Published in final edited form as:

J Epidemiol Community Health. 2015 November ; 69(11): 1035–1039. doi:10.1136/jech-2014-204079.

Overdiagnosis in lung cancer screening: why modelling is essential

Kevin ten Haaf and Harry J de Koning

Department of Public Health, Erasmus MC, University Medical Center Rotterdam, Rotterdam, The Netherlands

INTRODUCTION

Screening for diseases is an important part of healthcare, as the detection of disease at an early stage may improve the chance of successful treatment of the disease. Although high-quality screening programmes may provide substantial benefits, one of the major harms associated with screening is overdiagnosis. Overdiagnosis refers to the event that a disease is diagnosed that would not have been clinically detected; for example, screening may detect a slowly progressing tumour that would not have caused symptoms during the patient's life-time. Overdiagnosis may lead to serious consequences, such as unnecessary treatments. Furthermore, overdiagnosis will lead to biased survival outcomes for screen-detected cases.¹

Notable harms of CT lung cancer screening include the high proportion of false-positive results and radiation exposure.² The decision of the USPSTF to recommend lung cancer screening with CT for persons up to age 80 has also raised concerns on the magnitude of overdiagnosis.² The Medicare Evidence Development and Coverage Advisory Committee commented that the USPSTF's decision to extend the upper age of screening from 74 to 80 years 'was based upon modelling only, with no empirical data'.³

The opinion that modelling cannot provide additional information beyond that of a clinical trial is a common misconception. Models are often criticised for depending on assumptions with regard to the processes of carcinogenesis and progression of cancer.⁴ However, the same could be said of statistical tests used to evaluate the results of clinical trials, many of which are also based on implicit assumptions. Similarly to other statistical methods, if the underlying assumptions are properly stated, models can be valuable tools in bridging the gaps between the evidence provided by clinical trials and the evidence needed for the

Correspondence to Kevin ten Haaf, Department of Public Health, Erasmus MC, University Medical Center Rotterdam, P.O. Box 2040, Rotterdam 3000 CA, The Netherlands; k.tenhaaf@erasmusmc.nl.

Contributors

KtH and HJdK were responsible for conception and design, analysis and interpretation of the data, drafting of the article, critical revision of the article for important intellectual content and final approval of the article.

Competing interests

KtH and HJdK are members of the Cancer Intervention and Surveillance Modelling Network (CISNET) Lung working group. HJdK is the principal investigator of the Dutch-Belgian Lung Cancer Screening Trial (Nederlands-Leuvens Longkanker Screenings onderzoek; the NELSON trial). KtH is a researcher affiliated with the NELSON trial.

Provenance and peer review

Commissioned; externally peer reviewed.

development of clinical guidelines.⁵ In fact, modelling is essential to derive detailed estimates of the benefits and harms of screening beyond the time period of the clinical trial, especially for overdiagnosis.

Analyses using an excess-incidence approach by Patz suggest that over 18% of all lung cancers detected by CT in the National Lung Screening Trial (NLST) were overdiagnosed.⁶ However, five microsimulation models, which were calibrated to various data sources including the NLST, suggest that less than 10% of lung cancers detected by screening would be overdiagnosed in a 1950 US birth cohort adhering to the annual screening policy recommended by the USPSTF.⁷

Previous reports have discussed different methods to derive overdiagnosis.⁸⁹ In this report, we will address the difficulties of deriving estimates on overdiagnosis from clinical trials alone. Furthermore, we will detail how models can aid in extrapolating information from clinical trials to screening programmes.

