

RESEARCH ARTICLE

Refining and Validating a Two-stage and Web-based Cancer Risk Assessment Tool for Village Doctors in China

Xing-Rong Shen^{1&}, Jing Chai^{1&}, Rui Feng², Tong-Zhu Liu³, Gui-Xian Tong¹, Jing Cheng¹, Kai-Chun Li⁴, Shao-Yu Xie⁴, Yong Shi⁴, De-Bin Wang^{1,5*}

Abstract

The big gap between efficacy of population level prevention and expectations due to heterogeneity and complexity of cancer etiologic factors calls for selective yet personalized interventions based on effective risk assessment. This paper documents our research protocol aimed at refining and validating a two-stage and web-based cancer risk assessment tool, from a tentative one in use by an ongoing project, capable of identifying individuals at elevated risk for one or more types of the 80% leading cancers in rural China with adequate sensitivity and specificity and featuring low cost, easy application and cultural and technical sensitivity for farmers and village doctors. The protocol adopted a modified population-based case control design using 72,000 non-patients as controls, 2,200 cancer patients as cases, and another 600 patients as cases for external validation. Factors taken into account comprised 8 domains including diet and nutrition, risk behaviors, family history, precancerous diseases, related medical procedures, exposure to environment hazards, mood and feelings, physical activities and anthropologic and biologic factors. Modeling stresses explored various methodologies like empirical analysis, logistic regression, neuro-network analysis, decision theory and both internal and external validation using concordance statistics, predictive values, etc..

Keywords: Cancer - risk index - prediction model - case control - population-based

Asian Pac J Cancer Prev, **15** (24), 10683-10690

Introduction

Cancers have become one of the most serious threats to human health worldwide (Popat et al., 2013). Steadily growing new cases, high mortality rate combined with lack of radical cures have made prevention and early diagnosis priority strategies stemming the epidemic (Jemal, 2012; Caplan, 2014; Tarraga-Lopez et al., 2014). Tremendous efforts have been invested on public education (e.g., disseminating prevention information via various mass media) (Levano et al., 2014; Seven et al., 2014), screening service, and drug prevention (e.g., use of tamoxifen) and treatment of precancerous conditions (e.g., polyps, *Helicobacter pylori* infection) (Gao et al., 2013; Hady et al., 2013; Lansdorp-Vogelaar et al., 2013). However, there exists a big gap between actual implementation of preventions and expectations (Gupta et al., 2005). Although public education is most cost-effective in communicating knowledge about cancer, its benefit is restricted since general knowledge does not necessarily follow desired behavior. Similarly, screening for high-risk groups and some drugs and treatment prevention are highly efficacious under research conditions, yet these measures are seldom in use in routine practices (Shi,

2009); even used, the effectiveness often turned out to be far from expected (Wang et al., 2007; Zhai, 2012; Honein-Abouhaidar et al., 2014). Lack of personalized behavior intervention may have played an important role underlying this discrepancy (Ozanne et al., 2014). Given the extreme complexity and heterogeneity of the factors determining cancer-related behaviors, general or non-tailored education and service promotion fails easily in initiating or maintaining desired prevention practices (Feng et al., 2014). The nexus of complex factors make it hard for ordinary residents to perceive cause-effect relationships between prevention measures and cancer onset and harms. This greatly weakens their motivation for implementing the measures. In addition, effectively changing the outcomes of a complicated behavior determinant system requires integrating multiple measures and continuous efforts in a synergetic way, which is the disadvantage of general "education" and often beyond the ability of ordinary people.

Personalized intervention against cancer faces various difficulties (Feng et al., 2013). One challenge originates from the intrinsic nature of the epidemic. Cancer happens at about 300 per 100,000 a year on average (He et al., 2012). Such a incidence rate suggests that individual-

¹School of Health Service Management, Anhui Medical University, ²Department of Literature Review and Analysis, Library of Anhui Medical University, ³Anhui Provincial Cancer Hospital, ⁴Luan Center for Disease Prevention and Control, ⁵Collaboration Center for Cancer Control, First Affiliated Hospital of Anhui Medical University, Hefei, Anhui, China ⁶Equal contributors *For correspondence: dbwang@vip.sina.com

based prevention against cancer targeting at non-selective subjects may not be cost-effective since the number needed to treat (NNT) is too big (Bender et al., 2007). It also leads to perception of low susceptibility by ordinary residents since only less than “a few out of thousands” could get cancer for a whole year (Patel et al., 2012). These issues may be solved by assessment tools capable of distinguishing high from low risk individuals. Risk indices or other forms of prediction rules have become widely used in clinical practice to assist medical decision-making when caring for patients with clinical disease, and to counsel patients regarding the likely courses of their diseases (Colditz et al., 2000). Applications of such indices in the prevention of chronic diseases (including specific cancers) are also emerging. The first application in this regard traces back to the Framingham Heart Study in 1976, which has constructed a prediction model to estimate the future risk of coronary heart disease and guide cholesterol-lowering therapy (Grundey et al., 1998). The best known and most widely applied cancer risk prediction model is developed by Gail et al., which uses a woman’s current age and panel of risk factors to assess her risk of breast cancer (Gail et al., 1989). Similar models for other types of cancers e.g. lung cancer (Spitz et al., 2007), gastric cancer (Shimoyama et al., 2000), prostate cancer (Eastham et al., 1999), colorectal cancer (Imperiale et al., 2003) etc. are also available from the literature. However, no parallel prediction rule has been developed for overall risk of cancer, except for the Harvard Cancer Risk Index (Kim et al., 2004).

