

Expanding TNM for lung cancer through machine learning

Matthew Hueman¹ | Huan Wang² | Zhenqiu Liu³ | Donald Henson^{4†} |
Cuong Nguyen⁵ | Dean Park⁶ | Li Sheng⁷ | Dechang Chen⁴ 

¹Department of Surgical Oncology, John P. Murtha Cancer Center, Walter Reed National Military Medical Center, Bethesda, Maryland, USA

²Department of Biostatistics, George Washington University, Washington, District of Columbia, USA

³Department of Public Health Sciences, Penn State Cancer Institute, Hershey, Pennsylvania, USA

⁴Department of Preventive Medicine & Biostatistics, F. Edward Hébert School of Medicine, Uniformed Services University of the Health Sciences, Bethesda, Maryland, USA

⁵Department of Pathology, Uniformed Services University of the Health Sciences, Bethesda, Maryland, USA

⁶Department of Hematology-Oncology, John P. Murtha Cancer Center, Walter Reed National Military Medical Center, Bethesda, Maryland, USA

⁷Department of Mathematics, Drexel University, Philadelphia, Pennsylvania, USA

Correspondence

Dechang Chen, Department of Preventive Medicine & Biostatistics, F. Edward Hébert School of Medicine, Uniformed Services University of the Health Sciences, 4301 Jones Bridge Rd, Bethesda, MD 20814, USA.

Email: dechang.chen@usuhs.edu

Funding information

John P. Murtha Cancer Center Research Program, Grant/Award Number: Using Dendrograms to Create Prognostic Systems; Pennsylvania State University, Grant/Award Number: Four Diamonds Fund from Penn State University

Abstract

Background: Expanding the tumor, lymph node, metastasis (TNM) staging system by accommodating new prognostic and predictive factors for cancer will improve patient stratification and survival prediction. Here, we introduce machine learning for incorporating additional prognostic factors into the conventional TNM for stratifying patients with lung cancer and evaluating survival.

Methods: Data were extracted from SEER. A total of 77 953 patients were analyzed using factors including primary tumor (T), regional lymph node (N), distant metastasis (M), age, and histology type. Ensemble algorithm for clustering cancer data (EACCD) and C-index were applied to generate prognostic groups and expand the current staging system.

Results: With T, N, and M, EACCD stratified patients into 11 groups, resulting in a significantly higher accuracy in survival prediction than the 10 AJCC stages (C-index = 0.7346 vs. 0.7247, increase in C-index = 0.0099, 95% CI: 0.0091–0.0106, p-value = 9.2×10^{-147}). There nevertheless remained a strong association between the EACCD grouping and AJCC staging (rank correlation = 0.9289; p-value = 6.7×10^{-22}). A further analysis demonstrated that age and histological tumor could be integrated with the TNM. Data were stratified into 12 prognostic groups with an even higher prediction accuracy (C-index = 0.7468 vs. 0.7247, increase in C-index = 0.0221, 95% CI: 0.0212–0.0231, p-value $< 5 \times 10^{-324}$).

Conclusions: EACCD can be successfully applied to integrate additional factors with T, N, M for lung cancer patients.

KEYWORDS

C-index, lung cancer, machine learning, staging, survival

INTRODUCTION

Cancer staging systems play an essential role in cancer medicine. They are used to develop prognosis, determine appropriate treatments, evaluate clinical trials, and convey clinical experiences. Lung cancer is classified according to the TNM staging system, based on anatomic factors of

tumor extent, nodal status, and metastatic spread.¹ The TNM provides basic information for tumor evaluation, treatment, and prognosis. However, lung cancer is no longer characterized by the anatomic extent of disease, but by a combination of various factors that can be clinical, biological, molecular, or genetic. Therefore, new and important factors need to be integrated in order to build prognostic systems that can improve evaluation and management decisions for lung cancer patients. Unfortunately, additional

†Deceased.

prognostic factors cannot be easily incorporated into the TNM staging system because the system is a result of consensus across many different areas.

Cox regression modeling^{2,3} and tree modeling^{4,5} are two major approaches that allow expansion of the TNM by integrating additional factors. Although Cox regression modeling can achieve a high accuracy in survival prediction, the risk groups extracted from the output (e.g., the nomogram) usually have a lower accuracy of survival prediction than the original model. Traditional survival tree modeling, which can be used to explicitly define prognostic groups, does not provide a high prediction accuracy in general.

In this study we describe a machine learning approach using the ensemble algorithm for clustering cancer data (EACCD)⁶⁻¹⁷ to create prognostic systems for lung cancer. EACCD can adapt to any type and number of prognostic factors and generate systems that can be viewed as expansions of the TNM staging system. Variables/factors can be integrated by the EACCD to generate prognostic systems for refinements in patient stratification and outcome prediction that are needed for patient care such as monitoring of large scale therapeutic trials. We demonstrate the method by building two prognostic systems. One system, based on primary tumor (T), regional lymph node (N), and distant metastasis (M) was primarily employed to compare our approach with the AJCC. The second system, based on T, N, M, age (A), and histological type (H), expanded the traditional staging system based on T, N, M only. These prognostic systems from EACCD provide well-defined patient stratification and high accuracy of survival prediction.

