**Using Dendrograms to Create Prognostic Systems for Multiple Prognostic Factors in Cancer: An Expansion of the TNM**

By

Donald E. Henson, MD

Dechang Chen, Ph.D

Arnold Schwartz, Ph.D, MD

Dengyuan Wu

Abstract

**Background:** Cancer is no longer considered an anatomic disease. It incorporates genomics, proteomics, signaling pathways, and biological factors. Therefore, an expansion of the TNM seems necessary to accommodate additional prognostic factors in order to provide more accurate estimations of recurrence and death.

**Methods:**.Dendrograms were created by an unsupervised ensemble-learning algorithm according to disease specific survival functions. Built on the basis of combinations of any number of prognostic factors, dendrograms had a tree structure that stratified patients. Cutting the dendrograms at different levels created prognostic systems that stratified patients into groups with related outcomes. Applications were given to women with invasive breast cancer from 1990-2000 obtained from NCI’s SEER Program.

**Results:** Three dendrograms were generated, one for tumor size, one for positive lymph node status, and one for both tumor size and positive node status. They showed how survival functions changed as the categories of prognostic factors varied. Prognostic systems were produced on the basis of tumor size and positive lymph node status. These prognostic systems largely agreed with the TNM system. Prognostic systems were also obtained according to histological grade, tumor size, positive nodes, and ER status, showing a new classification of patients when grade and ER status were added.

**Conclusions:** Dendrograms provide a relationship among survival and combinations of prognostic factors. Prognostic systems from the dendrograms expande the TNM without changing its definition. The prognostic systems based on more prognostic factors than the TNM are expected to provide more accurate survival estimations.

Introduction

The continuing discovery of new prognostic and predictive factors for cancer will require new types of prognostic systems in order to estimate more accurately recurrence and death. Expansion of the current TNM by adding prognostic factors is not straightforward because it is a bin model and, therefore, subject to the limitations of a bin model (Burke and Henson Ref). Consequently, new approaches for integrating multiple prognostic factors into a single statement of outcome are needed. In this report, we describe how to use the ensemble algorithm of clustering of cancer data (EACCD) (Supplement; Chen; Qi Ref) to create prognostic systems based on combinations of any number and any type of prognostic factors. The process involves generating dendrograms and then cutting the dendrograms. Using this ensemble learning algorithm we show that it is possible to expand the TNM to create new prognostic systems with additional factors without changing the TNM definitions. Herein, we present our demonstration for creating these systems using breast cancer as the example. The key to the prognostic system is creation of dendrograms based on survival functions and multiple prognostic factors.

Materials and Methods

Data Source

The source of data was the file “Case Listings” in NCI’s SEER (Surveillance, Epidemiology, and End Results) Program (SEER Ref). Data were obtained from 1990 through 2000, which allowed for 10 years of follow-up. The original data set contained 177,409 cases of unselected female breast cancer, but the selected data contained 99,951 cases. A case was selected if it contained complete records on the survival time, censoring status, as well as the following four variables: T (tumor size), N (nodal status), G (histological grade), ER (ER status). The variables T, N, G, and ER had 3, 4, 3, and 2 categories, respectively, as shown in Table 1. Histological types included infiltrating ductal carcinomas and comedo carcinomas. In situ carcinomas were excluded. Because of a small number (3%), cases of histological grade 4 were merged with grade 3 since their survival rates are similar (Henson1991 Ref). Inflammatory breast cancers and other T4 tumors were not considered. SEER does not include subdivisions of the primary stage grouping, for instance, the “m” category. Cases with metastatic tumors to other visceral organs and sites were excluded and only M0 cancers were taken into account.

Methods

The clustering algorithm EACCD was composed in the programming language “R” (R Ref). The main output of the algorithm is a tree-structured dendrogram that stratifies patients according to disease specific survival functions. EACCD is used to cluster combinations of prognostic factors. In this context, a combination is a subset of the data that corresponds to one category of each selected prognostic factors. For example, for breast cancer categories T0 and N0 when used together produce a combination, denoted by T1N0, which represents a subset of breast cancer patients whose category of the tumor size is T1 and category of the lymph node status is N0. In this report we use notations of categories of factors to code combinations. Computationally, the dendrogram is constructed from bottom to top by comparing survival functions of paired prognostic clusters.

In creating dendrograms, EACCD relies on three steps. In the first step, an initial dissimilarity between two combinations of prognostic factors is computed by using a statistic employed to test if a difference exists in survival between two survival functions associated with the two combinations. The second step is used to obtain the learnt dissimilarity between two combinations. This learnt dissimilarity comes from repeated use of partitioning approaches based on the initial dissimilarity. For the third step, the algorithm performs hierarchical clustering on combinations by using the learnt dissimilarity. This step generates a tree-structured dendrogram that stratifies patients.

