution for the two groups. This is done in figure 7 where it can be seen that there is no substantial difference between the groups.

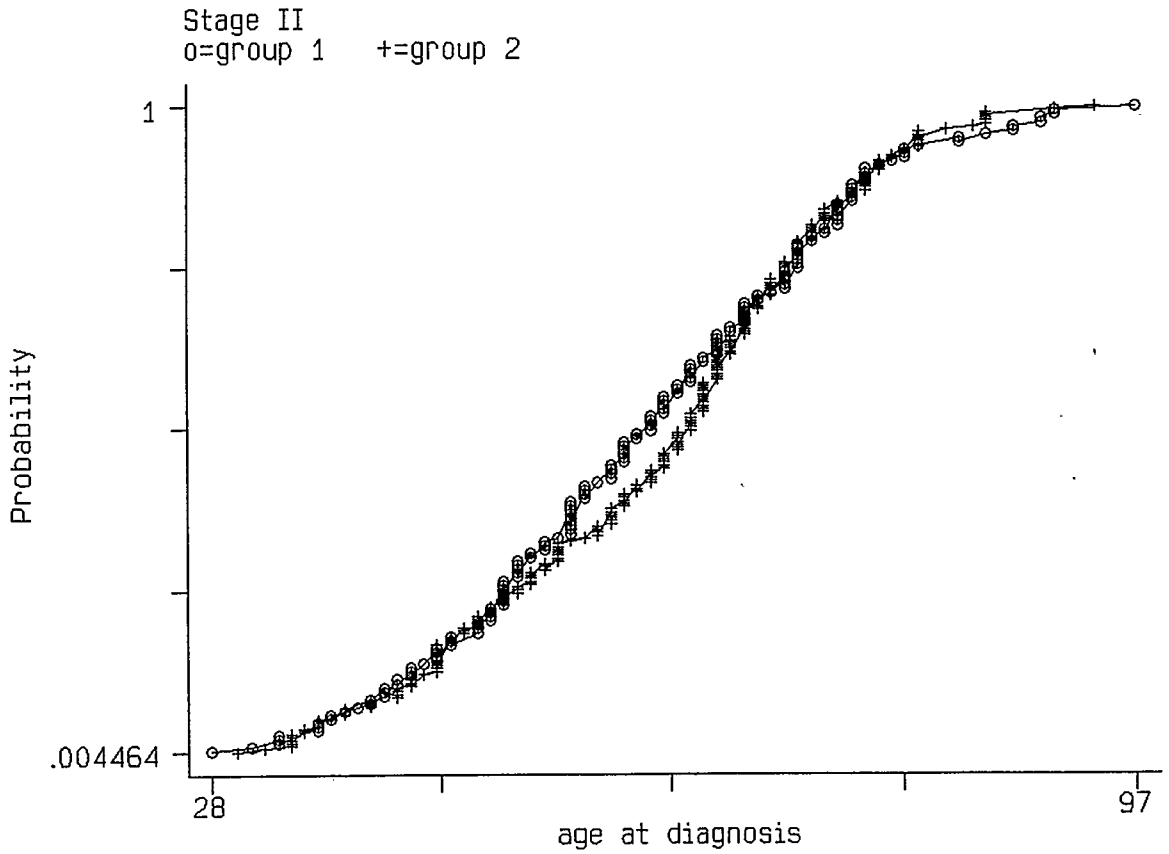

Figure 7 Cumulative age distribution for Stage II patients

For those individuals who had a baseline bone scan, survival was assessed in relation to the scan result. When all three outcomes of scan were used, it was apparent that those with an abnormal scan had a much poorer five-year survival (see figure 8). Those with an equivocal result were similar to those whose scan result was normal. Performing a logrank test revealed that there was no statistically significant difference between normal and equivocal and so these two categories were collapsed into one for further analysis. The results for this are shown in figure 9.