Potential biases in estimating overdiagnosis in clinical trials

The NLST randomised 53 454 current smokers and former smokers (quit less than 15 years) between the ages of 55–74 with a minimum smoking exposure of 30 pack-years to receive either three CT or chest radiography (CXR) screens.¹⁰ The NLST demonstrated a relative reduction in lung cancer mortality of 20% for CT compared to CXR (16% after extending the cut-off date for mortality analyses).¹⁰¹¹ However, the magnitude of overdiagnosis in the NLST cannot be easily ascertained from the results of the trial alone. Assuming the excess number of lung cancers in the CT arm compared to the CXR arm are due to overdiagnosis (excess-incidence approach) to derive the magnitude of overdiagnosis in the NLST, as suggested by Patz, is biased for two reasons.⁶

First, both arms of the NLST were screened, either with CT or CXR. While the Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial (PLCO) indicated that CXR screening does not reduce lung cancer mortality, the possibility of overdiagnosis in the CXR arm of the NLST is still present.¹² Therefore, the CXR arm of the NLST does not provide an unbiased baseline incidence of lung cancer in the absence of screening.⁸ Thus, ascertaining the magnitude of overdiagnosis in the CT arm of the NLST through an excess-incidence analysis may underestimate the level of overdiagnosis.

Second, the median follow-up duration of NLST participants was limited to approximately 6.5 years.¹⁰ A follow-up duration of sufficient length is essential to account for the effects of lead-time.⁹ Lead-time is the time interval between the detection of the cancer by screening and its clinical presentation (the time of diagnosis in the absence of screening), shown in figure 1. A higher proportion of cancers were detected at an early stage in the CT arm compared to the CXR arm, due to the higher sensitivity of CT.¹⁰¹³ This suggests that the lead-time of CT is longer than that of CXR. Figure 2 (derived from Black *et al*¹⁴, table S2–2) shows the effects of the difference in lead-time between CT and CXR in the NLST. In the years during which screening occurs (years 1–3), the incidence of lung cancer in the CT arm is higher compared to the CXR arm; due to the longer lead-time of CT. Non-overdiagnosed lung cancers would have become symptomatic at a later date, but screening advances their

detection to an earlier moment in time. Thus, lung cancers of which the moment of detection was advanced will not be detected at their original moment of clinical presentation. As the moment of detection was advanced for more lung cancers in the CT arm, the relative number of lung cancers detected in the CXR arm is higher in the years after screening has ended (years 4–7), as shown in figure 2. As the base incidence of lung cancer in the absence of screening is similar for both arms, the difference in lung cancer incidence between the arms, due to a difference in lead-time, will dissipate over time. The remaining excess of lung cancers will then represent the number of overdiagnosed lung cancers in the CT arm compared to the CXR arm.

Patz suggests that the difference in the absolute number of lung cancers converges in the last years of the trial.⁶ However, this may be due to the low number of person-years observed in the last years of the trial, as noted in table S2–2 of Black *et al.*¹⁴ Black *et al.*¹⁴ corrects for the lower number of person-years by investigating the difference in lung cancer incidence by a 0.5-year interval and also suggests a convergence (table S2–2 and figure S2–1 in their report).⁶ However, when one observes the 95% CIs of the incidence rate difference between the CT and CXR arms of the NLST by a 1-year interval instead, shown in figure 2, the incidence rates are not suggested to have converged yet. Therefore, the follow-up duration of the NLST does not seem of sufficient length to account for the difference in lead-time between CT and CXR.

Limitations of extrapolating overdiagnosis estimates from clinical trials to screening programmes

The design and limited follow-up duration complicate deriving the magnitude of overdiagnosis in the NLST. However, even if this information could be easily derived, it would provide little information on the magnitude of overdiagnosis in a screening programme on the population level, due to a number of limitations.

The first limitation is the trial's fixed design. If the NLST had considered a different number of screening rounds or different intervals between screening rounds, the number of lung cancers detected by screening would have been different. Consequently, the proportion of overdiagnosed cases could have been different as well, which poses difficulties in extrapolating the results of the NLST to other designs (eg, different number of screening rounds and/or different intervals between screenings).

