Journal of Surgical Oncology 2013;108:304–311

The Contribution of Risk Prediction Models to Early Detection of Lung Cancer

JOHN K. FIELD, PhD, FRCPath,[1]* YING CHEN, PhD,[1] MICHAEL W. MARCUS, PhD,[1] FIONA E. MCRONALD, PhD,[1] OLAIDE Y. RAJI, PhD,[1] AND STEPHEN W. DUFFY, MSc[2]

[1]Roy Castle Lung Cancer Research Programme, Department of Molecular and Clinical Cancer Medicine, The University of Liverpool Cancer Research Centre, Liverpool, UK

[2]Wolfson Institute of Preventive Medicine, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, London, UK

Low-dose computed tomography screening is a strategy for early diagnosis of lung cancer. The success of such screening will be dependent upon identifying populations at sufficient risk in order to maximise the benefit-to-harm ratio of the intervention. To facilitate this, the lung cancer risk prediction community has established several risk models with good predictive performance. This review focuses on current progress in risk modelling for lung cancer prediction, with some views on future development.
J. Surg. Oncol. 2013;108:304–311. © 2013 Wiley Periodicals, Inc.

KEY WORDS: lung neoplasms; statistical model; risk assessment; screening; early diagnosis

LUNG CANCER

Lung cancer is the most common malignancy worldwide. It is estimated that 1.35 million people are diagnosed with this disease every year (12% of all cancer diagnoses), and there are approximately 1.18 million lung cancer deaths per annum (18% of all cancer deaths) [1]. Despite recent advances in treatment, lung cancer remains incurable in the majority of cases, since the disease is often diagnosed at an advanced stage, when surgical resection is unlikely to be an option. With conventional cytotoxic agents, long-term survival is rarely achieved in clinically advanced tumour stages [2]. The 5-year survival rate for all stages of lung cancer ranges from 6% (UK) to 15% (USA), in contrast to 70% 5-year survival for stage IA tumours. This suggests that earlier diagnosis and treatment of lung cancer would vastly improve outcome and reduce mortality [3,4].

SCREENING FOR LUNG CANCER USING LOW-DOSE COMPUTED TOMOGRAPHY

Low-dose computed tomography (LDCT) screening is a strategy for early diagnosis of lung cancer. In practice, the success of lung cancer screening will be dependent on successfully identifying a sufficiently high proportion of early-stage cases from the population with predicted high risk. The first large-scale early detection trial was the National Lung Screening Trial (NLST), a USA-based CT screening trial (with three LDCT screening rounds) of over 50,000 heavy-smokers aged 55–74. This trial was finished early as it showed a statistically significant 20.3% relative reduction in lung cancer mortality in the LDCT arm [5]. However, as cost-effectiveness data have not yet been released by NLST, it is not yet clear how LDCT screening compares economically with other lung cancer prevention/early detection strategies. Several European CT-based randomised screening trials are currently underway or have recently reported their preliminary findings; the largest of these is the Dutch–Belgian NELSON (NEderlands Leuvens Screening ONderzoek) trial [6]. A number of smaller studies include the ItaLung and Dante Trials in Italy [7,8], the French pilot study, Depiscan [9] and the Danish lung cancer screening trial [10]. The United Kingdom Lung Cancer Screening Trial (UKLS) is in the process of completing recruitment of 4,000 subjects for its pilot phase. Individuals in this Randomised Controlled Trial receive either one round of LDCT, or no screening [11]. UKLS differs in several respects from other lung cancer screening trials: firstly by employing a single screen design, and secondly due to its selection of eligible participants via a formal risk prediction model rather than by smoking history and age alone. Calculation of individuals' lung cancer risk is achieved via the Liverpool Lung Project (LLP) risk prediction model, which incorporates factors such as family history of lung cancer, smoking and medical history and occupational exposure to asbestos [12]. People with a high LLP risk of lung cancer ($\geq$5% predicted absolute risk within incident 5-year period) are considered at sufficient risk to be included in the UKLS trial. The LLP model is a robust algorithm that has been validated on two international case–control populations (Harvard and EUELC) and one independent cohort (LLPPC) [13]. Results from UKLS should aid in defining the most appropriate participants for future lung CT screening, and should enable assessment of the optimum number of screening rounds, by comparing outcomes from UKLS with those from other multi-screening round trials. This latter point is important both in terms of minimising potential harms by CT-associated radiation exposure, and maximising cost-effectiveness of CT screening [14]. If a lung cancer mortality benefit is confirmed at both clinical and cost-effectiveness levels, there would be a strong argument for healthcare providers to offer thoracic LDCT screening to suitable individuals on a service-basis.

Grant sponsor: Roy Castle Lung Cancer Foundation

No potential conflicts of interest were disclosed.

*Correspondence to: Professor John K. Field, Roy Castle Lung Cancer Research Programme, Department of Molecular and Clinical Cancer Medicine, The University of Liverpool Cancer Research Centre, The University of Liverpool, 200 London Road, Liverpool L3 9TA, UK. Fax: 44-151-794-8989. E-mail: j.k.field@liv.ac.uk

Received 06 December 2012; Accepted 28 June 2013

DOI 10.1002/jso.23384

Published online 29 August 2013 in Wiley Online Library (wileyonlinelibrary.com).

© 2013 Wiley Periodicals, Inc.