METHODS

Data source

Disease-specific survival data with a primary diagnosis of lung cancer during 2010 to 2012 were obtained from 18 databases of the Surveillance, Epidemiology, and End Results Program (SEER) of the National Cancer Institute.¹⁸ This restriction on the year of diagnosis ensured a minimum five-year follow-up, since current release of SEER includes case reports up to the end of calendar year of 2017. As detailed below, patients diagnosed prior to 2010 were not included in our analysis. SEER cause-specific death classification variable¹⁹ was used to capture all deaths related to lung cancer. Survival time was measured in months.

Defining factors

SEER does not provide T, N, M categories in the eighth edition of the AJCC Cancer Staging Manual. Therefore, we used the derived AJCC seventh edition of the T, N, M variables²⁰ to match the T, N, M levels for the eighth AJCC cancer staging systems. Specifically, with the SEER Collaborative Stage (CS) data collection system, we made the following reclassification of T: (1) classify seventh

(edition) of T1a with CS tumor size²⁰ ≤ 10 mm as eighth of T1a; (2) classify seventh of T1a with CS tumor size between 10 and 20 mm as 8th of T1b; (3) classify seventh of T1b as eighth of T1c; (4) classify seventh of T2a with CS tumor size between 30 and 40 mm as eighth of T2a; (5) classify seventh of T2a with CS tumor size between 40 and 50 mm as eighth of T2b; (6) classify as eighth of T3 (i) seventh of T2b and (ii) seventh of T3 with CS tumor size ≤ 70 mm; (7) classify as eighth of T4 (i) seventh of T3 with CS tumor size >70 mm and (ii) seventh of T4.

Because SEER started to include the derived AJCC-7 T, N, M variables in 2010, patients with the year of diagnosis earlier than 2010 were not included in our analysis. This study investigated five factors: T, N, M, A, and H. Seven levels (T1a, T1b, T1c, T2a, T2b, T3, and T4) were used for T; four levels (N0, N1, N2, and N3) for N; and two levels (M0, M1) for M. Age and histological type were studied in this study since they are considered critical factors in survival prediction.^{2,3,5} Factor A had two levels: A0 (<70) and A1 (≥ 70). This cutoff, representing the lower boundary of senescence, was suggested by Gridelli et al.²¹ and was also used by Tanvetyanon et al.² We note that 70 is also the median age for lung cancer in the SEER data. We studied four main histological types: squamous cell carcinoma (H1), small cell carcinoma (H2), adenocarcinoma (H3), and large cell carcinoma (H4). These four types were defined according to WHO histological classification of tumors of the lung.²² Table S1 lists the detailed definition of T, N, M, A, and H.

Data management

Starting from the SEER lung cancer data with a primary diagnosis during 2010 to 2012, we selected all cases with complete information on the following factors/variables: T, N, M, A, H, survival time, and SEER cause-specific death classification variable. Further selection of cases was made in terms of combinations of factors. We define a combination as a subset of the data corresponding to one level of each factor and we use levels of factors to denote combinations (e.g., T1N1M0A0H1 represents a subset of patients with T = T1, N = N1, M = M0, A = A0, H = H1). Due to the statistical techniques employed in EACCD, we required each combination to contain a sufficient number of patients in order to optimize robustness of results. We retained each combination of T, N, M, A, and H that contained a minimum of 50 cases. The resulting dataset contained 227 combinations of T, N, M, A, and H (77 953 cases, Figure 1 and Table 1). The median follow-up of patients in the dataset was 70 months by the reverse Kaplan–Meier method.²³

EACCD

The EACCD (Supplementary Appendix A) is a machine learning algorithm for clustering combinations. It first defines initial dissimilarities between two combinations, then

