Short title: Overdiagnosis and lung cancer screening

The impact of overdiagnosis on the selection of efficient lung cancer screening strategies

Summer S. Han[1,2], Kevin ten Haaf[3], William D. Hazelton[4], Vidit N. Munshi[5], Jihyoun Jeon[6], Saadet A. Erdogan[1], Colden Johanson[5], Pamela M. McMahon[5], Rafael Meza[6], Chung Yin Kong[5], Eric J. Feuer[7], Harry J. de Koning[3] and Sylvia K. Plevritis[2]

Corresponding Author: Sylvia K. Plevritis, Department of Radiology, Stanford University, 318 Campus Dr., Stanford, CA 94305, U.S.A; Phone 650-498-5261; Fax 650-723-5795; Email: sylvia.plevritis@stanford.edu

Authors' affiliations:
[1]Department of Medicine, Stanford University, Palo Alto, CA USA; [2]Department of Radiology, Stanford University, Palo Alto, CA USA [3]Department of Public Health, Erasmus MC, Rotterdam, Netherlands; [4]Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA USA; [5]Department of Radiology, Massachusetts General Hospital, Boston, MA USA; [6]Department of Epidemiology, University of Michigan, Ann Arbor, MI USA; [7]Division of Cancer Prevention, National Cancer Institute, Bethesda, MD USA

Keywords: lung cancer, computed tomography, lung cancer screening, simulation model, overdiagnosis, USPSTF, microsimulation, health policy

Novelty and Impact:

We use a comparative modeling approach to evaluate efficient CT screening strategies for lung cancer (LC) in the U.S. population using a range of metrics that integrate overdiagnosis. Overdiagnosis affects the selection of efficient screening strategies specifically the screening stopping age. While screening through age 80 is efficient in reducing LC mortality irrespective of overdiagnosis, stopping screening at a younger age of 75 provides a greater efficiency in reducing LC deaths per overdiagnosed case.

Abstract

The U.S. Preventive Services Task Force (USPSTF) recently updated their national lung screening guidelines and recommended low-dose computed tomography (LDCT) for lung cancer (LC) screening through age 80. However, the risk of overdiagnosis among older populations is a concern. Using four comparative models from the Cancer Intervention and Surveillance Modeling Network, we evaluate the overdiagnosis of the screening program recommended by USPSTF in the U.S. 1950 birth cohort. We estimate the number of LC deaths averted by screening (D) per overdiagnosed case (O), yielding the ratio D/O, to quantify the trade-off between the harms and benefits of LDCT. We analyze 576 hypothetical screening strategies that vary by age, smoking, and screening frequency and evaluate efficient screening strategies that maximize the D/O ratio and other metrics including D and life-years gained (LYG) per overdiagnosed case. The estimated D/O ratio for the USPSTF screening program is 2.85 (model range: 1.5-4.5) in the 1950 birth cohort, implying LDCT can prevent ~3 LC deaths per overdiagnosed case. This D/O ratio increases by 22% when the program stops screening at an earlier age 75 instead of 80. Efficiency frontier analysis shows that while the most efficient screening strategies that maximize the mortality reduction (D) irrespective of overdiagnosis screen through age 80, screening strategies that stop at age 75 versus 80 produce greater efficiency in increasing life-years gained per overdiagnosed case. Given the risk of overdiagnosis with lung cancer screening, the stopping age of screening merits further consideration when balancing benefits and harms.

1. Introduction

The National Lung Screening Trial (NLST) recently demonstrated that low-dose computed tomography (LDCT) is effective in reducing lung cancer (LC) mortality[1]. However, overdiagnosis of LC in LDCT screening is a significant concern[2-4]. Overdiagnosis is defined as the screen-detected disease that in the absence of screening, would not have become clinically evident within one's lifetime[5]. Overdiagnosis can lead to unnecessary treatment and costs and negatively impact well-being and life expectancy[5].

The Cancer Intervention and Surveillance Modeling Network (CISNET) is an National Cancer Institute (NCI) sponsored consortium that uses a comparative statistical modeling approach to estimate the population-level impact of cancer screening. In prior work, the CISNET lung cancer screening models were used to evaluate the comparative effectiveness of 576 screening strategies that varied by smoking history, age, and screening frequency[6-8]. These analyses were used by the U.S. Preventive Services Task Force (USPSTF) as secondary evidence to support the recent recommendation to annually screen persons aged 55 to 80 with the same smoking criteria as the NLST[9]. One notable aspect of the USPSTF recommendation is the increased stopping age to 80 from 74 compared to the NLST. Although several harms (including overdiagnosis) were considered by the USPSTF, harms were not explicitly incorporated when ranking the efficient scenarios provided by CISNET; instead these scenarios were selected by maximizing the LC mortality reduction, i.e. the number of LC deaths (D) prevented due to screening over the number of CT screening examinations[7].

While the lung cancer screening guidelines by USPSTF recommended LDCT targeting the age group of 55 to 80, there is still considerable debate over the potential benefits and harms of screening in the older population. The Centers for Medicare & Medicaid Services (CMS) issued a national coverage determination for Medicare coverage of LDCT screening for individuals aged 55 to 77 (www.cms.gov) whereas the USPSTF recommends screening up to age 80. Other guidelines such as those proposed by the American Cancer Society (ACS) recommended CT screening for individuals ages 55 to 74 [10]. Hence, the stopping age of screening varies widely across different guidelines (ages 74, 77, 80) while the starting age (age 55) and smoking criteria (30 pack-years and 15 years since cessation) are consistent across recommendations. Not surprisingly, the age of LC patients also has shown to be associated with increased risk of postoperative complications: patients aged 50-69 have a 3-fold higher risk for life-threatening complications compared to patients aged <50, while the risk is a 9-fold higher for patients aged >70[11]. Given the importance of the effect of screening age on the potential harms and the divergence on the recommended stopping for lung screening, it is essential to evaluate the optimal stopping age of lung cancer screening by more directly accounting for screening-associated harms.