IDENTIFICATION OF HIGH-RISK INDIVIDUALS

The term 'risk' can be thought of as the inherent chance among healthy individuals of developing lung cancer during a prospective period of time. Risk factors are those characteristics of an individual's biology or lifestyle that could affect their probability of developing lung cancer. In order to maximise the benefit-to-harm ratio of lung cancer screening (both clinically and financially), it is important to identify which individuals are at sufficiently high risk of the disease, and to target screening to these people. This process involves identification of risk factors, quantitative summary of overall risk and selection of a suitable cut-off value for CT screening. Since lung cancer is mainly attributable to cigarette smoking and occurs amongst elderly populations [15], the selection criteria for eligible participants in the current two largest randomised controlled trials (NLST and NELSON) were based on smoking history and age [5,6]. However, the risk of lung cancer is also influenced by other factors, including environmental exposures (e.g., cooking fumes, ionising radiation, radon gas, asbestos, occupational exposures to carcinogens), current or previous medical conditions (e.g., COPD, other cancer, pneumonia), family history of lung cancer and genetic susceptibility (e.g., SNPs at 5p15, 6p21, 15q25) [16,17]. Approximately 10–20% of lung cancer patients have no history of tobacco smoking [17]. It has been observed that, as the smoking rate declines, non-tobacco risk factors take on a greater proportional role in risk prediction [18]. Furthermore, inclusion of non-tobacco-related risk factors for screening criteria is important in terms of social equality, particularly in countries such as the UK where the health care system is government-funded. It would be politically problematic to offer publicly-funded medical interventions solely to heavy smokers, when non-/light smokers (although in a smaller proportion) may be at equally high risk due to other environmental and genetic factors and their interactions. People employed in certain occupations may also need to be offered screening.

MAJOR RISK PREDICTION MODELS

Accurate selection of high-risk individuals for lung cancer screening requires robust methods for risk prediction. The discriminative performance of a risk model depends not only on the identification of individual risk factors, but also on how these risk variables operate in the presence/absence of other variables, how accurately these factors can be measured and the appropriateness of the population and statistical techniques used for modelling [19]. However, the main practical application of a risk prediction model is its use by non-specialists for selection of suitable high-risk people for lung cancer screening/intervention. Thus, in addition to being technically detailed and accurate, a risk model needs to be sufficiently user-friendly to be applied in the general population and/or primary care setting. In practical terms, this means that the risk variables should be straightforward to elicit and the algorithm should be simple to run.

An early risk index for lung cancer was developed by Colditz et al. [20]: this was based on consensus data from literature reviews. Parameters included were smoking characteristics, family history, air pollution and fruit and vegetable intake. However, the model is considered a general guide, rather than providing precise estimation of risk at an individual level. In 2002, an index table (based on age and smoking history) was developed for use in identifying risk groups for inclusion in randomised controlled trials for lung cancer CT screening [21]. This index stratifies the population into discrete subsets with respect to age and duration of smoking, forming a 2-way table with a risk score for each cell. The method has the advantage of ease of use and interpretation; however it will inevitably fail to identify individuals at high risk due to factors other than age and smoking.

Bach and colleagues used prospective cohort data for smokers in the carotene and retinol efficacy trial (CARET) to develop a method to distinguish the risk variations in lung cancer amongst smokers; this was the first time modern statistical modelling was incorporated into a risk model for lung cancer [22]. In their model, the predictors included age, gender, smoking history and exposure to asbestos. The model was externally validated in the alpha-tocopherol, beta-carotene cancer prevention (ATBC) Study control arm [23]. Spitz et al. [24] used a case–control population to develop their prediction tool. Models were built separately for never-, former- and current-smokers and the predictor variables were expanded significantly: (these included exposure to environmental tobacco smoke, family history of lung/smoking-related cancer, dust and asbestos exposure, history of respiratory diseases and smoking characteristics). The risk prediction model produced by the LLP was based on epidemiological data from lung cancer cases and age- and sex-matched population-based controls [12]. This multivariate model included smoking duration, history of pneumonia, occupational exposure to asbestos, prior diagnosis of malignant tumour (other than lung) and family history of lung cancer as risk predictors and was combined with age- and sex-standardised lung cancer incidence data to calculate an estimation of absolute risk for each individual. The model has been externally validated in three independent populations from the UK, Europe and North America [13]. More recently, the likely utility of the LLP model for stratifying people for lung cancer LDCT screening has also been demonstrated [13] using decision-theory methods (including relative utility curve analysis and decision curve analysis) [25–28]. These analyses have assessed the risk model's performance from a clinical perspective rather than using a pure statistical measure.

The discriminatory power and accuracy of the above lung cancer risk models were compared using a case–control population recruited at the Harvard School of Public Health and Massachusetts General Hospital. Based on this independent dataset, it was shown 1) that the positive predictive values were highest with the Spitz model, whereas the negative predictive values were highest with the LLP model; and 2) that the Spitz and Bach models had lower sensitivity but better specificity compared to the LLP model [29]. The clinical utility of the Spitz, Bach and LLP models is presented in Figure 1 (adapted from Ref [29]). The LLP model was much better at identifying lung cancer: using 2.5% (absolute risk within 5 years) as the minimum cut-off value, the model identified 66.7% of total cases with 33.4% of false-positive results, whereas at a higher cut-off point for example, 7.5%, the model was still able to identify 31.2% of total cases while limiting the rate of false-positive results to 7.7%. In the Spitz and Bach models, although results were better in terms of the false-positive rate, the percentage of cases identified was considerably lower based on the reported cut-off values (see Fig. 1). The relatively low sensitivity of the Bach and Spitz models is possibly due to the former not including risk factors other than smoking history and asbestos exposure and the latter under-estimating the overall impact of smoking by using smoking status as a matching variable. The relatively low specificity of the LLP model could be attributed to it not distinguishing between ex- and current-smokers and using smoking duration as the only smoking parameter. It is known that, in ex-smokers, smoking cessation period (time since quitting smoking) is significantly associated with reduction in lung cancer risk [30]. Nevertheless, a large and representative population is required for better assessment of the performance of these risk prediction models.