The second limitation is the investigated population. Recent studies indicate that the risk of lung cancer (death) varies, even across participants of the NLST.^{15,16} In addition, mortality for causes other than lung cancer varies greatly by smoking behaviour and age.¹⁷ Compared to the portion of the general US population that meets the entry criteria of the NLST, the participants of the NLST were younger (26.6% of the NLST participants were older than 65 compared to 35.5% of the eligible US population) and less likely to be current smokers (48.2% of NLST participants compared to 57.1% of the eligible US population).¹⁸ Consequently, the average person eligible for a US screening programme utilising the same entry criteria as the NLST may be younger and more likely to be a current smoker compared to the average participant of the NLST. As a result, the average risk of lung cancer (death) and mortality for causes other than lung cancer may differ between the two groups.^{15–17}

Therefore, the magnitude of overdiagnosis in a population-based screening programme could differ from that in the NLST.

Thus, even if information on the magnitude of overdiagnosis in the NLST can be ascertained, this information cannot be easily applied to different designs and populations. While the probability of overdiagnosis is constantly present, the rate of overdiagnosis is subject to different aspects of the screening programme and the screened population, such as screening frequency and age. However, microsimulation models can aid in extrapolating the information obtained from the NLST to different designs and populations.

Estimating overdiagnosis through microsimulation modeling

Microsimulation models can simulate a person's entire life-history in the absence and presence of screening, which allows one to determine which cancers are overdiagnosed within the simulation. Each individual's probability of developing cancer due to biological processes and/or exposure to carcinogens is modelled, as well as the individual's probability of dying from the disease or other causes. However, to estimate the magnitude of overdiagnosis in lung cancer screening for different designs and populations, a model must meet a number of requirements.

First, the model must be able to provide an estimate of the baseline incidence of the disease for the population of interest in the absence of screening.⁸ In modelling lung cancer, a smoking dose-response module, such as the Two-Stage Clonal Expansion model, is often used to determine the baseline incidence.¹⁹ Smoking dose-response modules use age, smoking history and other risk-factors to estimate a person's risk of lung carcinogenesis.¹⁹²⁰ When a model incorporates a smoking dose-response module calibrated to a wide range of risk profiles, including never-smokers, the model can be applied to any population.¹⁹²⁰ The Microsimulation SCreening ANalysis (MISCAN) Lung model demonstrates this by reproducing the incidence of lung cancer in populations with different risk profiles, such as the NLST and PLCO.¹³²¹ The assumptions, calibration process and sensitivity analyses of the MISCAN-Lung model used in this investigation were detailed previously.¹³²¹

Furthermore, the model must explicitly consider the natural history (preclinical progression) and screen-detectability of lung cancer. This is essential to estimate the lead-time achieved by the screening test and the potential for overdiagnosis.⁸ The model should take the statistical dependence between the natural history of the disease and the sensitivity of the screening test into account, which may cause uncertainty in the parameter estimates: a model with a low estimate for the sensitivity and a long preclinical duration may provide a similar fit compared to a model with a high estimate for the sensitivity and a short preclinical duration.²² Microsimulation models can derive this information through synthesising data from the NLST with data from clinical trials with non-screened control arms, such as the PLCO, which provides essential information on the natural history of lung cancer in the absence of screening.¹²¹³²¹ By calibrating to the number of cancers detected by screening round and the number of interval cancers per year, models can estimate the natural history and screen-detectability of lung cancer.¹³ This information allows the extrapolation of the findings of clinical trials to screening policies with different designs, for example, variations in number of screens and intervals between screens. In a previous

investigation, we derived estimates for the natural history and screen-detectability of lung cancer by histological type, stage and gender.¹³ Overall, our estimates suggested a greater window of opportunity for lung cancer screening compared to previous research.¹³

Figure 2 indicates that the incidence rate difference of the CT arm compared to the CXR arm estimated by MISCAN-Lung lies between the 95% CIs for the entirety of the observed follow-up duration of the trial. However, in contrast to other reports, the estimates of MISCAN-Lung suggest that the incidence rate difference between the two arms does not converge at the end of the observed follow-up duration of the trial (year 7).^{6,14} Instead, the incidence rate difference between the two arms is suggested to converge in year 12, 9 years after screening has ended and 5 years after the observed follow-up duration of the trial.

