Accuracy of Clinicians and Models for Estimating the Probability That a Pulmonary Nodule Is Malignant

Alex A. Balekian¹, Gerard A. Silvestri², Suzanne M. Simkovich², Peter J. Mestaz³, Gillian D. Sanders⁴, Jamie Daniel⁴, Jackie Porcel⁵, and Michael K. Gould⁵

¹Division of Pulmonary and Critical Care Medicine, Keck School of Medicine of the University of Southern California, Los Angeles, California; ²Division of Pulmonary and Critical Care Medicine, Medical University of South Carolina, Charleston, South Carolina; ³Office of Research Oversight, Midwest Region, Department of Veterans Affairs, Hines, Illinois; ⁴Duke Clinical Research Institute, Duke University, Durham, North Carolina; and ⁵Department of Research and Evaluation, Kaiser Permanente Southern California, Pasadena, California

Abstract

Rationale: Management of pulmonary nodules depends critically on the probability of malignancy. Models to estimate probability have been developed and validated, but most clinicians rely on judgment.

Objectives: The aim of this study was to compare the accuracy of clinical judgment with that of two prediction models.

Methods: Physician participants reviewed up to five clinical vignettes, selected at random from a larger pool of 35 vignettes, all based on actual patients with lung nodules of known final diagnosis. Vignettes included clinical information and a representative slice from computed tomography. Clinicians estimated the probability of malignancy for each vignette. To examine agreement with models, we calculated intraclass correlation coefficients (ICC) and kappa statistics. To examine accuracy, we compared areas under the receiver operator characteristic curve (AUC).

Measurements and Main Results: Thirty-six participants completed 179 vignettes, 47% of which described patients with malignant nodules. Agreement between participants and models was fair for the Mayo Clinic model (ICC, 0.37; 95% confidence interval [CI], 0.23–0.50) and moderate for the Veterans Affairs model (ICC, 0.46; 95% CI, 0.34–0.57). There was no difference in accuracy between participants (AUC, 0.70; 95% CI, 0.62–0.77) and the Mayo Clinic model (AUC, 0.71; 95% CI, 0.62–0.80; $P = 0.90$) or the Veterans Affairs model (AUC, 0.72; 95% CI, 0.64–0.80; $P = 0.54$).

Conclusions: In this vignette-based study, clinical judgment and models appeared to have similar accuracy for lung nodule characterization, but agreement between judgment and the models was modest, suggesting that qualitative and quantitative approaches may provide complementary information.

Keywords: solitary pulmonary nodule; diagnosis; prediction models

(Received in original form May 2, 2013; accepted in final form August 11, 2013)

Supported by National Cancer Institute grant CA117840.

The views expressed in this manuscript are those of the authors and not necessarily those of the National Cancer Institute.

Author contributions: Study conception and design (G.A.S., G.D.S., M.K.G.), obtaining funding (G.D.S., M.K.G.), data collection (G.A.S., S.M.S., P.J.M., G.D.S., J.D., M.K.G.), data analysis (A.A.B., P.J.M., J.P., M.K.G.), manuscript preparation (A.A.B., M.K.G.), manuscript review and revision for important content (A.A.B., G.A.S., S.M.S., P.J.M., G.D.S., J.D., J.P., M.K.G.).

Correspondence and requests for reprints should be addressed to Michael K. Gould, M.D., M.S., Department of Research and Evaluation, Kaiser Permanente Southern California, 100 South Los Robles Avenue, Pasadena, CA 91101. E-mail: Michael.k.gould@kp.org

Ann Am Thorac Soc Vol 10, No 6, pp 629–635, Dec 2013

Copyright © 2013 by the American Thoracic Society

DOI: 10.1513/AnnalsATS.201305-107OC

Internet address: www.atsjournals.org

The discovery of asymptomatic pulmonary nodules is becoming more frequent. Nearly one in five cardiac computed tomography (CT) scans has been reported to show an incidental pulmonary nodule (1). In addition, the National Lung Screening Trial reported that 39% of high-risk smokers who underwent annual CT screening had

a least one positive screening examination (2). Because 40% of Americans are current or former smokers (3), many of whom will meet eligibility criteria for CT screening, the number of pulmonary nodules that will need to be further evaluated is substantial.