In this study, CISNET re-examines efficient screening strategies for lung cancer using a range of metrics that incorporate overdiagnosis. One useful metric for assessing the impact of overdiagnosis is the ratio between LC deaths prevented due to screening (D) and overdiagnosed cases (O), represented by D/O. This metric has been previously used to quantify the trade-off between the harms and benefits of screening[5]. Other measures are also considered such as life-years gained (LYG) due to screening and the LYG per overdiagnosed case (LYG/O). We use four independent CISNET lung models to estimate LC overdiagnosis for 576 alternative CT screening scenarios that vary by smoking, age, and screening frequency in the general U.S. population. Included is the direct comparison between the USPSTF recommended scenario and the NLST-like scenario (i.e. ACS-like scenario), which only differ in the screening stopping age (80 vs. 75). We evaluate screening strategies that optimize a range of metrics integrating overdiagnosis, comparing their outcomes to those based on LC mortality reduction (D) alone. These findings can provide insights into the impact of incorporating overdiagnosis on the selection of efficient lung screening programs, providing a more balanced consideration of screening benefits and harms.

2. Methods

CISNET models

Four CISNET LC screening models were independently developed based on different sets of assumptions and mathematical model structures at the following institutions: Erasmus Medical Center; Fred Hutchinson Cancer Research Center; the Massachusetts General Hospital, and Stanford University. The common model components are: age-specific LC risk in the absence of screening, natural history model for tumor growth and progression, screening component for predicting detection age and stage of LC in the presence of screening, diagnostic workup component for following up lung nodules; and corresponding LC mortality and death from other causes[6-8](See **Supplemental Table 1**).

Each model was calibrated and validated using the data from NLST and PLCO (Prostate, Lung, Colorectal, and Ovarian Cancer Screening)[6, 12] to obtain estimates on screening-related parameters such as tumor size thresholds for diagnostic follow-up. Each model reproduced the observed incidence and mortality of LC (stratified by cancer stage at diagnosis, histology, sex, and detection mode) in both arms of these trials[6].

Target population and screening scenarios

The models were used to simulate life histories of the U.S. cohort born in 1950, whose smoking histories and other-cause mortalities were generated using the Smoking History Generator[13]. We chose the 1950 birth cohort because it was considered in the USPSTF report[9]. We evaluated a total of 576 screening scenarios, varying the frequency of screening (annual, biennial or triennial), starting age (45, 50, 55 or 60), stopping age (75, 80 or 85), minimum pack-years of smoking (25, 30, 35 or 40) and maximum years since quitting smoking (5, 10, 15 or 20) as considered in our previous reports[7, 8].

Each model was run for each of 576 screening strategies, assuming perfect screening compliance. For each strategy, each model produced several population-level outcomes, including number of LC deaths, number of LDCT screening examinations, number of prevented LC deaths (D) and life-years gained (LYG) due to screenings compared to a no-screening scenario. All counts were normalized per 100,000 persons in the cohort, who are followed up from age 45 to 90. False positives, radiation-related harms, and follow-up examinations were also previously quantified by some of the models[7].

Quantification of overdiagnosis

A patient is defined as overdiagnosed if their LC is detected in the screening scenario, but the tumor would not have been clinically detected before death from other causes in the no-screen scenario. We calculated a measure of overdiagnosis as the probability that a lung cancer detected by screening is an overdiagnosis. Using each simulation model, the risk of overdiagnosis was calculated as the number of overdiagnosed cases divided by the number of screen-detected cases, i.e., the proportion of screen-detected cases that are overdiagnosed. We compared overdiagnosis risk of 576 strategies stratified by screening starting/stopping age, smoking (pack-year and year-since-quit), gender or histology. To compare the estimated median overdiagnosis risk by groups, we applied the non-parametric Kruskal-Wallis (K-W) test. D/O was calculated by dividing the number of LC deaths prevented due to screening by the number of overdiagnosed cases.

Selection of consensus scenarios

For each model, we selected a series of scenarios that maximize the D/O ratio over the number of screening examinations by identifying the convex hull on a scatter plot between the D/O ratio (y-axis) and the number of CT screens (x-axis)[7]. An efficient frontier is defined as this convex hull, that is, a curve that connects a set of scenarios that maximize the y-values over x-values. A scenario was labeled as an “efficient scenario” if it is among the top 25% closest scenarios to the efficiency frontier. A set of consensus scenarios was identified by choosing scenarios that are defined as an “efficient scenario” by at least three out of the four models.

We selected consensus scenarios by maximizing the D/O ratio and, separately D, as a function of the number of screening examinations, by sex. In selecting efficient and consensus scenarios for both metrics, we focused on annual screening scenarios with starting age ≥ 55 and stopping age ≤ 80, since they are considered to be the most feasible for implementation. When analyzing the findings, we focused on the scenarios that are near the NLST and USPSTF scenarios, associated with the number of CT screens ranging between 250,000 and 350,000 (per 100,000 persons in the cohort) for males and 160,000 and 260,000 for females.

In a sensitivity analysis, we considered several alternative metrics that incorporate overdiagnosis to examine how the selection of efficient and consensus scenarios is affected by using different metrics. First, we considered LYG instead of D. While LYG takes into account different life expectancy among LC cases when measuring the benefit of screening, it does not explicitly incorporate overdiagnosis. Therefore, we also considered LYG/O as a metric for selecting efficient screening strategies. Another alternative metric that incorporate overdiagnosis is defined as D-O, which is the net prevented LC deaths subtracted by the number of overdiagnosed cases; this metric measures the benefit of screening (D) penalized by overdiagnosis (O). The last metric that was considered is defined as D/(O/S), the number of LC deaths prevented per overdiagnosis risk, where S is the number of screen-detected cases.

3. Results

The calibration results of the four models using the NLST data are shown in **Supplemental Figure 1**. This figure shows that the four models reliably reproduce the outcomes of the NLST data, in which the model-based estimates for excess LC incidence rate in the CT arm compared to the chest x-ray arm are consistent with the reported value of 18.5% (95% confidence interval 5.5%-30.6%) based on the NLST data[2].

