

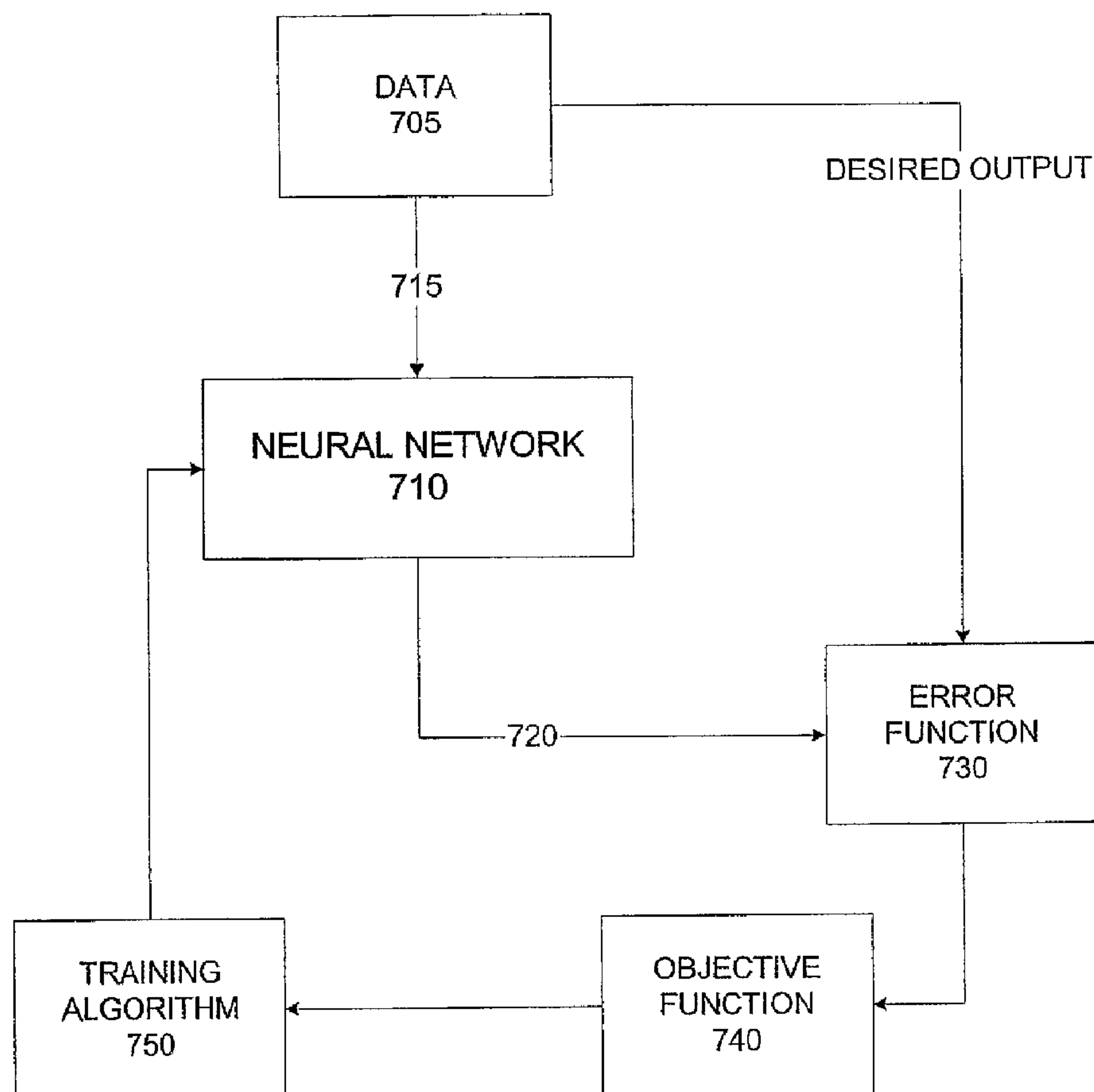


(86) Date de dépôt PCT/PCT Filing Date: 2005/02/25
(87) Date publication PCT/PCT Publication Date: 2005/09/15
(85) Entrée phase nationale/National Entry: 2006/08/25
(86) N° demande PCT/PCT Application No.: US 2005/006241
(87) N° publication PCT/PCT Publication No.: 2005/086068
(30) Priorités/Priorities: 2004/02/27 (US60/548,322);
2004/06/04 (US60/577,051)

(51) Cl.Int./Int.Cl. *G06F 19/00* (2006.01)
(71) Demandeur/Applicant:
AUREON BIOSCIENCES CORPORATION, US
(72) Inventeurs/Inventors:
SAIDI, OLIVIER, US;
VERBEL, DAVID A., US;
YAN, LIAN, US
(74) Agent: RIDOUT & MAYBEE LLP

(54) Titre : PROCÉDES ET SYSTÈMES DE PRÉVISION D'UN ÉVÉNEMENT

(54) Title: METHODS AND SYSTEMS FOR PREDICTING OCCURRENCE OF AN EVENT



(57) Abrégé/Abstract:

Embodiments of the present invention are directed to methods and systems for training a neural network having weighted connections for classification of data, as well as embodiments corresponding to the use of such a neural network for the

(57) Abrégé(suite)/Abstract(continued):

classification of data, including, for example, prediction of an event (e.g., disease). The method may include inputting input training data into the neural network, processing, by the neural network, the input training data to produce an output, determining an error between the output and a desired output corresponding to the input training data, rating the performance neural network using an objective function, wherein the objective function comprises a function C substantially in accordance with an approximation of the concordance index and adapting the weighted connections of the neural network based upon results of the objective function.

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property
Organization
International Bureau



(43) International Publication Date
15 September 2005 (15.09.2005)

PCT

(10) International Publication Number
WO 2005/086068 A2

(51) International Patent Classification⁷: **G06F 19/00**

(21) International Application Number:
PCT/US2005/006241

(22) International Filing Date: 25 February 2005 (25.02.2005)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/548,322 27 February 2004 (27.02.2004) US
60/577,051 4 June 2004 (04.06.2004) US

(71) Applicant (for all designated States except US): **AUREON BIOSCIENCES CORPORATION** [US/US]; 28 Wells Avenue, Building #3, Yonkers, NY 10701 (US).

(72) Inventors; and

(75) Inventors/Applicants (for US only): **SAIDI, Olivier**

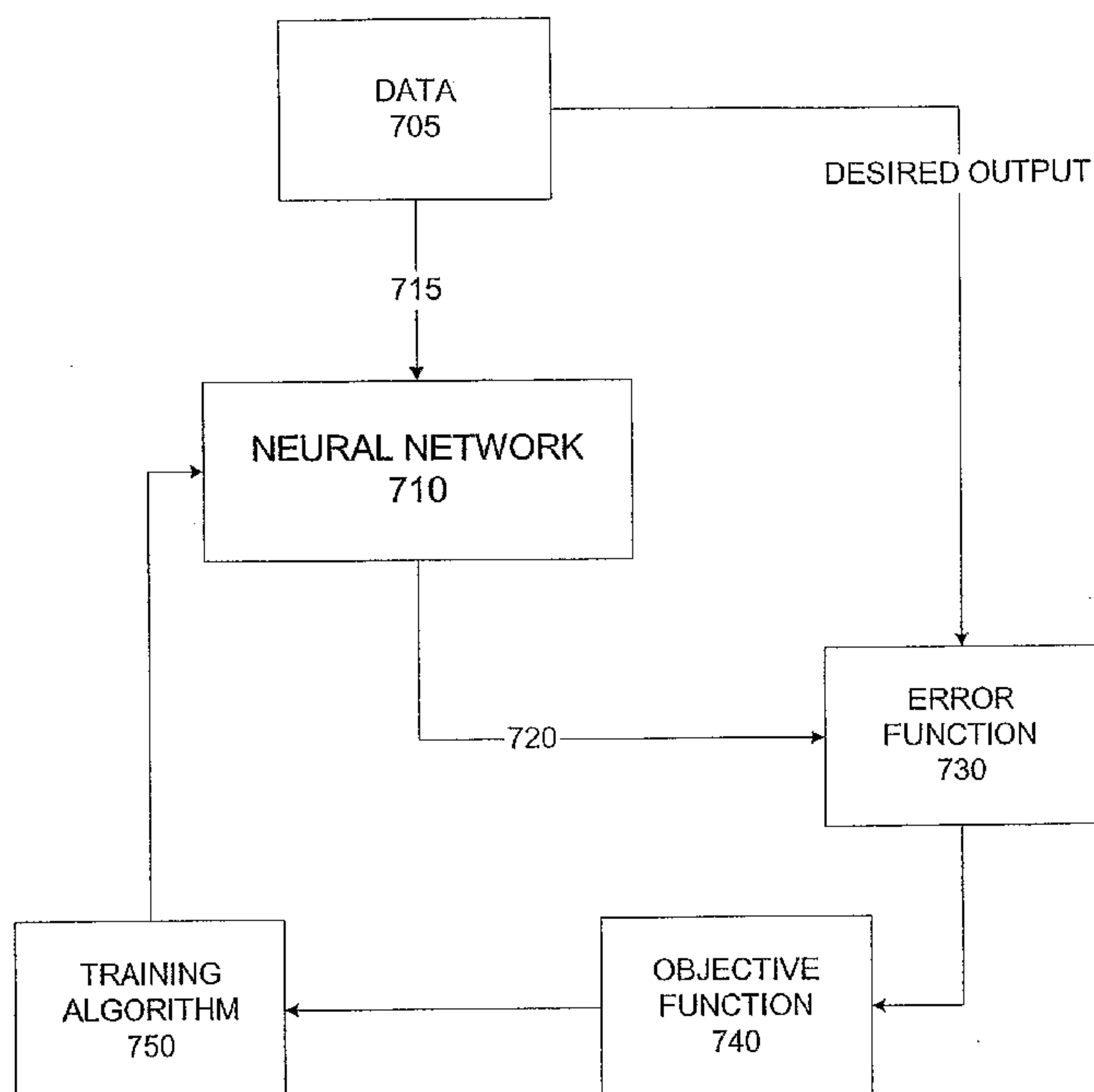
[FR/US]; 33 West Putnam Avenue, Apt #2, Greenwich, CT 06830 (US). **VERBEL, David, A.** [US/US]; 412 E. 46th Street, Apt 5B, New York, NY 10021 (US). **YAN, Lian** [CN/US]; 318 Willoway Street, Chester Springs, PA 19425 (US).