The prostate, lung, colorectal and ovarian cancer screening trial (PLCO) used prospective data to build their risk prediction model consisting of two versions: 1) for the general population and 2) for a sub-cohort of ever-smokers [31,32]. In addition to the population-based design and a large sample size, a strength of their models is the use of spline functions, a statistical method used to provide flexibility to modelling non-linear effects in regression models. In their study, they included spline effects of age, pack-years, smoking duration and smoking quit time; and also incorporated level of education, body mass index, COPD, family history of lung cancer, chest X-ray in the past

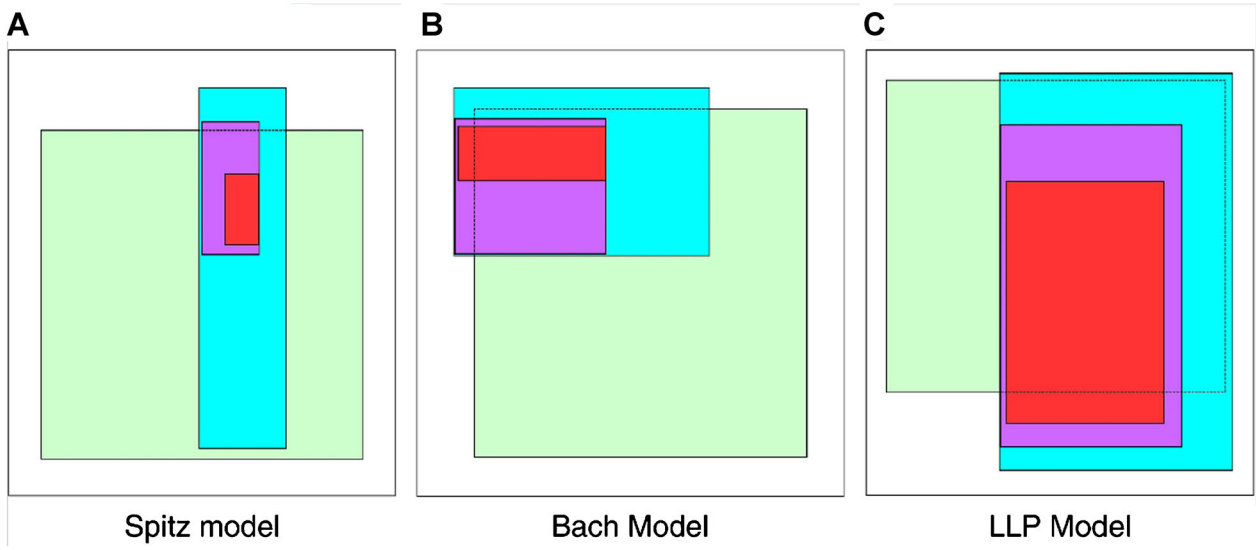

Fig. 1. Clinical utility of the Spitz, Bach and LLP models. Scaled rectangle diagrams for (**A**) the Spitz, (**B**) Bach and (**C**) LLP risk models at defined levels of lung cancer risk. For each model: white represents all controls (with <2.5% risk according to the risk model) and green represents all cases (with <2.5% risk according to the risk model). Blue represents all individuals who the risk model predicts have between 2.5% and 5.0% risk. Purple represents all individuals who the risk model predicts have between 5.0% and 7.5% risk. Red represents all individuals who the risk model predicts have at least 7.5% risk. Image originally published in Ref. [29]. Copyright © 2010 Cancer Research UK.

3 years and current smoking status [32]. The approach of applying flexible modelling to continuous variables such as smoking duration is statistically beneficial, as it ensures good resolution of continuous predictors and avoids loss of information associated with categorisation. However, risk prediction models with such complicated modelling procedures can be cumbersome to apply. Furthermore, the PLCO model placed more emphasis on the results of a case–control dataset using logistic regression analysis (rather than the Cox proportional model, a time-to-event analysis). Under the design of a prospective cohort recruiting healthy individuals, the study is likely to provide valuable information on all the necessary covariates to predict individuals' absolute risk based on standard survival methods.

Tammemagi et al. [33] have modified and updated the smoker-only version of the PLCO model to make it directly applicable to the NLST data. The amended model, $PLCO_{M2012}$, included age, race/ethnic group, education, body mass index, COPD, personal history of cancer, family history of lung cancer and smoking status (current vs. former), intensity, duration and quit time as predictive variables. It used a simplified evaluation for non-linear effects, applied logistic regression modelling to calculate the probability of developing lung cancer over a period of 6 years and reported a ROC-AUC = 0.803 and 0.797 in the training and validation sets, respectively [33]. It was shown that the use of a risk prediction model (the $PLCO_{M2012}$ model, ROC-AUC = 0.701 for the NLST participants) was more sensitive than the NLST criteria for lung cancer detection in LDCT screening (positive predictive value 4.0% vs. 3.4%) [33]. This suggests that the current USA lung cancer screening guidelines may need to reconsider the entry criteria for LDCT screening, focussing in particular on the feasibility of utilising a comprehensive risk model.

Hoggart et al. [34] used prospective data from the European prospective investigation into cancer and nutrition (EPIC) cohort to build their prediction model for lung cancer. They applied parametric survival statistics to model the time-dependent nature of the prospective data. However, the final model captures only the risks from age and smoking history and is only applicable to ever-smokers, although other risk variables were also considered during the development phase. In comparison to the Bach model (the earlier smoking-based model), EPIC outperformed Bach in their sampled subjects. The group's attempt to develop a risk model for the never-smoker population was unsuccessful; this could be explained by the amount of missing data and the measurement of some exposures in EPIC [34]. This latter point reflects the difficulties inherent in attempting to develop accurate lung cancer risk prediction models from large cohort data not initially designed for this purpose. This is problematic with regards to recording and measurement of both essential and potential risk predictors of lung cancer. Table I summarizes the main characteristics of the major lung cancer risk prediction models.