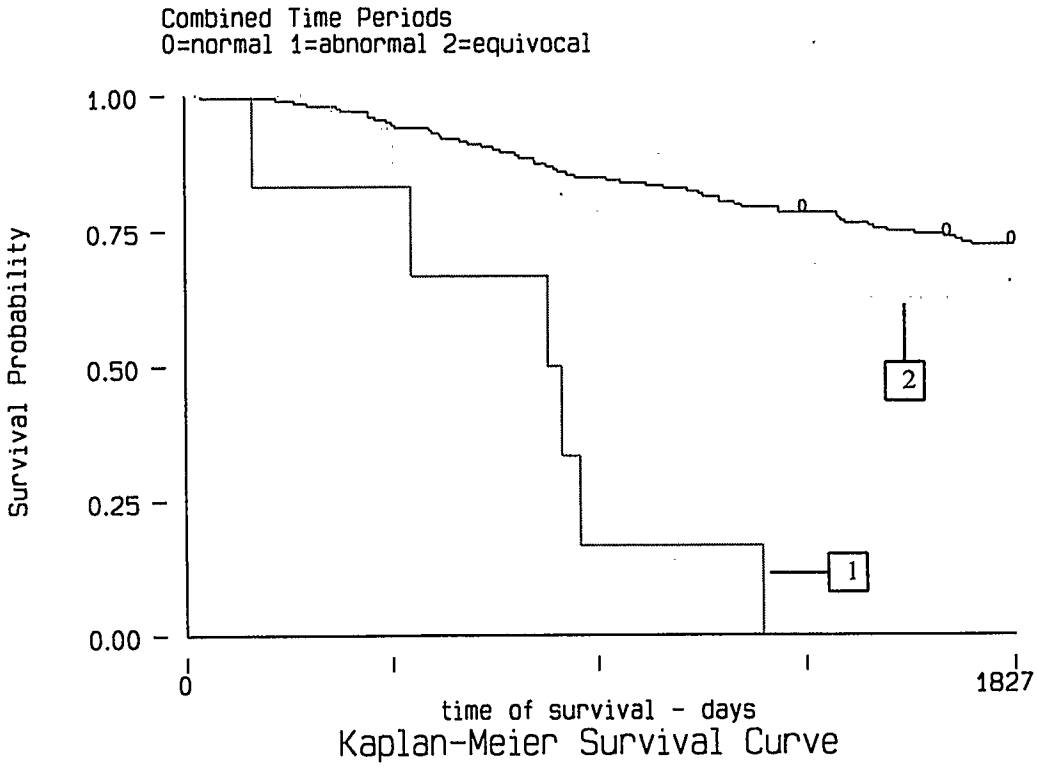

Figure 8 Survival curves by scan result - 3 levels.
Logrank test results: normal vs. equivocal chi2=1.53 p=0.22
normal vs. abnormal chi2=32 p=0.00
equivocal vs. abnormal chi2=13.4 p=0.00

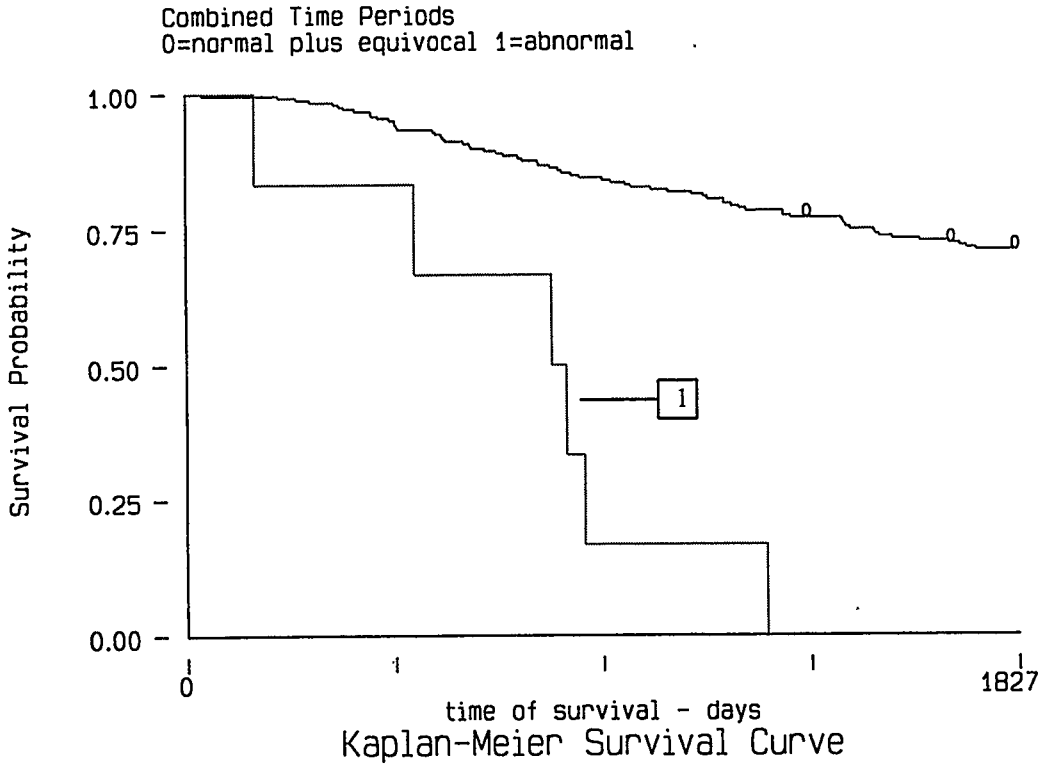

Figure 9 Survival curves by scan result - 2 levels
Logrank test result: chi2=30.4 p=0.00

