JAMA Network Open

Invited Commentary | Oncology

Prognosis and Treatment of Non–Small Cell Lung Cancer in the Age of Deep Learning

Frederick Matthew Howard, MD; Alexander T. Pearson, MD, PhD

+ Related article

Author affiliations and article information are listed at the end of this article.

Lung cancer is both the most common and the most deadly cancer, with more than 2 million cases diagnosed worldwide in 2018 per Global Cancer Observatory estimates and with non–small cell lung cancer (NSCLC) accounting for the great majority of cases. The 8th edition of the American Joint Committee on Cancer TNM stage groupings represents the most well-validated prognostic metric for NSCLC. However, a variety factors unaccounted for by these TNM stage groupings affect outcome, from patient-specific factors such as performance status, age, and socioeconomic status to tumor characteristics such as grade, lymphovascular invasion, programmed cell death 1 ligand 1 expression, and the presence of molecular driver variants. Integrating the various clinical and pathologic characteristics of each case to provide an accurate prognosis is challenging in the absence of easy-to-use and comprehensive predictive models.

She and colleagues[1] approach this problem using a novel deep learning proportional hazards model, DeepSurv.[2] They compiled data from the Surveillance, Epidemiology, and End Results (SEER) database for a subset of 16 140 patients with NSCLC who received a diagnosis between 2010 and 2015, the vast majority of whom underwent surgical resection for stage IIIA or earlier disease. A DeepSurv survival model incorporating the demographic characteristics of the patients, markers of tumor extent (such as TNM stage), and type of surgical resection was trained on 80% of the patients selected from the SEER database. This model outperformed use of the TNM staging system alone in both the remaining 20% of the SEER data (C statistic = 0.739 vs 0.706; $P < .001$) and an independent validation set of 1182 patients with resected lung cancer treated at Shanghai Pulmonary Hospital in China (C statistic = 0.742 vs 0.706; $P < .001$). These results are in line with the accuracy of neural network models of recurrence of resected NSCLC[3] and provide additional evidence of the increasing benefit of deep learning prognostic models over standard approaches.

After a seminal randomized clinical trial by the Lung Cancer Study Group demonstrated superior survival for patients with T1N0 lung cancer who underwent a lobectomy as opposed to a limited resection,[4] the role of sublobar resection has been hotly debated. Numerous studies have attempted to clarify this issue, including a prior SEER database analysis that found wedge resection and segmentectomy were associated with worse overall survival than lobectomy.[5] No subgroup of patients could be identified for whom limited resection would provide equivalent results; a benefit for lobectomy was seen even in tumors smaller than 1 cm in size. However, prior retrospective studies on this topic are limited by the use of linear models, in which covariates have a consistent association with survival independent of other model inputs. One principle advantage of DeepSurv and other nonlinear survival methods is the inherent ability to model the interaction between covariates without the need for explicit specification (such as interaction terms or repeated analysis in multiple subgroups).[2] By modeling the interdependence of patient- and disease-specific factors with treatment received, DeepSurv can provide a case-by-case prediction of survival with different therapeutic modalities, leading to truly personalized treatment decisions. She et al[1] apply these principles to identify patients who may safely undergo sublobar resections without compromising outcomes. She et al[1] trained 2 DeepSurv models, one on 10 766 of the aforementioned SEER patients who received lobectomy and another on 1444 patients undergoing sublobar resection. Lobectomy resection was recommended for patients who had longer predicted survival with the former model than the latter. Of the 3064 patients in the held-out validation subset of the SEER data set, as well as the 1142 patients examined in an external single-center cohort data set, survival was prolonged for patients receiving treatment according to these recommendations (with a hazard ratio of 2.99 for the SEER data set and 2.14 for an external single-center cohort data set [both $P < .001$]). Most importantly, survival was equivalent for lobectomy vs a more limited surgery if the model had recommend sublobar resection, which suggests that the model successfully identified patients for whom sublobar resection can be safely performed.