(74) Agent: **SNELL, Peter, E.**; Mintz, Levin, Cohn, Ferris, Glovsky and Popeo, P.C., Chrysler Center, 666 Third Avenue, New York, NY 10017 (US).

(81) Designated States (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NA, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, SM, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, YU, ZA, ZM, ZW.

[Continued on next page]

(54) Title: METHODS AND SYSTEMS FOR PREDICTING OCCURRENCE OF AN EVENT



(57) Abstract: Embodiments of the present invention are directed to methods and systems for training a neural network having weighted connections for classification of data, as well as embodiments corresponding to the use of such a neural network for the classification of data, including, for example, prediction of an event (e.g., disease). The method may include inputting input training data into the neural network, processing, by the neural network, the input training data to produce an output, determining an error between the output and a desired output corresponding to the input training data, rating the performance neural network using an objective function, wherein the objective function comprises a function C substantially in accordance with an approximation of the concordance index and adapting the weighted connections of the neural network based upon results of the objective function.

WO 2005/086068 A2

WO 2005/086068 A2

(84) Designated States (*unless otherwise indicated, for every kind of regional protection available*): ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IS, IT, LT, LU, MC, NL, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:

— *without international search report and to be republished upon receipt of that report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

METHODS AND SYSTEMS FOR PREDICTING OCCURRENCE OF AN EVENT

5 FIELD OF THE INVENTION

Embodiments of the invention relate to event occurrence analysis and in particular eventual event occurrence analysis that can make use of both censored data and uncensored data for predicting a recurrence of a disease, for example, including cancer.

10 BACKGROUND OF THE INVENTION

There are many instances in which it is desirable to predict the likelihood of an event occurring within a certain amount of time or the amount of time until an event is likely to occur. Indeed, event prediction type data, including censored data, is one of the most common types of data used in bioscience (for example). Predicting the occurrence of an event can help people plan for the occurrence of the event. For example, it is desirable to predict the time to recurrence of diseases or other health issues, such as cancer, or environmental events (e.g., earthquakes, hurricanes).

Censored data comprises incomplete data in which it is unknown when an event occurred/recurred. For example, in training a model to predict the recurrence of cancer in a patient, the training data would preferably include censored data comprising patient data in which no recurrence of cancer came about in particular patients. This is because most medical data includes both censored and non-censored data, and increasing the amount of data available to train a predictive model can increase the reliability and predictive power of the model. Censored data indicates whether the outcome under observation, e.g., recurrence of cancer, has occurred (for example) within a patient's follow-up visit time: if the recurrence of cancer has not been observed at a patient's follow-up visit, this patient's data is censored. In predicting recurrence of cancer (in patients who have been considered cured, for example), data for many patients may be censored. Such censored observation provides incomplete information about the outcome, since the event may eventually occur after the

follow-up visit, which should be taken into account by a predictive model. However, the current most accurate learning models, particularly machine learning techniques involving neural networks and support vector machines, do not make use of such censored data.

It would be highly desirable when training a predictive model to have as much data
5 from as many sources as possible. Thus, for example, for disease related events, it is generally desirable to have data from as many patients as possible, and as much data from each patient as possible. With such data, however, come difficulties in how to process censored data.

Typically, traditional survival analysis, e.g., the Cox proportional hazards model,
10 uses censored data. However, in general, the reliability of the Cox model deteriorates if the number of features is greater than the number of events divided by 10 or 20 [1]. For example, in one study included as an example for the present invention, the dataset consisted of only 130 patients, each of which was represented by a vector of 25 features. For such data, the Cox model could not be successfully derived from this dataset until the feature
15 dimensionality was reduced.

Neural networks have been shown to be able to outperform traditional statistical models, due to neural networks' capacity to model nonlinearities. However, in order to be successful, a neural network typically requires a large number of samples in the training set. Generally, several approaches have been used in applying survival data in neural networks.
20 One approach is to model the hazard or survival function as a neural network structure. For example, constructing the survival curve by a hazard function modeled by a neural network, for which the i th output is the estimated hazard at the discretized time interval i . Others have used the discretized time interval as an additional input to a neural network to model the survival probability. Still others have used several separately trained networks, each used to
25 model the hazard function at a different time interval.

Still, in order to effectively use machine learning algorithms, treatment of censored data is crucial. Simply omitting the censored observations or treating them as non-recurring samples bias the resulting model and, thus, should be avoided. Kaplan-Meier estimates of event probability have been used as target values during training for patients who had short

follow-up times and did not have the event recurred. Although this algorithm takes into account, to some extent, both follow-up time and censoring, it still fails to make complete use of available information. For instance, it treats two recurred patients as the same regardless of their survival time.

5

SUMMARY OF THE INVENTION

The present invention presents systems and methods to improve prediction of an occurrence of an event that may use both censored and non-censored data. The treatment of censored data in the present invention allows for the effective use of machine learning
10 algorithms, e.g., neural networks, for the analysis of survival data to produce an effective predictive model, for example. In particular, some embodiments of the present invention may be used to indicate how likely an event will eventually occur.

In particular, some embodiments of the present invention are directed to supervised learning models which use an objective function to improve the predictive accuracy of the
15 model. One such novel objective function may be an approximation of the concordance index (CI). Using an approximation to the CI allows the model to make use of the information from both censored and non-censored observations, thereby reducing event occurrence prediction into simpler predication/classification problem. Accordingly, such an approximation to the CI for the objective function may be a derivative of the CI.

20 Embodiments of the present invention have particular application to the prediction of the occurrence and, in particular, the recurrence of cancer after a patient has been initially cured (e.g., following removal and/or treatment of the cancer). Being able to make a more precise prediction on cancer recurrence allows for a more effective follow-up therapy to be administered to the patient. Accordingly, embodiments of the invention may be able to
25 determine a prognostic estimate for an individual patient. Comparing embodiments of the present invention with the traditional Cox proportional hazards model (and several other algorithms based on neural networks and support vector machines, see below), the present invention may achieve improvement in being able to identify high-risk and low-risk groups of patients.

Accordingly, in a first group of embodiments of the present invention, a supervised model (e.g., predictive, prognostic) for a neural network is provided and may include one or more of the following components: a neural network having weighted connections, an input for receiving training data and an output, an error module for determining an error between
5 output of the neural network and a desired output from the training data, an objective function for providing a rating of the performance of the model, the objective function comprising a function C substantially in accordance with a derivative of the concordance index and a training algorithm for adapting the weighted connections of the neural network in accordance with the results of the objective function. The input training data may include
10 one or more features (clinical, molecular, image) for each of a plurality of known patients. Output data may include, for example, a predictive or prognostic determination for each of the patients.

In a second group of embodiments of the invention, a method for predicting recurrence of cancer in a patient is provided and may comprise estimating the probability that
15 cancer will recur within a shorter period of time in a patient with a higher prognostic score than a recurrence time in a patient with a lower prognostic score. Estimating may include conducting pair-wise comparisons between prognostic scores for patients i and j, using a neural network trained using an objective function comprising a function C substantially in accordance with a derivative of the concordance index.

20 In a third group of embodiments of the present invention, a method for training a neural network having weighted connections for classification of data is provided and may include inputting input training data into the neural network, processing, by the neural network, the input training data to produce an output, determining an error between the output and a desired output corresponding to the input training data, rating the performance
25 neural network using an objective function, wherein the objective function comprises a function C substantially in accordance with a derivative of the concordance index and adapting the weighted connections of the neural network based upon results of the objective function.

In any of the above groups of embodiments, the function C may comprise

$$C = \frac{\sum_{(i,j) \in \Omega} R(\hat{t}_i, \hat{t}_j)}{|\Omega|},$$

wherein Ω comprises pairs of data (i, j) meeting one or more predetermined conditions, or

$$C = \frac{\sum_{(i,j) \in \Omega} (\hat{t}_i - \hat{t}_j) \cdot R(\hat{t}_i - \hat{t}_j)}{D},$$

where

$$D = \sum_{(i,j) \in \Omega} (\hat{t}_i - \hat{t}_j),$$

may be a normalization factor, and Ω may comprise pairs of data (i, j) meeting one or more predetermined conditions. In the above equation, t_i and t_j represent estimates by the neural network of how long it will take for patients i and j to recur, respectively.

Moreover, any of the above groups of embodiments may be used, or modified to be used, to determine a prediction of the recurrence of disease (e.g., cancer). In such cases, the pairs of data may comprise patients $\{i, j\}$ and the predetermined conditions may comprise at least: both patients i and j have experienced recurrence and the recurrence time t_i of patient i is shorter than the recurrence time t_j of patient j ; or only patient i has experienced recurrence and t_i is shorter than patient j 's follow-up visit time t_j .

Other groups of embodiments may further include computer systems, computer readable media and computer application programs for carrying out any one or more of particular embodiments of the groups of embodiments listed above.

The present invention will become even clearer with reference to the attached figures, brief descriptions of which are provided below, and the following detailed description.

BRIEF DESCRIPTION OF THE DRAWINGS

Fig. 1 is a chart illustrating a comparison of $R(\hat{t}_i, \hat{t}_j)$ with $I(\hat{t}_i, \hat{t}_j)$; the horizontal axis being $\hat{t}_i - \hat{t}_j$, $\gamma = 0.1$ and $n = 2$ in $R(\hat{t}_i, \hat{t}_j)$.

Fig. 2 is a chart illustrating a comparison of survival curves for high-risk and low-risk patient groups between NN_{ci} and NN_{km} .

Fig. 3 is a chart illustrating a comparison of survival curves for high-risk and low-risk patient groups between NN_{ci} and NN_{ae} .