The final part of the analysis was the examination of the stage II data in order to determine which variables had a significant impact on survival. A Cox's proportional hazard model was used. The initial fit incorporated all of the variables (Table 11). Further refinement through repeated application and evaluation of the model resulted in a final model (Table 12). For purposes of comparison, the output from a similar model without a term for menopausal status is also provided (Table 13). To help the interpretation of this output, it should be recalled that the hazard ratios do not represent absolute values but are comparisons to a baseline group. The categorical variables were numerically coded (see Appendix 1) and were introduced into the statistical model as dummy variables. For each category, the output in Tables 11-13 shows only those dummy variables other than the baseline one, i.e., each categorical variable can be represented by one less dummy variable than the number of categories. For example in Table 11, the receptor category has 3 categories which can be represented by 2 dummy variables. The baseline category is considered to have no increased hazard and therefore does not need to be shown.

TABLE 11 STATISTICAL OUTPUT FOR FULL MAIN EFFECTS MODEL

Variable	Hazard Ratio	Standard Error	z	Pr|z|	95% CI
age	**1.043**	**0.016**	**2.81**	**0.005**	**1.013-1.074**
postmen	**0.409**	**0.200**	**-1.83**	**0.067**	**0.157-1.064**
? menop	0.854	0.709	-0.19	0.849	0.168-4.350
scan done	1	.	.	.	.
abn scan	**3.613**	**1.984**	**0.02**	**0.019**	**1.232-10.599**
equiv scan	1.673	0.558	0.12	0.123	0.869-3.218
abn alkp	1.322	0.414	0.37	0.373	0.716-2.441
? alkp	1.596	0.636	1.17	0.241	0.731-3.488
neg recep	**1.672**	**0.485**	**1.77**	**0.077**	**0.946-2.954**
? recep	1.279	0.409	0.77	0.441	0.683-2.395
hist 2	6.99E-16	4.08E-18	0.00	1.000	.
hist 3	0.505	0.231	-1.49	0.136	0.206-1.240
hist 5	0.848	0.736	-0.19	0.849	0.155-4.644
infilt duct	**0.421**	**0.172**	**-2.12**	**0.034**	**0.189-0.937**
lobular	**0.218**	**0.178**	**-1.86**	**0.063**	**0.044-1.085**
hist 10	5.28E-16	1.77E-08	0	1.000	.
hist 11	1.04E-16	6.05E-09	0	1.000	.
hist 12	4.03E-16	2.35E-08	0	1.000	.
hist 14	3.59E-16	1.43E-08	0	1.000	.

(postmen=postmenopausal, ?menop=unavailable menopausal status, abn scan=abnormal bone scan, equiv=equivocal bone scan, abn alkp=abnormal alkaline phosphatase, ?alkp=unavailable alkaline phosphatase, neg recep=negative estrogen receptor, ?recep=unavailable estrogen receptor, hist #=different histologic types)

TABLE 12 STATISTICAL OUTPUT FOR FINAL MODEL

Variable	Hazard Ratio	Standard Error	z	Pr|z|	95% CI
age	1.075	0.023	3.35	0.001	1.03-1.121
abn scan	6.216	2.850	3.99	0.000	2.532-15.258
postmen	0.451	0.217	-1.66	0.097	0.176-1.156
neg recep	34.371	49.989	2.43	0.015	1.987-594.544
recep*age	0.953	0.021	-2.20	0.028	0.913-0.995

(recep*age=interaction term of age and estrogen receptor)

TABLE 13 STATISTICAL OUTPUT FOR MODEL WITHOUT MENOPAUSE

Variable	Hazard Ratio	Standard Error	z	Pr|z|	95% CI
age	1.059	0.218	2.80	0.005	1.017-1.103
abn scan	6.543	2.946	4.17	0.000	2.707-15.814
neg recep	44.732	69.696	2.44	0.015	2.110-948.137
recep*age	0.949	0.022	-2.21	0.027	0.906-0.994