Overdiagnosis risk across 576 LC screening strategies

The analysis of 576 screening scenarios indicates that the overdiagnosis risk is highly influenced by screening stopping age. Panel A-D in **Figure 1** displays the results for males, in which the screening programs with stopping age 85 have higher overdiagnosis rates (model median range: 5.51-16.44%) than the programs with lower stopping ages (model median range: 3.91%-10.72% for stopping age 75 and 4.73-13.71% for stopping age 80). These patterns are similar for females and across the four models with all p-values of the eight K-W tests less than 10^{-10}. The comparisons of overdiagnosis risks by screening frequency, starting age, and pack-years of smoking are presented in **Supplemental Materials S1** and **Supplemental Figures 2-4**. In these analyses, we find overdiagnosis is higher for more frequent screening, older starting age, and higher smoking pack-years. Among histologic subtypes, BAC (bronchioloalveolar carcinoma) has the highest overdiagnosis risk.

Comparisons of the USPSTF and NLST-like scenarios

Comparisons of the USPSTF and the NLST-like scenarios (**Figure 2** A-C) show that overdiagnosis risk is higher for the USPSTF scenario (mean: 11.9%; model range: 5.5%-23.2%) than the NLST-like scenario (mean: 9.7%; model range: 4.4%-17.6%) by 21.7%(model range for percentage increase: 10%-31.8%) due to the extended stopping age. This pattern is consistently observed across the models for each gender. The analysis of the D/O ratio (**Figure 2** D-F and **Supplemental Table 2**) shows that the USPSTF scenario prevents 2.85 LC deaths per overdiagnosed case (mean D/O ratio=2.85; model range: 1.5 - 4.5). Notably, the D/O ratio increased by 22% when the program stops screening at an earlier age 75 instead of 80 as shown in the NLST-like scenario (mean: 3.49, model range: 2.10-5.61). The range of D/O ratios of all 576 scenarios are shown in **Supplemental Table 3** by model and gender.

Consensus scenarios

The scenarios that maximize D over the numbers of CT screens among males are shown in **Figure 3A**, plotted for a representative model, with all other model results for both genders shown in **Supplemental Figure 5-6**. It is notable that all the consensus scenarios have stopping age 80 (instead of 75), which reflects the greater efficiency of programs that screen through older age 80, as opposed to 75, for reducing the number of LC deaths. Notably, the consensus scenarios include the USPSTF scenario (A-55-80-30-15). Other measures of benefits of LDCT for the consensus scenarios such as life-years and mortality reduction rate are shown in **Table 1**.

The consensus scenarios that maximize D/O over the number of CT screens are shown in **Figure 3B**. Interestingly, all the consensus scenarios maximizing D/O have stopping age 75, instead of 80, which suggests that programs that stop screening earlier are more efficient in reducing the number of LC deaths per overdiagnosed case. Notably, the NLST-like scenario (A-55-75-30-15) is included among the consensus scenarios selected by four models for each gender.

The gray colored scenarios in **Table 1** are consensus scenarios that overlap across genders within each metric for D and D/O. Overall, higher consensus was observed using the metric D/O across genders, where 80% of all the consensus scenarios (4 out of 5) are selected in both genders while for metric D, around 44% of all consensus scenarios (4 out of 9) appear in both genders.

Sensitivity analysis to outcomes metric

The selection of consensus scenarios using LYG instead of D is shown in **Figure 3C**, which is similar to the selection under D in the sense that both sets of consensus scenarios include the USPSTF scenario. However, a tendency was observed that using LYG (vs. D) penalizes screening through older ages; while all consensus scenarios chosen under D screen through 80, the consensus scenarios using LYG includes a scenario that stops screening at 75 (A-55-75-30-20). When overdiagnosis is taken into account by using LYG/O, however, the selection of consensus scenarios was remarkably similar to those using D/O (see **Figure 3D**).

Further sensitivity analyses using alternative outcomes metrics, namely D-O and D/(O/S), show that the consensus scenarios that incorporate overdiagnosis are consistent with the ones selected by maximizing D/O (**Figure 3** E-F). Most of the consensus scenarios have stopping age 75 (except for one scenario) and the NLST-like scenario is included among the consensus scenarios. Using the metric D-O (**Figure 3**E), four out of the five consensus scenarios are shown to be selected as consensus scenarios using the metric D/O(**Table 1**). Using the outcomes metric D/(O/S)(**Figure 3**F), four out of the five consensus scenarios also appear in the list selected using the metric D/O (**Table 1**).

4. Discussion

We presented a comparative model-based analysis of overdiagnosis in lung cancer screening by quantifying the trade-off between harms and benefits of LDCT. Our analysis shows that the lifetime screening program recommended by the USPSTF can prevent approximately 3 LC deaths by per overdiagnosed case (mean D/O ratio=2.85). The D/O ratio increases by 22% when the program stops screening at an earlier age 75 instead of 80, as shown in the NLST-like scenario (mean D/O ratio = 3.49). Given that the USPSTF scenario prevents more LC deaths than the NLST-like scenario (i.e. a larger value of D in the USPSTF scenario), the lower D/O of the USPSTF scenario implies that the number of overdiagnosed cases (O) increases more quickly than the number of LC deaths prevented (D) as screening is extended to the older ages. Overall, overdiagnosis was significantly associated with increased screening stopping age in our analysis of 576 hypothetical screening strategies ($P < 10^{-10}$).

The efficiency frontier analysis shows that the most efficient screening strategies that maximize the outcomes metrics incorporating overdiagnosis, namely D/O, LYS/O, D-O and D/(O/S), are consistently the strategies that stop screening at age 75 (which include the NLST-like scenario) compared to programs that screen through age 80. On the other hand, efficient programs chosen based on maximizing the number of LC deaths prevented (D) irrespective of overdiagnosis are the ones that screen through 80, which includes the USPSTF recommendation. While previous model-based studies considered mortality reduction (D) when identifying efficient screening strategies[14-16] including our earlier work[7, 8], we examined the impact of incorporating overdiagnosis on the selection of efficient scenarios by investigating various metrics that integrate overdiagnosis. Undoubtedly there are other useful metrics to more explicitly quantify the harms associated with overdiagnosis such as quality-adjusted life-years (QALY); such metrics were not used in the current study because we intended our analysis to be directly comparable to the recent CISNET analyses performed for the USPSTF[7, 9], which used D, not QALY, in ranking the efficient scenarios.