One of the first steps in pulmonary nodule evaluation is to estimate the

likelihood of cancer. Clinicians typically use their judgment and intuition to assess the risk of malignancy and guide the subsequent workup. For those who are more quantitatively inclined, clinical prediction models (or risk assessment models) have been developed to provide a more explicit, transparent, and reproducible assessment of

this risk. One commonly cited model was developed and internally validated at the Mayo Clinic (4) and subsequently validated in independent samples of U.S. veterans with lung nodules, a surgical population from an academic center in the southeastern United States, and patients from the Netherlands who underwent positron emission tomography (5–7). A similar model was developed and validated using samples of U.S. veterans with a higher prevalence of malignancy (8).

Whether models perform better than clinicians is uncertain. In one prior study, there was no difference in accuracy between the judgment of four expert physicians and the probabilities generated by the Mayo Clinic model, although the experts tended to overestimate risk for nodules identified as low risk by the model (9). To provide additional empirical evidence about this question, we compared the accuracy of two clinical prediction models with those of a larger number of expert and nonexpert clinicians.

Methods

We performed a prospective, vignette-based study of clinician decision making to compare the accuracy of clinician judgment versus published clinical prediction models for estimating the probability that a lung nodule is malignant.

Participants

The target population for this study included clinicians who care for patients with lung nodules. To recruit study participants, we contacted members of the Section on Thoracic Oncology of the American Thoracic Society by e-mail. A single reminder was sent to nonresponders. In addition, we recruited a separate sample of faculty members and fellows from the Departments of Medicine and Family Medicine at the Medical University of South Carolina. All participants provided informed consent. The study protocol was approved by Institutional Review Boards at Stanford University, the University of Southern California, and the Medical University of South Carolina.

Settings

We performed one set of assessments at the 2010 International Conference of the American Thoracic Society in New Orleans,

Louisiana. We performed another set of assessments in September 2011 at the Medical University of South Carolina.

Vignettes

Clinical vignettes are written case scenarios that summarize all of the important information about a specific patient with an acute or chronic medical problem. Because they are easily administered and inexpensive, vignettes have been used to measure processes of care in a variety of clinical settings (10, 11).

For this study, we developed a library of 35 clinically representative vignettes that were based on actual patients with pulmonary nodules in whom the final diagnosis was known. Each vignette provided relevant clinical information for participating clinicians to estimate the probability of cancer, including both patient characteristics (age, sex, smoking history, previous history of malignancy) and nodule characteristics (dimensions/size, location within the lung, nodule edge properties). Each vignette was accompanied by one or more representative slices from the corresponding patient's chest CT scan.

Vignettes were drawn from a sample of 151 consecutive patients with pulmonary nodules measuring between 7 mm and 30 mm and a known final diagnosis at Veterans

Affairs (VA) Palo Alto Health Care System during the years 2000 to 2006 (5). The mean age was 66.9 ± 10.1 years, 3% were women, 8% were black, and 3% were Hispanic. The prevalence of malignancy was 44%. The final diagnosis was established by histopathology or, in the case of benign causes, at least 2 years of clinical stability. When selecting cases for inclusion in the library of vignettes, we attempted to minimize ambiguity and to capture multiple sources of variability in patient and nodule characteristics. To protect the confidentiality of patients, all protected health information was removed in a Health Insurance Portability and Accountability Act-compliant manner. Figure 1 shows an example of a representative vignette.