V. DISCUSSION

1. Results

The first part of the study was to examine the use of routine baseline bone scans between two time periods which represented the beginning and the end of the 1980's. It was hypothesized that there would be a significant decrease in the use of these scans in stage I/II patients. Surprisingly, there was no significant change. In this case, it seemed worthwhile to estimate the power to identify a 25% change (55). For stage I, there was only a power of 22% to detect such a difference. However, in stage II, the estimate was 89% while for combined stages III/IV it was 61%. While one could question the value of a null finding in stage I, the more important issue is that any baseline scans were done at all as there has been a consistent and relatively longstanding recommendation for not doing such studies in this stage. In stage II, the lack of change is a fairly strong negative finding. This result could represent the conflicting opinion about the utility of routine baseline bone scans in this group even though the majority of reports in the literature seemed to favour not doing such studies. As for stages III/IV, it was not expected to find any change.

The main focus of the study was the survival of stage II patients. Survival analysis shows that the use of a baseline bone scan does not identify a group with reduced survival for either time period or when both time periods are combined. Again in the face of a null result, the question of power arises. Using a formula for power calculations for survival (56) (see Appendix 2), the power to detect a 25% change in survival for group one, group two and the combined group is 49%, 77% and 90%, respectively. The low

power in group one is a result of the lower number of patients in this group. Since there was no difference in survival and since the two groups were very similar, it seemed reasonable to use the combined group. This improves the power considerably.

When the survival of all stage II patients was analysed with respect to the result of the bone scan, there was a statistically significant reduction in survival for those patients who had an abnormal study result. However, it should be noted that this only represents a total of six individuals. These six cases are 2.5% of the total number of scans and only represent 8% of all of the deaths in stage II.

The choice of only two results for the bone scan was based on an examination of the analysis results. There was no statistical difference in survival between those individuals with an equivocal scan result and those with a normal scan result. This was also fairly clear from viewing the survival curves. It seemed logical to collapse these two into a single category. What was not shown was that the results for equivocal combined with abnormal did not change the final conclusions. The similarity of the results between the equivocal and normal groups is most likely due to the sensitive nature of the scan. The scan allows for detection of any area of increased bony remodelling which in the majority of patients in this study is more likely to reflect degenerative disease. If these scans were considered abnormal, they would most likely be false positives but further investigations would be necessary to confirm this.

With statistical modelling, it was obvious that the large main effects model was too large for the data set. However, it allowed for the selection of variables for a smaller, more parsimonious model. Repeated iterations of fitting a model led to a much smaller

model being proposed. The final model included age, menopausal status, bone scan result and estrogen receptor result. As mentioned previously, the scan result was collapsed into two categories. Both menopausal status and estrogen receptor were initially categorized into three categories. The actual number of unavailable menopausal results was small (twelve) and so these were dropped. However, there was a large number of receptor results which were unavailable and it did not seem appropriate to drop them. During analysis, it was clear that the hazard associated with this unavailable group was very similar to the negative receptor group. Thus, a choice was made to collapse the unavailable category into the negative receptor one.

The retention of menopausal status in the final model also warrants comment. It is obvious that the hazard associated with menopausal status is not significant at the 0.05 level. If one drops it from the model, the resultant estimates of hazard for the other variables only change slightly. From a statistical point of view it makes sense not to include menopausal status. However, the choice was made to keep it in the final model for historical reasons, as menopausal status is often noted as a risk factor for breast cancer (41-43).

Further refinement of the model was attempted by incorporating a higher order term for age to see if there was non-linearity in the data. This term was not statistically significant. Similarly, all two-way interactions of the variables were included in the model. Of these, only one appeared to be significant and was included in the final model. This was the interaction of age and estrogen receptor. Of importance is the fact that the term for whether a baseline scan was done or not is not included in the final model.

The final model suggests that an abnormal scan result is associated with an increased hazard when compared to a normal result. This effect is relatively strong as is the hazard associated with a negative estrogen receptor result. However, this latter hazard is not fixed and is modified by age. The overall effect is that the hazard or risk associated with a negative receptor result diminishes slightly with increasing age. The remaining factors have less striking hazards. It should not be surprising that there is a slight increase in the risk of dying with increasing age. The relatively increased risk associated with premenopausal status may in fact relate to the suggestion that women less than 35 years of age have a poorer prognosis (45). It is suggested that it is this correlation with age that makes the menopausal variable less significant in the fit. A final comment relates to the substantial change in the hazard ratio from the initial model to the final one. Part of the increase is likely related to the fact that there may be a correlation between various histologic types and receptor status so that when the histologic types are removed from the model any increase in hazard related to them may now be expressed in the receptor coefficient.. Also, the standard error is large which may suggest that there is some collinearity in the model. This could result in an elevation of the hazard ratio for the negative receptor status. Finally, the interaction of age and receptor status may reflect a lack of independence of these variables for when they are cross tabulated there is evidence that some cells in this table are nearly perfectly predicted (data not shown). This will also inflate the hazard associated with receptor status.