A noteworthy aspect in the analysis of the D/O ratios for 576 screening strategies is that in most scenarios (99%), the D/O values are larger than one (i.e. D/O>1) across the four models. This finding implies that the number of LC deaths prevented by screening is greater than the number of overdiagnosis cases over a wide range of screening scenarios. In comparison to screening programs for other cancers such as prostate cancer and breast cancer which have been estimated to have D/O values less than one (0.2 for prostate cancer[17] and 0.3 for breast[18]), our results suggest that the negative impact of screening could be lower for LC compared to screening for other cancers. However, morbidity and mortality associated with overdiagnosis of LC may be higher than those of other cancer hence a direct comparison of overdiagnosis-related outcomes across different cancers is not warranted.

Our model-based approach for analyzing overdiagnosis in lung cancer screening has several advantages compared to a trial data based approach. Recently an upper bound of the overdiagnosis risk of 18.5% for LC was estimated based on excess incidence using the NLST data[2]. This estimate is an upper bound because it quantifies the excess incidence observed in the CT arm compared to the chest x-ray arm after a short follow-up period (8 years from trial entry), and likely includes screen-detected cases that would have been clinically detected. Longer follow-up would be needed to observe “catch-up” cancers in the control arm to allow a more accurate estimate of overdiagnosis based on excess incidence [19, 20]. A trial-derived estimate of the overdiagnosis risk would be of limited generalizability even if based on a sufficient follow-up period because the estimate would be associated with the specific screening strategy of the trial. For example, most participants in NLST were screened annually for three years and aged between 55 and 74 with at least 30 pack-years smoking and less than 15 years since they quit smoking at the time of enrollment. However, different screening strategies (e.g. variations in smoking, stopping age or screening frequency) would likely yield different overdiagnosis risks. Given of all these challenges, a model-based approach is valuable for estimating overdiagnosis in a lifetime screening by providing insights into how different screening strategies affect overdiagnosis.

While our overdiagnosis analyses across numerous screening scenarios could not be performed without modeling, modeling has limitations. Firstly, we found that the absolute values of the overdiagnosis risks vary across the four models. This variation is due in part to the fact that the four models were developed independently based on different assumptions, datasets, and mathematical formulations. Given these differences, the model variation captures a range of uncertainty associated with model building that could not be captured by one model alone. However, despite this model variation, relative magnitudes of overdiagnosis risks and related statistics such as D/O were notably consistent across the models. For example, overdiagnosis was larger in the USPSTF scenario than in the NLST-like scenario in all four models, as was D/O in the NLST-like scenario compared to the USPSTF. Second, while we accounted for smoking-related effects on other cause mortality, we assumed that screening was performed on any individual who met the smoking and age criteria without explicit consideration of existing co-morbidities. A screen detected cancer patient with significant co-morbidities is more likely to be overdiagnosed than one without significant co-morbidities, because the patient with comorbidities has a higher competing risk of death. As additional data becomes available to associate other cause mortality with co-morbidities in screening eligibility, the risk of overdiagnosis will likely to decrease. Third, our study assumed perfect screening compliance but if screening compliance reduces at the older ages, then overdiagnosis risks will decrease. Fourth, our analysis is based on calibrations to the practice patterns of NLST; should practice patterns change, particularly for the management of small indeterminant nodules on CT, our D/O estimates would need to be modified. Finally, our study evaluated hypothetical screening strategies that vary by age, smoking, and screening frequency as considered in the USPSTF guidelines, but we did not vary other factors such as nodule size or screening results that may also have impacts on LC mortality or overdiagnosis. Our future research includes the evaluation of efficient diagnostic work up strategies by varying several factors for follow-up, such as nodule size, features, follow-up interval, use of biomarkers, and prior screening results to examine how these factors affect the benefits and harms of CT in the population setting.

In summary, our model-based analysis shows that incorporating overdiagnosis affects the selection of efficient screening strategies. Consistent results across four independent models indicate that our findings are robust. We conclude that while screening through age 80 is efficient in reducing LC mortality irrespective of overdiagnosis, stopping screening at a younger age of 75 provides a greater efficiency in reducing LC deaths and increasing life-years gained per overdiagnosed case, which merit further consideration when balancing the benefits and harms of screening.

Funding

NELSON – Netherlands-Leuven Lung Cancer Screening trial

The NELSON trial is supported by: ‘Zorg Onderzoek Nederland-Medische Wetenschappen’(ZonMw), ‘KWF Kankerbestrijding’, ‘Stichting Centraal Fonds Reserves van Voormalig Vrijwillige Ziekenfondsverzekeringen’ (RVVZ). Kevin ten Haaf is an investigator affiliated with the NELSON trial

Siemens Germany provided 4 digital workstations and LungCARE for the performance of 3D measurements.

‘Health Technology Assessment for CT Lung Cancer Screening in Canada’. Cancer Care Ontario, dr. Paszat. Grant.

Competing Interests

HJdK took part in a 1-day advisory meeting on biomarkers organized by M.D. Anderson/Health Sciences during the 16th World Conference on Lung Cancer.

HJdK and KtH received a grant from the University of Zurich to assess the cost-effectiveness of computed tomographic lung cancer screening in Switzerland

6. References

1. Aberle D, Adams A, Berg C, Black W, Clapp J, Fagerstrom R, Gareen I, Gatsonis C, Marcus P, Sicks J. Reduced lung-cancer mortality with low-dose computed tomographic screening. The New England journal of medicine 2011;365:395.

2. Patz EF, Pinsky P, Gatsonis C, Sicks JD, Kramer BS, Tammemägi MC, Chiles C, Black WC, Aberle DR. Overdiagnosis in low-dose computed tomography screening for lung cancer. JAMA internal medicine 2014;174:269-74.

