Selecting high-risk individuals for lung cancer screening; the use of risk prediction models *vs.* simplified eligibility criteria

Rudolf Kaaks[1,2], Anika Hüsing[1,2], Renée T. Fortner[1,2]

[1]Division of Cancer Epidemiology, German Cancer Research Center (DKFZ), Heidelberg, Germany; [2]Translational Lung Research Center Heidelberg (TLRC-H), Member of the German Center for Lung Research (DZL), Heidelberg, Germany

Correspondence to: Rudolf Kaaks. Division of Cancer Epidemiology, German Cancer Research Center (DKFZ), Heidelberg, Germany.
Email: r.kaaks@dkfz.de.

Provenance: This is a Guest Editorial commissioned by Section Editor Jianrong Zhang, MD (Department of Thoracic Surgery, First Affiliated Hospital of Guangzhou Medical University, Guangzhou Institute of Respiratory Disease, Guangzhou, China).

Comment on: Ten Haaf K, Jeon J, Tammemägi MC, *et al.* Risk prediction models for selection of lung cancer screening candidates: A retrospective validation study. PLoS Med 2017;14:e1002277.

Submitted Jun 22, 2017. Accepted for publication Jul 03, 2017.
doi: 10.21037/atm.2017.07.14
View this article at: http://dx.doi.org/10.21037/atm.2017.07.14

In 2011 the US National Lung Cancer Screening Trial (NLST) showed a 20% reduction in lung-cancer mortality among long-term heavy smokers screened by low-dose spiral computed tomography (LDCT) compared to standard X-ray diagnostics (1), and the US Prevention Services Task Force (USPSTF) recommends lung cancer screening for all individuals who meet the original NLST eligibility criteria of being 55 to 80 years of age, having smoked 30 or more pack years, and not having quit smoking more than 15 years ago. In Europe, among seven European lung cancer screening trials, jointly including less than 38,000 participants (2) *vs.* more than 53,000 in the NLST alone, preliminary results from four studies so far did not provide definitive confirmation of a mortality reduction (3-5), but individually these trials were small and lacked statistical power. A final evaluation of the effects on lung cancer mortality in at least three further European trials, including the larger Dutch-Belgian NELSON study (n=7,915 in the screening arm) (6), is expected within the next 1–2 years, as well as a pooled analysis of the European trials. As in the NLST, the eligibility criteria for the European trials were based on age and a simplified index of past cumulative smoking exposure, plus time since quitting for former smokers. The criteria, however, differed from those in the NLST in terms of specific age range, minimum lifetime smoking duration, cumulative smoking exposure (pack years), and maximum time since smoking cessation (7,8) and the criteria varied also across the European trials themselves.

At start of the trials, a major motivation for focusing screening on individuals meeting these eligibility criteria was cost. The vast majority of lung cancer cases occur among long-term smokers and at a more advanced age, and focusing on high-risk individuals allows drastic reduction in the number of individuals needed to be screened to identify one case, while still capturing the majority of individuals who develop lung cancer. A further motivation for focusing on high-risk individuals is that it enriches the screened population with participants who may have a benefit from screening (i.e., those who will actually have a screen-detectable lung cancer) while avoiding possible harms to individuals who are unlikely to have a lung tumor. The NLST and the various European trials have now provided extensive documentation not only of the potential benefit of screening, in form of mortality reduction and life years that may be gained by early lung cancer detection, but also of the possible harms. One major type of harm is false-positive diagnoses. In NLST and other trials, depending on the diagnostic criteria (e.g., nodule size) and procedures used for further diagnostic work-up, up to 25% of participants were classified as having suspicious nodules necessitating follow-up examinations, associated with psychologic stress and additional radiation exposure during diagnostic verification. Smaller but still significant

proportions of screening participants underwent more invasive examinations (bronchoscopy, biopsy, surgery), with substantial risks of complications, before being definitively classified with false-positive diagnoses (2). Another possible harm is over-diagnosis: the detection of tumors that, either because of unaggressive biologic behavior or because of a person's limited life expectancy, would not present symptomatically during this person's lifetime in the absence of screening. Although harder to estimate than the occurrences of false-positive diagnoses, modeling suggests that screening according to the USPSTF guidelines may result in about one over-diagnosis for three cancer deaths avoided (9). It is assumed that, by focusing screening on a high-risk population, the ratio of true-positive case detection to false-positive diagnoses and over-diagnosis will be generally improved.