2. Strengths and Weaknesses

There is always a concern with bias in this type of study. Most of the information sought was fairly objective and so did not require much in the way of interpretation. Therefore, it seems unlikely that there was a systematic observer error introduced into the data. While the information is incomplete for some variables, this is likely a reflection of the actual practice rather than truly lost or missing data. Transcription error is possible but this should be a random type of error. Although death was the end point, it remains uncertain what was the cause of death as this information was not sought. The implication was that it was related to breast cancer but no attempt was made to determine the actual cause of death. Certainly, not all deaths were caused by breast cancer as there were a substantial number of women in this study who were elderly and who may have died of "natural" causes. However, if one examines Tables 1-2, it can be seen that there does not appear to be a disproportionately large number of deaths in the oldest age category. Another problem with this type of study design is that there was no random allocation of patients with regard to whether or not they had a bone scan. Therefore, the possibility of selection bias exists.

Another area of potential weakness relates to the adequacy of the database. The possibility exists that not all cases of breast cancer in the catchment area were captured. If this were true then this could influence the generalizability and validity of the study. On an anecdotal basis, it seemed to me that the data collection process followed by the Cancer Centre was very good. This can be seen by the very few numbers of patients lost to follow-up. Additionally, the numbers of patients actually registered was very close to

information may be somewhat misleading. Granted that the result of the bone scan appears to predict survival in that individuals with an abnormal study have a decreased survival, but is this worthwhile information, at least at the time of diagnosis? These patients are asymptomatic at the time and the scan result merely labels them as metastatic for a longer period of time. Does this knowledge have any impact on treatment choices and are these treatments effective? The patients with a normal result may be falsely reassured since the bulk of deaths in these Stage II patients occurs in patients with a normal scan - 55/73 or 75%. For those patients with an equivocal result, there is the increased anxiety associated with having to undergo further investigation during which time they are concerned that they have metastatic disease. With these concerns for cost and benefit, it seems appropriate to suggest that further evaluation may be useful. A cost-utility study may be the best way to answer these types of questions. It may also help to identify and/or measure the impacts on providers and patients.

Although this study may be considered only a preliminary investigation, the results of the survival analysis suggest that a routine baseline bone scan in Stage II patients may not be useful. This result is consistent with many of the reports in the literature. However, further investigation may allow for a better understanding of the use of such routine studies. Such a well designed prospective study (based on the guidelines proposed by Guyatt et al (32,33)) may serve as a template for the evaluation of other diagnostic imaging tests. It could provide some insight into the reasons why providers persist in ordering tests when there appears to be no effect on patient outcome. This information could be useful in relation to the development of clinical practice guidelines.

results of this study should be generalizeable, at least to the population of the Prairie provinces.

3. Implications

What are the implications of this study? It came as quite a surprise that there were still some baseline scans done in stage I patients let alone stage II. This leads one to wonder what the reasons were for ordering these studies. For the stage II patients there were obviously no differences in survival for having had a baseline scan. With regard to costs, it can be estimated that almost $60000 was spent to find six abnormal studies which is a cost of almost $10000 per abnormal study. If one adds in the costs of doing scans in stage I, the total amount spent increases to $75000 (these figures are based on the current cost of a bone scan being approximately $250, i.e., the total number of scans (301) X $250). These costs are for the bone scan alone. The true cost is likely far larger if one considers other costs. There are costs to patients such as travelling time, accommodation, time off work, etc. Some other system costs include the cost of further investigations such as X-rays as well as opportunity costs such as costs incurred by other patients who may have not had a study done because resources were expended elsewhere. These are only some of the most obvious costs.

It has been assumed that there is no immediate impact of the baseline scan on therapy. From this analysis, it has been determined that there is no impact on survival. Therefore, it may be that the impact of the scan is at the level of the provider. This testing could influence anxiety or confidence in either provider or patient. However, the

what was expected for the population served by the Centre.