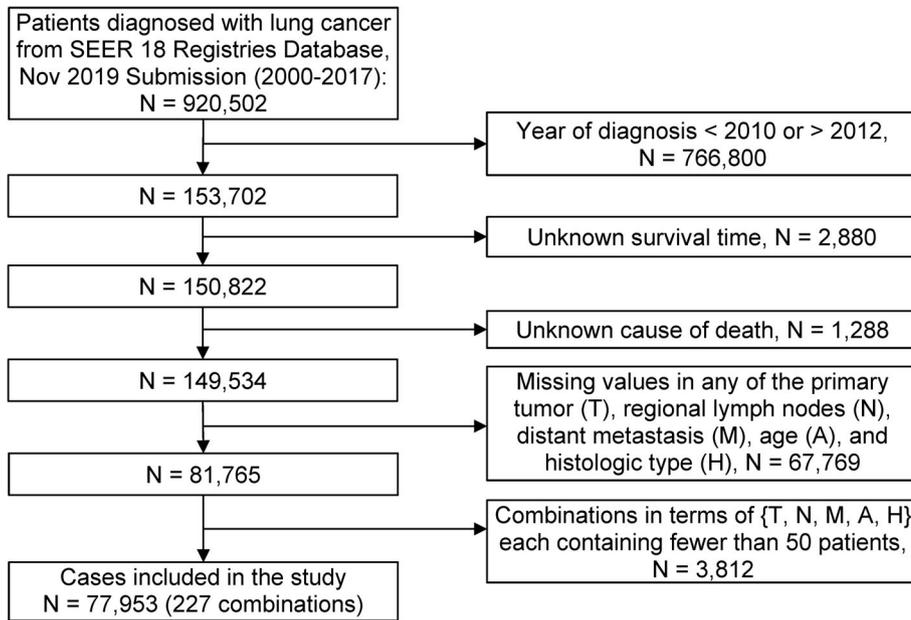


FIGURE 1 Clinical and demographic characteristics of the study cohort (N = 77 953)

TABLE 1 Clinical and demographic characteristics of the study cohort (N = 77 953)

	N	%
Primary tumor		
T1a	1708	2.2
T1b	10 088	12.9
T1c	8830	11.3
T2a	7939	10.2
T2b	4789	6.1
T3	16 636	21.3
T4	27 963	35.9
Regional lymph node		
N0	34 422	44.2
N1	6253	8.0
N2	28 355	36.4
N3	8923	11.4
Distant metastasis		
M0	43 936	56.4
M1	34 017	43.6
Age		
A0	39 547	50.7
A1	38 406	49.3
Histological type		
H1	22 306	28.6
H2	10 449	13.4
H3	44 437	57.0
H4	761	1.0

obtains learned dissimilarities by an ensemble learning process, and then performs hierarchical clustering analysis to cluster combinations. The output of the algorithm is a tree-structured

dendrogram, showing the relationship among survival of patients in different combinations. Several approaches are available for each step. In this study, the initial dissimilarity between two combinations was defined by the Mann–Whitney parameter²⁴ (Supplementary Appendix B); the ensemble learning process was based on the two-phase Partitioning Around Medoids algorithm;²⁵ and the minimax linkage method²⁶ was chosen for hierarchical clustering. This is the first time the Mann–Whitney parameter and the minimax linkage have been used together.

Prognostic systems

The dendrogram, obtained from the EACCD, can be cut horizontally to generate individual prognostic groups that serve the same role as the staging groups in the TNM. We cut the dendrogram in light of the C-index.²⁷ C-index serves as an estimate of the probability that a subject who died at an earlier time had a shorter predicted survival time than a subject who died at a later time. Because of the tradeoff between model simplicity and prediction accuracy, we chose the “optimal” number of prognostic groups n^* around the “knee” point of the C-index curve (the C-index vs. the number of prognostic groups).^{12,14–16} Survival curves for the prognostic groups were plotted by using Kaplan–Meier estimates.²⁸ The final prognostic system included the dendrogram, group assignment, C-index, and survival curves for the prognostic groups.

RESULTS

Prognostic system for T, N, M

Applying the EACCD to the data based on T, N, and M yielded the dendrogram in Figure 2(a). The C-index curve