In this report, we used the following setting to run EACCD for a collection of combinations. The log-rank test (Klein JP Ref) was used to determine the initial dissimilarity in the first step of the algorithm. In the second step, we used the Partitioning Around Medoids (PAM) (Kaufman Ref) as the only partition technique. This technique was repeatedly used for m=10000 times. For each time in using the PAM, the number of clusters was randomly selected from an interval [K1, K2], where K1=2 and K2=the total number of combinations. In the third step, the average linkage hierarchical clustering technique (Hastie Ref) was used. In addition, due to approximation of the chi-square distribution for the log-rank test statistic, we excluded all combinations that contained less than 50 patients. However, this threshold of 50 patients could be set at any reasonable level.

Results

Dendrograms

Dendrograms are the main output of EACCD. A dendrogram utilizes dissimilarity to measure the difference between two survival functions associated with two clusters (of combinations). A dissimilarity value ranges from 0 to 1, with a larger value of dissimilarity implying a larger difference between two survival functions. The dendrogram is created by beginning with individual combinations of prognostic factors as a distinct cluster and at each successive step merging two clusters that have the least dissimilarity into a larger single cluster.

Depending on the need, the 10-year survival rates of combinations, computed by the Kaplan-Meier procedure (Kaplan Ref), can be added into the dendrogram.

Figure 1 shows the adendrogram only for a single variable, namely tumor size. In viewing the dendrogram from bottom to top, it is seen that T2 and T3 (each viewed as a distinct cluster) are first merged at an approximate dissimilarity value of 0. Here the dissimilarity value of 0 means the survival function of T2 does not differ from the survival function of T1 in terms of the dissimilarity measurement. Then the cluster containing T2 and T3, denoted by {T2, T3}, is merged with T1 at an approximate dissimilarity value of 0.52. Here an approximate dissimilarity value of 0.52 means the survival function of the cluster {T2, T3} differs from the survival function of T1 by an approximate dissimilarity value of 0.52. The dendrogram in Figure 1 reveals a stratified relation for outcome when only T is used. Especially it shows thatT2 and T3 are more related to each other than they are to T1.

Figure 2 shows the dendrogram only for the number of positive lymph nodes. This dendrogram shows the following. The combinations N2 and N3 are first merged to form cluster {N2, N3}, where the survival function of N2 does not differ from the survival function of N3 according to the dissimilarity measurement. Then the cluster {N2, N3} are merged with N1 to form the cluster {N2, N3, N1} at an approximate dissimilarity value of 0.45. And finally, the cluster {N2, N3, N1} is merged with the combination N0 at an approximate dissimilarity value of 1, showing the large difference between the survival function of {N2, N3, N1} and the survival function of N0. The dendrogram in Figure 2 reveals a stratified relationship for outcome when only N is used.

To generate a dendrogram that reflects the TNM staging for breast cancer, the T and N variables from Figures 1 and 2 were combined to produce 12 combinations (3×4). The resulting dendrogram is shown in Figure 3. This dendrogram is created by first merging combination T2N3 with combination T3N2 and then at each next step merging two clusters with the least dissimilarity. The dendrogram in Figure 3 represents a more refined stratification of patients than those in Figures 1 and 2. This dendrogram indicates that a number of combinations of T and N are associated with similar patient survival rates. For instance, T1N1 has a similar 10-year survival as T2N0 and T2N2 has a similar outcome as T3N1.

As demonstrated above, a dendrogram presents a visual relationship among survival and combinations of prognostic factors. It shows how combinations are related in terms of survival.

Graphically,a dendrogram is a tree diagram, where leafs represent individual combinations of prognostic factors and two branches merged at a node with a value of dissimilarity.

Prognostic Systems

Combinations with similar survival functions can be grouped. These groups of combinations and their survival curves constitute a useful and defined prognostic system for cancer patients. Grouping combinations with “similar” survivals can be done by cutting the dendrogram at a specified height of the dissimilarity axis.

If we cut the dendrogram in Figure 3 at the first level or the first division, then we obtain two distinct prognostic groups (Figure 4), which is in contrast to the four stage groups from the AJCC for breast cancer. Note that all the combinations from the left group are stage III while all the combinations from the right group are stage I or stage II, although there is one stage IIIA from the right group, the only exception and possibly an misclassification of the TNM. If, on the other hand, we cut at the second level, then four prognostic groups are generated (Figure 5). Table 2 lists a detailed contribution of each combination to each prognostic group. The survival functions of the four prognostic groups (Figure 5) are estimated by the Kaplan-Meier procedure and plotted in Figure 6. The four survival functions are statistically different. The prognostic system here consists of these four groups and their survival curves. This system can then be used to estimate the survival of any new patient based on the categories of T and N.