Another challenge of personalized cancer intervention concerns the widespread lack of professional manpower in delivering tailored, continuous and thus relatively sophisticated counseling, demonstration, supervision etc. This especially applies to China, a nation that has a long history of separated disease prevention and treatment systems (Zhang et al., 1996) and suffers from severe shortage of preventive personnel especially at the frontier level in vast rural areas. Primary care givers, or village doctors in China, may provide an ideal solution to this challenge. They form the bulk of health manpower (over a million in China), enjoy the easiest access to community residents and know well the local sociocultural contexts (Ministry of Health of the People’s Republic of China, 2009).

Based on the above considerations, we started a project called eCROPS-CA (for brief introduction of the project, please refer to our registered trial at DOI 10.1186/ISRCTN33269053). As an acronym, eCROPS-CA summarizes an innovative intervention package against leading cancers in resource poor rural China consisting of 6 major components, i.e., electronic supports and supervision (e), counseling cancer prevention (C), recipe for objective behaviors (R), operational toolkit (O), performance-based incentives (P), and screening and assessment (S). Its goal is to demonstrate that eCROPS-CA is effective in preventing leading cancers and high risk individuals in the intervention arm will, compared to those in the delayed intervention condition, show a lower incidence of cancers, improved cancer-related KAP (knowledge, attitudes and practices) and psycho-

biophysical indicators, and increased use of cancer prevention service. eCROPS-CA utilizes a tentative yet detailed two-stage cancer risk assessment tool for use by village doctors that automatically produces a score for the specific individual under concern predicting his/her overall chance for developing any of the leading cancers in the future. The risk score serves to: a) identify high-risk farmers according to a cutoff score and thus deliver focused intervention; b) inform personalized and outcome-oriented behavior intervention; and c) raise awareness about cancer risk and leverage protection behavior. Developed via systematic literature review, consensus group processes and small scale piloting, the tentative tool merits further modification and validation.

This study aims at developing and validating a two-stage and web-based cancer risk assessment tool, out from the tentative one in use by eCROPS-CA, capable of identifying individuals at elevated risk for 80% leading forms of cancers (further referred to as leading cancers) in rural China with adequate sensitivity (over 75%) and specificity (over 65%) and featuring low cost, easy application and culturally and technically sensitive to farmers and village doctors in resource poor rural China.

Design and Methods

Data sources and design

The study adopts a population- or community-based case control design which draws controls from 36 intervention villages (including 18 intervention and 18 delayed-intervention villages) and cases from 36 townships containing the intervention villages.

As mentioned earlier, the study is an integral part of an ongoing umbrella project, eCROPS-CA. So, it uses two data sources eCROPS-CA generates, namely cancer risk assessment and cancer case survey. Cancer risk assessment happens in the first year of eCROPS-CA and applies to: a) all eligible farmers who live within the intervention (including delayed intervention) villages of the umbrella project and have not been diagnosed with any cancer; and b) cases of the leading cancers diagnosed during the first year among farmers within the observation villages (to be defined below). Cancer case survey proceeds in different time periods at different study sites. For the intervention villages, it starts at the beginning and lasts for the whole process of eCROPS-CA; while for the observation villages, it happens only in the first 1-2 years of the project. The survey aims at finding newly diagnosed cases of the leading cancers and soliciting information about all the variables included in the cancer risk assessment using the same questionnaire.

Study sample and recruitment

As an integral part, subjects of the study are determined by eCROPS-CA recruitment (Figure 1). Selection of intervention and observation villages proceeds in 5 steps. Step 1 classifies all the counties in Anhui, an inland province located in central China, into southern, northern and middle areas. Step 2 randomly selects 3 counties from each of these areas. Step 3 randomly draws 4 townships from each of the counties selected. Step 4 chooses 1 village

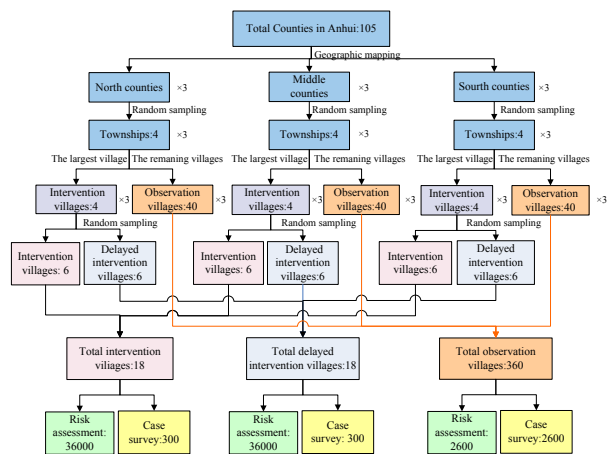


Figure 1. Study Subject Sampling and Randomization

(from each of the townships selected) with the largest number of farmers as intervention villages (36 villages in total) and treats the remaining villages as observation villages. Step 5 randomizes the intervention villages into two equal groups, i.e., 18 intervention and 18 delayed intervention villages.