Clinical Models

We examined the accuracy of two clinical models for estimating the probability of cancer, one developed at Mayo Clinic and one developed in an independent sample of veterans with lung nodules (4). The Mayo model was developed in a sample of 419 patients with noncalcified nodules measuring between 4 mm and 30 mm on chest radiography between 1984 and 1986. In this sample, the prevalence of malignancy was 23%. Independent

Figure 1. Representative clinical vignette for a patient with a benign pulmonary nodule. Vignette 30: Mr. AD is an 82-year-old man and a former smoker. He was found to have a 10-mm, noncalcified nodule in the left upper lobe on chest CT. What is your best estimate for the probability that this nodule is malignant? CT = computed tomography.

predictors of malignancy included older age (odds ratio [OR], 1.04 for each yr; 95% CI, 1.01–1.07), current or past smoking history (OR, 2.2; 95% CI, 1.17–4.16), past history of extrathoracic cancer more than 5 years before nodule detection (OR, 3.8; 95% CI, 1.39–10.49), nodule diameter (OR, 1.14 for each millimeter; 95% CI, 1.09–1.19), spiculated borders (OR, 2.8; 95% CI, 1.47–5.45), and upper lobe location (OR, 2.2; 95% CI, 1.27–3.79). The VA model was developed in a sample of 375 veterans (98% men) with a new nodule identified by chest X-ray between 1998 and 2001. The mean nodule size was 16 mm, and 54% of the nodules were malignant. In the VA model, independent predictors of malignancy included older age (OR, 2.2 per 10-yr increase; 95% CI, 1.7–2.8), current or past smoking history (OR, 7.9; 95% CI, 2.6–23.6), nodule diameter (OR, 1.1 per millimeter increase; 95% CI, 1.1–1.2), and (for former smokers) time since quitting (OR, 0.6 per 10-yr increase; 95% CI, 0.4–0.7). Neither of the models accounts for other nodule characteristics (e.g., attenuation) or other clues that may be available to readers of the chest CT scan. Both the Mayo and VA models have been internally and externally validated and found to have good accuracy for discriminating between malignant and benign nodules (range of *c* statistics, 0.73–0.80) (5–7).

Elicitation

Clinical vignettes were created with Microsoft PowerPoint software and displayed on a laptop computer. Each participant was asked to complete a series of five vignettes. After reviewing the clinical information and representative image(s), each participant was asked to provide an estimate of the probability of malignancy.

Data Analysis

The likelihood of malignancy, whether derived from the models or elicited from the clinicians, was expressed as a decimal ranging from 0 to 1.0. Because the data were not normally distributed, we used the nonparametric Wilcoxon signed rank test to compare means, and used Spearman rho to measure simple correlations between probability estimates of clinicians and each of the models. In addition, we created Bland-Altman plots and modified Bland-Altman plots, in which we plotted the difference between the estimates of

clinicians and those of models as a function of the estimate that was derived from the model (12). To assess agreement in probability estimates measured continuously, we calculated intraclass correlation coefficients. We also calculated weighted kappa statistics to assess agreement in probability estimates across clinically meaningful tertiles of risk categorized as low (<0.30), moderate (0.30–0.60), or high (>0.60) by clinicians (13).

We compared accuracy by creating receiver operating characteristic (ROC) curves, calculating the areas under the ROC curves, and using the method of DeLong and colleagues to compare these areas (14). We used multiple logistic regression to test whether either model provided additional information when added to the predictions of clinicians. Data analysis was performed using IBM SPSS statistics version 21 (Armonk, NY) and SAS version 9.2 (SAS Institute, Cary, NC).

Results

Eleven physician participants completed 54 vignettes at the American Thoracic Society (ATS) International Conference, and 25 additional physician participants completed 125 vignettes at the Medical University of South Carolina (MUSC). The 11 ATS participants were all experienced pulmonologists from academic medical centers, including 2 women and 9 men. At MUSC, the sample included 11 fellows in pulmonary and critical care medicine and 14 faculty members, including 7 from pulmonary medicine and 7 from primary care (internal medicine or family medicine).

Among the 179 vignettes sampled at random, 84 (47%) represented malignant nodules. The percentage of sampled vignettes with malignant nodules was similar at ATS (48%) and MUSC (46%). For 25 of 179 of the sampled vignettes, data were missing for past history of extrathoracic cancer, so the probability of malignancy could not be estimated by the Mayo Clinic model.