Accumulating evidence on both benefits and harms of lung cancer screening has stimulated research on the question for whom lung screening may work best, in view of optimizing net clinical benefit. Epidemiologic models for estimating individualized lung cancer risk (incidence, mortality) have been developed that, besides a more detailed modeling of risks in relation to age, account for more detailed smoking history in terms of lifetime duration, intensity and ages at start and quitting, while including also further risk predictors such as pre-existing lung diseases (COPD, pneumonia), family history of cancer or occupational exposures to asbestos, allergies and respiratory function tests (10). Models have been developed mostly, though not exclusively, within the context of prospective cohort studies, which have the advantage that besides fitting models for optimal risk discrimination they also allow models to be directly fitted to absolute disease risks. Major models developed in the context of prospective studies include a model by Bach and colleagues (11) developed within CARET trial, a two-stage clonal expansion model (TSCE) developed within the Nurses' Health Study and Health Professionals' Follow-up Studies (12), and a model within the Prostate, Lung, Colon and Ovarian cancer trial cohort ($PLCO_{m2012}$) (13), and further models were developed within the European EPIC cohort (14) and in the UK Biobank (15). Other models were based on case-control data for the development of a relative risk discrimination score, using population incidence data from local cancer registries to calculate an appropriate risk model intercept for calibrated estimates of individuals' absolute risks; this includes a model developed by Spitz *et al.* (16), in context of a large US case-control study, and by Cassidy *et al.*, developed in context of the Liverpool Lung Project (LLP model) (17).

For lung cancer prediction models to be useful in actual screening context, they need to provide sufficiently strong discrimination between individuals with high *vs.* low likelihood of being diagnosed with lung cancer in the following years. Also, in view of selecting a defined risk threshold for screening eligibility, models need to be well-calibrated in terms of absolute risk estimation. Both qualities—discrimination and calibration—should hold not only within specific cohorts or population contexts in which models were originally developed, but should also translate robustly to new screening settings in other populations. Within a given study cohort, statistical re-sampling methods such as bootstrapping may help adjust for model over-fitting and over-optimism in estimated diagnostic discrimination characteristics; however, such methods do not adjust for variability in unspecified risk determinants that may cause variation in discrimination capacity or absolute risk calibration across different population settings. Thus, to gain confidence that models will function properly in new populations, it is important that they be externally validated in independent population data sets. For validation of absolute risk estimates, external validation of risk prediction models is best performed within prospective population-based study cohorts or screening trials.

In *PLoS Medicine*, Ten Haaf and colleagues (18) recently reported on a comprehensive validation of seven different prediction models for lung cancer incidence, and two models for lung cancer mortality, using data of the control and intervention arms of NLST (>53,000 participants; 1,925 lung cancer cases between study entry and 6 years of follow-up) and of the PLCO trial (>80,000 ever-smoking participants; 1,463 lung cancer cases) to examine the performances of each model for prediction of individuals diagnosed with, or dying from, lung cancer within the first 6 years of prospective follow-up. For all models tested, the discrimination was substantially better in the PLCO (AUCs ranging from 0.74 to 0.81) than in the NLST datasets (AUCs ranging from 0.61 to 0.73)—a difference that can be explained by the greater heterogeneity in risk factor profiles in the PLCO cohorts (individuals not selected by smoking history) compared to the NLST (individuals with a history of heavy smoking only). ROC curve analyses showed best overall predictive discrimination performance for the $PLCO_{m2012}$, Bach and TSCE (incidence) models, with AUC >0.77 in PLCO and >0.68 in NLST. Comparing the predictive performance of lung cancer risk prediction

models with that of the NLST eligibility criteria, in the PLCO data the models generally provided better sensitivity than the NLST criteria, at equal specificity. For the $PLCO_{m2012}$ model, these latter findings confirm findings from earlier re-analyses of NLST and PLCO data by Tammemägi *et al.* (13). Regarding the estimation of absolute lung cancer incidence or mortality, models showed generally satisfactory calibration in terms of predicted numbers of cancer cases relative to the numbers actually observed.

Further to analyses of discrimination capacity (area under the ROC curve; sensitivity at given specificity) and overall model calibration, a useful approach to evaluating the performance of risk stratification criteria and models is decision curve analysis. Central to decision curve analysis is the concept of a risk threshold above which an individual may expect to have a greater benefit than possible harm (19). Here, the net benefit (NB) is defined as

$$\text{NB} = (\text{TP} - \text{FP} \times \text{weighting factor})/(\text{number of individuals assessed for screening eligibility}) \quad [1]$$

where TP is the count of true positives (i.e., persons identified as eligible for screening and who are indeed developing clinically manifest lung cancer), and FP is the count of false positives (i.e., persons classified as eligible for screening but not developing clinical lung cancer during their lifetime and hence merely at risk of possible harms). The weighting factor represents the relative weight of possible harms that false-positives may experience, as compared to the weight of expected benefits for the true positives, and can be directly related to an absolute risk threshold for screening eligibility:

$$\text{weighting factor} = \text{risk threshold}/(1\text{-risk threshold}) \quad [2]$$