RISK PREDICTION MODELS FOR NON-CAUCASIAN POPULATIONS

It is accepted that non-Caucasians share similar lung cancer risk factors with their Caucasian counterparts. However, the level of risk may differ depending on the type and intensity of exposure. Furthermore, individuals of certain ethnic groups may encounter unique exposures. Existing lung cancer risk prediction models were developed from Caucasian populations, and therefore may not be appropriate for predicting risk in non-Caucasian populations.

Spitz's group carried out a comprehensive epidemiological analysis of risk factors for lung cancer in African–Americans, and observed similarities and differences in distribution of risk factors compared to Caucasians within the same geographical region [35]. They hence developed a risk model specific for African–Americans. The model exhibited good discrimination (ROC-AUC = 0.75) for internal data and moderate discrimination (ROC-AUC = 0.63) for the external data group. For risk prediction in African–Americans, this model significantly outperformed a similar risk model developed primarily for Caucasians by the same research group [35]. This demonstrates the importance of developing ethnic-specific models.

Most recently, a Korean research group developed a prediction model using data collected from health examination on a very large population-based, male cohort. They were the first to include physical activity and fasting glucose level, in addition to variables including age, body mass index, and smoking status, intensity and age at initiation. Using Cox

TABLE I. The Main Characteristics of Lung Cancer Risk Prediction Models

Study factor	Bach	Spitz	Model LLP	PLCO	EPIC
Study design	Prospective cohort	Case–control[a]	Case–control[b]	Prospective cohort	Prospective cohort
Characteristics of population for modelling	14,254 Heavy current- or former-smokers and 3,918 asbestos-exposed current- or former-smokers (subjects in CARET). Age 45–69 at baseline [22]	1851 cases and 2001 age-, sex-, ethnicity- and smoking status-matched control (75 and 25% as training and validation sets, respectively). Mean ages (62 ± 11) and (60 ± 11) for cases and controls, respectively [24]	579 cases and 1157 age- and sex-matched population-based controls. Mean age (66 ± 9) for both cases and controls [12]	70,962 cancer-free population-based individuals (subjects in the control arm of PLCO). Age 55–74 at recruitment [32]	169,035 former- and current-smokers from the general population (subjects in EPIC). A random sample of 90% of each of the cases and controls as the training set, and the remaining 10% as validation set [34]
Risk factors included in model	Age, sex, asbestos exposure, and smoking duration, duration of abstinence (if applicable) and average amount per day (while smoking) [22]	Age, sex, exposure to dust, asbestos and environmental tobacco smoke, history of respiratory diseases, family history of lung and other smoking-related cancers, and age at smoking cessation (if applicable) and smoking pack-year [24]	Age, sex, history of pneumonia, asbestos exposure, malignant tumour, family history of lung cancer, and smoking duration [12]	Age, socioeconomic status (education), body mass index, family history of lung cancer, COPD, recent chest X-ray and smoking status, pack-year, duration and time since quitting smoking (if applicable) [32]	Age, smoking intensity (measured by average number of cigarettes smoked per day), age started smoking and smoking duration [34]
Risk estimation described	Absolute risk	Absolute risk	Absolute risk	Absolute risk	Absolute risk
Calibration in prospective cohort	Slightly under-estimated: E/O = 0.81 for 1-year prediction; E/O = 0.89 for 10-year prediction [23]	NA	Slightly under-estimated: E/O = 0.76 for overall subjects; E/O = 0.89 for subjects with the highest quartile of risk scores where screening might be initiated [13]	Excellent: E/O = 0.99 for all subjects; E/O = 0.98 for ever-smokers only [32]	Good: The points, according to estimated and observed deciles of risk, lie around the 45-degree line, indicating good calibration (exact value was not shown) [34]
Discriminatory power in modelling population	0.72 [22]	0.57, 0.63 and 0.58 for never-, former- and current-smokers, respectively (reported based on the validation set) [24]	0.70 [12]	0.86 and 0.81 for all subjects and ever-smokers only, respectively [32]	0.84 [34]
Discriminatory power in populations for external validation	0.69 in the placebo arm of ATBC study [23]; 0.66 in a case–control set from Harvard [29]	0.69 in a case–control set from Harvard [29]	0.67 in EUELC case–control set [13]; 0.69 in a case–control set from Harvard [29]; 0.76 in the second case–control set from Harvard [13]; 0.82 in the LLPPC cohort [13]	0.84 and 0.78 for all subjects and ever-smokers only, respectively in the PLCO intervention arm[c] [32]	NA
Expanded/modified model	NA	Incorporation of markers of DNA repair capacity [43]	Incorporation of SEZ6L SNP [44]	Incorporation of pulmonary function test and SDIC [42]; The $PLCO_{M2012}$ model [33]	NA
Proposed application of model	Identifying high-risk groups for prevention and screening	Identifying high-risk individuals for increased screening surveillance	Identifying high-risk individuals for screening and early detection	Identifying high-risk groups for prevention trials and screening programs	Identifying high-risk groups for monitoring, screening and prevention, and encouraging to quit smoking (if applicable)

[a]In combination with sex- and age-standardised lung cancer incidence rates, and with adjustment for smoking status-specific incidence rates.
[b]In combination with sex- and age-standardised lung cancer incidence rates.
[c]As PLCO is a randomised screening trial, validation in the opposite arm of trial may not be fully considered as external validation; E/O, estimated cases (probabilities)/observed cases (probabilities); NA, not applicable; SNP, single-nucleotide polymorphism; SDIC, sputum DNA image cytometry.

proportional hazards modelling, the model showed impressive performance in discrimination (C statistic = 0.871 in the validation set), although the dataset seemed to limit the researchers to include some more traditional predictive factors into their model [36]. Given the significant proportion of non-smoking lung cancer in Asian females, a successful model for women will be more promising and is awaited.