Figure 7 shows a dendrogram generated from four factors – histological grade, tumor size, positive nodes, and ER status, which form 72 combinations (3×3×4×2). Note that the dendrogram contains only 56 combinations because 16 combinations containing less than 50 patients were excluded. It is almost impossible, for instance, to accrue 50 cases of T1, N3, grade 1 breast cancer. It is observed that in Figure 7 combinations with different categories of prognostic factors can have similar survival rates. For example, the combination G3T1N3ER- (i.e.,grade 3, T1, N3, ER-) has the same 10-year survival rate (32%) as the combination G3T3N3ER+ (i.e., grade 3, T3, N3, ER+). Clearly, cutting the dendrogram at the first level creates 3 prognostic groups. When cut at the second level, the dendrogram segregates into eight prognostic groups (Figure 8). The eight corresponding survival curves are estimated by the Kaplan-Meier procedure and shown in Figure 9. All survival functions are statistically different (p-value is about 0). A prognostic system here consists of the eight prognostic groups and their corresponding eight survival curves. Although not shown, we have generated dendrograms for up to eight prognostic factors.

As more prognostic factors are used, the number of combinations will increase. Cutting the dendrogram can aggregate a larger number of combinations into a smaller number of more manageable prognostic groups. A prognostic system could be considered as a “compressed” version of a dendrogram. Dendrograms can be cut at any level regardless of the number of combinations. Cutting at a lower value of dissimilarity usually leads to a larger number of prognostic groups. Moreover, any branch of a dendrogram can be cut so that survival rates of sub-branches can be investigated.

Discussion

Cluster analysis is a well-known technique that targets at grouping objects such that objects from the same group are more similar to each other than to those from other groups. This technique has been used in many domains, such as geography, marking, image processing, biology, medicine, psychology, psychiatry, and so on (Everitt Ref). The algorithm EACCD developed to establish prognostic groups for cancer patients presents a new method in cluster analysis, which takes into account censored survival times of cancer patients. This is in contrast to traditional algorithms for clustering data that cannot be directly applied to censored data. This algorithm can take as input any type of prognostic factor (e.g., continuous, ordinal, or nominal) and has no limit on the number of factors or variables. If a factor is continuous, such as tumor size, discretizing the factor is necessary before applying the algorithm.

We have shown that dendrograms from EACCD can be used to provide an overall view of the relationship among survival functions as categories of prognostic factors are changed. We have also shown that that can be used to produce prognostic systems. There exist other applications of dendrograms. They can reveal the complexity of interaction among prognostic factors which cannot be predicted by using the individual contributions of factors. Figures 1 and 2 show dendrograms for T and N respectively. Figure 3 shows the combinations of T and N as observed in Figures1 and 2. Figure 3, which cannot be predicted from Figures 1 and 2, demonstrates the interaction of the two factors in relation to survival. Therefore, even with only two factors, the interaction of prognostic factors can be complex and their combined contributions difficult to predict on the basis of the contribution of individual factors. In addition, with a dendrogram, the effect of one prognostic factor on survival can be visually checked while varying the factor but fixing remaining factors at constant categories. For instance, we have shown that the histological grade in breast cancer remains a prognostic factor regardless of the number of involved lymph nodes (Schwartz, submitted).

While a dendrogram focuses on combinations, a prognostic system encompasses groups of patients each consisting of at least one combination. In general, the number of groups in a prognostic system is much smaller than the number of combinations in a dendrogram. Correspondingly, the number of survival functions considered in a prognostic system is much smaller than that in a dendrogram. For instance, in the dendrogram of four factors of T, N, grade, and ER (Figure 7), there are 56 survival functions each one corresponding to a combination. However, in the prognostic system described in Figure 8, there are only 8 survival rates available for comparison (Figure 9). In many cases, a clinical trial for instance, a prognostic system might be more convenient to use than a dendrogram, since it contains grouped patients with related outcomes.

We view a prognostic system from a dendrogram as an expansion of the TNM, where the prognostic groups can in principle fulfill the role of staging. Such an expansion is likely to be more accurate than using the TNM alone because more prognostic factors have been added into the expansion which help further subsets the patient population into more homogeneous and specific survival groups. Integrating additional factors, such as age and histological tumor type, may significantly change management by pushing the patient into a different stage group resulting in a different treatment. (Henson, submitted). Moreover, this expansion is expected to have a significant application for personalized medicine because disease specific outcomes for individual combinations of prognostic factors are given so that outcome predictions for new cancer patients can be based on previous patients who have had similar combinations of prognostic factors.