The distributions of predicted probabilities were significantly different from normal for participants ($P < 0.01$), the Mayo Clinic model ($P < 0.01$), and the VA model ($P < 0.01$). The median (interquartile range) values for the estimated probabilities were 0.50 (0.21–0.75)

for participants, 0.43 (0.15–0.53) for the Mayo Clinic model, and 0.42 (0.26–0.69) for the VA model. Estimated probabilities differed significantly between participants and models ($P = 0.045$). Two-way comparisons revealed significant differences between participants and the Mayo Clinic model ($P = 0.003$) but not between participants and the VA model ($P = 0.23$).

Simple correlations between the estimated probabilities of participants and those derived from models were similar for the Mayo model ($r = 0.41$, $P < 0.001$) and the VA model ($r = 0.47$, $P < 0.001$). However, the intraclass correlation between clinicians and models appeared to be greater for the VA model (0.46; 95% CI, 0.34–0.57) than for the Mayo Clinic model (0.37; 95% CI, 0.23–0.50), indicating moderate (VA model) to fair (Mayo Clinic model) agreement. Agreement between participants and the models also appeared to be greater for the VA model across tertiles of risk (low, medium, high). The weighted kappa for the VA model was 0.31 (95% CI, 0.20–0.42), indicating fair agreement, whereas the weighted kappa for the Mayo model was 0.19 (95% CI, 0.07–0.30), indicating slight agreement. Modified Bland-Altman plots showed that, relative to the models, clinicians frequently overestimated the probability of malignancy at lower predicted probabilities and somewhat less frequently underestimated the probability of malignancy at higher predicted values (Figure 2).

Participants and models had similar accuracies for identifying malignancy. For 154 vignettes with complete data for the Mayo Clinic model, the areas under the ROC curve (AUC) for all participating clinicians and the Mayo Clinic model were 0.72 (95% CI, 0.63–0.80) and 0.71 (95% CI, 0.62–0.80), respectively ($P = 0.90$). For the full set of 179 vignettes, the AUC for all participants and the VA model were 0.70 (95% CI, 0.62–0.77) and 0.72 (95% CI, 0.64–0.80), respectively ($P = 0.54$). Figure 3 compares ROC curves for participants, the Mayo Clinic model, and the VA model for the 154 vignettes with complete data.

In subgroup analyses, we found that the accuracy of clinicians did not differ from that of the Mayo Clinic model for any group of clinicians, including specialists ($P = 0.96$), nonfellows ($P = 0.97$), or clinical experts, defined as a specialist who was not

Figure 2. Modified Bland-Altman plots showing the difference between clinical probability estimates generated by physician participants and the Mayo Clinic model (A) or the Veterans Affairs (VA) model (B), plotted as a function of the probability derived from the models. Plots show that compared with models, physicians tend to overestimate the risk of cancer at lower predicted probabilities and underestimate the risk of cancer at higher predicted probabilities.

a pulmonary fellow ($P = 0.97$). Likewise, there were no differences in accuracy between any group of clinicians and the VA model (Table 1). Although the accuracy of nonexperts appeared to be slightly lower than that of experts, this was probably related to vignette selection, as the accuracy of models also appeared to be lower for these vignettes.

In multiple logistic regression models, predictions of both clinicians and models were independently associated with the presence of cancer (Table 2). For each 10-percentage point increase in the probability of cancer as estimated by the Mayo model, the odds of cancer were 28% higher (OR,

1.28; 95% CI, 1.09–1.50). For each 10-percentage point increase in the probability of cancer as estimated by the VA model, the odds of cancer were 32% higher (OR, 1.32; 95% CI, 1.14–1.52). Thus, models provided additional information on top of clinician judgment.

Discussion

In this study, we found that clinicians and validated prediction models had similar accuracy for characterizing pulmonary nodules. However, accuracy was only fair to good for both clinicians and models.

Furthermore, agreement between clinicians and the Mayo Clinic model was only fair, whereas agreement between clinicians and the VA model was moderate, suggesting that models and clinical judgment may provide complementary information. Finally, clinicians tended to overestimate the risk of malignancy for nodules identified as low risk by the models, a conservative strategy that may help to avoid missed opportunities for surgical cure.