The ability to select patients for less extensive surgery without compromising outcome has practice-changing implications, but several obstacles have stymied the use of machine learning tools for patient care. She et al[1] have already crossed 1 hurdle by designing an intuitive interface for clinicians, with graphical survival curves that can facilitate shared decision-making. Another inherent challenge with deep learning is the lack of transparency of the inner workings of the model.[6] Whereas associations in linear models can be assessed for biologic plausibility, the factors that lead the DeepSurv model to recommend sublobar resection are undefined, and thus predictions must be held to a high degree of scrutiny before implementation. Several methods can provide insight into the inner workings of neural networks.[7] Analysis of variable importance can identify features with the strongest correlation to the outcome in question. Single variable sensitivity analysis could demonstrate plausible associations between features and outcome—such as an increased preference for sublobar resection in the smallest size tumors or a decreased benefit of lobectomy for patients of advanced age. Furthermore, preexisting bias within the training data set may lead to bias in the treatment recommendations.[6] For example, if socioeconomic factors are associated with certain patients receiving sublobar resections by more skilled surgeons, the model may incorrectly assume that such patients are the only to benefit from resection. The most rigorous way to validate recommendations would be through a trial in which patients with discordant physician and model recommendations are randomized to receive lobectomy or sublobar resection, but such a trial may not be feasible. Alternatively, continuous study of real-world use would build confidence in the accuracy of the model. Although excitement must be tempered by the need for further study, the work of She et al[1] has laid the groundwork for the future of personalized prognosis and treatment of early stage NSCLC.

ARTICLE INFORMATION

Published: June 3, 2020. doi:10.1001/jamanetworkopen.2020.6368

Corresponding Author: Alexander T. Pearson, MD, PhD, Department of Medicine, Section of Hematology/Oncology, The University of Chicago, 5841 S Maryland Ave, MC 2115, Chicago, IL 60637 (apearson5@medicine.bsd.uchicago.edu).

Author Affiliations: Department of Medicine, Section of Hematology/Oncology, The University of Chicago, Chicago, Illinois.

Conflict of Interest Disclosures: Dr Pearson reported being on the advisory board for Prelude Therapeutics.

REFERENCES

1. She Y, Jin Z, Wu J, et al. Development and validation of a deep learning model for non–small cell lung cancer survival. *JAMA Netw Open*. 2020;3(6):e205842. doi:10.1001/jamanetworkopen.2020.5842

2. Katzman JL, Shaham U, Cloninger A, Bates J, Jiang T, Kluger Y. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Med Res Methodol*. 2018;18(1):24. doi:10.1186/s12874-018-0482-1

3. Lee B, Chun SH, Hong JH, et al. DeepBTS: prediction of recurrence-free survival of non-small cell lung cancer using a time-binned deep neural network. *Sci Rep*. 2020;10(1):1952. doi:10.1038/s41598-020-58722-z

4. Ginsberg RJ, Rubinstein LV; Lung Cancer Study Group. Randomized trial of lobectomy versus limited resection for T1 N0 non-small cell lung cancer. *Ann Thorac Surg*. 1995;60(3):615-622. doi:10.1016/0003-4975(95)00537-U

5. Dai C, Shen J, Ren Y, et al. Choice of surgical procedure for patients with non–small-cell lung cancer ≤ 1 cm or > 1 to 2 cm among lobectomy, segmentectomy, and wedge resection: a population-based study. *J Clin Oncol*. 2016;34(26):3175-3182. doi:10.1200/JCO.2015.64.6729

6. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med*. 2019;25(1):44-56. doi:10.1038/s41591-018-0300-7

7. Zhang Z, Beck MW, Winkler DA, Huang B, Sibanda W, Goyal H; written on behalf of AME Big-Data Clinical Trial Collaborative Group. Opening the black box of neural networks: methods for interpreting neural network models in clinical applications. *Ann Transl Med*. 2018;6(11):216. doi:10.21037/atm.2018.05.32