In summary, dendrograms from EACCD are useful in relating prognostic factors to survival of patients. They not only validate staging systems but also generate prognostic systems.

the prognostic system essentially serves the same purpose as the TNM, but incorporates additional prognostic factors that are likely to add benefit by providing more accurate estimations of recurrence and death.

References

1. [Aldecoa R, Ignacio M. Jerarca: Efficient Analysis of Complex Networks Using Hierarchical Clustering](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2904377/?tool=pmcentrez). PLoS One. 2010; 5(7): e11585. Published online 2010 July.4.doi: 10.1371/journal.pone.0011585
2. Burke HB, Henson DE. Criteria for prognostic factors and for an enhanced prognostic system. Cancer 1993;72:3131-3135.
3. Cancer Staging Manual, American Joint Committee on Cancer. Editors: Edge SB, Byrd DR, Compton CC, Fritz AG, Greene FL, Trotti A III, 7th Edition, Springer 2010.
4. Chen D., Xing K., Henson D., Sheng L., Schwartz A. M., and Cheng X. Developing Prognostic Systems of Cancer Patients by Ensemble Clustering .*Journal of Biomedicine and Biotechnology,* Volume 2009, Article ID 632786, 7 pages, doi:10.1155/2009/632786.
5. [Everitt](http://www.amazon.com/s/ref=ntt_athr_dp_sr_1?_encoding=UTF8&field-author=Brian%20S.%20Everitt&search-alias=books&sort=relevancerank) BS , [Landau](http://www.amazon.com/s/ref=ntt_athr_dp_sr_2?_encoding=UTF8&field-author=Sabine%20Landau&search-alias=books&sort=relevancerank) S,  [Leese](http://www.amazon.com/s/ref=ntt_athr_dp_sr_3?_encoding=UTF8&field-author=Morven%20Leese&search-alias=books&sort=relevancerank) M, [Stahl](http://www.amazon.com/s/ref=ntt_athr_dp_sr_4?_encoding=UTF8&field-author=Daniel%20Stahl&search-alias=books&sort=relevancerank) D, *Cluster Analysis*, 5th Edition, 2011 John Wiley & Sons, Ltd
6. Gospodarowicz M, Mackillop W, O’Sullivan B, Henson D, Hutter RV, Wittekind C. Prognostic factors in clinical decision making. Cancer (suppl) 2001;91:1688-1695.
7. [Gospodarowicz M](http://www.ncbi.nlm.nih.gov/pubmed?term=%22Gospodarowicz%20M%22%5BAuthor%5D), [Benedet L](http://www.ncbi.nlm.nih.gov/pubmed?term=%22Benedet%20L%22%5BAuthor%5D), [Hutter RV](http://www.ncbi.nlm.nih.gov/pubmed?term=%22Hutter%20RV%22%5BAuthor%5D), et al. History and international developments in cancer staging. [Cancer Prev Control.](http://www.ncbi.nlm.nih.gov/pubmed/10470455)1998 6:262-8.
8. Hastie T, Tibshirani R, Friedman J, *The Elements of Statistical Learning*, Springer, New York, NY, USA, 2001
9. Henson DE, Ries L, Freedman LS, Carriaga M. Relationship among outcome, stage of disease, and histologic grade for 22,616 cases of breast cancer. Cancer 68;2142-2149:1991.
10. Henson DE, Schwartz AM, Chen D, Wu D. The clinical implications of integrating additional prognostic factors into the TNM.(Submitted for publication)
11. Kaplan, E. L and P. Meier. 1958. Nonparametric estimation from incomplete observations. *J. Am. Stat. Assoc*. 53:457-481.
12. Kaufman L, Rousseeuw P. *Finding Groups in Data: An Introduction to Cluster Analysis*, John Wiley& Sons, New York, NY, USA, 1990.
13. Klein, J. P. and M. L. Moeschberger. 2005. Survival Analysis: Techniques for Censored and Truncated Data, 2nd edition. New York: Springer Verlag.
14. Odong TL, van Heerwaarden J, Jansen J, van Hintum TJL, van Eeuwijk FA. [Determination of genetic structure of germplasm collections: are traditional hierarchical clustering methods appropriate for molecular marker data?](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3114091/?tool=pmcentrez)TheorAppl Genet. 2011 July; 123: 195–205. Published online 2011 April 7.doi: 10.1007/s00122-011-1576-x
15. Park PJ, Manjourides J, Bonetti M, Pagano M. [A permutation test for determining significance of clusters with applications to spatial and gene expression data](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3023458/?tool=pmcentrez). Comput Stat Data Anal 2009;53:4290-4300.
16. Park YH, Lee SJ, Choi YL, Le JE, Nam SJ, Yang JH Ko EY, Han Bk, Ahn JS, IM YH. Clinical relevance of TNM staging system according to breast cancer subtypes. Ann Oncol 2011;Jan 17 EPUB.
17. Qi R, Wu D, Sheng L, Henson D, Schwartz AM, Xu E, Xing K, Chen D. On an ensemble algorithm for clustering cancer patient data. BMC Systems Biology, in press.
18. SEER (http://seer.cancer.gov).
19. Surveillance Research Program, National Cancer Institute SEER\*Stat software (www.seer.cancer.gov/seerstat) version 8.0.4.
20. The R Project for Statistical Computing (http://www.r-project.org).
21. [Tien YJ](http://www.ncbi.nlm.nih.gov/pubmed?term=%22Tien%20YJ%22%5BAuthor%5D), [Lee YS](http://www.ncbi.nlm.nih.gov/pubmed?term=%22Lee%20YS%22%5BAuthor%5D), [Wu HM](http://www.ncbi.nlm.nih.gov/pubmed?term=%22Wu%20HM%22%5BAuthor%5D), [Chen CH](http://www.ncbi.nlm.nih.gov/pubmed?term=%22Chen%20CH%22%5BAuthor%5D).Methods for simultaneously identifying coherent local clusters with smooth global patterns in gene expression profiles. BMC Bioinformatics2008;9:155.