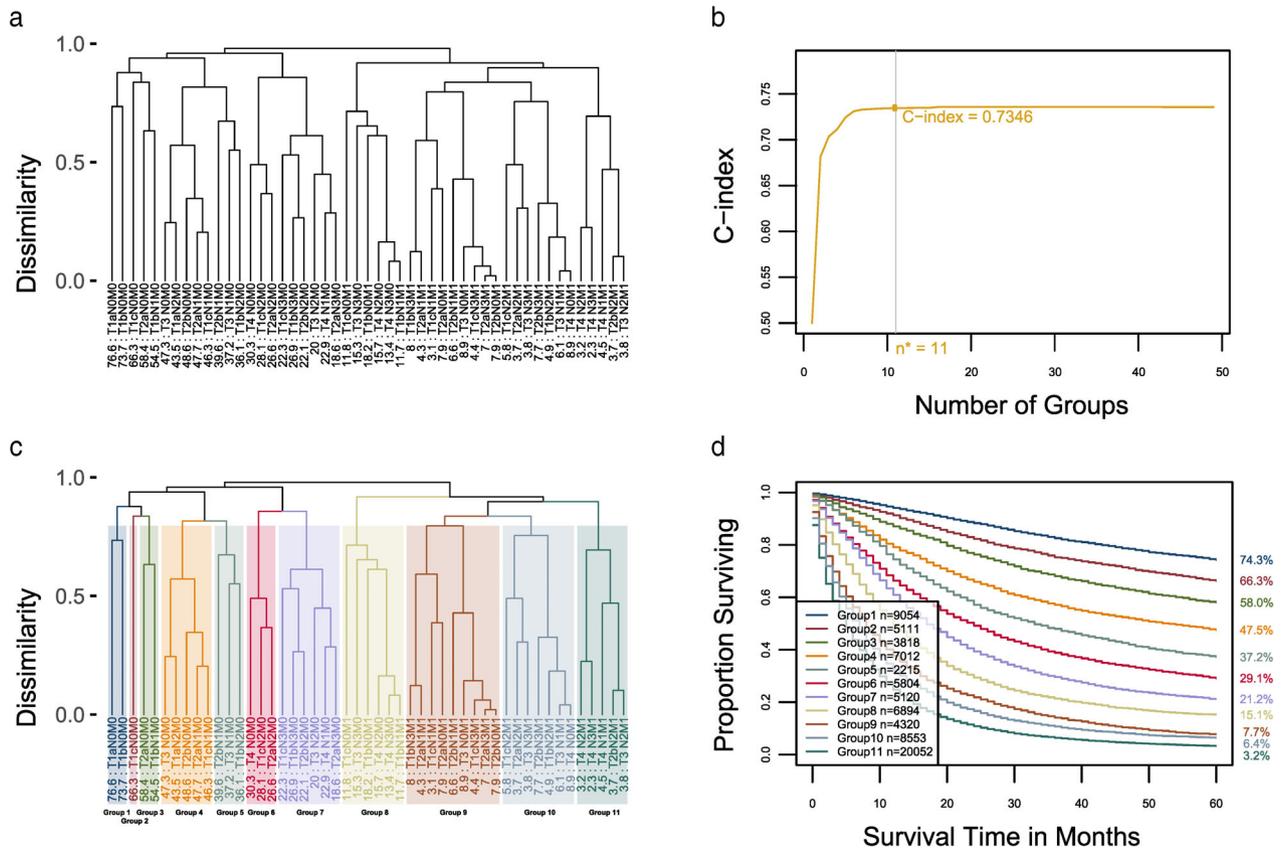


FIGURE 2 Creating ensemble algorithm for clustering cancer data (EACCD) prognostic groups on T, N, and M. (a) Dendrogram from running EACCD. A five-year cancer-specific survival rate in percentage is provided below each combination. (b) C-index curve based on the dendrogram in panel (a). The knee point of the curve corresponds to 11 groups and a C-index value of 0.7346. (c) Cutting the dendrogram in panel (a) according to $n^* = 11$ suggested in panel (b) creates 11 prognostic groups. Group numbers are listed on the bottom of the dendrogram. (d) Lung cancer-specific survival of 11 prognostic groups in panel (c). Five-year cancer-specific survival rates are listed on the right side

(Figure 2(b)) was used to find the optimal number of prognostic groups n^* . The knee point of the curve corresponds to 11 groups (C-index = 0.7346), which suggested $n^* = 11$ (the C-index slowly increases from 7 to 11 groups). Cutting the dendrogram into $n^* = 11$ groups is shown in rectangles (Figure 2(c)). The survival curves for these 11 groups are seen in Figure 2(d). For convenience, the definition for all 11 groups is restated in the fourth column of Table S2. The resulting prognostic system for T, N, and M includes the dendrogram with cutting (Figure 2(c)), the groups in the fourth column of Table S2, and the survival curves (Figure 2(d)). This system contains 11 groups: group 1, group 2 ... group 11 whose risk increases as the group number increases.

For comparison, the eighth edition AJCC divides the data into 10 groups. Details are seen in the fifth column of Table S2 and Figure 3 (we did not include stage 0 and we treated stage M1a/b/c as M1). Calculation shows that the AJCC staging system has a C-index of 0.7247. The p -value of the C-index based test²⁹ for testing differences between the prediction accuracy of the above EACCD prognostic system (11 groups, C-index = 0.7346) and the AJCC staging system TNM (10 groups, C-index 0.7247) was 9.2×10^{-147} .

This shows that the EACCD system has a significantly higher survival prediction accuracy than the AJCC system.

Prognostic system for T, N, M, A, and H

Before building the EACCD prognostic system for T, N, M, A, and H, we assessed the performance in survival prediction of the models for the following three sets of factors: {T, N, M, A}, {T, N, M, H}, and {T, N, M, A, H}, as compared with the model based on {T, N, M}. This is done by examining the C-index curves for all these four scenarios (Figure 4). For more than three groups, the curves for {T, N, M, A} and {T, N, M, H} are higher than the curve for {T, N, M}. Therefore, adding A or H to {T, N, M} increases the C-index and thus improves the prediction accuracy. The curve of {T, N, M, A, H} is the highest among all four curves, implying that adding both A and H to {T, N, M} leads to the biggest improvement on the prediction accuracy of {T, N, M}.