Fig. 4 is a chart illustrating a comparison of survival curves for high-risk and low-risk patient groups between NN_{ci} and SVR_c .

Fig. 5 is a chart illustrating a comparison of survival curves for high-risk and low-risk patient groups between NN_{ci} and the Cox model.

Fig. 6 is a nomogram based on the score from the NN_{ci} model.

Fig. 7 is a general schematic of a supervised learning model for a neural network.

DETAILED DESCRIPTION OF THE EMBODIMENTS

Neural networks are a form of artificial intelligence which functions to imitate the way a human brain works. A neural network typically comprises a plurality of weighted, interconnected processing elements, to emulate neurons of a brain. The organization of the processing elements and the weights of the connections between the elements determine the output. Generally, neural networks may be effective for predicting events, but typically require large amounts of data to do so.

A supervised model (e.g., predictive, prognostic) for a neural network is used to train a neural network to solve particular problems, for example, to predict an occurrence of an event. The supervised learning model trains the neural network by providing it with inputs and desired outputs. These inputs and desired outputs are referred to as training data. The difference between what is actually output by the network and the desired output is then used by a training algorithm in the model to change the weights of the connections between the processing elements to obtain the desired output.

Thus, the goal of the training a neural network is to find a function that, when training input data is applied to it, yields the desired output for the given input from the training data. It can then be used to input data to seek output in accordance with the found function. As shown in Fig. 7, a supervised learning model 700 generally includes a neural network 710 having weighted connections, an input 715 for receiving training data 705 (input training data) and an output 720, an error function 730 for determining an error between output of the neural network and a desired output from the training data (output training data), an objective function 740 for providing a rating of the performance of the model, the objective function comprising a function C substantially in accordance with a derivative of the concordance index and a training algorithm 750 (i.e., optimization method) for adapting the weighted connections of the neural network in accordance with the results of the objective function.

Prediction of an Event: Cancer Recurrence Example

The present description is in reference to training a neural network using an objective function according to embodiment of the present invention. The neural network may comprise a typical multilayer perception (MLP) with softmax outputs and a single hidden layer and direct connection between the input and output layers. Please note, however, that the invention is not so limited to such a network, and may be applied to any neural network platform/arrangement.

Embodiments of the present invention enable improved prediction models by using an approximation (e.g., a derivative) of the concordance index (CI), as the objective function in training a neural network in a supervised learning model. Once trained, such models (and systems/methods which operate under such a model) may be used to estimate the probability that, of a pair of randomly chosen comparable patients, cancer will recur within a shorter period of time in one patient with a higher prognostic score than the other patient. In that regard, such use of a derivative of the CI as the objective function in supervised learning model during training allows such models to make use of the training information from both censored (event has not recurred) and non-censored (event has occurred) observations.

The Concordance Index

The concordance index may be expressed in the form:

$$CI = \frac{\sum_{(i,j) \in \Omega} I(\hat{t}_i, \hat{t}_j)}{|\Omega|}, \quad (1)$$

5 where

$$I(\hat{t}_i, \hat{t}_j) = \begin{cases} 1: \hat{t}_i > \hat{t}_j \\ 0: otherwise \end{cases}, \quad (2)$$

and may be based on pair-wise comparisons between the prognostic estimates $\hat{t}_i$ and $\hat{t}_j$ for patients i and j , respectively. Here, Ω consists of all the pairs of patients $\{i, j\}$ who meet any
10 of the following conditions:

- both patients i and j experienced recurrence, and the recurrence time t_i of patient i is shorter than patient j 's recurrence time t_j ; or
- only patient i experienced recurrence and t_i is shorter than patient j 's follow-up visit time t_j .

15 In other words, the numerator of the CI represents the number of times that the patient predicted to recur earlier by the neural network actually does recur earlier. The denominator is the total number of pairs of patients who meet the predetermined conditions.

Though the CI has long been used as a performance indicator for survival analysis, it has never been used as an objective function in a supervised learning model for a neural
20 network to allow complete use of information from both censored and non-censored observations, as in embodiments of the present invention. The difficulty of using the CI as a training objective function in the past is that it is non-differentiable (i.e., the step function cannot be differentiated) and cannot be optimized by gradient-based methods. The present invention overcomes this obstacle by using an approximation of the CI as the objective
25 function.

As briefly stated above, the accuracy of a predictive model can be determined by using the CI: when the CI is maximized, the model is more accurate. Thus, by preferably substantially maximizing the CI, the performance of a model is maximized. Accordingly, a differentiable approximation to the step function in Eq. (2) has been proposed in [2] to directly optimize the AUC. In [2], several alternative approximations to eq. 2 are discussed in the context of approximating the Wilcoxon-Mann-Whitney statistic, which is different from the concordance index [3]. An intuitive choice of the discussed approximations is the sigmoid function:

$$S(\hat{t}_i, \hat{t}_j) = \frac{1}{1 + e^{-\beta(\hat{t}_i - \hat{t}_j)}}, \quad (4)$$

where $\beta > 0$. However, this is found to be less effective than the following function:

$$R(\hat{t}_i, \hat{t}_j) = \begin{cases} (-(\hat{t}_i - \hat{t}_j - \gamma))^n : \hat{t}_i - \hat{t}_j < \gamma \\ 0 : otherwise \end{cases}, \quad (5)$$

where $0 < \gamma \leq 1$ and $n > 1$. $R(\hat{t}_i, \hat{t}_j)$ can be regarded as an approximation to $I(-\hat{t}_i, -\hat{t}_j)$. An

example of $R(\hat{t}_i, \hat{t}_j)$ with $I(\hat{t}_i, \hat{t}_j)$ is shown in Fig. 1. Thus, in order to maximize the CI in Eq. (1), a prognostic model is trained by minimizing the objective function:

$$C = \frac{\sum_{(i,j) \in \Omega} R(\hat{t}_i, \hat{t}_j)}{|\Omega|}. \quad (6)$$

Empirically, it has been found that a weighted version of C in the following form, according to some embodiments, achieves improved results:

$$C_w = \frac{\sum_{(i,j) \in \Omega} -(\hat{t}_i - \hat{t}_j) \bullet R(\hat{t}_i, \hat{t}_j)}{D}, \quad (7)$$

where

$$D = \sum_{(i,j) \in \Omega} -(\hat{t}_i - \hat{t}_j), \quad (8)$$

is the normalization factor. Here each $R(\hat{t}_i, \hat{t}_j)$ is weighted by the difference between $\hat{t}_i$ and $\hat{t}_j$. The process of minimizing the C_ω (or C) seeks to move each pair of samples in Ω to satisfy $\hat{t}_i - \hat{t}_j > \gamma$ and thus to make $I(\hat{t}_i, \hat{t}_j) = 1$ in Eq. (1).

When the difference between the outputs of a pair in Ω is larger than the margin γ , this pair of samples will stop contributing to the objective function. This mechanism effectively overcomes over-fitting of the data during training of the model and makes the optimization preferably focus on only moving more pairs of samples in Ω to satisfy $\hat{t}_i - \hat{t}_j > \gamma$. The influence of the training samples is adaptively adjusted according to the pair-wise comparisons during training. Note that the positive margin γ in R is preferable for improved generalization performance. In other words, the parameters of the neural network are adjusted during training by calculating the CI after all the patient data has been entered. The neural network then adjusts the parameters with the goal of minimizing the objective function and thus maximizing the CI. As used above, over-fitting generally refers to the complexity of the neural network. Specifically, if the network is too complex, the network will react to “noisy” data. Overfitting is risky in that it can easily lead to predictions that are far beyond the range of the training data.

Embodiments of the present invention, described above and further explained with reference to the example below, allow a supervised learning model to make complete use of information (censored and non-censored data) from survival data (for example) using a derivative of the CI as the objective function in training the neural network. Using embodiments of the present invention, improvement may be achieved in identifying high-risk and low-risk groups of patients in a cancer recurrence predictive model (for example). Moreover, embodiments of the present invention may be applicable for use in the

approximation in a more general problem of learning ranks, where an objective function based on a step function can be formed.

The above description and example given below is considered as illustrative only of the principles of the various embodiments of the invention. Further, since numerous
5 modifications and changes will readily occur to those skilled in the art, the invention is not limited to the exact construction and operation shown and described, and accordingly, all suitable modifications and equivalents may be resorted to, falling within the scope of the invention.

10 Prostate Cancer Study

Prostate cancer (PCa) is a malignant tumor that usually begins in the outer-most part of the prostate and is the most common form of cancer found in American men. More than 180,000 men in the U.S. will be diagnosed with prostate cancer this year, and more than 30,000 will die of the disease. While the number of men diagnosed with prostate cancer
15 remains high, survival rates have been steadily improving primarily due to early detection. Eighty-nine percent of the men diagnosed with the disease will survive at least five years, while 63% will survive 10 years or longer.

The American Urological Association and the American Cancer Society recommend annual screening for men ages 50 to 70. The most effective screening tests available include
20 a blood test for an enzyme called prostate-specific antigen (PSA) which is produced by the prostate gland and the employment of a digital rectal exam (DRE). Elevated PSA levels (> 4ng/ml or greater) may indicate prostate cancer. However, increases in PSA are also reported in benign conditions such as prostatitis and a pathologic enlargement of the prostate known as benign proliferative hyperplasia (BPH). The standard of care once PCa is
25 suspected is to obtain a biopsy, typically a sextant (six-part) biopsy to assess presence or absence of disease.

The most common treatment for localized or confined PCa, in men under the age 70 who do not have other health complications is a radical prostatectomy, i.e., surgery to remove the prostate gland, seminal vesicles, vas deferens and some surrounding tissue. After surgery,

the PSA levels in the blood should be reduced to 0.2ng/ml or less. If the PSA levels begin to rise at any time after treatment (also known as a biochemical recurrence BCR), a local or distant recurrence may be suspected, and will necessitate restaging the cancer, as well as a discussion of possible salvage therapies with the patient including radiation or hormonal therapy, experimental protocols or observation [4].