3. Dammas S, Patz Jr EF, Goodman PC. Identification of small lung nodules at autopsy: implications for lung cancer screening and overdiagnosis bias. Lung Cancer 2001;33:11-16.

4. Black WC. Overdiagnosis: an underrecognized cause of confusion and harm in cancer screening. Journal of the National Cancer Institute 2000;92:1280-82.

5. Welch HG, Black WC. Overdiagnosis in cancer. Journal of the National Cancer Institute 2010;102:605-13.

6. Meza R, Haaf Kt, Kong CY, Erdogan A, Black W, Tammemagi M, Choi SE, Jeon J, Han S, Munshi V, Rosmalen JMv, Pinsky P, et al. Comparative Analysis of Five Lung Cancer Natural History and Screening Models that Reproduce Outcomes of the NLST and PLCO Trials. Cancer 2014.

7. de Koning HJ, Meza R, Plevritis SK, ten Haaf K, Munshi VN, Jeon J, Erdogan SA, Kong CY, Han SS, van Rosmalen J. Benefits and harms of computed tomography lung cancer screening strategies: a comparative modeling study for the US Preventive Services Task Force. Annals of internal medicine 2014.

8. McMahon PM, Meza R, Plevritis SK, Black WC, Tammemagi CM, Erdogan A, ten Haaf K, Hazelton W, Holford TR, Jeon J. 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.

9. Moyer VA. Screening for lung cancer: US Preventive Services Task Force recommendation statement. Annals of internal medicine 2014.

10. Wender R, Fontham ET, Barrera E, Colditz GA, Church TR, Ettinger DS, Etzioni R, Flowers CR, Scott Gazelle G, Kelsey DK. American Cancer Society lung cancer screening guidelines. CA: A Cancer Journal for Clinicians 2013;63:106-17.

11. Yano T, Yokoyama H, Fukuyama Y, Takai E, Mizutani K, Ichinose Y. The current status of postoperative complications and risk factors after a pulmonary resection for primary lung cancer: a multivariate analysis. European journal of cardio-thoracic surgery 1997;11:445-49.

12. Oken MM, Hocking WG, Kvale PA, Andriole GL, Buys SS, Church TR, Crawford ED, Fouad MN, Isaacs C, Reding DJ. Screening by chest radiograph and lung cancer mortality. JAMA: the journal of the American Medical Association 2011;306:1865-73.

13. Holford TR, Clark L. Development of the counterfactual smoking histories used to assess the effects of tobacco control. Risk Analysis 2012;32:S39-S50.

14. Mandelblatt JS, Cronin KA, Bailey S, Berry DA, de Koning HJ, Draisma G, Huang H, Lee SJ, Munsell M, Plevritis SK. Effects of mammography screening under different screening schedules: model estimates of potential benefits and harms. Annals of internal medicine 2009;151:738-47.

15. Zauber AG, Lansdorp-Vogelaar I, Knudsen AB, Wilschut J, van Ballegooijen M, Kuntz KM. Evaluating test strategies for colorectal cancer screening: a decision analysis for the US Preventive Services Task Force. Annals of internal medicine 2008;149:659-69.

16. van den Akker-van ME, van Ballegooijen M, van Oortmarssen GJ, Boer R, Habbema JDF. Cost-effectiveness of cervical cancer screening: comparison of screening policies. Journal of the National Cancer Institute 2002;94:193-204.

17. Gulati R, Gore JL, Etzioni R. Comparative effectiveness of alternative prostate-specific antigen–based prostate cancer screening strategies: model estimates of potential benefits and harms. Annals of internal medicine 2013;158:145-53.

18. Screening IUPoBC. The benefits and harms of breast cancer screening: an independent review. Lancet 2012;380:1778.

19. Marcus PM, Bergstralh EJ, Zweig MH, Harris A, Offord KP, Fontana RS. Extended lung cancer incidence follow-up in the Mayo Lung Project and overdiagnosis. Journal of the National Cancer Institute 2006;98:748-56.

20. Etzioni R, Xia J, Hubbard R, Weiss NS, Gulati R. A reality check for overdiagnosis estimates associated with breast cancer screening. Journal of the National Cancer Institute 2014;106:dju315.

Figure legends

Figure 1 Overdiagnosis risk (%) of 576 scenarios by stopping age of screening programs for each model and gender. Overdiagnosis risk is calculated as the number of overdiagnosed cases divided by the number of screen-detected cases. The number for each box represents a median of overdiagnosis risk of screening programs with given stopping age. “KW” denotes Kruskal-Wallis (K-W) test.

Figure 2 Comparisons of overdiagnosis risks and D/O ratios (the number of LC deaths prevented per overdiagnosed case) of the USPSTF and the NLST-like scenarios, by gender and both genders combined.