Applying the EACCD to the data based on T, N, M, A, and H yielded the dendrogram in Figure 5(a). The optimal number of prognostic groups $n^* = 12$ with a corresponding

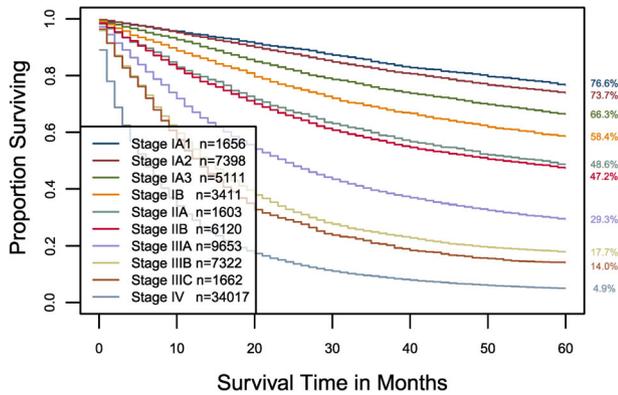


FIGURE 3 Cancer-specific survival of AJCC stages defined in the fifth column in Table S2. The five-year cancer-specific survival rates for 10 stages are listed on the right side of the figure

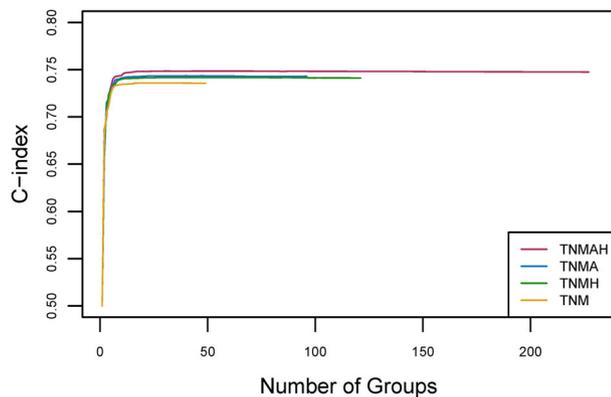


FIGURE 4 C-index curves based on different factors

C-index of 0.7468 is indicated in Figure 5(b). Therefore, we cut the dendrogram into $n^* = 12$ groups (rectangles in Figure 5(a)). Accordingly, the survival curves for the 12 prognostic groups can be plotted (Figure 5(c)). A detailed definition for all 12 groups is listed in Table S3.

DISCUSSION

Comparison with the TNM

The EACCD prognostic system based on TNM can be compared with the AJCC staging system in terms of both stratification and prediction. Previously, we showed that the EACCD system based on TNM (11 prognostic groups with C-index = 0.7346) has a significantly higher survival prediction accuracy than the AJCC staging system (10 stage groups with C-index = 0.7247). Below we compare the two systems by examining how patients are stratified.

In fact, there is a strong positive association between AJCC staging and EACCD grouping. Table 2 presents the distribution of patients of each of 10 AJCC stages over the 11 EACCD groups. The upper right and lower left corners of the table are filled with 0. Approximately, the higher stage

the patient is assigned to by the AJCC system, the higher risk group the patient is assigned to by the EACCD, and vice versa. Indeed, the assignment to ordered stages and the assignment to ordered prognostic groups have a large Spearman’s rank correlation coefficient³⁰ of 0.9289 with a p -value of 6.7×10^{-22} .

In summary, in predicting survival, the EACCD prognostic system on {T, N, M} has a significantly higher accuracy than the AJCC staging system TNM; in stratifying patients, the EACCD grouping and AJCC staging are strongly positively associated.

We note that the EACCD prognostic system on {T, N, M} can be further compared with the AJCC staging system TNM through validation datasets, preferably from sources other than SEER. For instance, when an appropriate validation set is available, EACCD and AJCC can be compared in terms of survival curves and values of C-index on the validation set.

Effect of factor levels on survival

The EACCD prognostic system on {T, N, M, A, H} allowed us to examine the effect of levels of individual factors on survival. To simplify the analysis, we considered the following two risk categories: low risk (groups 1–5) and high risk (groups 6–12) (this partition was suggested by the survival of prognostic groups shown in Figure 5(c)). A graph is used to show how patients associated with a factor level are distributed across the two risk categories (Figure 5(d)).

The first row shows that as the T levels become more aggressive, patients are more likely to be classified into the high risk category.

The second and third rows reveal that patients with N0, or N1, or M0 status tend to have favorable survival while those with N2, or N3, or M1 have unfavorable survival.

The fourth row indicates that A0 and A1 curves are similar, both showing an approximately equal distribution in the two risk categories. This is the marginal effect of age, given the two risk categories. Earlier age was shown to be an important prognostic factor. These suggest that age should be considered in conjunction with other factors when informing prognosis.

The fifth row details the distribution of patients associated with each histological type. Patients with H1, H3, and H4 are approximately even distributed across the low and high risk categories, suggesting that squamous cell carcinoma, adenocarcinoma, and large cell carcinoma are not prone to high or low risk when these levels are presented alone. In comparison, small cell carcinoma shows a strong tendency towards high risk. (The H2 curve is increasing, with a small percentage of patients at low risk and a majority at high risk.) This finding reconfirms that small cell carcinoma and non-small cell carcinoma play different roles in prognosis.^{31,32}

The above analysis shows how a factor level is associated with risk. Although these observations have been previously reported in the literature, this is the first time that these