|  |  |  |
| --- | --- | --- |
| Prognostic Factor | Categories | Notation |
| Tumor Size | ≤ 2 cm | T1 |
| >2 &≤5 cm | T2 |
| >5 cm | T3 |
| Nodal Status | no positivenodes | N0 |
| 1-3 nodes positive | N1 |
| 4-9 nodes positive | N2 |
| >10 nodes positive | N3 |
| Histological Grade | 1 | G1 |
| 2 | G2 |
| 3 or 4 | G3 |
| ER status | positive | ER+ |
| negative | ER- |
|  | | |

Table 1.Categories of the prognostic factors for SEER breast cancer cases from 1990-2000. Categories of the tumor size and those of the nodal status are from the AJCC.

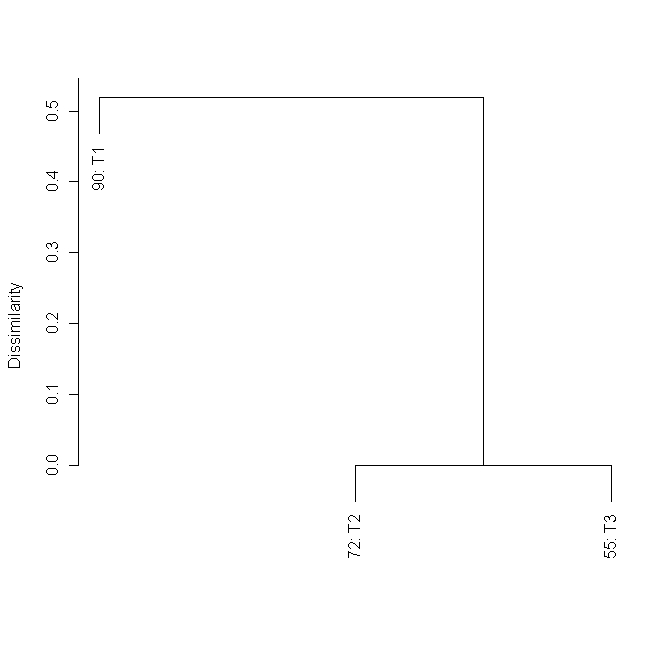


Figure 1.Dendrogram for breast cancer using combinations defined by the tumor size T (Table 1). A combination is given beneath each leaf. The number before the colon (:) represents the 10 year survival (in percentage) of the combination.