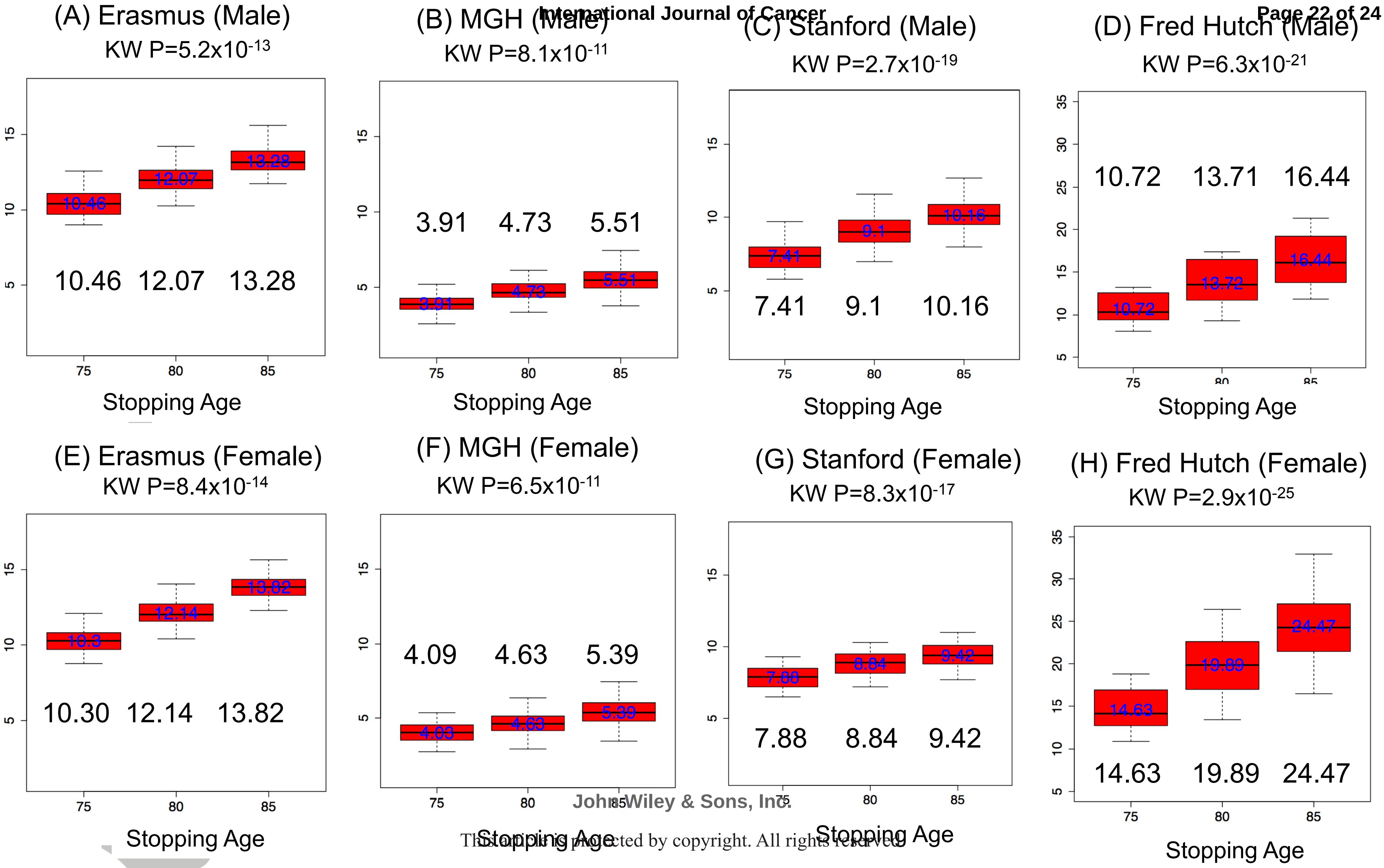
Note: Overdiagnosis risk is calculated as the number of overdiagnosed cases divided by the number of screen-detected cases.

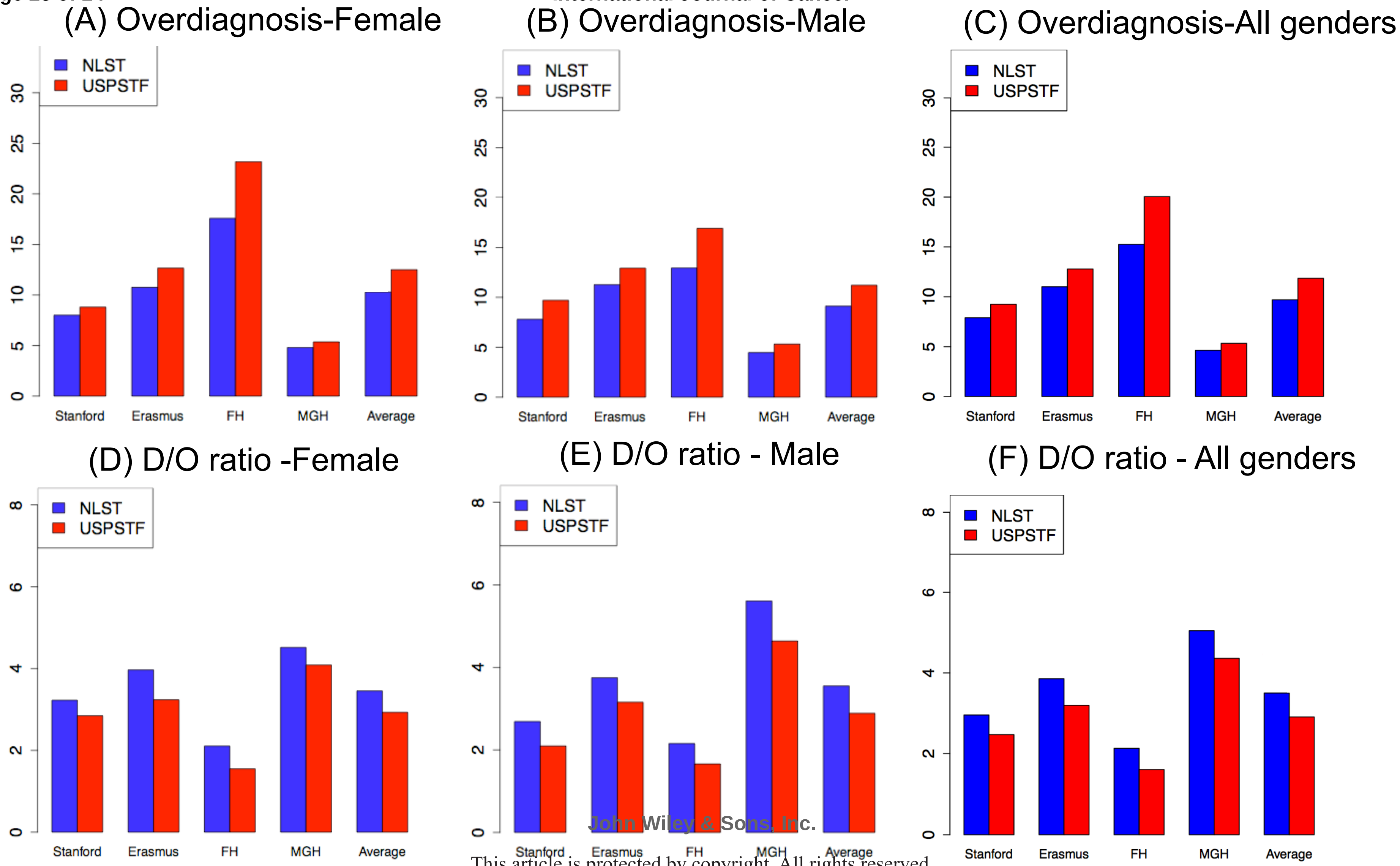
Figure 3 Consensus screening scenarios chosen for males by maximizing: (i) D, the number of LC deaths prevented (A); (ii) D/O, the number of LC deaths prevented per overdiagnosed case (B); (iii) life-years gained (LYG) (C); (iv) life-year gained per overdiagnosed case (LYG/O)(D); (v) D-O, net LC deaths prevented subtracting overdiagnosed cases (E); (vi) and D/(O/S), the number of LC deaths prevented per overdiagnosis risk (F), where S is the number of screen-detected cases. In each figure, we show the outcomes under several screening strategies that vary by age and smoking eligibility criteria. Each dot represents a specific screening strategy, with selected scenarios highlighted in color. Here, the x-axis is the number of CT screens that need to be performed under each strategy. Panel A shows the number of LC deaths avoided versus no-screening (D, y-axis) under the given strategy. Panels B-F show alternative outcome metrics: LC deaths avoided per overdiagnosed case (D/O, panel B), life-years gained (LYG, panel C), life-years gained per overdiagnosed case (LYG/O, panel D), the net prevented LC deaths subtracted by the number of overdiagnosed cases (D-O, panel E), and the number of LC deaths prevented per overdiagnosis risk (D/(O/S), panel F). Within each metric, a consensus scenario was identified by choosing a scenario that is defined as an “efficient scenario” (i.e. top 25% closest scenarios to the efficient frontier) by at least three out of the four models under each metric. The consensus scenarios are listed in the legend box and highlighted for a representative model. For each panel, the NLST-like and the USPSTF scenarios are plotted for reference purposes, regardless on whether or not they are included in the consensus list. The results for females are shown in Supplemental Figure 5-6.