If the actual screening method and any additional diagnostic work-up have high efficacy and only minimal adverse effects, and if in addition early detection brings a meaningful survival benefit, the risk threshold for participation in a screening program can be set at a low level. As an example, if the expected clinical benefit of screening participation for a person actually developing lung cancer is weighted (considered "worth") 48 times the average harm incurred by a screening participant free of lung cancer, the minimum risk prediction threshold for screening eligibility should be set at 2%. By contrast, low diagnostic accuracy of screening and diagnostic follow-up investigations, a low likelihood that early tumor detection will bring a meaningful survival benefit, or a relatively high frequency or seriousness of harms occurring among screening participants free of lung cancer, all translate into higher risk threshold for screening eligibility. Decision curves visualize the theoretical NB over a range of risk thresholds, allowing one to discern whether and at which risk thresholds a model can be clinically useful. Performing such analyses, ten Haaf and colleagues found that prediction models (especially, $PLCO_{m2012}$, Bach, TSCE) outperformed the NLST eligibility criteria, with a positive NB over a substantial range of absolute risk thresholds.

Using the data of a smaller cohort—the German component of European EPIC study (20,700 ever smokers; 92 incident cases of lung cancer within the first 5 years of follow-up)—we recently performed a similar external validation of risk models, comparing the performances of four risk models ($PLCO_{m2012}$, Bach, LLP and Spitz) with the eligibility criteria used in NLST or in the various European screening trials. Our findings were very similar to those by ten Haaf *et al.* All four models showed good predictive discrimination with AUC estimates between 0.78 and 0.81, similar to ten Haaf's estimates within the PLCO data sets. In addition, all but the Spitz model provided well-calibrated risk estimates (ratio of predicted to observed incident case numbers close to 1.0 *vs.* 3.75 for the Spitz model). The $PLCO_{m2012}$ model, in particular, showed systematically better predictive sensitivity for future lung cancer occurrences than the eligibility criteria of NLST or of the European trials. Finally, as in the analyses by ten Haaf and colleagues, decision curve analyses documented a uniformly greater NB for the $PLCO_{m2012}$, Bach and LLP models as compared to any of trial eligibility criteria, with positive NB estimates over a broad range of risk thresholds. The much inferior performance of the Spitz model in decision curve analyses was largely explained by gross miscalibration of its absolute risk estimates with regard to actual observations in the German EPIC cohort.

Given the accumulating evidence that lung cancer risk prediction models such as $PLCO_{m2012}$, TSCE or the Bach model provide good diagnostic discrimination and well-calibrated absolute risk estimates, and that they may outperform simpler trial eligibility criteria based on age and pack-years of smoking, a central question is, what absolute risk threshold should be used to select individuals for lung cancer screening. From a perspective of decision analysis, as described above, the risk threshold is directly related to the relative weighting factor of long-term expected benefits to the possible harms of a diagnostic or any other medical procedure. In context of screening, NB may be best defined by the expected gain in life years for participants

who are indeed developing a detectable tumor bound to become clinically manifest, minus the potential harms for participants who in reality are not developing clinically manifest lung cancer but who may suffer from consequences of false-positive screening diagnosis or over-diagnosis. The information of such long-term benefits and harms generally is not available in the prospective studies (non-trial data) that have been used for risk model development or validation. However, information on overall benefits *vs.* harms as observed in prospective screening trials could be integrated into simulation models for estimation of the expected NB at different eligibility (lung cancer risk) prediction thresholds. Doing so would require several assumptions. One is that the benefits and harms of screening primarily depend on an individual's lung cancer risk, independently of the combination of determinants underlying the model (20). A second is that the relative (weighted) balance between actual benefits (e.g., life years gained) and harms (frequency and seriousness of consequences to false positive diagnosis, over-diagnosis) will be invariant to the selected risk threshold chosen as eligibility criterion for screening. The question is whether these assumptions truly hold. Conceivably, the risk threshold may be related not only to an individual's probability of actually developing lung cancer, but also clinical and molecular characteristics of tumors and a patient's probability of survival. Likewise, the level of risk threshold chosen may also be related to numbers and sub-categories of false-positive diagnoses (e.g., with or without need for biopsies or surgery) or the likelihood of a cancer patient being over-diagnosed (e.g., in view of overall residual life expectancy).