The fair accuracy of both clinicians and models suggests that there may be room for improvement in risk assessment. Although uncommonly used in clinical settings to date, future advances in computer-aided diagnosis (CAD) using artificial neural networks may eventually supplant the simple models studied in this analysis (15).

The greater agreement between clinicians and the VA model may reflect the inclusion of vignettes that were developed using a sample of U.S. veterans as well as the lower prevalence of malignancy in the sample used to develop the Mayo Clinic model. More importantly, the limited agreement between clinicians and either model suggests that the combination of clinical judgment and quantitative prediction may prove superior to either when used alone. Although this hypothesis requires testing in subsequent studies of risk assessment, we believe that the rich additional information about patient and nodule characteristics that is available to the clinician can improve prediction when superimposed on the systematic and data-driven information provided by the models. Limited support for this hypothesis comes from a Chinese study of CAD using an artificial neural network. In this study, the addition of CAD to clinical judgment improved the predictions of both senior and more junior radiologists (16). Conversely, the quantitative information provided by models can help the clinician to refine his or her intuitive judgment. The results of our multiple logistic regression analyses supported this hypothesis by showing that both clinicians and models were independent predictors of malignancy.

Another study compared the predictions of the Mayo Clinic model with those of four expert clinicians from the Mayo Clinic, including a chest radiologist, a pulmonologist, a general internist, and a thoracic surgeon (9). Each clinician reviewed clinical information and plain

Figure 3. Receiver operating characteristic (ROC) curves representing the accuracy of predictions by physician participants (*black*), the Mayo Clinic model (*green*), and the Veterans Affairs (VA) model (*red*). Areas under the ROC curve were 0.70 (95% confidence interval [CI], 0.62–0.78) for participants, 0.71 (95% CI, 0.62–0.80) for the Mayo Clinic model, and 0.72 (95% CI, 0.64–0.81) for the VA model. Differences in accuracy between participants and the Mayo Clinic model ($P = 0.46$) and between participants and the VA model ($P = 0.30$) were not statistically significant.

chest radiographs from 100 case vignettes, including 49 patients with malignant nodules and 51 patients with benign nodules, and made a prediction of malignancy on a 10-point scale. In this

study, there was no difference in accuracy between clinicians and the Mayo Clinic model. However, both the model and clinicians appeared to be more accurate in this study than what we observed, with

a reported AUC of 0.79 for the model and a range of AUCs between 0.82 and 0.87 for the clinicians. The better accuracy could be related to either the greater expertise of the clinicians or perhaps to the inclusion of cases and images that were less ambiguous. Similar to our finding, this analysis also reported that clinicians overestimated the probability of malignancy among cases in which the calculated risk of cancer was low.

A third study from Japan compared the accuracy of a novel risk assessment model with that of a single experienced chest radiologist (17). This model combined three radiographic findings (calcification, spiculation, and bronchus sign) and two serologic markers (C-reactive protein and carcinoembryonic antigen). Although the model was highly accurate in the development or training set (AUC, 0.94), the chest radiologist (AUC, 0.91) was more accurate than the model (AUC, 0.86) in a separate validation data set. In retrospect, the model may have benefited from the inclusion of demographic and other patient characteristics (although this information was not available to the radiologist either). Given the very high prevalence of malignancy in both the development set (75%) and validation set (89%), the results