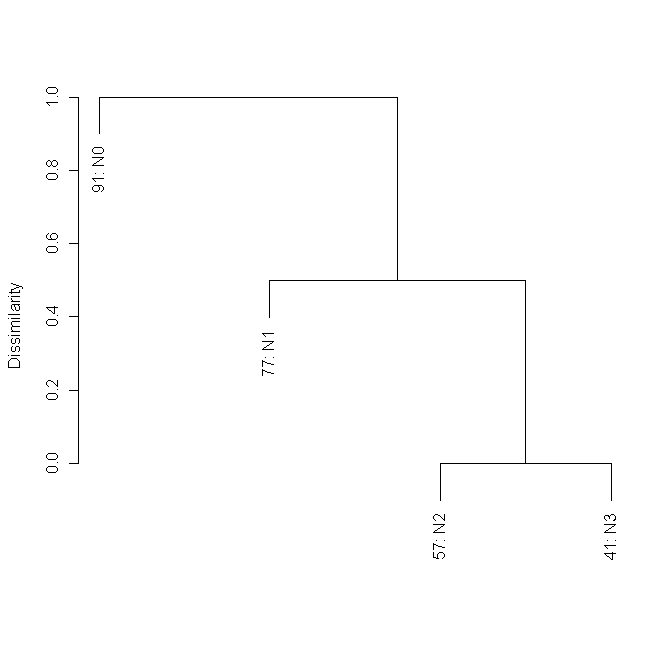


Figure 2.Dendrogram for breast cancer using combinations defined by the nodal status N (Table 1). A combination is given beneath each leaf. The number before the colon (:) represents the 10 year survival (in percentage) of the combination.

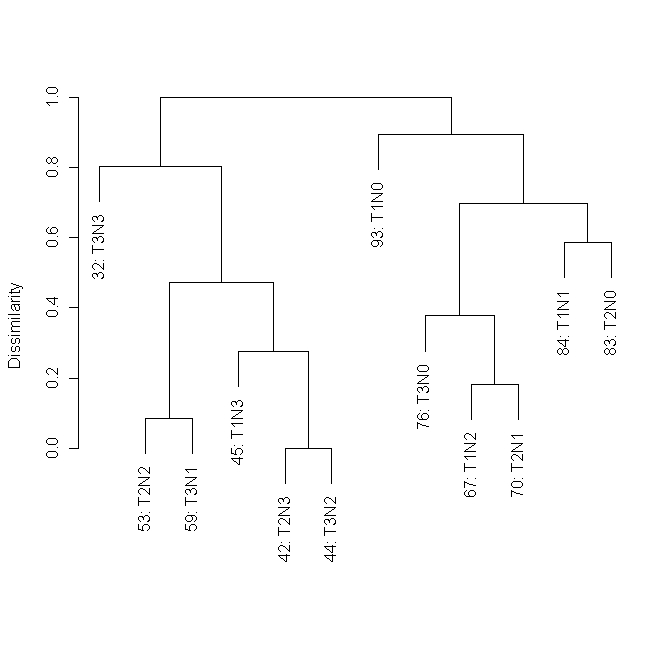


Figure 3.Dendrogram for breast cancer using combinations defined by the tumor size T and nodal status N (Table 1). A combination is given beneath each leaf. The number before the colon (:) represents the 10 year survival (in percentage) of the combination.

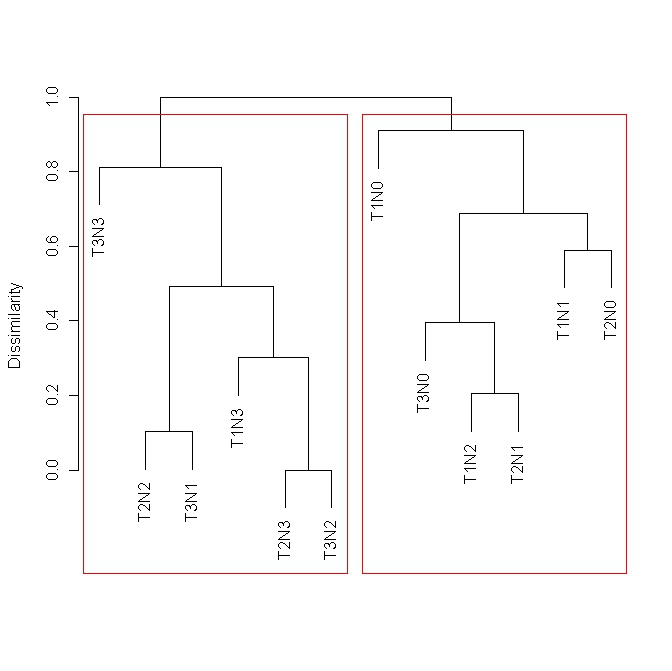
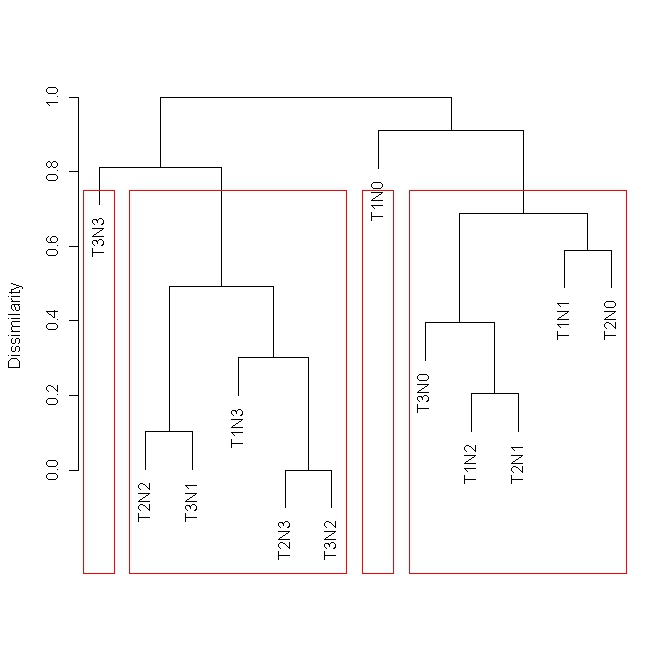