The LLP is currently involved in recruiting participants for developing a lung risk prediction tool for people of Chinese ethnicity. This is particularly important given that only 18% of lung cancer in Chinese females can be attributed to tobacco smoking [17]. Some unique risk factors such as indoor coal burning, cooking fumes and infections may play important roles. In addition, differences in genetic susceptibility to lung cancer may also contribute significantly [37]. It is also known that the distribution of histological types of lung cancer varies according to ethnic group; this provides further evidence for the existence of distinct aetiologies.

CLINICAL EXPANSION OF RISK PREDICTION MODEL

'Real-time' clinical information is superior to self-reported responses to questionnaire data or medical records, more relevant to an individual's current status and would alleviate the effect of recall bias on estimated risks. Specific respiratory risk factors (e.g., COPD [38]) have been highlighted as relevant in the development of lung cancer. Objective measurement of reduction in pulmonary function (e.g., detecting COPD based on FEV_1/FVC and/or FEV_1%) has been speculated to be a helpful measurement in improving risk assessment for lung cancer [39]. Several studies have also reported that individuals with a diagnosis of moderate or worse sputum cytological atypia may be at high risk for future development of lung cancer [40,41]. Prindiville et al. [41] reported that sputum cytology can identify those at substantially higher risk for lung cancer, with a cumulative incidence of 10% in 3 years and 20% in 6 years. Incorporation of pulmonary function (FEV_1%) and sputum DNA image cytometry (SDIC) into risk prediction modelling for lung cancer was pioneered by the developers of the PLCO model. Using individuals (ever-smokers, age >40 and pack-year ≥20 at baseline) enrolled and followed prospectively for the development of lung cancer in trials at the British Columbia Cancer Agency, they demonstrated that pulmonary function measurement data modestly improved the PLCO model in predicting lung cancer (ROC-AUC improvement = 0.049) [42]. However, the effect of SDIC was shown to be negligible [42].

BIOLOGICAL EXPANSION OF RISK PREDICTION MODEL

Much of lung cancer risk is explained by known epidemiological factors. Lung cancer risk prediction models have therefore tended to concentrate on these parameters (age, smoking, history of respiratory diseases, etc.) However, the current models are not perfect in terms of their discriminatory performance. In particular, it has been recognised that they have poor predictive accuracy in patients with early onset of lung cancer [29]. The occurrence of lung cancer at a young age is more likely to be due to a genetic predisposition. There is potential to improve prediction of inter-individual risk by incorporating factors/markers associated with genetic variation. Initial attempts included 1) the incorporation of two markers of DNA repair capacity into the Spitz model [43] and 2) the incorporation of a genetic susceptibility polymorphism (a SNP in SEZ6L) into the LLP model [44]. Both incorporations enhanced prediction performance. However, the improvement was only modest in absolute terms, compared to the baseline models (ROC-AUC improvement ≤0.05) [43,44].

The recent advances in molecular and genetic epidemiological studies have expanded the potential for genomics-based prediction. Lung cancer-associated SNPs across the whole human genome have been discovered through large-scale genome-wide association studies (GWAS). Although individual polymorphisms confer only a modest effect upon risk (typical allelic odds ratio <1.3) [45], it is speculated that, in combination, they may provide significant risk stratification. It was shown that, on average, 80 (independent) common genetic variants with odds ratios of 1.25 each were sufficient to build a model useful for detection of high-risk people with a ROC-AUC >0.80 [46]. However, major lung cancer susceptibility SNPs are concentrated at only three loci [17]. Within each locus, common SNPs are likely in linkage disequilibrium (thus may be not independent of each other). It is hoped that large meta-analyses of existing and future GWAS will uncover additional susceptibility loci/SNPs. However, it is also likely that loci/SNPs that are discovered in the future will tend to have a smaller effect size. Based on the results for selected cancers (although lung cancer is not included), Park et al. [45] recently speculated that the utility of susceptibility SNPs (including those both discovered and yet to be discovered) for risk prediction is limited. However, there is potential for improvement of existing and future models by introducing novel statistical approaches that incorporate gene–gene/environment interactions or pathway-based analysis [24,46].

In addition to DNA polymorphisms, other numerous bio-markers have been identified and validated in case–control studies, using specimens such as serum, plasma, bronchial lavage, induced sputum, bronchial washing, and tissue [47,48]: it is anticipated that many more will also be discovered. However, due to fewer opportunities to access biomaterials from large prospective cohort studies, most current markers are of diagnostic utility (e.g., diagnostic efficiency [48]), and their value in risk prediction has not been assessed. For a biomarker to be a promising candidate for risk prediction, it would need to be measureable in non-invasive samples (e.g., saliva, mouthwash, buccal/nasal brushings, sputum, breath test) for clinical utility in the general population/primary care settings.

Discovery of genetic and epigenetic predictors of lung cancer assists the improvement of risk prediction models. However, the incorporation of these markers into risk models (alongside major epidemiologic risk factors) will be a major challenge in terms of cost and feasibility. Improvement of discriminative performance by a single biomarker may be limited [42–44]. As lung cancer is a heterogeneous disease, certain biomarkers associated with particular pathways may only add predictive power for a small proportion of cases. Statistical modelling (for example, including a number of SNPs and several epigenetic markers as well as epidemiologic variables), with consideration of interactions on different levels, will be a complex task. Furthermore, models involving biomarkers require very careful validation for two reasons: 1) issues with over-fitting when developing complex models with a large number of markers, and 2) inter-laboratory variation in assays used to quantify these markers [49]. Finally, any test requiring bio-materials is costly both in terms of time and money, and needs a certain level of technical expertise in carrying out assays. These factors may restrict the use of biomarkers for the purposes of population-level risk prediction. However, the step-wise lung cancer risk model decision cascade, as described previously by the LLP group, may be an applicable strategy (see Fig. 2, adapted from Ref [47]). This separates the first-line risk assessment (using only easily obtainable epidemiological and clinical factors) from further involvement requiring specimens and bio-assays. The latter are only employed in cases where it is necessary to carry out finer discrimination.