It would seem appropriate to mention other possible types of error which can be seen in studies of survival, especially as related to screening or early diagnosis. While this study was not an assessment of a screening program, it is necessary to be aware of some of the bias that can be seen when survival is analysed. The main type of bias is that of lead time where there appears to be an improvement in survival but in reality there is only an increase in the amount of time the patient is known to have disease(57). One way of assessing this is to calculate the age specific mortality for those patients who do or do not have the test. If there is lead time bias, then there will be no difference in age specific mortality. The other kind of bias to be aware of is that of length where there may be a preference for identifying those patients with a better prognosis because the disease is " slower " with regard to its progression(57). It may be possible that the difference in survival that was found as a result of the scan result could be related to these biases.

A strength of this study is that the ascertainment of patients was very high with very few individuals lost to follow-up. These patients are therefore reflective of the underlying patient population since there was no sampling done for the years selected. As well, the number of referring physicians was relatively high which adds to the argument that this is a true reflection of actual practice. The number of patients was large enough to establish adequate power, especially for stage II. The results of the analyses were consistent between the survival analysis and the fitting of the proportional hazards model. In addition, the study was ethical in that there was respect for persons (no personal identification or risk of harm) and a potential for benefit to society. Finally, the

A comment related to the ordering of routine scans and patient age at diagnosis is necessary (Table 10). On an anecdotal basis, there was supposed to be a bias for ordering more scans for patients aged 45 years or less, presumably since these individuals have a poorer prognosis along with an inclination to be more "aggressive" with younger patients. No increase in scan use was found for this age group. What is of greater concern is that there were substantially fewer scans done on individuals aged 75 years or more. This possible age discrimination may warrant further investigation. If, in fact, it does exist then the reasons for its occurrence should be determined. Should the reason(s) be a result of a bias against the elderly, then this may need to be brought out into the open and discussed. This is especially important as this segment of the population is growing rapidly and will have a substantial impact on the health care system.

REFERENCES

1. Harris JR, Hellman S, Henderson IC, Kinne DW, editors. Breast Diseases. 2nd ed. Philadelphia: J.B.Lippincott, 1991.

2. Burkett FE, Scanlon EF, Garces RM, Khandekar JD. The value of bone scans in the management of patients with carcinoma of the breast. Surg Gynecol Obstet 1979;149(4):523-5.

3. Lindholm A, Lundell L, Martenson B, Thulin A. Skeletal scintigraphy in the initial assessment of women with breast cancer. Acta Chir Scand 1979;145(2):65-71.

4. Baker RR, Holmes ER III, Alderson PO, Khouri NF, Wagner HN Jr. An evaluation of bone scans as screening procedures for occult metastases in primary breast cancer. Ann Surg 1977;186(3):363-8.

5. Campbell DJ, Banks AJ, Oates GD. The value of preliminary bone scanning in staging and assessing the prognosis of breast cancer. Br J Surg 1976;63(10):811-6.

6. Charkes ND, Malmud LS, Caswell T, Goldman L, Hall J, Lauby V, et al. Preoperative bone scans: use in women with early breast cancer. JAMA 1975;233(6):516-8.

7. Citrin DL, Bessent RG, Greig WR, McKellar NJ, Furnival C, Blumgart LH. The application of the ^{99m}Tc phosphate bone scan to the study of breast cancer. Br J Surg 1975;62(3):201-4.

8. Galasko CSB. The significance of occult skeletal metastases, detected by skeletal scintigraphy, in patients with otherwise apparent "early" mammary carcinoma. Br J Surg 1975;62(9):694-6.

9. Lee YN. Bone scanning in patients with early breast carcinoma:should it be a routine staging procedure? Cancer 1981;47(3):486-95.

10. Evans DM, Wright DJ. The role of bone and liver scans in surveying patients with breast cancer for metastatic disease. Am Surg 1987;53(10):603-5.

11. Pedrazzini A, Gelber R, Isley M, Castiglione M, Goldhirsch A. First repeated bone scan in the observation of patients with operable breast cancer. J Clin Oncol 1986;4(3):389-94.

12. Butzelaar RM, van Donegen JA, de Graaf PW, van der Schoot JB. Bone scintigraphy in patients with operable breast cancer stages I and II. Final conclusion after five-year follow-up. Eur J Cancer Clin Oncol 1984;20(7):877-80.

13. Piffer S, Amichetti M, Valentini A. Skeletal scintigraphy and physical examination in the staging of early breast cancer. Acta Oncol 1988;27(1):21-4.

14. Khansur T, Haick A, Patel B, Balducci L, Vance R, Thigpen T. Evaluation of bone scan as a screening work-up in primary and local regional recurrence of breast cancer. Am J Clin Oncol 1987;10(2):167-79.

15. Strender LE, Lagergren C, Wallgren A, Liljevall S. Role of bone scans in the initial assessment of operable patients with breast carcinoma. Acta Radiol [Oncol] 1981;20(3):187-91.