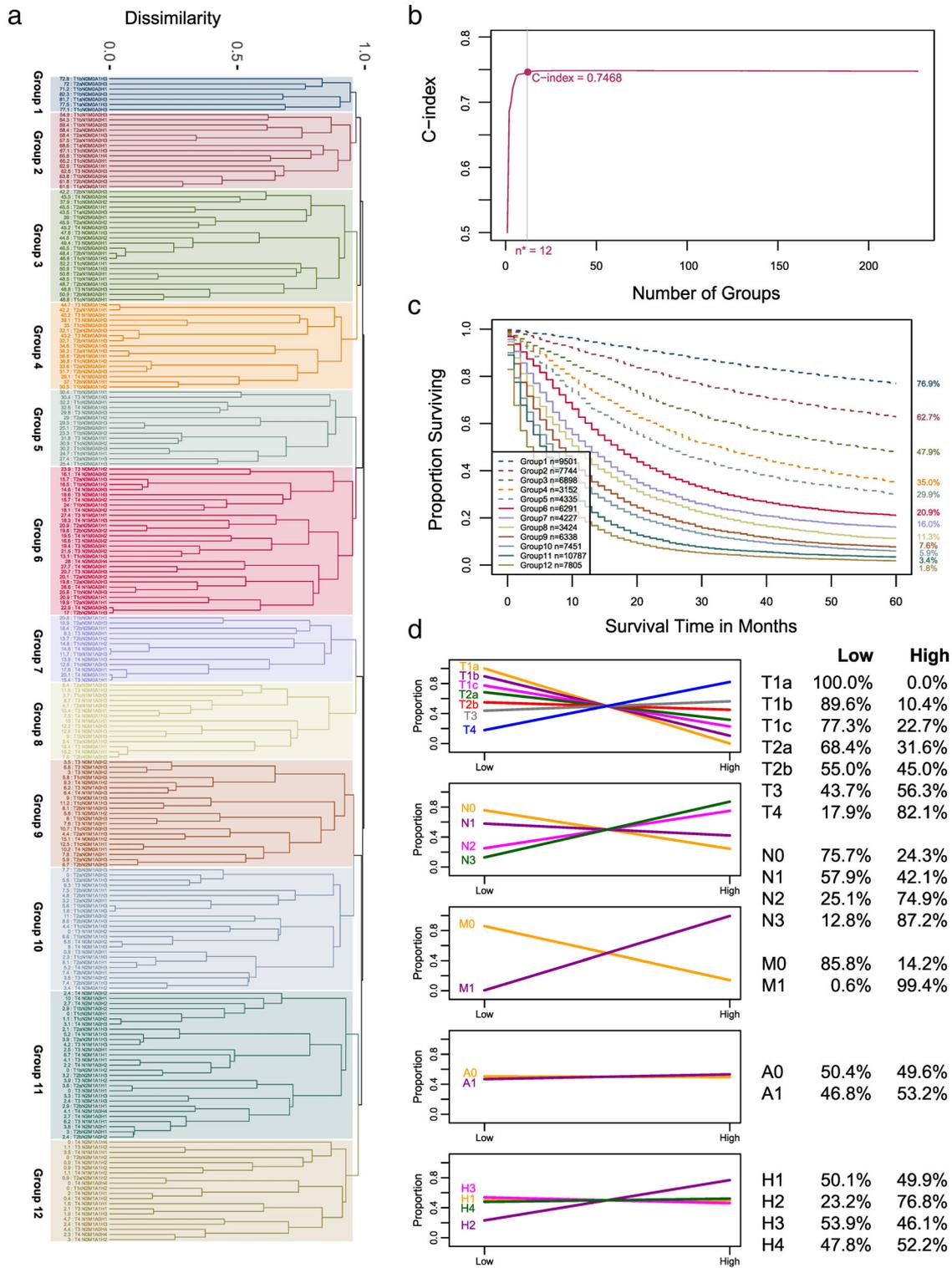


FIGURE 5 Ensemble algorithm for clustering cancer data (EACCD) prognostic groups on T, N, M, A, and H. (a) Dendrogram and cutting the dendrogram (shown in rectangles). A five-year cancer-specific survival rate in percentage is provided to the left of each combination. Cutting the dendrogram according to $n^* = 12$ in panel (b) creates 12 prognostic groups. Listed on the left of the dendrogram are group numbers. (b) C-index curve based on the dendrogram in panel (a). The knee point of the curve corresponds to 12 groups and a C-index value of 0.7468. (c) Lung cancer-specific survival of 12 prognostic groups in panel (a). The five-year cancer-specific survival rates for 12 groups are listed on the right side of the figure. (d) Distributions of patients over risk categories. In each row, one factor is concerned, and for each level of the factor, the distribution of patients (two proportions at two risk categories) is presented in two ways: Plot on the left and tabulation on the right

TABLE 2 Contingency table between ensemble algorithm for clustering cancer data (EACCD) grouping and AJCC staging on the basis of T, N, M