Note: Each legend box shows the scenarios selected by consensus across the four models and annotated as Frequency–Start Age (y)–Stop Age (y)–Pack- Years–Years Since Quitting.

Table 1 Consensus scenarios chosen by maximizing the number of prevented LC deaths (D); and the number of prevented LC deaths per overdiagnosed case (D/O). A consensus scenario was identified by choosing a scenario that is defined as an “efficient scenario” (i.e. top 25% closest scenarios to the efficient frontier) by at least three out of the four models. For the numbers in each cell below, model average values were used. The USPSTF scenario and NLST-like scenarios are highlighted in red and blue, respectively. Grayed scenarios are the ones that overlap between genders within each metric.

	Metric	Efficient Scenario Frequency–Start Age (y)–Stop Age (y)–Pack- Years–Years Since Quitting	# of CT scans	Overdiagnosis (%)	# of overdiagnosed cases (O)	# of prevented LC deaths (D)	Mortality Reduction (%)	Life-years saved	D/O
Female	D	A-55-80-40-25	166177	13.82	241	453	14.6	5983	2.56
		A-60-80-20-10	186932	12.29	224	456	14.7	5889	2.74
		A-60-80-30-20	189433	13.55	264	480	15.7	6212	2.55
		A-60-80-30-25	208614	13.69	290	510	16.9	6593	2.47
		A-60-80-20-15	227049	12.81	272	527	17.6	6833	2.65
		A-55-80-30-15	**232461**	**12.5**	**239**	**528**	**17.7**	**7342**	**2.94**
		A-60-80-10-15	261556	12.77	281	551	18.3	7209	2.69
	D/O	A-55-75-30-10	186549	9.98	145	422	13.6	6194	3.56
		A-55-75-30-15	**214158**	**10.28**	**166**	**466**	**15.3**	**6831**	**3.45**
		A-55-75-30-20	235702	10.39	180	495	16.5	7340	3.48
		A-55-75-30-25	250305	10.49	189	512	17.2	7616	3.46
		A-55-75-20-10	253105	9.66	169	504	16.9	7350	3.63
Male	D	A-55-80-40-25	260832	11.78	256	526	14.2	7700	2.70
		A-60-80-30-20	261778	11.88	277	528	14.2	7122	2.55
		A-55-80-30-10	286878	11.2	250	547	14.7	7936	2.87
		A-60-80-30-25	287521	11.89	294	563	15.2	7616	2.58
		A-60-80-20-20	307380	11.77	288	561	15.3	7779	2.57
		A-55-80-30-15	**326549**	**11.21**	**270**	**597**	**16.3**	**8698**	**2.90**
	D/O	A-55-75-30-10	267730	9.13	180	498	13.2	7513	3.55
		A-55-75-30-15	**301853**	**9.12**	**192**	**536**	**14.4**	**8170**	**3.54**
		A-55-75-20-10	312252	9.09	186	532	14.3	8164	3.57
		A-55-75-30-20	330807	9.02	200	566	15.4	8679	3.60


(A) Erasmus (Male)
KW P=5.2x10^{-13}
10.46
12.07
13.28
5
10
15
75
80
85
Stopping Age
(B) MGH (Male)
KW P=8.1x10^{-11}
3.91
4.73
5.51
5
10
15
75
80
85
Stopping Age
(C) Stanford (Male)
KW P=2.7x10^{-19}
7.41
9.1
10.16
5
10
15
75
80
85
Stopping Age
(D) Fred Hutch (Male)
KW P=6.3x10^{-21}
10.72
13.71
16.44
13.72
5
10
15
20
25
30
35
75
80
85
Stopping Age
(E) Erasmus (Female)
KW P=8.4x10^{-14}
10.30
12.14
13.82
10.3
5
10
15
75
80
85
Stopping Age
(F) MGH (Female)
KW P=6.5x10^{-11}
4.09
4.63
5.39
4.03
5
10
15
75
80
85
Stopping Age
(G) Stanford (Female)
KW P=8.3x10^{-17}
7.88
8.84
9.42
5
10
15
75
80
85
Stopping Age
(H) Fred Hutch (Female)
KW P=2.9x10^{-25}
14.63
19.89
24.47
5
10
15
20
25
30
35
75
80
85
Stopping Age