Table 1. Comparisons of accuracy between clinicians and models

Comparison	No. of Vignettes	AUC (95% CI) for Clinicians	AUC (95% CI) for Model	P Value for Comparison
All clinicians				
vs. Mayo model	154	0.72 (0.63–0.80)	0.71 (0.62–0.80)	0.90
vs. VA model	179	0.70 (0.62–0.77)	0.72 (0.64–0.80)	0.54
Specialists				
vs. Mayo model	129	0.73 (0.64–0.81)	0.72 (0.63–0.82)	0.96
vs. VA model	144	0.72 (0.63–0.80)	0.75 (0.66–0.83)	0.47
Primary care				
vs. Mayo model	25	0.53 (0.23–0.82)	0.55 (0.22–0.88)	0.90
vs. VA model	35	0.55 (0.32–0.77)	0.58 (0.36–0.81)	0.77
Pulmonary fellows				
vs. Mayo model	43	0.69 (0.54–0.85)	0.72 (0.55–0.88)	0.79
vs. VA model	55	0.70 (0.55–0.84)	0.75 (0.62–0.89)	0.39
Nonfellows				
vs. Mayo model	111	0.71 (0.61–0.82)	0.71 (0.60–0.82)	0.97
vs. VA model	124	0.70 (0.60–0.79)	0.71 (0.61–0.81)	0.78
Experts (specialist and not a pulmonary fellow)				
vs. Mayo model	86	0.74 (0.63–0.84)	0.73 (0.62–0.85)	0.97
vs. VA model	89	0.74 (0.63–0.84)	0.75 (0.64–0.86)	0.80
Nonexperts (primary care or pulmonary fellow)				
vs. Mayo model	68	0.69 (0.56–0.82)	0.68 (0.53–0.83)	0.87
vs. VA model	90	0.66 (0.54–0.78)	0.69 (0.58–0.81)	0.60

Definition of abbreviations: AUC = area under the receiver operating curve; CI = confidence interval; VA = Veterans Affairs.

Table 2. Results of multiple logistic regression analysis, adjusted for specialty and trainee status

	OR	95% CI	P Value
Model 1			
Clinicians	1.28	1.10–1.49	0.001
Mayo model	1.30	1.10–1.53	0.002
Model 2			
Clinicians	1.17	1.02–1.34	0.03
VA model	1.32	1.14–1.54	<0.0001

Definition of abbreviations: CI = confidence interval; OR = odds ratio; VA = Veterans Affairs.

of this study might be difficult to generalize to other settings.

Our study is the first to describe agreement and compare accuracy between clinician judgment and two validated models, both of which are easily applied in clinical practice. In addition, we recruited a relatively large number of physician participants who represented both different specialties and levels of training. In contrast to previous studies, we found that clinicians from various disciplines and with widely varying experience have comparable accuracy to validated clinical models.

Another important insight is that compared with models, physicians often overestimate the probability of malignancy,

especially at the low end of predicted risk. The mechanism underlying this overestimation of risk is unclear, but the tendency for physicians to “overcall” benign nodules and err on the side of not missing a potentially curable malignancy is reasonable. However, the tests required to evaluate these nodules—additional scans, invasive biopsies, and surgical resections—can be expensive, are not without risk, and may be unnecessary. This may be especially true in the context of lung cancer screening, in which the prevalence of malignancy is typically low and the potential for harm from unnecessary invasive testing is greater. It is tempting to speculate that a low-probability estimate from a model may help strike the right balance, by serving as a reality check against the understandable concern that “cancer is the answer.”

Our study has several important limitations. First, although we enrolled a diverse group of clinicians, they were not sampled at random, and most practiced in academic settings and therefore were not representative of practicing clinicians. Second, the vignettes provided clinicians with a limited amount of information that was largely identical to the information used by the models, with the important exception that clinicians had access to the rich information contained in a representative slice from the patient’s chest CT scan. With access to more detailed

information about geographic location, past medical history, symptoms, and exposures, one might expect that clinicians would be even more accurate. Third, vignettes were based on patients with lung nodules measuring between 7 mm and 30 mm in widest diameter, so our findings do not apply to the characterization of smaller nodules. Finally, it should be noted that the assessment of clinical probability is only a first step in the evaluation of pulmonary nodules. To date, no studies have examined the effect of probability assessment on subsequent test selection, resource use, costs, or clinical outcomes.

In conclusion, this vignette-based study showed that there was limited agreement between clinicians and models about the characterization of lung nodules and that accuracy was only fair to good for each of them in isolation. Additional studies are needed to clarify the incremental value of prediction models above and beyond clinical judgment and to examine whether they improve care and outcomes for patients with pulmonary nodules. ■

Author disclosures are available with the text of this article at www.atsjournals.org.

Acknowledgment: The authors thank In-Lu Amy Liu, M.S., for assistance with statistical analyses.