Through modelling, one can investigate the excess-incidence in both arms of the NLST compared to hypothetical non-screened arms, as shown in figure 3. As expected, in the years in which screening occurs (years 1–3) the excess incidence is higher for both arms, compared to their hypothetical non-screened counterparts. The estimated excess-incidence of the CT arm is higher than the CXR arm, indicating that CT is more sensitive than CXR: MISCAN-Lung estimates the excess-incidence in the CT arm compared to the CXR arm, estimated as the excess number of lung cancers in the CT arm compared to the CXR arm divided by the number of screen-detected cancers in the CT arm, at year 8 to be 12.5% compared to the reported 18.5% (95%CI 5.4 to 30.6%).⁶ The estimated difference in yearly lung cancer incidence between the CXR arm and the hypothetical non-screened arm dissipates at approximately year 10, 7 years after screening has ended. For the CT arm, this occurs at approximately year 12, 9 years after screening has ended, which suggests a longer lead-time for CT compared to CXR. This information allows the model to estimate the amount of overdiagnosis, defined as the number of cancers that would not have been detected if screening had not occurred, divided by the number of cancers detected by screening that occurred in both arms of the NLST. MISCAN-Lung estimates that 6.75% of all screen-detected cases in the CXR arm are overdiagnosed compared to 8.62% of all screen-detected cases in the CT arm. These percentages are relatively low, as approximately 75% of the participants of the NLST were aged younger than 65.¹⁰ Furthermore, the participants of the NLST only received three screens, and over 50% of the detected cancers in the CT-arm were detected in persons aged under 65.²³ Therefore, the potential for overdiagnosis was relatively low in the NLST.

The final requirement to extrapolate the results of clinical trials to different populations is that the model must be able to utilise data specific to the investigated population. Such information includes smoking behaviour and mortality for causes other than lung cancer by smoking behaviour, gender and age. For example, information on these aspects for a 1950 US birth cohort has been used in modelling analyses to inform the USPSTF on its recommendation for lung cancer screening.⁷

CONCLUSION

Clinical trials such as the NLST are essential to provide information on the efficacy of lung cancer screening.¹⁰ However, information on overdiagnosis is difficult to ascertain from

clinical trials alone, as the follow-up duration may be insufficient and, in trials without an unscreened control group, an unbiased baseline incidence of the disease in the absence of screening may not be available. Furthermore, an estimate for overdiagnosis based on a clinical trial only provides information with regard to the design and population investigated in that trial. Furthermore, the risk of lung cancer and (smoking-related) mortality from causes other than lung cancer can vary substantially across individuals, which must be taken into account when one considers implementing a lung cancer screening programme.^{15–17}

This report shows that, while important, the results of the NLST by itself provide limited information on the magnitude of overdiagnosis in future lung cancer screening programmes. Microsimulation models can provide estimates on overdiagnosis for lung cancer screening programmes with designs and populations different from those considered in the NLST, which is essential for policy makers.^{7132124–26}

Furthermore, modelling can provide insights in situations where information from trials or observational studies alone is insufficient. For example, in colorectal screening, a colonoscopy can remove precancerous lesions (adenomas), preventing the occurrence of colorectal cancer altogether. As a result, if one compares a screened group with a non-screened control group, the control group may have a higher incidence compared to the screened group. While overdiagnosis will still occur in the screened group, the magnitude cannot be ascertained with an excess-incidence approach.

However, sophisticated microsimulation models require detailed data, which can often only be obtained from clinical trials such as the NLST and PLCO, to provide accurate estimates.¹³²¹ Like any analysis, the assumptions and validity of the model should be clearly detailed.⁴ Furthermore, one should take into account that models have other limitations, such as uncertainty in incidence trends and drifts in screening efficacy.²⁷ As lung cancer screening is implemented across the USA, more data will become available on the effects of lung cancer screening in the general population, such as the effect of nodule cut points. Besides influencing the sensitivity of CT screening, it may also affect overdiagnosis. This information can be used to further improve the estimates on the magnitude of overdiagnosis in lung cancer screening.