Thus, the ability to predict which patients will have a BCR would be very important to urologists and oncologists in managing the course of future treatment. A number of prostate cancer nomograms which combine clinical and/or pathologic factors to predict an individual patients probability of disease recurrence or survival have been published, e.g., [5] [6] [7]. The postoperative nomogram developed by Kattan et al. [5] is widely used by clinicians and allows a prediction of the probability of disease recurrence for patients who have received radical prostatectomy as treatment for prostate cancer.

The postoperative nomogram used Cox proportional hazards regression analysis to model the clinical and pathologic data and disease follow-up for men treated with radical prostatectomy by a single surgeon. Prognostic variables included pretreatment serum prostate-specific antigen level, radical prostatectomy Gleason sum, prostatic capsular invasion, surgical margin status, seminal vesicle invasion, and lymph node status. Treatment failure was recorded when there was clinical evidence of disease recurrence, a rising serum prostate-specific antigen level, or initiation of adjuvant therapy.

Despite the widespread use of the postoperative nomogram and its reasonable predictive accuracy, better tools are needed to predict an individual patients probability of disease recurrence after radical prostatectomy. Systems pathology or biology is a new discipline that is positioned to significantly impact biological discovery processes. This emerging approach attempts to facilitate discovery by systematic integration of technologies, gathering information at multiple levels (instead of only one) and examining complex interactions which results in a superior output of data and information, thereby enhancing our understanding of biological function and chemico-biological interactions [8].

The number of features generated by these technologies can be larger than standard survival methods can handle. Thus, the underlying hypothesis of this study is that an

improved predictive model for disease recurrence after radical prostatectomy can be derived from a novel integrated or systems pathology approach, that will use neural networks to handle the expanded multidimensional sources of data input, including clinical and pathological variables (variables used in original nomogram plus additional clinical
5 variables). Molecular biomarker data can also be derived for use from IHC analyses of tissue microarrays, as well as results of machine vision image analysis which quantitate histopathological features of H&E slides.

It has been found that clinical, histopathological, immunohistochemical (IHC), and bio-imaging data can be used to predict prostate cancer BCR. In order to achieve this
10 objective, a cohort of 539 patients who underwent radical prostatectomy at a single hospital in the US was studied. 16 clinical and histopathological features were collected, which include patient age, race, Gleason grade and score, and other pre- and post-operative parameters. In addition, high-density tissue microarrays (TMAs) were constructed from the patients' prostatectomy specimens. A single hematoxylin and eosin-stained (H&E) slide for
15 each patient was used for image analysis, while the remaining sections made from the paraffin- embedded tissue blocks were used to conduct IHC studies of selected biomarkers in the laboratory. Data generated by the IHC studies included the number of cells which stained positive for a particular biomarker, if any, and the level of intensity at which the cell(s) stained positive for the biomarker. 43 IHC features from 12 biomarkers were obtained and
20 studied. Images of the H&E slides were captured via a light microscope at 20X magnification using a SPOT Insight QE Color Digital Camera (KAI2000).

Using a proprietary image analysis system, pathologically meaningful objects were identified and various statistical features associated with these objects were generated. Such objects include spectral-based characteristics (channel means, standard deviations, etc.),
25 position, size, perimeter, shape (asymmetry, compactness, elliptic fit, etc.), and relationships to neighboring objects (contrast). In the end, 496 bio-imaging features were produced.

This study was restricted to those patients who had non-missing data for each of the above three domains (clinical and histopathological, IHC, and bioimaging). Thus, the effective sample size consisted of only 130 patients. For these patients, the time from the

surgery to the most recent follow-up visit ranged from 1 month to 133 months. Patients who had measurable prostate-specific antigen (PSA) at this visit were considered to have recurrent prostate cancer. If a patient did not recur as of this last visit, or the patient outcome was unknown as of their most recent visit (e.g. due to loss to follow-up), then the patient outcome was considered censored, specifically right-censored.

Twenty (20) patients experienced PSA recurrence among the 130 patients, while the remaining patients were censored. Thus, the available sample was very small and heavily censored. By consulting domain experts and using an in-house domain specific feature selection procedure, which combines greedy forward selection and backward elimination based on the relevant importance of feature groups given by domain experts, the final feature set was reduced to 25 features.

The new training algorithm/model according to embodiments of the present invention, denoted as NN_{ci} , was compared with four other algorithms over the data collected. The first compared model used an algorithm based on the Kaplan-Meier estimate of recurrence probability, and is used as the target value for the patients who had short follow-up times and did not experience recurrence. The MLP network trained by this algorithm will be referred to as NN_{km} . The patients who had follow-up times longer than 7 years and remained disease free are assumed to be successfully cured and a target value of 0 is assigned to such patients. Those patients who had recurrence at the follow-up visit have a target value of 1. Like the new algorithm of the present invention, this model should output a higher score for a higher-risk patient.

The second compared model uses an algorithm that trains an MLP model to regress over the survival/censoring time. It uses an asymmetric squared error function for the censored patients, which becomes zero when the model output is larger than the censoring time. This model is referred to as NN_{ae} . Unlike NN_{ci} and NN_{km} , a higher-risk patient should have a lower score, an estimate proportional to the survival time, in this model.

In addition, the new training model is compared to a support vector machine regressor with an asymmetric penalty function, which incurs a smaller penalty when the model output is larger than the target value (survival time) and a larger penalty when the output is smaller

than the target value. This model is referred to as SVR_c , and should output a higher score for a lower-risk patient.

A last model compared to the present invention is the Cox proportional hazard model. The Cox model outputs a prognostic hazard score, which is a function of a linear combination of the covariates (input features). The higher the score is, the more risk the model predicts the patient would have.

The empirical results are based on leave-one-out cross-validation. For all models, model selection was conducted based on cross-validation over the training data for fold 1, and the same model setting was used for all the folds. Generally, a fold is a partition of an original dataset. For example, if the original dataset had only 10 patients and a leave-one-out cross-validation was performed, there would be 10 folds of size 1 where one patient is excluded for testing and the other 9 for training. As another example, 5 folds of size 2 are present in 5-fold cross-validation of 10 patients. All the neural networks based models have 5 hidden units after the model selection. For NN_{ci} , γ was chosen to be 0.01. It has been found that the results in terms of the CI value are more sensitive to γ than the AUC metric. n is typically set as 3. The RBF kernel is used in SVR_c . To obtain a Cox model, the number of covariates was reduced to 23, since the dataset was too small (i.e., the number of event/recurrences in the dataset was too small to handle the number of variables in question).

The performance is measured in two ways. The first measure is the Concordance Index, which evaluates the model's general predictive accuracy for estimating the probability that a patient with the higher prognostic score will recur within a shorter time than a patient with a lower score. Table 1 (below) shows the Concordance Index values for all the models. Not surprisingly, the new model according to embodiments of the present invention, NN_{ci} , achieved the largest CI value over the cross validation results.

NN_{ci}	NN_{km}	NN_{ae}	SVR_c	Cox
0.8178	0.5411	0.7375	0.6206	0.7037

The survival curves for both the predicted high-risk and low-risk patients are shown in Figs. 2 to 5 for all the models. Survival curves are constructed for both high-risk and low-risk patients by Kaplan-Meier estimates based on the scores from each model. In each figure, the new model is shown with one of the four other models. A sign of a model successfully distinguishing between the high risk and low risk patient groups is the distance that the high-risk and low-risk survival curves are apart – the farther apart the two curves are, the more successful the model is at distinguishing between the two groups. As is evident in these figures, the high-risk and low-risk survival curves are farthest apart for the NN_{ci} model, indicating a more successful model than the comparison models.

For NN_{ci}, NN_{km}, and the Cox model, the high-risk group is defined as the patients in the highest quartile of the scores, and the low-risk group consists of the patients in the lowest quartile of the scores. However, for both NN_{ae} and SVR_c, the high-risk patients are in the lowest quartile of the scores, and the low-risk patients fall in the highest quartile of the scores. A log-rank test [9] was also conducted for each pair of survival curves of high-risk and low-risk groups within each model, with a “p” value indicating the ability of the model to accurately distinguish high-risk and low-risk patient groups - the lower the p value, the better the ability of the model to distinguish between the two groups.

Table 2 summarizes the p values for all the methods. As shown, the p value for the new model NN_{ci} is much smaller (< 0.0001) than all the other models, giving further evidence of the new model's improvement in identifying high-risk and low-risk patient groups.

NN _{ci}	NN _{km}	NN _{ae}	SVR _c	Cox
< 0.0001	0.15	0.01	0.10	0.01

Recalibration

In order to make the score from the model have a clinically meaningful interpretation, the scores may be recalibrated to probabilities of remaining free of cancer recurrence in the next 7 years following surgery. The probability is estimated using the cumulative hazard function, also known as the Nelson-Aalen estimator. The Nelson-Aalen estimator incorporates both the baseline hazard rate and the hazard function, estimated via partial likelihood maximization using the Newton-Raphson method [10]. Based on these probability estimates, a nomogram was generated which is shown in Fig. 6. For ease of use, the original scores from the model are multiplied by 1000 in this nomogram.

To use the nomogram, a straight line is drawn up to the points axis from a patient's NN score to determine how many points toward recurrence the patient received. This process is repeated for each feature, and the points added together to equal the Total Points. In the example illustrated in Figure 6, with a single feature, which is the NN score, the Points and Total Points axes are identical. Accordingly, a straight line would be drawn down from the Total Points axis to the corresponding probability of the patient remaining recurrence-free for 7 years following the surgery, assuming the patient does not die of another cause first.

REFERENCES

The following references are all hereby incorporated by reference herein in their entireties.

1. F. E. Harrell. *Regression Modeling Strategies with Applications to Linear Models, Logistic Regression, and Survival Analysis*. Springer, New York, 2001.
2. L. Yan *et al.*, "Optimizing classifier performance via an approximation function to the Wilcoxon-mann-whitney statistic," *Proc. of 20th Int'l Conf. Machine Learning*, pages 848-855, 2003.
3. F.E. Harrell, *Regression Modeling Strategies*, Springer-Verlag 2001, pp.247 and 493.
4. H. Gronberg. Prostate cancer epidemiology. *Lancet*, 361:859-864, 2003.
5. M. W. Kattan, T. M. Wheeler, and P. T. Scardino. Postoperative nomogram for disease recurrence after radical prostatectomy for prostate cancer. *Journal of Clin Oncol*, 17:1499-1507, 1999.