References

- Burt JR, Iribarren C, Fair JM, Norton LC, Mahbouba M, Rubin GD, Hlatky MA, Go AS, Fortmann SP, Atherosclerotic Disease VF, et al.; Atherosclerotic Disease, Vascular Function, and Genetic Epidemiology (ADVANCE) Study. Incidental findings on cardiac multidetector row computed tomography among healthy older adults: prevalence and clinical correlates. *Arch Intern Med* 2008; 168:756–761.
- Aberle DR, Adams AM, Berg CD, Black WC, Clapp JD, Fagerstrom RM, Gareen IF, Gatsonis C, Marcus PM, Sicks JD; National Lung Screening Trial Research Team. Reduced lung-cancer mortality with low-dose computed tomographic screening. *N Engl J Med* 2011; 365:395–409.
- Summary Health Statistics for U.S. Adults. National Health Interview Survey, 2011. Vital and Health Statistics Series 10, No. 256. Hyattsville, MD: U.S. Department of Health and Human Services; 2011.
- Swensen SJ, Silverstein MD, Ilstrup DM, Schleck CD, Edell ES. The probability of malignancy in solitary pulmonary nodules. Application to small radiologically indeterminate nodules. *Arch Intern Med* 1997; 157:849–855.
- Schultz EM, Sanders GD, Trotter PR, Patz EF Jr, Silvestri GA, Owens DK, Gould MK. Validation of two models to estimate the probability of malignancy in patients with solitary pulmonary nodules. *Thorax* 2008;63:335–341.
- Isbell JM, Deppen S, Putnam JB Jr, Nesbitt JC, Lambright ES, Dawes A, Massion PP, Speroff T, Jones DR, Grogan EL. Existing general population models inaccurately predict lung cancer risk in patients referred for surgical evaluation. *Ann Thorac Surg* 2011;91:227–233. [Discussion, p. 233.]
- Herder GJ, van Tinteren H, Golding RP, Kostense PJ, Comans EF, Smit EF, Hoekstra OS. Clinical prediction model to characterize pulmonary nodules: validation and added value of 18F-fluorodeoxyglucose positron emission tomography. *Chest* 2005; 128:2490–2496.
- Gould MK, Ananth L, Barnett PG, Veterans Affairs SCSG; Veterans Affairs SNAP Cooperative Study Group. A clinical model to estimate the pretest probability of lung cancer in patients with solitary pulmonary nodules. *Chest* 2007;131:383–388.
- Swensen SJ, Silverstein MD, Edell ES, Trastek VF, Aughenbaugh GL, Ilstrup DM, Schleck CD. Solitary pulmonary nodules: clinical prediction model versus physicians. *Mayo Clin Proc* 1999;74:319–329.
- Carey TS, Garrett J. Patterns of ordering diagnostic tests for patients with acute low back pain. The North Carolina Back Pain Project. *Ann Intern Med* 1996;125:807–814.
- Peabody JW, Luck J, Glassman P, Dresselhaus TR, Lee M. Comparison of vignettes, standardized patients, and chart abstraction: a prospective validation study of 3 methods for measuring quality. *JAMA* 2000;283:1715–1722.
- Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986;1: 307–310.

- 13 Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol Bull* 1968;70:213–220.
- 14 DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 1988;44:837–845.
- 15 Shiraishi J, Li Q, Appelbaum D, Doi K. Computer-aided diagnosis and artificial intelligence in clinical imaging. *Semin Nucl Med* 2011;41:449–462.
- 16 Chen H, Wang XH, Ma DQ, Ma BR. Neural network-based computer-aided diagnosis in distinguishing malignant from benign solitary pulmonary nodules by computed tomography. *Chin Med J (Engl)* 2007;120:1211–1215.
- 17 Yonemori K, Tateishi U, Uno H, Yonemori Y, Tsuta K, Takeuchi M, Matsuno Y, Fujiwara Y, Asamura H, Kusumoto M. Development and validation of diagnostic prediction model for solitary pulmonary nodules. *Respirology* 2007;12:856–862.