Acknowledgments

Funding

This publication was indirectly supported by Grant 5U01CA152956-04 from the National Cancer Institute as part of the Cancer Intervention and Surveillance Modelling Network (CISNET).

References

1. Ruano-Ravina A, Heleno B, Fernández-Villar A. Lung cancer screening with low-dose CT (LDCT), or when a public health intervention is beyond the patient's benefit. *J Epidemiol Community Health*. 2015; 69:99–100. [PubMed: 25138391]
2. Moyer VA. Screening for lung cancer: U.S. preventive services task force recommendation statement. *Ann Intern Med*. 2014; 160:330–8. [PubMed: 24378917]
3. Centers for Medicare & Medicaid Services. Proposed Decision Memo for Screening for Lung Cancer with Low Dose Computed Tomography (LDCT). 2014. (CAG-00439N)<http://www.cms.gov/>

medicare-coverage-database/details/nca-proposed-decision-memo.aspx?NCAId=274 (accessed 20 Nov 2014)

4. Carter JL, Coletti RJ, Harris RP. Quantifying and monitoring overdiagnosis in cancer screening: a systematic review of methods. *BMJ*. 2015; 350
5. Habbema JDF, Wilt TJ, Etzioni R, et al. Models in the development of clinical practice guidelines models in the development of clinical practice guidelines. *Ann Intern Med*. 2014; 161:812–18. [PubMed: 25437409]
6. Patz EF Jr, Pinsky P, Gatsonis C, et al. Overdiagnosis in low-dose computed tomography screening for lung cancer. *JAMA Intern Med*. 2014; 174:269–74. [PubMed: 24322569]
7. de Koning HJ, Meza R, Plevritis SK, et al. Benefits and Harms of computed tomography lung cancer screening strategies: a comparative modeling study for the U.S. preventive services task force. *Ann Intern Med*. 2014; 160:311–20. [PubMed: 24379002]
8. Etzioni R, Gulati R, Mallinger L, et al. Influence of study features and methods on overdiagnosis estimates in breast and prostate cancer screening. *Ann Intern Med*. 2013; 158:831–8. [PubMed: 23732716]
9. de Gelder R, Heijnsdijk EAM, van Ravesteijn NT, et al. Interpreting overdiagnosis estimates in population-based mammography screening. *Epidemiol Rev*. 2011; 33:111–21. [PubMed: 21709144]
10. 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. [PubMed: 21714641]
11. Pinsky PF, Church TR, Izmirlian G, et al. The national lung screening trial: results stratified by demographics, smoking history, and lung cancer histology. *Cancer*. 2013; 119:3976–83. [PubMed: 24037918]
12. Oken MM, Hocking WG, Kvale PA, et al. Screening by chest radiograph and lung cancer mortality: the prostate, lung, colorectal, and ovarian (PLCO) randomized trial. *JAMA*. 2011; 306:1865–73. [PubMed: 22031728]
13. ten Haaf K, van Rosmalen J, de Koning HJ. Lung cancer detectability by test, histology, stage, and gender: estimates from the NLST and the PLCO trials. *Cancer Epidemiol Biomarkers Prev*. 2015; 24:154–61. [PubMed: 25312998]
14. Black WC, Gareen IF, Soneji SS, et al. Cost-effectiveness of CT screening in the national lung screening trial. *N Engl J Med*. 2014; 371:1793–802. [PubMed: 25372087]
15. Kovalchik SA, Tammemagi M, Berg CD, et al. Targeting of low-dose CT screening according to the risk of lung-cancer death. *N Engl J Med*. 2013; 369:245–54. [PubMed: 23863051]
16. Tammemagi MC, Katki HA, Hocking WG, et al. Selection criteria for lung-cancer screening. *N Engl J Med*. 2013; 368:728–36. [PubMed: 23425165]
17. Rosenberg MA, Feuer EJ, Yu B, et al. Chapter 3: cohort life tables by smoking status, removing lung cancer as a cause of death. *Risk Anal*. 2012; 32:S25–38. [PubMed: 22882890]
18. Aberle DR, Adams AM, Berg CD, et al. Baseline characteristics of participants in the randomized national lung screening trial. *J Natl Cancer Inst*. 2010; 102:1771–9. [PubMed: 21119104]
19. Meza R, Hazelton W, Colditz G, 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. [PubMed: 18058248]
20. Hazelton WD, Luebeck EG, Heidenreich WF, et al. Analysis of a historical cohort of Chinese tin miners with arsenic, radon, cigarette smoke, and pipe smoke exposures using the biologically based two-stage clonal expansion model. *Radiat Res*. 2001; 156:78–94. [PubMed: 11418076]
21. Meza R, ten Haaf K, Kong CY, et al. Comparative analysis of 5 lung cancer natural history and screening models that reproduce outcomes of the NLST and PLCO trials. *Cancer*. 2014; 120:1713–24. [PubMed: 24577803]
22. van Ballegooijen M, Rutter CM, Knudsen AB, et al. Clarifying differences in natural history between models of screening. *Med Decis Making*. 2011; 31:540–9. [PubMed: 21673187]
23. Pinsky PF, Gierada DS, Hocking W, et al. National lung screening trial findings by age: medicare-eligible versus under-65 population National lung screening trial findings by age. *Ann Intern Med*. 2014; 161:627–33. [PubMed: 25199624]