6. M. Graefen, P. I. Karakiewicz, I. Cagiannos, and et al. Validation study of the accuracy of a postoperative nomogram for recurrence after radical prostatectomy for localized prostate cancer. *Journal of Clin Oncol*, 20:951-956, 2002.
7. L. Hood. Systems biology: integrating technology, biology, and computation. *Mech Ageing Dev*, 124:9-16, 2003.
8. E. Davidov, J. Holland, E. Marple, and S. Naylor. Advancing drug discovery through systems biology. *Drug Discov Today*, 8:175-183, 2003.
9. J. D. Kalbfleisch and R. L. Prentice. *The Statistical Analysis of Failure Time Data*. John Wiley & Sons, New York, 1980.
10. J. P. Klein and M. L. Moeschberger. *Survival Analysis: Techniques for Censored and Truncated Data*. Springer, New York, 1997.

WHAT IS CLAIMED IS:

1. A supervised prognostic model for a neural network comprising:
 - a neural network having weighted connections, an input for receiving training data and an output;
 - an error module for determining an error between output of the neural network and a desired output from the training data;
 - an objective function for providing a rating of the performance of the learning model, the objective function comprising a function C substantially in accordance with an approximation of the concordance index; and
 - a training algorithm for adapting the weighted connections of the neural network in accordance with the results of the objective function.
2. The supervised model according to claim 1, wherein the function C is substantially in accordance with a derivative of the concordance index.
3. The supervised model according to claim 1, wherein the function C allows for the use of censored training data.
4. The supervised model according to claim 1, wherein the function C comprises

$$C = \frac{\sum_{(i,j) \in \Omega} R(\hat{t}_i, \hat{t}_j)}{|\Omega|},$$
 wherein Ω comprises pairs of data (i, j) meeting one or more predetermined conditions.
5. The supervised model according to claim 1, wherein the function C comprises

$$C_{\omega} = \frac{\sum_{(i,j) \in \Omega} -(\hat{t}_i - \hat{t}_j) \cdot R(\hat{t}_i, \hat{t}_j)}{D},$$
 wherein $D = \sum_{(i,j) \in \Omega} -(\hat{t}_i - \hat{t}_j)$ is a normalization factor, and Ω comprises pairs of data (i, j) meeting one or more predetermined conditions.
6. The supervised model according to claim 4, wherein the model is used to predict recurrence of cancer, and wherein the pairs of data comprise patients {i, j} and wherein the predetermined conditions comprise at least:

both patients i and j have experienced recurrence and the recurrence time t_i of patient i is shorter than the recurrence time t_j of patient j ; and
patient i has experienced recurrence and t_i is shorter than patient j 's follow-up visit time t_j .

7. The supervised model according to claim 5, wherein the model is used to predict recurrence of cancer and wherein the pairs of data comprise patients $\{i, j\}$ and wherein the predetermined conditions comprise at least:

both patients i and j have experienced recurrence and the recurrence time t_i of patient i is shorter than the recurrence time t_j of patient j ; and
patient i has experienced recurrence and t_i is shorter than patient j 's follow-up visit time t_j .

8. A method for predicting recurrence of cancer in a patient comprising:

estimating the probability that cancer will recur within a shorter period of time in a patient with a higher prognostic score than a patient with a lower prognostic score, wherein estimating comprises conducting pair-wise comparisons between prognostic scores for patients i and j , using a neural network trained using an objective function comprising a function C substantially in accordance with an approximation of the concordance index.

9. The method according to claim 8, wherein using a neural network trained using an objective function comprises using a neural network trained using an objective function comprising a function C substantially in accordance with a derivative of the concordance index.
10. The method according to claim 8, wherein using a neural network trained using an objective function comprises using a neural network trained using an objective function comprising a function C that allows for the use of censored data in the training.

11. The method according to claim 8, wherein the function C comprises

$$C = \frac{\sum_{(i,j) \in \Omega} R(\hat{t}_i, \hat{t}_j)}{|\Omega|}, \text{ wherein } \Omega \text{ comprises pairs of patients } \{i, j\} \text{ meeting one or}$$

more predetermined conditions.

12. The method according to claim 8, wherein the function C comprises

$$C_{\omega} = \frac{\sum_{(i,j) \in \Omega} -(\hat{t}_i - \hat{t}_j) \bullet R(\hat{t}_i, \hat{t}_j)}{D}, \text{ wherein } D = \sum_{(i,j) \in \Omega} -(\hat{t}_i - \hat{t}_j) \text{ is a normalization}$$

factor, and Ω comprises pairs of data (i, j) meeting one or more predetermined conditions.

13. The method according to claim 11, wherein the predetermined conditions comprise at least:

both patients i and j have experienced recurrence and the recurrence time t_i of patient i is shorter than the recurrence time t_j of patient j; and
patient i has experienced recurrence and t_i is shorter than patient j's follow-up visit time t_j .

14. The method according to claim 12, wherein the predetermined conditions comprise at least:

both patients i and j have experienced recurrence and the recurrence time t_i of patient i is shorter than the recurrence time t_j of patient j; and
patient i has experienced recurrence and t_i is shorter than patient j's follow-up visit time t_j .

15. A method for training a neural network having weighted connections for classification of data comprising:

inputting input training data into the neural network;
processing, by the neural network, the input training data to produce an output;
determining an error between the output and a desired output corresponding to the input training data;

rating the performance of the neural network using an objective function, wherein the objective function comprises a function C substantially in accordance with an approximation of the concordance index; and
 adapting the weighted connections of the neural network based upon results of the objective function.

16. The method according to claim 15, wherein rating the performance of the neural network using an objective function comprises rating the performance of the neural network using an objective function comprising a function C substantially in accordance with a derivative of the concordance index.
17. The method according to claim 15, wherein rating the performance of the neural network using an objective function comprises rating the performance of the neural network using an objective function comprising a function C that allows for the use of censored data in the training.
18. The method according to claim 15, wherein the function C comprises

$$C = \frac{\sum_{(i,j) \in \Omega} R(\hat{t}_i, \hat{t}_j)}{|\Omega|}, \text{ and wherein } \Omega \text{ comprises pairs of data } (i, j) \text{ meeting one or}$$

more predetermined conditions.

19. The method according to claim 15, wherein the function C comprises:

$$C_{\omega} = \frac{\sum_{(i,j) \in \Omega} -(\hat{t}_i - \hat{t}_j) \cdot R(\hat{t}_i, \hat{t}_j)}{D}, \text{ wherein } D = \sum_{(i,j) \in \Omega} -(\hat{t}_i - \hat{t}_j) \text{ is a normalization}$$

factor, and wherein Ω comprises pairs of data (i, j) meeting one or more predetermined conditions.

20. The method according to claim 18, wherein the neural network is used to predict the recurrence of cancer and wherein the pairs of data comprise patients $\{i, j\}$ and wherein the predetermined conditions comprise at least:

both patients i and j have experienced recurrence and the recurrence time t_i of patient i is shorter than the recurrence time t_j of patient j ; and

patient i has experienced recurrence and t_i is shorter than patient j's follow-up visit time t_j .

21. The method according to claim 19, wherein the neural network is used to predict the recurrence of cancer and wherein the pairs of data comprise patients $\{i, j\}$ and wherein the predetermined conditions comprise at least:
 - both patients i and j have experienced recurrence and the recurrence time t_i of patient i is shorter than the recurrence time t_j of patient j; and
 - patient i has experienced recurrence and t_i is shorter than patient j's follow-up visit time t_j .
22. Computer readable media comprising computer instructions for allowing a computer system to perform a method for predicting recurrence of cancer in a patient, the method comprising estimating the probability that cancer will recur within a shorter period of time in a patient with a higher prognostic score than a patient with a lower prognostic score, wherein estimating comprises conducting pair-wise comparisons between prognostic scores for patients i and j, using a neural network trained using an objective function comprising a function C substantially in accordance with an approximation of the concordance index.
23. The computer readable media of claim 22, wherein using a neural network comprises using a neural network trained using an objective function comprising a function C substantially in accordance with a derivative of the concordance index.
24. The computer readable media of claim 22, wherein using a neural network comprises using a neural network trained using an objective function comprising a function C that allows for the use of censored data in the training.
25. The computer readable media according to claim 22, wherein the function C comprises $C = \frac{\sum_{(i,j) \in \Omega} R(\hat{t}_i, \hat{t}_j)}{|\Omega|}$, wherein Ω comprises pairs of patients $\{i, j\}$ meeting one or more predetermined conditions.

26. The computer readable media according to claim 22, wherein the function C comprises $C_{\omega} = \frac{\sum_{(i,j) \in \Omega} -(\hat{t}_i - \hat{t}_j) \cdot R(\hat{t}_i, \hat{t}_j)}{D}$, wherein $D = \sum_{(i,j) \in \Omega} -(\hat{t}_i - \hat{t}_j)$ is a normalization factor, and Ω comprises pairs of data (i, j) meeting one or more predetermined conditions.
27. The computer readable media according to claim 25, wherein the predetermined conditions comprise at least:
- both patients i and j have experienced recurrence and the recurrence time t_i of patient i is shorter than the recurrence time t_j of patient j; and
 - patient i has experienced recurrence and t_i is shorter than patient j's follow-up visit time t_j .
28. The computer readable media according to claim 26, wherein the predetermined conditions comprise at least:
- both patients i and j have experienced recurrence and the recurrence time t_i of patient i is shorter than the recurrence time t_j of patient j; and
 - patient i has experienced recurrence and t_i is shorter than patient j's follow-up visit time t_j .
29. Computer readable media comprising computer instructions for training a neural network having weighted connections, the method comprising:
- inputting input training data into the neural network;
 - processing, by the neural network, the input training data to produce an output;
 - determining an error between the output and a desired output corresponding to the input training data;
 - rating the performance of the neural network using an objective function, wherein the objective function comprises a function C substantially in accordance with an approximation of the concordance index; and

adapting the weighted connections of the neural network based upon results of the objective function.