APPLICATION AND MISCELLANEOUS ISSUES

In the UK, the cancer reform strategy recently established a National awareness and early diagnosis initiative (NAEDI) with a view to promote early diagnosis of cancer [50]. Lung cancer risk prediction models are primarily used to discriminate between individuals of

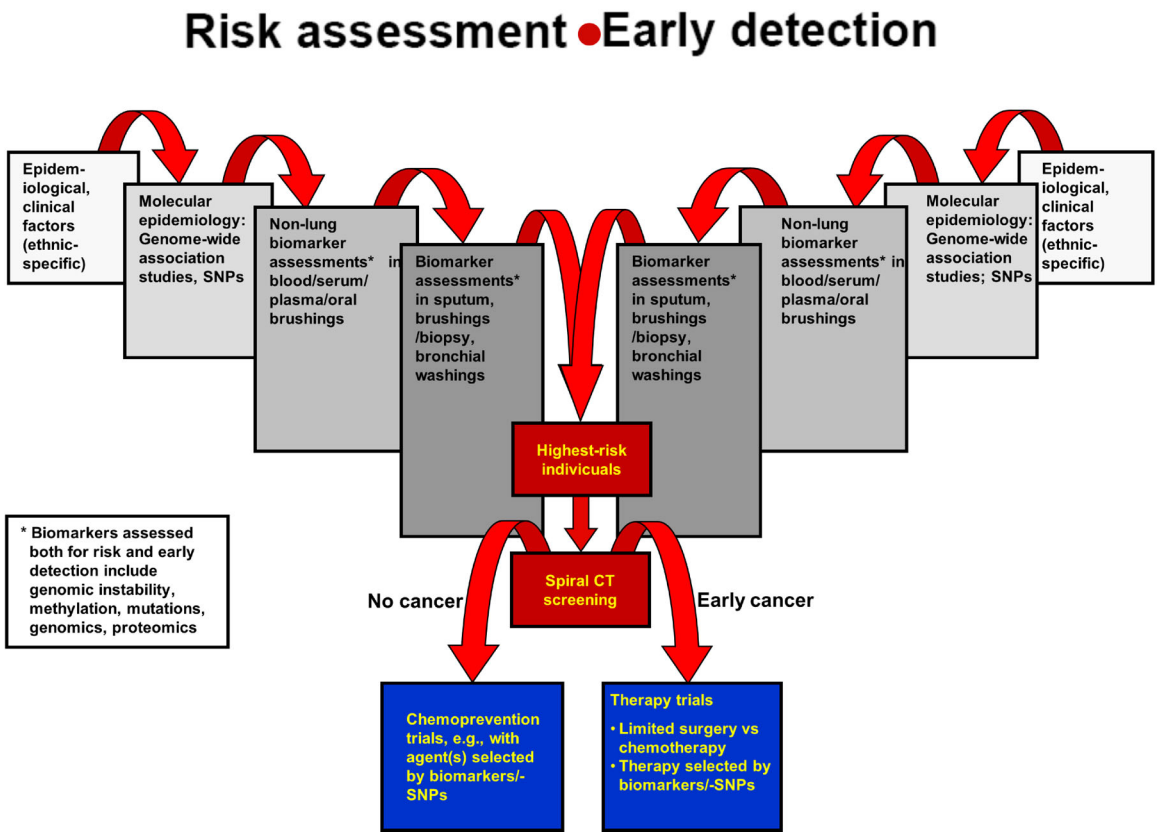

Fig. 2. Lung cancer risk model decision cascade. Mirror images of shared aspects of the cascades of lung cancer risk assessment (left) and early detection (right) are shown. These cascades move from clinical and epidemiologic assessments to molecular epidemiologic assessments to biomarker assessments in non-lung samples and finally to biomarker assessments in lung samples. They merge in the middle with the identification of the highest-risk individuals who need computed tomography (CT) screening. Individuals with a positive CT scan will be entered into clinical workup and treatment protocols. Those with a negative CT scan are clearly at a high risk and should be considered for prevention research studies. SNP, singlenucleotide polymorphism. Image originally published in Ref. [47]. Copyright © 2008 American Association of Cancer Research.

different risks, and to identify individuals who are at risk of developing lung cancer prior to developing symptoms. High-risk people could undergo a programme of screening surveillance and may also consider chemoprevention. However, a single average risk threshold (risk cut-off value) for further investigation/prevention (e.g., LDCT screening, chemo-intervention) for a population has not yet been established. To identify such a cut-off, adequate data on harms, benefits and actual outcomes in randomised controlled trials is required; this will involve careful evaluation of clinical, financial and many other aspects. Unlike cardiovascular disease, for which a 10-year risk of 20% has been agreed to classify individuals as high-risk [51], no consensus is available in cancer screening [52]. Prospective updates from NLST, NELSON, UKLS and other trials for lung cancer screening are expected to help the lung cancer research community to standardise the risk threshold at which to recommend population-based CT screening for lung cancer [11].

Risk prediction tools may be a useful adjunct for planning, designing and conducting smaller, more powerful and 'smarter' prevention trials by enriching the number of observed events [22]. The incorporation of the LLP risk model into the design of a CT screening trial and a population screening intervention programme has been discussed. It has been demonstrated that increasing the minimum 5-year absolute risk criterion of individuals to be selected in a trial from 1.5–2.5% reduces the required sample size by approximately one-third [18]. However, risk models may over-emphasise a particular group of risk factors. As a result, the enriched disease events may be more associated with particular aetiological mechanisms, and thus not be entirely representative of the general population. It is therefore possible that any outcomes of further trials (such as those for drug interventions), may be applicable only to the screened group and less so to the general population. Thus, the development and improvement of risk models needs to consider as many major risk factors as possible.