24. McMahon PM, Meza R, Plevritis SK, et al. Comparing benefits from many possible computed tomography lung cancer screening programs: extrapolating from the national lung screening trial using comparative modeling. PLoS ONE. 2014; 9:e99978. [PubMed: 24979231]
25. Jeon J, Meza R, Krapcho M, et al. Chapter 5: actual and counterfactual smoking prevalence rates in the U.S. population via microsimulation. Risk Anal. 2012; 32:S51–68. [PubMed: 22882892]
26. Holford TR, Levy DT, McKay LA, et al. Patterns of birth Cohort-specific smoking histories, 1965–2009. Am J Prev Med. 2014; 46:e31–7. [PubMed: 24439359]
27. Kramer BS, Elmore JG. Projecting the benefits and harms of mammography using statistical models: proof or proofiness? J Natl Cancer Inst. 2015; 107

Figure 1.

Lead-time and overdiagnosis. The two halves of the figure depict two scenarios: in the scenario depicted in the upper half of the figure, the person has developed a cancer that would have become symptomatic before the person would have died due to causes other than cancer. However, due to screening, this cancer is detected earlier. In the scenario depicted in the lower half of the figure, the person has developed a cancer that would not have become symptomatic before the person would have died due to causes other than cancer. However, due to screening, this cancer is detected, while it would not have been diagnosed without screening. The lead-time in both scenarios represents the time between the detection of the cancer by screening and the moment that the cancer would have become symptomatic.

Figure 2.

Observed and estimated lung cancer incidence rate difference (per 1000 person-years) in the CT arm of the National Lung Screening Trial (NLST) compared to the chest radiography (CXR) arm by 1-year interval. Observed data derived from table S2–2 of Black *et al.*¹⁴ Error bars denote 95% CIs for the incidence rate difference. The MICrosimulation SCReening ANALysis (MISCAN) Lung model was used to estimate the excess incidence of the CT arm compared to the CXR arm of the NLST and extrapolate beyond the follow-up duration of the trial.

Figure 3.

Estimated lung cancer incidence rate difference (per 1000 person-years) of the CT and chest radiography (CXR) arms of the National Lung Screening Trial (NLST) compared to hypothetical non-screened arms by 1-year interval. The Microsimulation SCreening ANalysis (MISCAN) Lung model was used to estimate the excess incidence of the CT and CXR arms of the NLST compared to hypothetical arms in which screening did not occur.