Figure 4.Same dendrogram as in Figure 3but cut at the first level or division.The dendrogram splits into 2 distinct T,N prognostic groups or clusters instead of the 4 as in the AJCC.



3

1

4

2

Figure 5. Cutting the dendrogram (Figure 3) at the second level or division reveals four prognostic groups labeled 1-4. Cuts at lower levels reveal more prognostic groups.

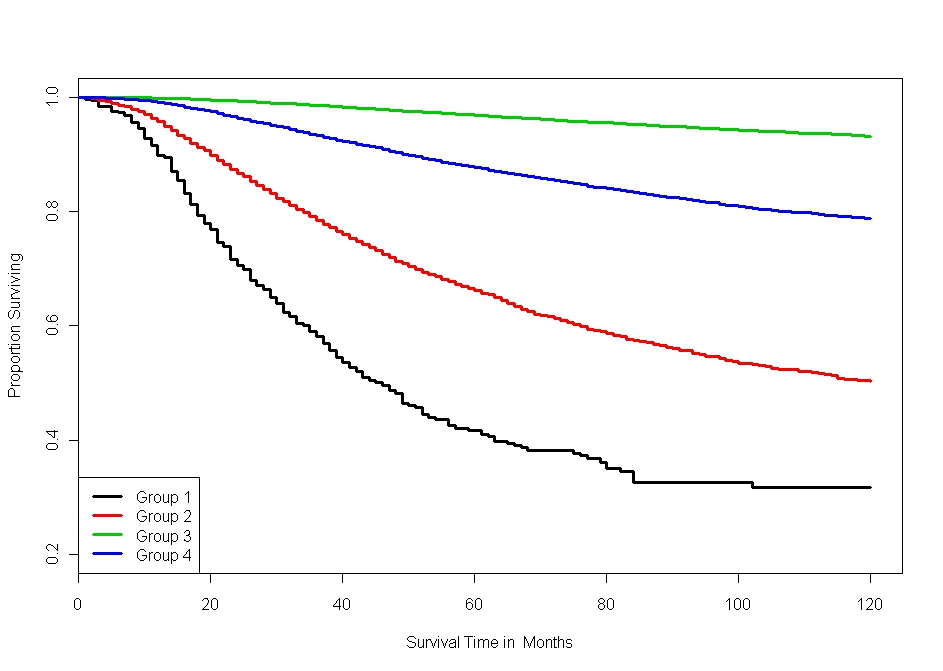


Figure 6.10-year survival rates for the four prognostic groups generated in Figure 5. Numbers refer to the prognostic groupsin Figure 5. All rates are significantly different.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | | | |
|  | Prognostic Group | | | |
| Combination | 1 | 2 | 3 | 4 |
| 11 | 0 | 0 | 50580 | 0 |
| 12 | 0 | 0 | 0 | 11925 |
| 13 | 0 | 0 | 0 | 3036 |
| 14 | 0 | 697 | 0 | 0 |
| 21 | 0 | 0 | 0 | 14567 |
| 22 | 0 | 0 | 0 | 9343 |
| 23 | 0 | 4768 | 0 | 0 |
| 24 | 0 | 1560 | 0 | 0 |
| 31 | 0 | 0 | 0 | 872 |
| 32 | 0 | 974 | 0 | 0 |
| 33 | 0 | 112 | 0 | 0 |
| 34 | 525 | 0 | 0 | 0 |
| Table 2. Number of breast cancer cases in each prognostic group in Figure 5. Note that prognostics groups 1 and 3 each had cases from only one combination. (This table needs to be checked again. Or simply indicate Group 1 consists of what combinations. Size should be given in the table and the survival plot.) | | | | |