In addition to potential clinical utility, risk prediction tools based on scientific modelling are also of value in refining understanding of risk factors. This can lead to educational initiatives in terms of awareness of relevant risks and how to avoid them, especially when this information is made available on a publicly accessible platform (e.g., the internet). In this way, there may be social benefits; for example increasing the successful rate of smoking cessation, and public awareness of occupational safety. However, the actual impact of this has yet to be evaluated. Most recently, an online version of the LLP lung risk calculator with a user-friendly interface, 'MyLungRisk', has been made available to the public for self-assessment at www.mylungrisk.org.

CONCLUSION

Lung cancer kills the vast majority of people that it afflicts, mainly due to late presentation of the disease. This has driven the need for screening programmes, in order to identify cancer cases as early as possible and prior to symptomatic presentation. The success of lung cancer screening will be dependent on successfully identifying a sufficiently high proportion of early-stage cases from the population. To

target high-risk individuals, the lung cancer risk prediction community has established several risk models with good predictive performance. These models may serve as a useful adjunct to both investigators and patients, perhaps playing a role similar to risk prediction models for colon cancer and breast cancer. Since some major risk factors for lung cancer are avoidable, risk prediction models for lung cancer also have the potential to serve as educational tools.

ACKNOWLEDGMENT

Y.C. is funded by the Roy Castle Lung Cancer Foundation.

REFERENCES

1. Parkin DM, Bray F, Ferlay J, et al.: Global cancer statistics, 2002. CA Cancer J Clin 2005;55:74–108.
2. Schiller JH, Harrington D, Belani CP, et al.: Eastern cooperative oncology group: Comparison of four chemotherapy regimens for advanced non-small-cell lung cancer. N Engl J Med 2002;346: 92–98.
3. van Rens MT, de la Riviere AB, Elbers HR, et al.: Prognostic assessment of 2,361 patients who underwent pulmonary resection for non-small cell lung cancer, stage I, II, IIIA. Chest 2000;117: 374–379.
4. Ganti AK, Mulshine JL: Lung cancer screening: Panacea or pipe dream? Ann Oncol 2005;16Suppl 2:ii 215–219.
5. The National Lung Screening Trial Research Team. Reduced lung-cancer mortality with low-dose computed tomographic screening. N Engl J Med 2011;365:395–409.
6. van Iersel CA, de Koning HJ, Draisma G, et al.: Risk-based selection from the general population in a screening trial: Selection criteria, recruitment and power for the Dutch-Belgian randomised lung cancer multi-slice CT screening trial (NELSON). Int J Cancer 2007;120:868–874.
7. Lopes Pegne A, Picozzi G, Mascalchi M: Design, recruitment, and baseline results of the ITALUNG trial for lung cancer screening with low-dose CT. Lung Cancer 2009;64:34–40.
8. Infante M, Cavuto S, Lutman FR, et al. A randomized study of lung cancer screening with spiral computed tomography: Three-year results from the DANTE trial. Am J Respir Crit Care Med 2009;180:445–453.
9. Blanchon T, Brechot JM, Grenier PA, et al.: Baseline results of the Depiscan study: A French randomized pilot trial of lung cancer screening comparing low dose CT scan (LDCT) and chest X-ray (CXR). Lung Cancer 2007;58:50–58.
10. Pedersen J, Ashraf H, Dirksen A, et al.: The Danish randomized lung cancer CT screening trial–overal design and results of the prevalence round. J Thorac Oncol 2009;5:608–614.
11. Baldwin DR, Duffy SW, Wald NJ, et al.: UK lung screen (UKLS) nodule management protocol: Modelling of a single screen randomised controlled trial of low-dose CT screening for lung cancer. Thorax 2011;66:308–313.
12. Cassidy A, Myles JP, van Tongeren M, et al.: The LLP risk model: An individual risk prediction model for lung cancer. Br J Cancer 2008;98:270–276.
13. Raji OY, Duffy SW, Agbaje OF, et al.: Predictive accuracy of the liverpool lung project risk model for stratifying patients for computed tomography screening for lung cancer. Ann Intern Med 2012;157:242–250.
14. Vansteenkiste J, Dooms C, Mascaux C, et al.: Screening and early detection of lung cancer. Ann Oncol 2012;23:x320–x327.
15. Centers for Disease Control and Prevention (CDC). Annual smoking-attributable mortality, years of potential life lost, and productivity losses: United States, 1997–2001. Morb Mortal Wkly Rep 2005;54:625–628.
16. McCarthy WJ, Meza R, Jeon J, et al.: Lung cancer in never smokers: Epidemiology and risk prediction models. Risk Anal 2012;32: S69–S84.
17. Brennan P, Hainaut P, Boffetta P: Genetics of lung-cancer susceptibility. Lancet Oncol 2011;12:399–408.
18. Duffy SW, Raji OY, Agbaje OF, et al.: Use of lung cancer risk models in planning research and service programs in CT screening for lung cancer. Expert Rev Anticancer Ther 2009;9:1467–1472.
19. Cassidy A, Duffy SW, Myles JP, et al.: Lung cancer risk prediction: A tool for early detection. Int J Cancer 2006;120:1–6.
20. Colditz GA, Atwood KA, Emmons E, et al.: Harvard report on cancer prevention volume 4: Harvard cancer risk index. Risk index working group, Harvard center for cancer prevention. Cancer Causes Control 2000;11:477–488.
21. Van Klaveren RJ, de Koning HJ, Mulshine J, et al.: Lung cancer screening by spiral CT. What is the optimal target population for screening trials? Lung Cancer 2002;38:243–252.
22. Bach PB, Kattan MW, Thornquist MD, et al.: Variations in lung cancer risk among smokers. JNCI 2003;95:470–478.
23. Cronin KA, Gail MH, Zou Z, et al.: Validation of a model of lung cancer risk prediction among smokers. J Natl Cancer Inst 2006;98:637–640.
24. Spitz MR, Hong WK, Amos CI, et al.: A risk model for prediction of lung cancer. J Natl Cancer Inst 2007;99:715–726.
25. Vickers AJ, Cronin AM, Elkin EB, et al.: Extensions to decision curve analysis, a novel method for evaluating diagnostic tests, prediction models and molecular markers. BMC Med Inform Decis Mak 2008;8:53.
26. Vickers AJ, Elkin EB: Decision curve analysis: A novel method for evaluating prediction models. Med Decis Making 2006;26: 565–574.
27. Baker SG: Putting risk prediction in perspective: Relative utility curves. J Natl Cancer Inst 2009;101:1538–1542.
28. Baker SG, van Calster B, Steyerberg EW: Evaluating a new marker for risk prediction using the test tradeoff: An update. Int J Biostat 2012;8:pii: /j/ijb.2012.8.issue-1/1557-4679.1395/1557-4679.1395.xml. doi: 10.1515/1557-4679.1395. See also "http://www.ncbi.nlm.nih.gov/pubmed/22499728".
29. D'Amelio AM,Jr Cassidy A, Asomaning K, et al.: Comparison of discriminatory power and accuracy of three lung cancer risk models. Br J Cancer 2010;103:423–429.
30. Peto R, Darby S, Deo H, et al.: Smoking, smoking cessation, and lung cancer in the UK since 1950: Combination of national statistics with two case–control studies. BMJ 2000;321:323–329.
31. Gohagan JK, Prorok PC, Hayes RB, et al.: Prostate, lung, colorectal and ovarian cancer screening trial project team. The prostate, lung, colorectal and ovarian (PLCO) cancer screening trial of the National cancer institute: History, organization, and status. Control Clin Trials 2000;21:251S–272S.
32. Tammemagi CM, Pinsky PF, Caporaso NE, et al.: Lung cancer risk prediction: Prostate, lung, colorectal and ovarian cancer screening trial models and validation. J Natl Cancer Inst 2011;103:1058–1068.
33. Tammemagi MC, Katki HA, Hocking WG, et al.: Selection criteria for lung-cancer screening. N Engl J Med 2013;368:728–736.
34. Hoggart C, Brennan P, Tjonneland A, et al.: A risk model for lung cancer incidence. Cancer Prev Res 2012;5:834–846.
35. Etzel CJ, Kachroo S, Liu M, et al.: Development and validation of a lung cancer risk prediction model for African–Americans. Cancer Prev Res 2008;1:255–265.
36. Park S, Nam BH, Yang HR, et al.: Individualized risk prediction model for lung cancer in Korean men. PLoS ONE 2013;8:e54823.
37. Zhou W, Christiani DC: East meets West: Ethnic differences in epidemiology and clinical behaviours of lung cancer between East Asians and Caucasians. Clin J Cancer 2011;30:287–292.
38. Punturieri A, Szabo E, Croxton TL, et al.: Lung cancer and chronic obstructive pulmonary disease: Needs and opportunities for integrated research. J Natl Cancer Inst 2009;101:554–559.
39. Field JK, Raji OY: The potential for using risk models in future lung cancer screening trials. F1000 Med Rep 2010;2:doi:pii: 38 See also "http://www.ncbi.nlm.nih.gov/pubmed/20948847".
40. Tockman MS, Gupta PK, Myers JD, et al.: Sensitive and specific monoclonal antibody recognition of human lung cancer antigen on preserved sputum cells: A new approach to early lung cancer detection. J Clin Oncol 1988;6:1685–1693.
41. Prindiville SA, Byers T, Hirsch FR, et al.: Sputum cytological atypia as a predictor of incident lung cancer in a cohort of heavy smokers