16. Spencer GR, Khan M, Bird C, Seymour R, Brown TR, Collins CD. Is bone scanning of value in patients with breast cancer? Acta Chir Scand 1981;147(4):247-8.

17. Wilson GS, Rich MA, Brennan MJ. Evaluation of bone scan in preoperative clinical staging of breast cancer. Arch Surg 1980;115(4):415-9.

18. Feig SA. Diagnostic imaging in breast cancer. Crit Rev Diagn Imaging 1987;27(1):1-16.

19. Coleman RE, Rubens RD, Fogelman I. Reappraisal of the baseline bone scan in breast cancer. J Nucl Med 1988;29(6):1045-9.

20. Ahmed A, Glynne-Jones R, Ell PJ. Skeletal scintigraphy in carcinoma of the breast--a ten year retrospective study of 389 patients. Nucl Med Commun 1990;11(6):421-6.

21. Rosselli Del Turco M, Palli D, Cariddi A, Ciatto S, Pacini P, Distante V. Intensive diagnostic follow-up after treatment of primary breast cancer. A randomized trial. National Research Council Project on breast cancer follow-up. JAMA 1994;271(20):1593-7.

22. Sorensen MK, Willingham C, Broadwater JR. Managing care in breast cancer staging: routine bone scan is not indicated. [Abstract] Proc Annu Meet Am Soc Clin Oncol 1994;13:A81

23. Hannisdal E, Gundersen S, Kvaloy S, Lindegaard MW, Aas M, Finnanger AM, et al. Follow-up of breast cancer patients stage I-II: a baseline strategy. Eur J Cancer 1993;29A(7):992-7.

24. Kennedy H, Kennedy N, Barclay M, Horobin M. Cost efficiency of bone scans in breast cancer. Clin Oncol (R Coll Radiol) 1991;3(2):73-7.

25. Schunemann H, Langecker PJ, Ellgas W, Leonhardt A, Merkl H. Value of bone scanning in the follow-up of breast cancer patients. A study of 1000 patients. J Cancer Res Clin Oncol 1990;116(5):486-91.

26. Pace BW, Tinker MA. Follow-up of patients with breast cancer. Clin Obstet Gynecol 1994;37(4):998-1002.

27. Kinne DW. Staging and follow-up of breast cancer patients. Cancer 1991;67(4 Suppl):1196-8.

28. Jacobson AF, Stomper PC, Jochelson MS, Ascoli DM, Henderson IC, Kaplan WD. Association between number and sites of new bone scan abnormalities and presence of skeletal metastases in patients with breast cancer. J Nucl Med 1990;31(4):387-92.

29. Al-Nahhas AM, McCready VR. The role of radionuclide studies in breast cancer. Nucl Med Commun 1993;14:192-6.

30. Hendee WR. Technology assessment in medicine: methods, status and trends. Med Prog Technol 1991;17:69-75.

31. Kent DL, Larson EB. Disease, level of impact, and quality of research methods. Three dimensions of clinical efficacy assessment applied to magnetic resonance imaging. Invest Radiol 1992;27(3):245-54.

32. Guyatt GH, Tugwell PX, Feeny DH, Haynes RB, Drummond M. A framework for clinical evaluation of diagnostic technologies. Can Med Assoc J 1986;134:587-94.

33. Guyatt GH, Drummond M, Feeny DH, Tugwell PX, Stoddart G, Haynes RB, et al. Guidelines for the clinical and economic evaluation of health care technologies. Soc Sci Med 1986;22(4):393-408.

34. Goin JE, Hermann GA. The clinical efficacy of diagnostic imaging evaluation studies. Problems, paradigms, and prescriptions. Invest Radiol 1991;26(5):507-11.

35. Wells PN, Garrett JA, Jackson PC. Assessment criteria for diagnostic imaging technologies. Med Prog Technol 1991;17(2):93-101.

36. Chalmers TC, Hewett P, Reitman D, Sacks HS. Selection and evaluation of empirical research in technology assessment. Int J Technol Assess Health Care 1989;5(4):521-36.

37. Sheps SB, Schechter MT. The assessment of diagnostic tests. A survey of current medical research. JAMA 1984;252:2418-22.

38. Jaeschke R, Sackett DL. Research methods for obtaining primary evidence. Int J Technol Assess Health Care 1989;5(4):503-19.