AJCC\EACCD	1	2	3	4	5	6	7	8	9	10	11	Total
IA1	1656	0	0	0	0	0	0	0	0	0	0	1656
IA2	7398	0	0	0	0	0	0	0	0	0	0	7398
IA3	0	5111	0	0	0	0	0	0	0	0	0	5111
IB	0	0	3411	0	0	0	0	0	0	0	0	3411
IIA	0	0	0	1603	0	0	0	0	0	0	0	1603
IIB	0	0	407	5357	356	0	0	0	0	0	0	6120
IIIA	0	0	0	52	1859	5804	1938	0	0	0	0	9653
IIIB	0	0	0	0	0	0	3182	4140	0	0	0	7322
IIIC	0	0	0	0	0	0	0	1662	0	0	0	1662
IV	0	0	0	0	0	0	0	1092	4320	8553	20 052	34 017
Total	9054	5111	3818	7012	2215	5804	5120	6894	4320	8553	20 052	77 953

factor levels have been integrated together and explicitly highlighted in the ordered risk groups of the prognostic system TNMAH created in this study.

Limitations of analyses

Cancer-specific survival data were used in this study. Although the SEER cause-specific death classification is determined by taking into account other elements (e.g., tumor sequence, site of the original cancer diagnosis, and comorbidities), death certificate errors can be problematic in estimating the cause-specific survival. Another limitation is that the EACCD requires a relatively large size for each combination to produce robust estimates of survival. This report includes combinations with at least 50 cases. This may exclude some “rare” but interesting combinations. Improved estimates of survival can be achieved with a larger cutoff. Clearly, this requirement on the size of combinations will be met automatically when more data becomes available. Finally, due to the current restriction of SEER data, we derived the AJCC eighth edition of the T, N, M variables from those of the seventh edition, which could introduce some bias to this present study.

In conclusion, here we describe a machine learning approach based on EACCD and C-index to refine the TNM system for lung cancer by integrating additional prognostic factors. We demonstrated the approach by using the SEER lung cancer data to create a prognostic system based on T, N, and M, which classifies patients in a way strongly positively correlated with the AJCC TNM staging system but has a higher accuracy for predicting survival. Using SEER, we created one computational prognostic system based on T, N, M, A, and H, which expanded (with additional factors) and improved (with a higher accuracy of survival prediction) the TNM for lung cancers. Results have shown that the machine learning approach takes into account both prediction and stratification and is analogous to the AJCC scheme in generating stages.

ACKNOWLEDGMENTS

This work was partially supported by grants “Using Dendrograms to Create Prognostic Systems for Cancer” and “Creating Prognostic Systems for Cancer” sponsored by John P. Murtha Cancer Center Research Program and grant “Four Diamonds Fund from Penn State University” sponsored by Penn State University.

Disclaimer: The contents, views or opinions expressed in this publication or presentation are those of the authors and do not necessarily reflect official policy or position of Uniformed Services University of the Health Sciences, the Department of Defense (DoD), or Departments of the Army, Navy, or Air Force. Mention of trade names, commercial products, or organizations does not imply endorsement by the U.S. Government.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ORCID

Dechang Chen  <https://orcid.org/0000-0001-5787-3441>

REFERENCES

- Amin MB, Edge S, Greene F, Byrd DR, Brookland RK, Washington MK, et al. AJCC cancer staging manual. 8th ed. New York: Springer International Publishing; 2017. <https://www.springer.com/gp/book/9783319406176>.
- Tanvetyanon T, Finley DJ, Fabian T, Riquet M, Voltolini L, Kocaturk C, et al. Prognostic nomogram to predict survival after surgery for synchronous multiple lung cancers in multiple lobes. *J Thorac Oncol*. 2015;10(2):338–45.
- Liang W, Zhang L, Jiang G, Wang Q, Liu L, Liu D, et al. Development and validation of a nomogram for predicting survival in patients with resected non-small-cell lung cancer. *J Clin Oncol*. 2015;33(8):861–9.
- Sculier JP, Chansky K, Crowley JJ, van Meerbeek J, Goldstraw P. The impact of additional prognostic factors on survival and their relationship with the anatomical extent of disease expressed by the 6th edition of the TNM classification of malignant tumors and the proposals for the 7th edition. *J Thorac Oncol*. 2008;3(5):457–66.
- Chansky K, Sculier JP, Crowley JJ, Giroux D, van Meerbeek J, Goldstraw P. The International Association for the Study of Lung