(A) Overdiagnosis-Female

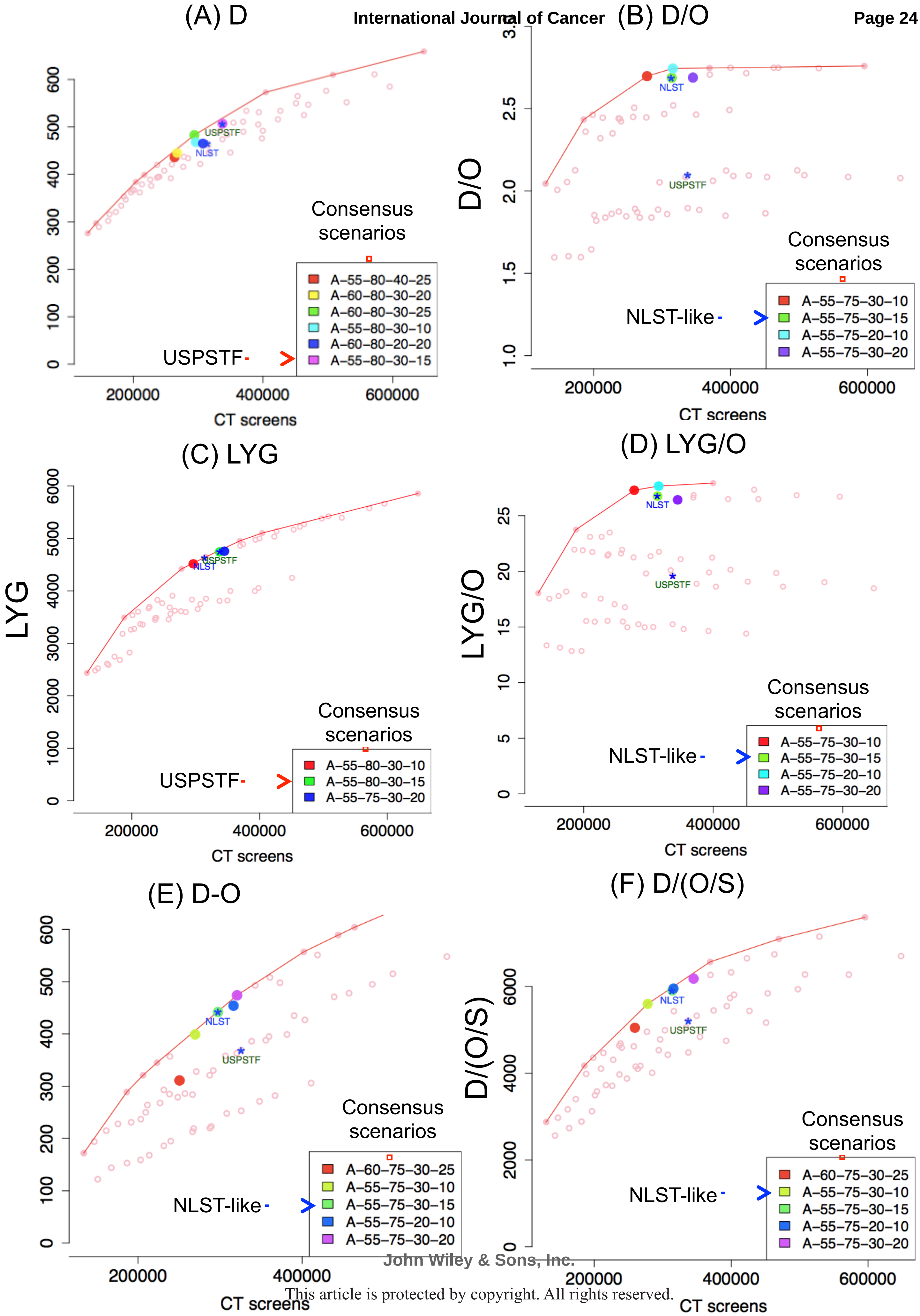
(B) Overdiagnosis-Male

(C) Overdiagnosis-All genders

(D) D/O ratio -Female

(E) D/O ratio - Male

(F) D/O ratio - All genders


(A) D
600
500
400
300
200
100
0
200000
400000
600000
CT screens
USPSTF
NLST
Consensus scenarios
A–55–80–40–25
A–60–80–30–20
A–60–80–30–25
A–55–80–30–10
A–60–80–20–20
A–55–80–30–15
USPSTF
(B) D/O
D/O
3.0
2.5
2.0
1.5
1.0
200000
400000
600000
CT screens
NLST
USPSTF
Consensus scenarios
A–55–75–30–10
A–55–75–30–15
A–55–75–20–10
A–55–75–30–20
NLST-like
(C) LYG
LYG
6000
5000
4000
3000
2000
1000
0
200000
400000
600000
CT screens
USPSTF
NLST
Consensus scenarios
A–55–80–30–10
A–55–80–30–15
A–55–75–30–20
USPSTF
(D) LYG/O
LYG/O
25
20
15
10
5
0
200000
400000
600000
CT screens
NLST
USPSTF
Consensus scenarios
A–55–75–30–10
A–55–75–30–15
A–55–75–20–10
A–55–75–30–20
NLST-like
(E) D-O
600
500
400
300
200
100
0
200000
400000
CT screens
NLST
USPSTF
Consensus scenarios
A–60–75–30–25
A–55–75–30–10
A–55–75–30–15
A–55–75–20–10
A–55–75–30–20
NLST-like
(F) D/(O/S)
D/(O/S)
6000
4000
2000
0
200000
400000
600000
CT screens
NLST
USPSTF
Consensus scenarios
A–60–75–30–25
A–55–75–30–10
A–55–75–30–15
A–55–75–20–10
A–55–75–30–20
NLST-like