with airflow obstruction. Cancer Epidemiol Biomarkers Prev 2003; 12:987–993.
42. Tammemagi MC, Lam SC, McWilliams AM, et al.: Incremental value of pulmonary function and sputum DNA image cytometry in lung cancer risk prediction. Cancer Prev Res 2011;4:552–561.
43. Spitz MR, Etzel CJ, Dong Q, et al.: An expanded risk prediction model for lung cancer. Cancer Prev Res 2008;1:250–254.
44. Raji OY, Agbaje OF, Duffy SW, et al.: Incorporation of a genetic factor into an epidemiologic model for prediction of individual risk of lung cancer: The liverpool lung project. Cancer Prev Res 2010;3:664–669.
45. Park JH, Gail MH, Greene MH, et al.: Potential usefulness of single nucleotide polymorphisms to identify persons at high cancer risk: An evaluation of seven common cancers. J Clin Oncol 2012;30: 2157–2162.
46. Janssens AC, Aulchenko YS, Elefante S, et al.: Predictive testing for complex diseases using multiple genes: Fact or fiction? Genet Med 2006;8:395–400.
47. Field JK: Lung cancer risk models come of age. Cancer Prev Res 2008;1:226–228.
48. Nikolaidis G, Raji OY, Markopoulou S, et al.: DNA methylation biomarkers off improved diagnostic efficiency in lung cancer. Cancer Res 2012;72:5692–5701.
49. Taylor JMG, Ankerst DP, Andridge RR: Validation of biomarker-based risk prediction models. Clin Cancer Res 2008;14:5977–6983.
50. Richards MA: The National awareness and early diagnosis initiative in England: Assembling the evidence. Br J Cancer 2009;101:s1–s4.
51. Expert panel on detection, evaluation, and treatment of high blood cholesterol in adults. Executive summary of the third report of the National Cholesterol Education Program (NCEP) expert panel on detection, evaluation, and treatment of high blood cholesterol in adults (adult treatment panel III). JAMA 2001;285:2486–2497.See this web link "http://www.ncbi.nlm.nih.gov/pubmed/11368702".
52. Pepe MS, Janes HE: Gauging the performance of SNPs, biomarkers, and clinical factors for predicting risk of breast cancer (Editorial). J Natl Cancer Inst 2008;100:978–979.