30. The computer readable media according to claim 29, wherein rating the performance of the neural network comprises rating the performance of the neural network using an objective function, wherein the objective function comprises a function C substantially in accordance with a derivative of the concordance index.
31. The computer readable media according to claim 29, wherein rating the performance of the neural network comprises rating the performance of the neural network using an objective function, wherein the objective function comprises a function C that allows for the use of censored data in the training data.
32. The computer readable media according to claim 29, wherein the function C comprises $C = \frac{\sum_{(i,j) \in \Omega} R(\hat{t}_i, \hat{t}_j)}{|\Omega|}$, and wherein Ω comprises pairs of data (i, j) meeting one or more predetermined conditions.
33. The computer readable media according to claim 29, wherein the function C comprises: $C_{\omega} = \frac{\sum_{(i,j) \in \Omega} -(\hat{t}_i - \hat{t}_j) \bullet R(\hat{t}_i, \hat{t}_j)}{D}$, wherein $D = \sum_{(i,j) \in \Omega} -(\hat{t}_i - \hat{t}_j)$ is a normalization factor, and wherein Ω comprises pairs of data (i, j) meeting one or more predetermined conditions.
34. The computer readable media according to claim 32, wherein the neural network is trained to predict recurrence of cancer, the pairs of data comprise patients $\{i, j\}$ and wherein the predetermined conditions comprise at least:

both patients i and j have experienced recurrence and the recurrence time t_i of patient i is shorter than the recurrence time t_j of patient j ; and
 patient i experienced recurrence and t_i is shorter than patient j 's follow-up visit time t_j .

35. The computer readable media according to claim 33, wherein the neural network is trained to predict recurrence of cancer, the pairs of data comprises patients $\{i, j\}$ and wherein the predetermined conditions comprise at least:

both patients i and j have experienced recurrence and the recurrence time t_i of patient i is shorter than the recurrence time t_j of patient j ; and
patient i experienced recurrence and t_i is shorter than patient j 's follow-up visit time t_j .

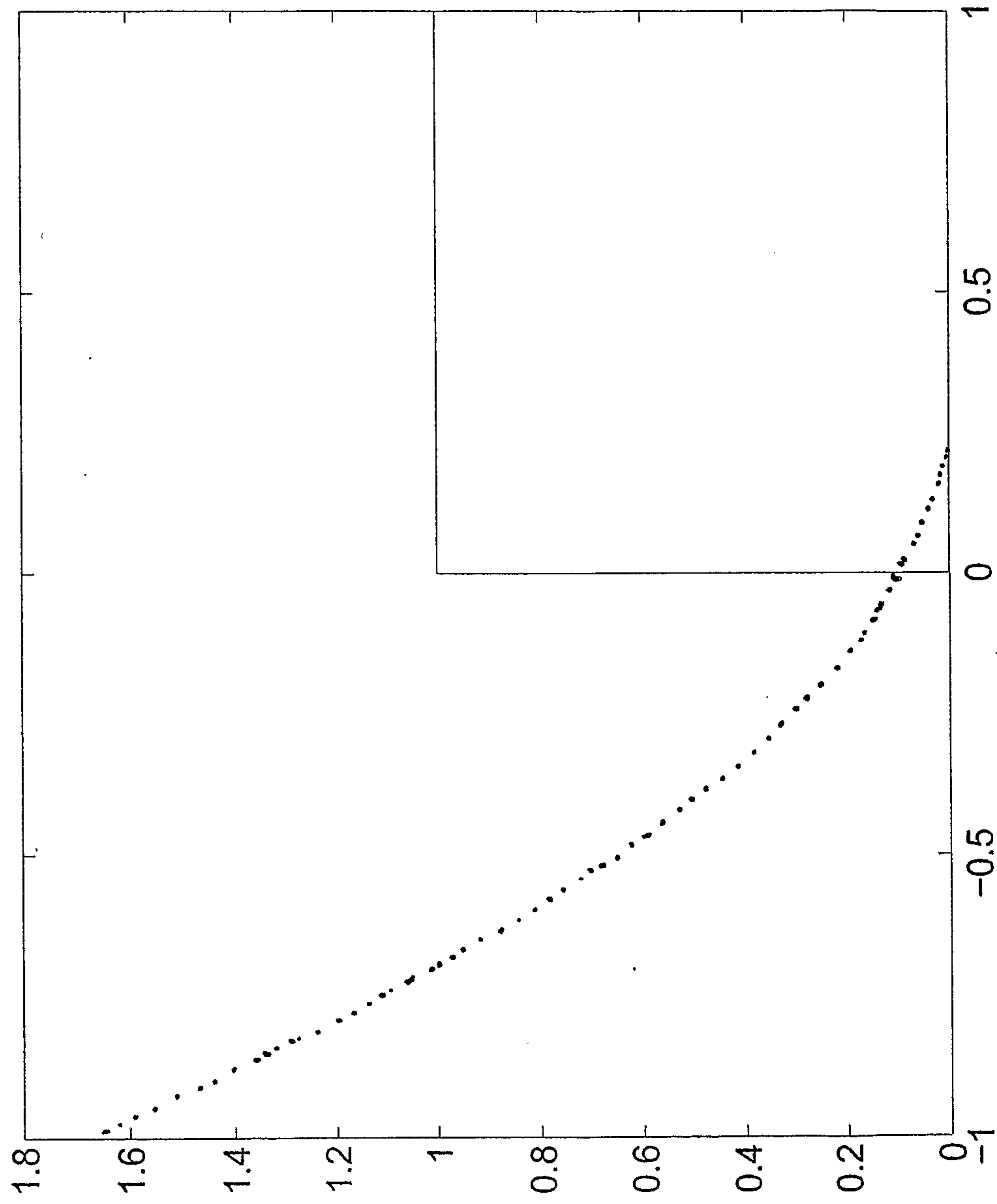


Figure 1: $R(\hat{t}_i, \hat{t}_j)$ compares with $I(\hat{t}_i, \hat{t}_j)$. The horizontal axis is $\hat{t}_i - \hat{t}_j$. $\gamma = 0.1$ and $n = 2$ in $R(\hat{t}_i, \hat{t}_j)$.

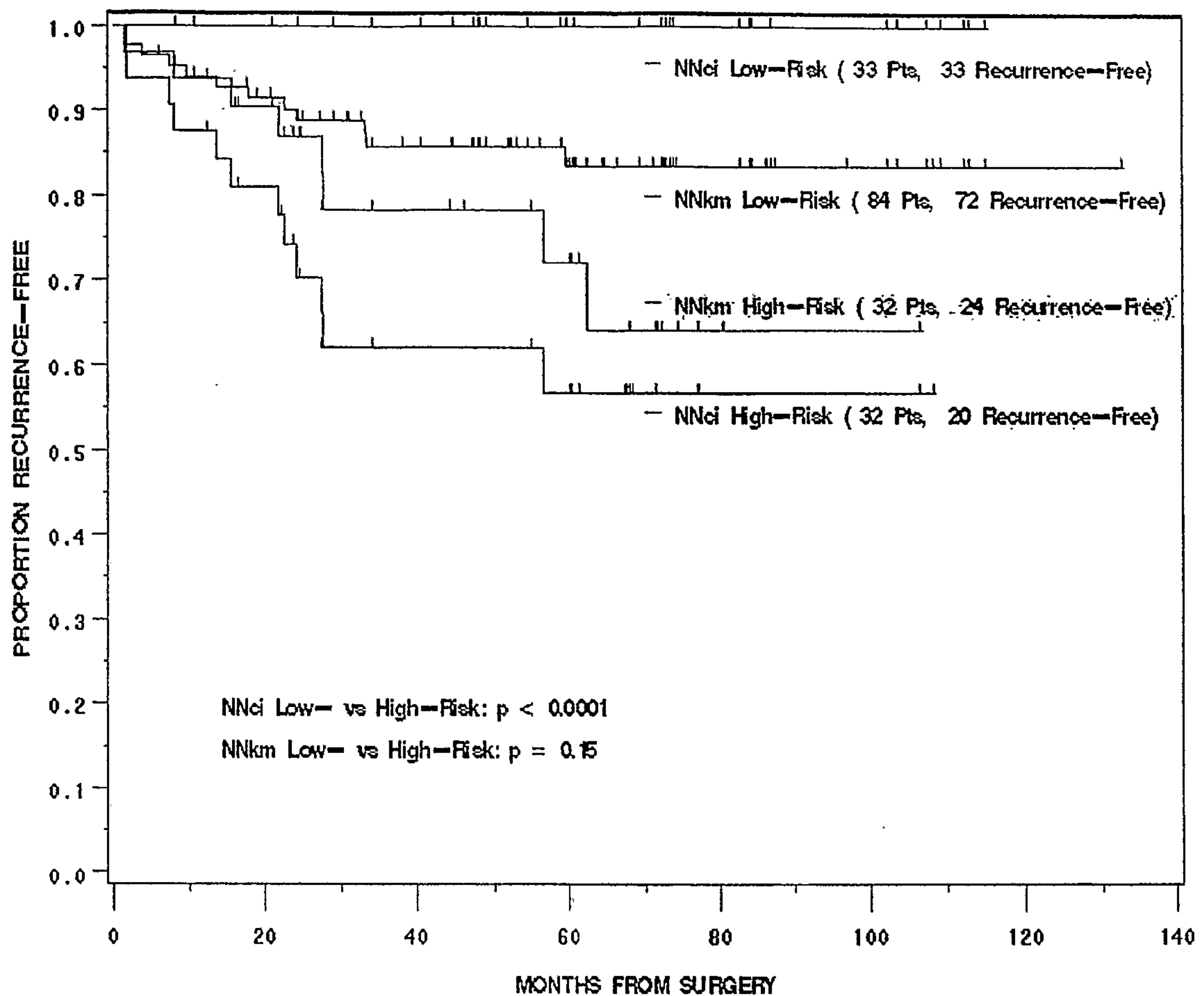


Figure 2: The comparison of survival curves for the high-risk and low-risk patient groups between NN_{ci} and NN_{km} . Note that the low risk group of NN_{km} consists of 84 patients because of tied scores.