39. Torrance GW, Feeny D. Utilities and quality-adjusted life years. Int J Technol Assess Health Care 1989;5(4):559-75.

40. Parker SL, Tong T, Bolden S, Wingo PA. Cancer statistics, 1996. CA Cancer J Clin 1996;46(1):5-27.

41. Colditz GA. Epidemiology of breast cancer. Findings from the Nurses' Health Study. Cancer 1993;71(4 Suppl):1480-9.

42. McPherson K, Steel CM, Dixon JM. ABC of Breast Diseases: Breast cancer - epidemiology, risk factors, and genetics. BMJ 1994;309: 1003-6.

43. Henderson IC. Risk factors for breast cancer development. Cancer 1993;71(6 Suppl):2127-40.

44. Wold LE, Ingle JN, Pisansky TM, Johnson RE, Donohue JH. Prognostic factors for patients with carcinoma of the breast. Mayo Clin Proc 1995;70(7):678-9.

45. Miller WR, Ellis IO, Sainsbury JRC, Dixon JM. ABC of Breast Diseases: Prognostic factors. BMJ 1994;309: 1573-6.

46. Berg JW, Hutter RVP. Breast cancer. Cancer 1995;75(1 Suppl):257-69.

47. International Union Against Cancer. TNM Classification of Malignant Tumours. 4th ed. New York: Springer-Verlag, 1987.

48. American Joint Committee on Cancer. Manual for Staging of Cancer. 3rd ed. Philadelphia: J.B.Lippincott, 1988.

49. Fordham EW, Ali A.Skeletal imaging in malignant disease. In: Gottschalk A, Hoffer PB, Potchen EJ, editors. Diagnostic Nuclear Medicine. 2nd ed. Baltimore: Williams & Wilkins, 1988:1011-32.

50. Lee ET. Nonparametric methods of estimating survival functions. In: Statistical methods for survival data analysis. Belmont: Lifetime Learning Publications, 1980: 75-121.

51. Lee ET. Nonparametric methods for comparing survival distributions. In: Statistical methods for survival analysis. Belmont: Lifetime Learning Publications, 1980:122-56.

52. Lee ET.Identification of prognostic factors related to survival time. In: Statistical methods for survival data analysis. Belmont: Lifetime Learning Publications, 1980: 298-337.

53. Dean AG, Dean JA, Coulombier D, Brendel KA, Smith DC, Burton AH, et al. Epi Info, version 6: a word processing, database, and statistics program for epidemiology on microcomputers. Atlanta: Centers for Disease Control and Prevention, 1994.

54. StataCorp. Stata statistical software: release 4.0. College Station: Stata Corporation, 1995.

55. Kelsey JL, Thompson WD, Evans AS. Methods in observational epidemiology. New York: Oxford University Press, 1986.

56. Lachin JM, Foulkes MA. Evaluation of sample size and power for analyses of survival with allowance for non-uniform patient entry, losses to follow-up, noncompliance, and stratification. Biometrics 1986;42(3):507-19.

57. Morrison AS. Screening in chronic disease. New York: Oxford University Press, 1992.

APPENDIX 1 Data Input for Study using Epi Info (53)

ID <idnum >

STAGE#

DATE {DIAGNOSIS} <dd/mm/yy>

{BONE SCAN} DONE <Y>

RESULT #

0=NORMAL 1= ABNORMAL 2=EQUIVOCAL

{ALK PHOS} RESULT#

0=NORMAL 1=ABNORMAL 2=UNAVAILABLE

AGE##

{BONE PAIN}_____

{HISTOLOGY} _____

HISTOLOGIC GRADE#

RECEPTOR#

0=POSITIVE 1=NEGATIVE 2=UNAVAILABLE

{MENOPAUSAL}#

0=PRE 1=POST 2=UNAVAILABE

DATE OF LAST {VISIT} <dd/mm/yy>

STATUS LAST VISIT

0=ALIVE 1=DEAD

ON PROTOCOL#

0=NO 1=YES

APPENDIX 2 Formula for Calculating Power/Sample Size for Survival Analysis (56)

$$\sqrt{N}\ |\lambda_e - \lambda_c| = (Z_{\alpha/2} + Z_\beta)\ [\phi(\lambda_e)Q_e^{-1} + \phi(\lambda_c)Q_c^{-1}]^{1/2}$$

$$\phi(\lambda) = \lambda^2[1 - \frac{e^{-\lambda(T-R)} - e^{-\lambda T}}{\lambda R}^{-1}]$$

where N=total sample size

$Q_e = n_e/N$ n_e*=number exposed*

$Q_c = n_c/N$ n_c*=number of controls*

R=2 (accrural)

T=5 (duration of follow-up)

Note that T and R are measured in years