- Cancer staging project: prognostic factors and pathologic TNM stage in surgically managed non-small cell lung cancer. *J Thorac Oncol*. 2009;4(7):792–801.
6. Chen D, Xing K, Henson D, Sheng L, Schwartz AM, Cheng X. Developing prognostic systems of cancer patients by ensemble clustering. *BioMed Res Int*. 2009;2009. <https://doi.org/10.1155/2009/632786>.
 7. Wu D, Yang C, Wong S, Meyerle J, Zhang B, Chen D. An examination of TNM staging of melanoma by a machine learning algorithm. In 2012 international conference on computerized healthcare (ICCH), (pp. 120–126). IEEE. 2012.
 8. Qi R, Wu D, Sheng L, Henson D, Schwartz A, Xu E, et al. On an ensemble algorithm for clustering cancer patient data. *BMC Syst Biol*. 2013;7(4):S9.
 9. Chen D, Hueman MT, Henson DE, Schwartz AM. An algorithm for expanding the TNM staging system. *Future Oncol*. 2016;12(8):1015–24.
 10. Chen D, Wang H, Sheng L, Hueman MT, Henson DE, Schwartz AM, et al. An algorithm for creating prognostic systems for cancer. *J Med Syst*. 2016;40(7):160.
 11. Wang H, Chen D, Hueman MT, Sheng L, Henson DE. Clustering big cancer data by effect sizes. In proceedings of the second IEEE/ACM international conference on connected health: applications, systems and engineering technologies, (pp. 58–63). IEEE Press; 2017. <https://ieeexplore.ieee.org/abstract/document/8010615>.
 12. Hueman MT, Wang H, Yang CQ, Sheng L, Henson DE, Schwartz AM, et al. Creating prognostic systems for cancer patients: a demonstration using breast cancer. *Cancer Med*. 2018;7(8):3611–21.
 13. Wang H, Hueman M, Pan Q, Henson D, Schwartz A, Sheng L, et al. Creating prognostic systems by the Mann-Whitney parameter. In 2018 IEEE/ACM international conference on connected health: applications, systems and engineering technologies (CHASE), (pp. 33–39). IEEE; 2018. <https://ieeexplore.ieee.org/abstract/document/8648686>.
 14. Yang C, Gardiner L, Wang H, Hueman M, Chen D. Creating prognostic systems for well differentiated thyroid cancer using machine learning. *Front Endocrinol*. 2019;10:288.
 15. Hueman M, Wang H, Henson D, Chen D. Expanding the TNM for cancers of the colon and rectum using machine learning: a demonstration. *ESMO Open*. 2019;4(3):e000518.
 16. Wang H. Development of prognostic systems for cancer patients. Doctoral dissertation, The George Washington University; 2020. <https://search.proquest.com/openview/3add97db64a77ddab4b1ceca94b44ec/1?pq-origsite=gscholar&cbl=18750&diss=y>
 17. Praiss AM, Huang Y, Clair CM, Tergas AI, Melamed A, Khoury-Collado F, et al. Using machine learning to create prognostic systems for endometrial cancer. *Gynecol Oncol*. 2020;159(3):744–50.
 18. Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) Research Data (2000–2017), National Cancer Institute, DCCPS, Surveillance Research Program. 2020.
 19. SEER Cause-specific Death Classification. 2019. <https://seer.cancer.gov/causespecific/>. Accessed 15 August 2019.
 20. SEER Research Data Record Description. 2020. <https://seer.cancer.gov/data-software/documentation/seerstat/nov2019/TextData.FileDescription.pdf> Accessed 27 December 2020.
 21. Gridelli C, Maione P, Colantuoni G, Rossi A. Chemotherapy of non-small cell lung cancer in elderly patients. *Curr Med Chem*. 2002;9(16):1487–95.
 22. Travis WD, Brambilla E, Müller-Hermelink HK, Harris CC, editors. Pathology and genetics of tumours of the lung, pleura, thymus and heart. World Health Organization Classification of Tumours. Lyon: IARC Press; 2004. <https://patologi.com/who%20lung.pdf>.
 23. Schemper M, Smith TL. A note on quantifying follow-up in studies of failure time. *Control Clin Trials*. 1996;17(4):343–6.
 24. Mann HB, Whitney DR. On a test of whether one of two random variables is stochastically larger than the other. *Ann Math Stat*. 1947;18:50–60.
 25. Kaufman L, Rousseeuw PJ. Finding groups in data: an introduction to cluster analysis. Hoboken, New Jersey: John Wiley & Sons; 1990. http://www.ru.ac.bd/stat/wp-content/uploads/sites/25/2019/03/503_09_Kaufman_Finding-Groups-in-Data-An-Introduction-to-Cluster-Analysis.pdf.
 26. Bien J, Tibshirani R. Hierarchical clustering with prototypes via mini-max linkage. *J Am Stat Assoc*. 2011;106(495):1075–84.
 27. Harrell FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med*. 1996;15(4):361–87.
 28. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *J Am Stat Assoc*. 1958;53(282):457–81.
 29. Kang L, Chen W, Petrick NA, Gallas BD. Comparing two correlated C indices with right-censored survival outcome: a one-shot nonparametric approach. *Stat Med*. 2015;34(4):685–703.
 30. Daniel WW. Biostatistics: a foundation for analysis in the health sciences. 7th ed. New York: John Wiley & Sons; 1999.
 31. van Meerbeeck JP, Fennell DA, de Ruyscher DK. Small-cell lung cancer. *Lancet*. 2011;378(9804):1741–55.
 32. Byers LA, Rudin CM. Small cell lung cancer: where do we go from here? *Cancer*. 2015;121(5):664–72.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Hueman M, Wang H, Liu Z, et al. Expanding TNM for lung cancer through machine learning. *Thorac Cancer*. 2021;12:1423–1430. <https://doi.org/10.1111/1759-7714.13926>