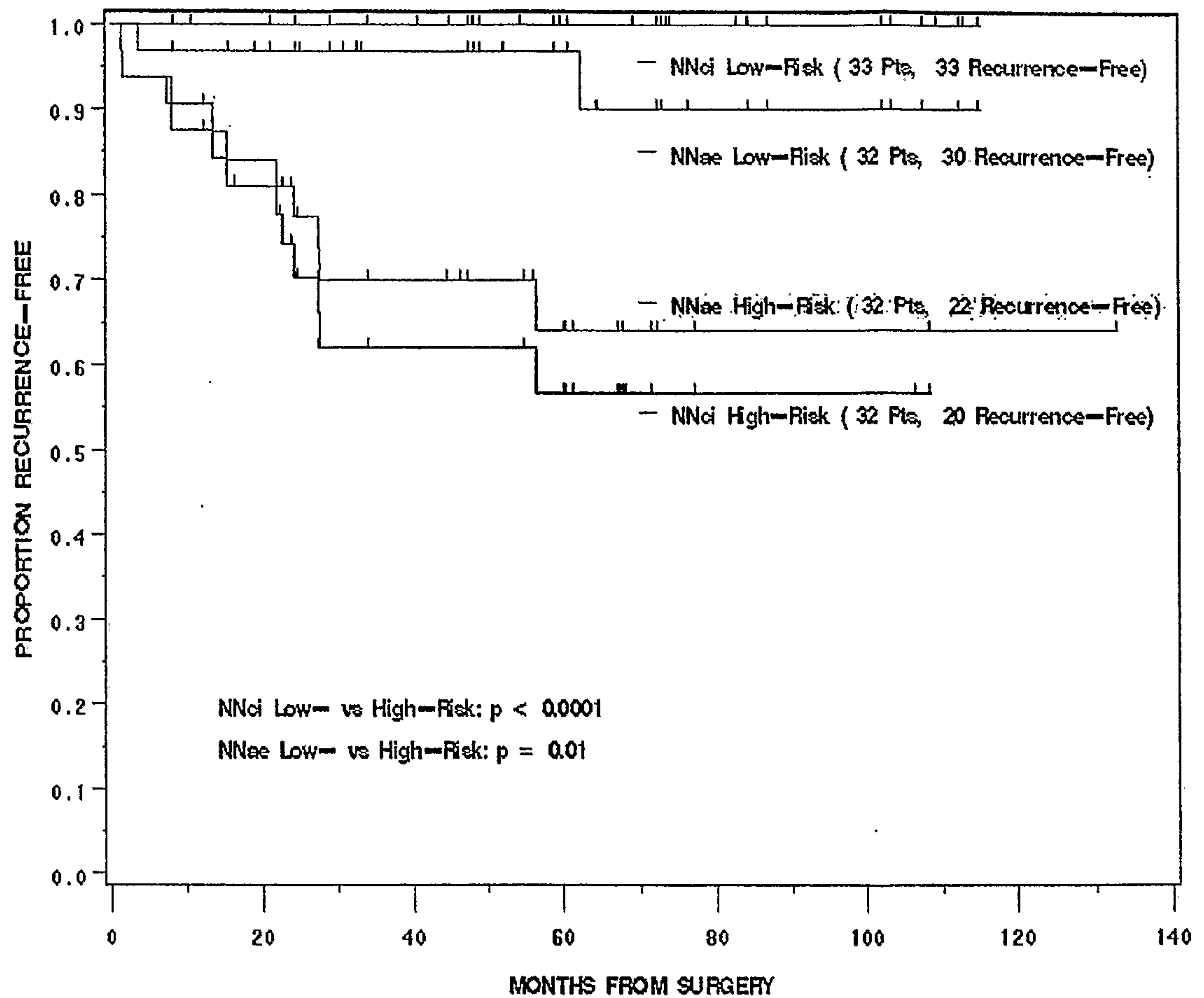


Figure 3: The comparison of survival curves for the high-risk and low-risk patient groups between NN_{ci} and NN_{ae}.

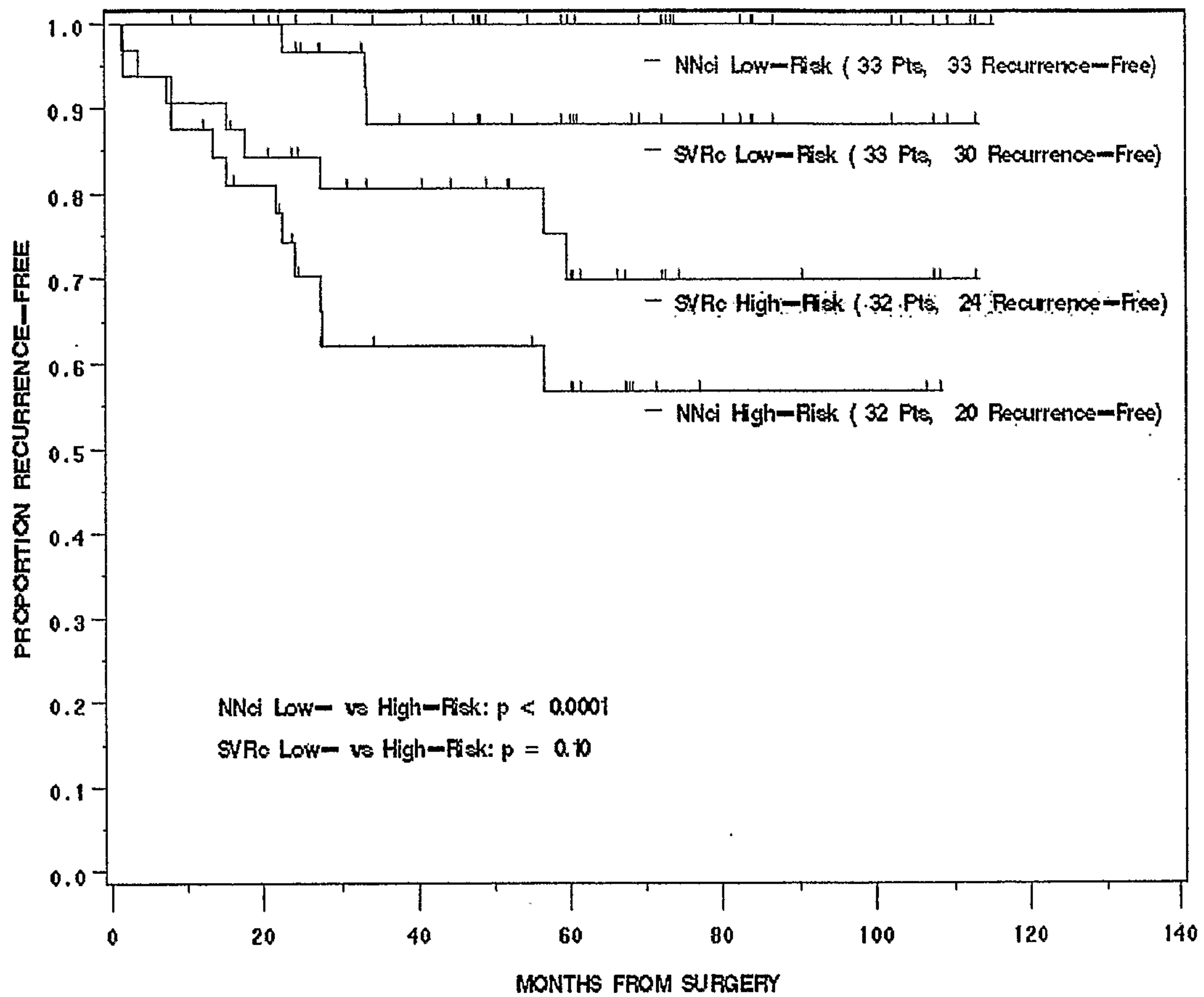


Figure 4: The comparison of survival curves for the high-risk and low-risk patient groups between NN_{ci} and SVR_c.