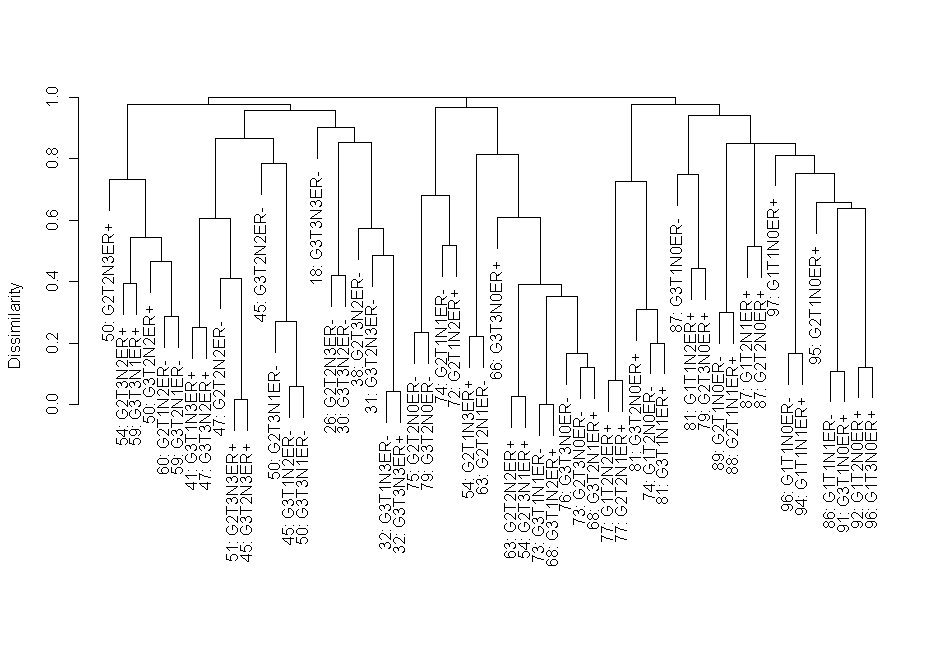
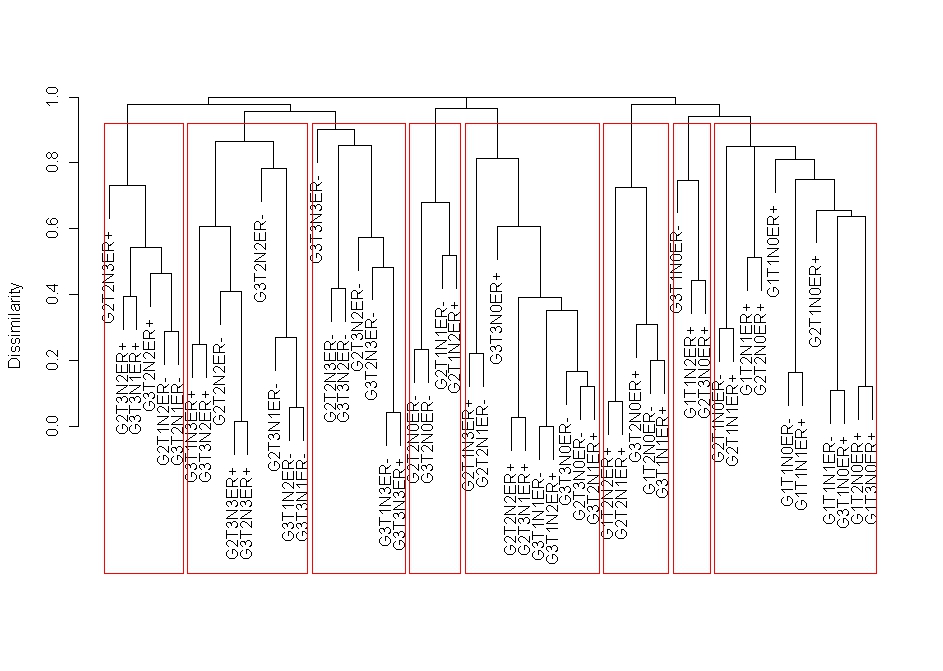


Figure 7.Dendrogram of breast cancer using combinations defined by four prognostic factors – grade, T, N, and ER status (Table 1). There are 72 combinations, but only 56 in the diagram since some combinationscontained less than 50 patients and were excluded. A combination is given beneath each leaf. The number before the colon (:) represents the 10 year survival (in percentage) of the combination.

Figure 8. Cutting the dendrogram illustrated in Figure 7 at level 2 creates 8 defined prognostic groups. Survival rates have been removed from the dendrogram. Only the codes for combinations remain.

7

6

4

8

5

3

2

1



Figure 9. 10-year survival rates for the 8 prognostic groups generated in Figure 7 with 4 prognostic factors – T, N, ER status, and histological grade. All rates are significantly different.