

**A Clustering-Based Approach to
Predict Outcome in Cancer
Patients**

Is medical practice a science?

What kind of science is it?

**Predictions in the Science of
Medicine**

<u>Prediction</u>	<u>Predictive Accuracy</u>
1. Risk of Disease	<<100%
2. Diagnosis	>>99%
3. Outcome	<<100%

**Predictive methods--very
important in medicine**

Current Prediction of Outcome in Cancer

- Achieved by classifying the extent of disease
- Extent – Most powerful predictive factor
- Extent -- classified into three variables T, N, M

Background

TNM

TNM defines extent of disease. It involves 3 variables:

- T the extent of the primary tumor
- N the absence or presence and extent of regional lymph node metastasis
- M the absence or presence of distant metastasis

Levels of these 3 variables can be combined. When combined and associated with survival, these 3 variables form the TNM staging system.

Example of Liver Cancer

Stage Grouping

I	T1	N0	M0
II	T2	N0	M0
IIIA	T3	N0	M0
IIIB	T4	N0	M0
IIIC	Any T	N1	M0
IV	Any T	Any N	M1

Definition of T Categories for Liver

- T1 Solitary tumor without vascular invasion
- T2 Solitary tumor with vascular invasion or multiple tumors, non more than 5 cm
- T3 Multiple tumors more than 5 cm or tumor involving a major branch of the portal or hepatic vein(s)
- T4 Tumor(s) with direct invasion of adjacent organs other than the gallbladder or with perforation of visceral peritoneum

N0 No regional lymph node metastasis
N1 Regional lymph node metastasis

M0 No distant metastasis
M1 Distant metastasis

Why Expand TNM

- Advances in molecular medicine, imaging, and therapeutics are now forcing a reconsideration of the TNM in order to accommodate additional prognostic factors
- TNM does not take into account demographic factors which influence outcome
- The TNM has only 3 factors (or variables), and cannot be expanded since a bin model
- The expanding role of early detection and screening reduces the benefit of TNM staging

Goal of Expansion

Our approach --- discover useful additional prognostic factors by grouping patients into subgroups such that:

- two survival experiences corresponding to two patients from the same subgroup should be close to each other
- two survival experiences corresponding to two patients from different subgroups should differ significantly

Cancer Data

Special issues include:

- Censoring
- Covariate type
- Large data size

Cancer data usually contain a high percentage of censored observations.

For example, a lung cancer data set from SEER that contains records of lung cancer patients (black and white with known stage and grade status) from 13 states who were diagnosed from 1988 through 2002 comprises 34% records with censored survival times.

Covariate Type & Large Data Size

- A typical cancer data set contains different types of covariates (factors, variables, etc.). Covariates can be continuous (e.g., tumor size), ordinal (e.g., grade), or nominal scale (e.g., sex).
- A cancer data set can be very large. For example, the lung cancer data set from SEER that contains records of lung cancer patients from 13 states who were diagnosed from 1973 through 2002 has a size of more than 500,000 patients, each patient having 81 measurements made on 81 covariates.

Algorithm

Survival Function

Let T be the lifetime (after diagnosis) of an individual.

The survival function is defined to be

$$S(t) = P(T > t)$$

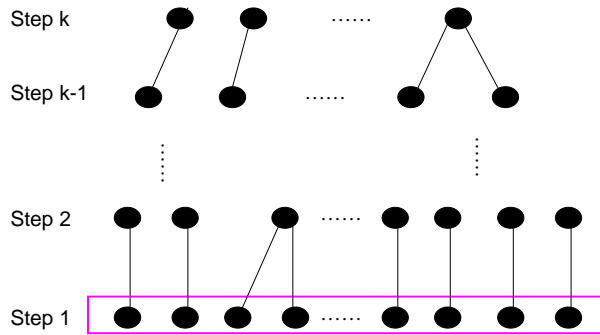
$S(t)$ represents the probability of survival to some t .

Survival function – estimated by KM

Clustering Algorithm

- starts with the partition of data using the combinations of levels of factors
- then the two most similar groups are merged at each subsequent step

Graphical Description of Algorithm



Result

We applied our algorithm to the breast cancer data

- 193,312 cases
- Patients from 13 areas in US
- Diagnosed from 1973 through 2003

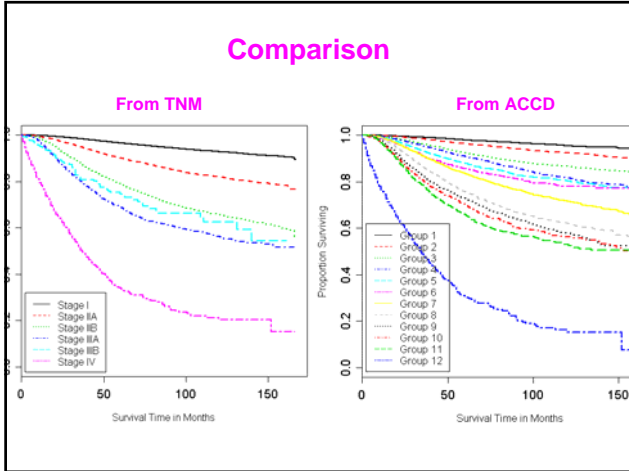
A Simple Example

- In running algorithm, we used the following factors from SEER EOD code
 - tumor size, 7 levels
 - tumor extension, 5 levels
 - lymph nodes, 2 levels
- Cases with unknown values on the 3 factors are filtered. The cleaned dataset contains 47 combinations, 191,941 cases in total.

Main Findings

After using our algorithm

- 29 groups are formed initially
- Combined into 12 groups each with a size larger than 1000, which include 185,829 cases total



Conclusions

- **A useful predictive system requires the integration of multiple clinical factors related to outcome**
- **Cluster analysis can incorporate any number of prognostic factors and can identify useful factors.**