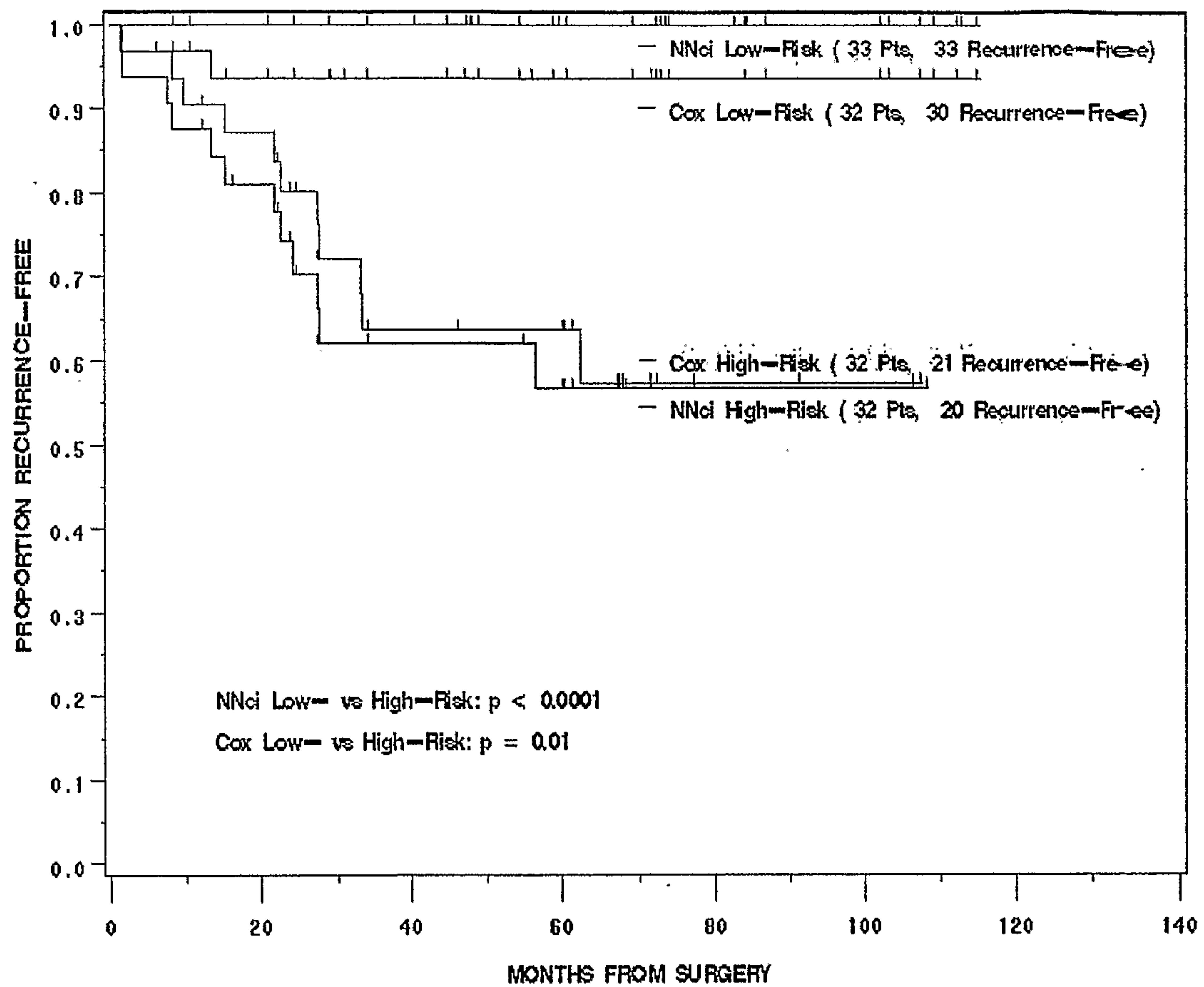


Figure 5: The comparison of survival curves for the high-risk and low-risk patient groups between NN_{ci} and the Cox model.

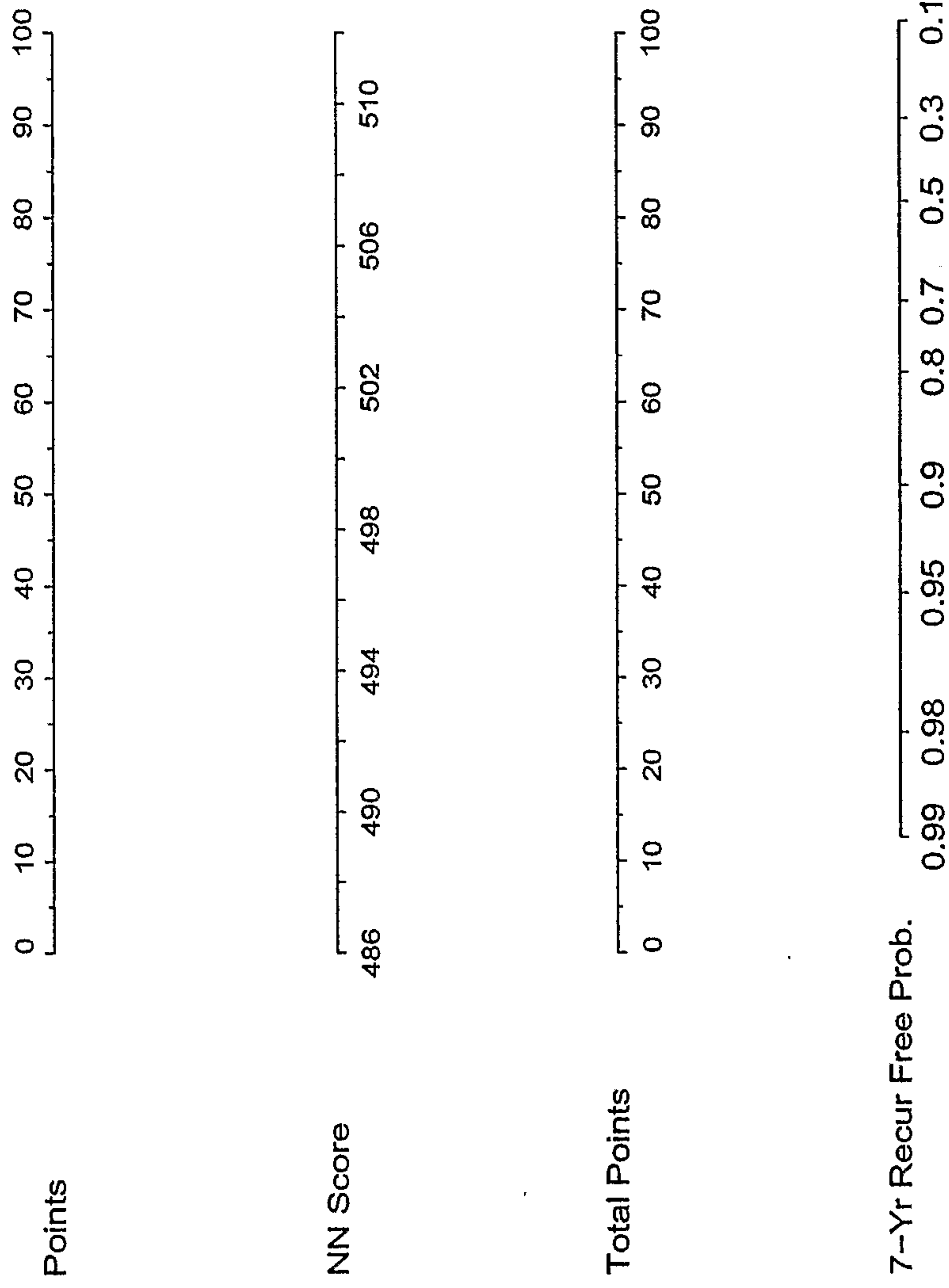


Figure 6: A nomogram based on the score from the NN_{ci} model.

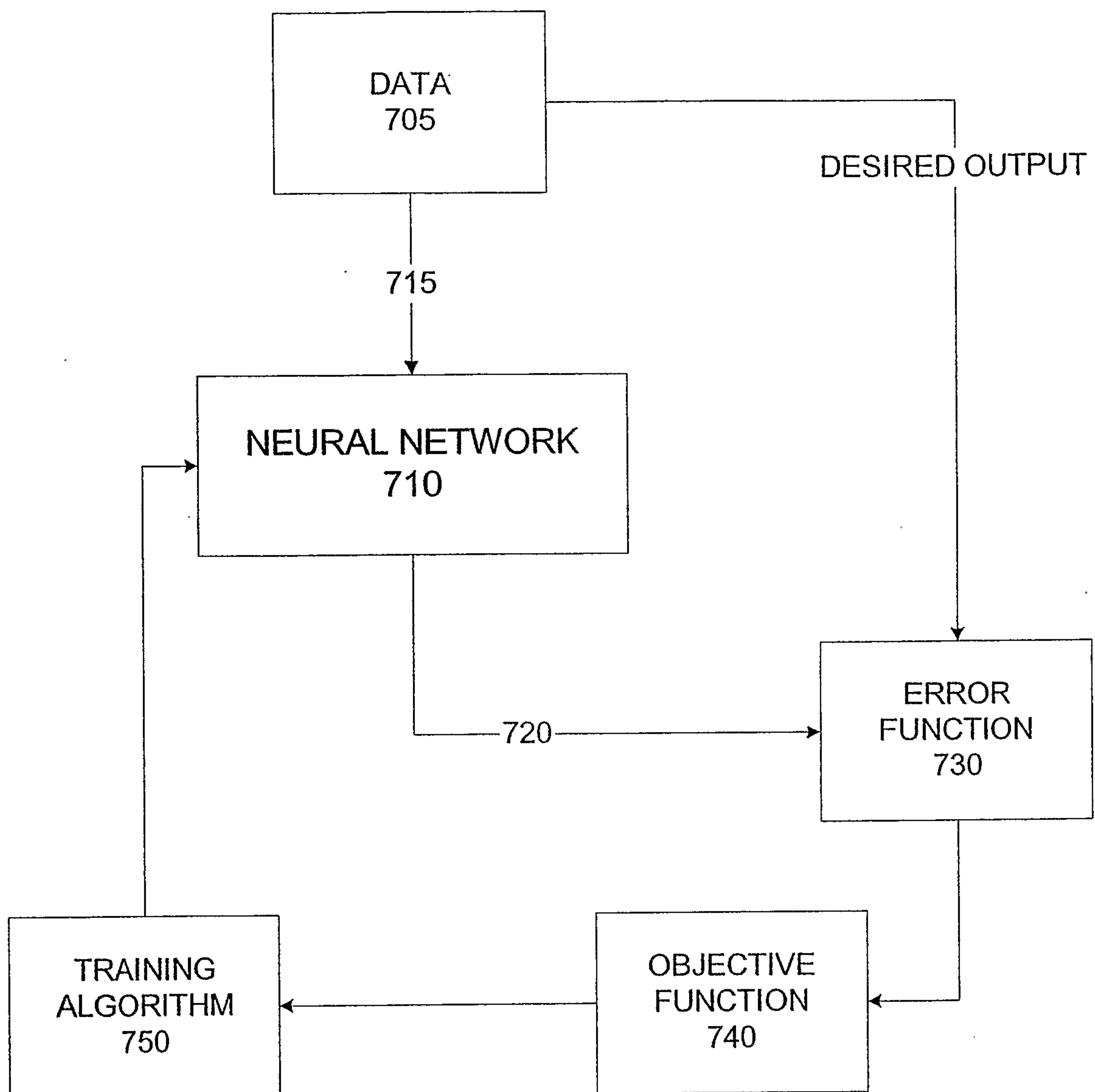


FIG. 7