In conclusion, evidence is accumulating that the selection of individuals for lung cancer screening using individual risk prediction may be superior to using selection criteria based on age and pack-years alone. Studies on the external validation indicate that existing risk prediction models such as $PLCO_{m2012}$, Bach and TSCE may have good general performance in terms of discrimination and absolute risk calibration. Validation of $PLCO_{m2012}$ and other models in still further prospective cohort settings may help dissipate remaining concerns about the generalizability of model calibration, even though population based cohort studies will never perfectly represent all possible populations to be screened. Finally, more detailed analysis of existing screening trial data should assess possible relationships of individuals' predicted risks as screening eligibility thresholds with actual observations of observed clinical benefits and harms, and their weighted balance, upon screening participation.

Acknowledgements

None.

Footnote

Conflicts of Interest: The authors have no conflicts of interest to declare.

References

1. Aberle DR, Adams AM, Berg CD, et al. Reduced lung-cancer mortality with low-dose computed tomographic screening. N Engl J Med 2011;365:395-409.
2. van der Aalst CM, Ten Haaf K, de Koning HJ. Lung cancer screening: latest developments and unanswered questions. Lancet Respir Med 2016;4:749-61.
3. Infante M, Sestini S, Galeone C, et al. Lung cancer screening with low-dose spiral computed tomography: evidence from a pooled analysis of two Italian randomized trials. Eur J Cancer Prev 2017;26:324-9.
4. Wille MM, Dirksen A, Ashraf H, et al. Results of the Randomized Danish Lung Cancer Screening Trial with Focus on High-Risk Profiling. Am J Respir Crit Care Med 2016;193:542-51.
5. Paci E, Puliti D, Lopes Pegna A, et al. Mortality, survival and incidence rates in the ITALUNG randomised lung cancer screening trial. Thorax 2017;72:825-31.
6. Yousaf-Khan U, van der Aalst C, de Jong PA, et al. Final screening round of the NELSON lung cancer screening trial: the effect of a 2.5-year screening interval. Thorax 2017;72:48-56.
7. Silva M, Pastorino U, Sverzellati N. Lung cancer screening with low-dose CT in Europe: strength and weakness of diverse independent screening trials. Clin Radiol 2017;72:389-400.
8. Li K, Husing A, Sookthai D, et al. Selecting High-Risk Individuals for Lung Cancer Screening: A Prospective Evaluation of Existing Risk Models and Eligibility Criteria in the German EPIC Cohort. Cancer Prev Res (Phila) 2015;8:777-85.
9. Han SS, Ten Haaf K, Hazelton WD, et al. The impact of overdiagnosis on the selection of efficient lung cancer screening strategies. Int J Cancer 2017;140:2436-43.
10. Tammemägi MC. Application of risk prediction models to lung cancer screening: a review. J Thorac Imaging

2015;30:88-100.
11. Bach PB, Kattan MW, Thornquist MD, et al. Variations in lung cancer risk among smokers. J Natl Cancer Inst 2003;95:470-8.
12. Meza R, Hazelton WD, Colditz GA, et al. Analysis of lung cancer incidence in the Nurses' Health and the Health Professionals' Follow-Up Studies using a multistage carcinogenesis model. Cancer Causes Control 2008;19:317-28.
13. Tammemägi MC, Katki HA, Hocking WG, et al. Selection criteria for lung-cancer screening. N Engl J Med 2013;368:728-36.
14. Hoggart C, Brennan P, Tjonneland A, et al. A risk model for lung cancer incidence. Cancer Prev Res (Phila) 2012;5:834-46.
15. Muller DC, Johansson M, Brennan P. Lung Cancer Risk Prediction Model Incorporating Lung Function: Development and Validation in the UK Biobank Prospective Cohort Study. J Clin Oncol 2017;35:861-9.
16. Spitz MR, Hong WK, Amos CI, et al. A risk model for prediction of lung cancer. J Natl Cancer Inst 2007;99:715-26.
17. Cassidy A, Myles JP, van Tongeren M, et al. The LLP risk model: an individual risk prediction model for lung cancer. Br J Cancer 2008;98:270-6.
18. Ten Haaf K, Jeon J, Tammemägi MC, et al. Risk prediction models for selection of lung cancer screening candidates: A retrospective validation study. PLoS Med 2017;14:e1002277.
19. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making 2006;26:565-74.
20. Katki HA, Kovalchik SA, Berg CD, et al. Development and Validation of Risk Models to Select Ever-Smokers for CT Lung Cancer Screening. JAMA 2016;315:2300-11.

Cite this article as: Kaaks R, Hüsing A, Fortner RT. Selecting high-risk individuals for lung cancer screening; the use of risk prediction models *vs.* simplified eligibility criteria. Ann Transl Med 2017;5(20):406. doi: 10.21037/atm.2017.07.14