All the village doctors working for the observation villages determined above are requested to monitor and recommend eligible cancer patients to the local township health centers starting from the beginning of eCROPS-CA until a preset numbers of cases for specific leading cancers (200 for each type of cancers) have reached. The eligibility here defines patients who: a) are 35 to 70 years old; b) live in the selected villages for over 6 months in the past year; c) have diagnosed with one of the leading cancers by a county or higher level hospital within the past month. A trained physician from each of the township health centers checks the eligibility of each patient recommended and performs the cancer case survey as well as cancer risk assessment.

Similarly, a trained village doctor from each of the intervention villages is responsible for recruiting eligible visiting farmer patients and performing the cancer risk assessment at the village clinics in the first year of project implementation. Inclusion criteria for participation in the risk assessment include men and women who: a) are 35 years or older; b) live in the intervention (including delayed intervention) villages for over 6 months in the past year. Farmers who have already diagnosed with cancer (s) or have mental illness or serious illness or disability are excluded. This trained village doctor also monitors, for the whole project period of eCROPS-CA, all the farmers within his/her village who have completed the cancer risk assessment, identifies newly diagnosed cases of the leading cancers among them, and administers the case survey to any cases found.

Given the above criteria, recruitment procedures and our knowledge of local population and cancer prevalence, anticipated subjects comprise: a) 72, 000 non-patient participants from the intervention villages (controls); b) 2600 patient participants from the observation villages (cases for model building); and c) 600 cancer patients from the intervention villages identified via eCROPS-CA follow up evaluations (cases for external model validation). The

number of controls (72, 000) is determined by eCROPS-CA since this study takes the advantage of its umbrella project; while the number of cases (200 per cancer) is a rough estimation of cases required to serve our intention to detect statistically significant odds ratios (ORs) of each of the leading cancers for all the variables included in the questionnaires using conventional values of $\beta=0.10$ and $\alpha=0.05$.

Content and format of instrument

Data collection for purpose of this study employs a cancer case form and a cancer risk questionnaire. The cancer case form applies to any of the leading cancers and collects data about: a) name of hospital where the cancer was diagnosed; b) methods (especially histological methods) used by the hospital for diagnosing the case; and c) type, time, and stage of the cancer diagnosed. Leading cancers include gastric, esophagus, trachea/bronchus/lung, liver, colon/rectum, bladder, lymphoid, kidney/unspecified urinary organs, pancreas, breast, cervix, ovary, and prostate cancer. Nine of them are common cancers among males and twelve of them, common cancers among females.

As summarized in Table 1, the cancer risk assessment questionnaire solicits information about 13 domains of potential etiological factors of the leading cancers. The items included in the questionnaire are designed as either structured questions or questions asking for specific numbers (e.g., age, year of first menstruation). For the purpose of producing a two-stage tool, the questionnaire is further divided into two parts, rapid and detailed risk assessment. The rapid risk assessment consists of 21 unconditional items and takes about 10 minutes to administer; while the detailed risk assessment, 194 conditional items and some 20 minutes to complete. By conditional, we mean that inclusion of an item in the detailed assessment depends on the responses to the previous rapid assessment. For example, the item about smoking dose only occurs in the detailed assessment for a certain individual when he/she has responded that he/she is a smoker in the rapid assessment (Table 2 provides sample items from both parts). Both the risk assessment questionnaire and case survey form had been pilot tested for wording and distribution of potential responses. Taking the example of responses to the question "how much alcohol did you drink per time", they were designed as "1-10g, 11-30g, 31-50g, 51-70g, and >70g" because our pilot study indicated that 20% of the responses fell into each of these categories.

Webpage-based assistance

In order to facilitate project implementation, eCROPS-CA uses extensive electronic support including a user-friendly cancer risk assessment and case survey tool. Written in C# language, the tool runs on a webpage-based system built with Microsoft Visual Studio 2008 and provides instant: a) display of questionnaire or form items; b) reminding of missing or illogic items; c) branching or skipping from items to items; d) recording of entered data; and e) calculation and presentation of resultant risk scores (Figure 2).

Table 1. List of Items Included in Rapid and Detailed Risk Assessment Questionnaires

Domain	Variables
Diet and nutrition	Intake of preserved food, smoked food, fried food, spicy food, leftovers, garlic, bean products, sea foods, fish and shrimp, milk, rice and wheat, vegetable, fruits, tea, roughage, livestock meat; preference of diet temperature, hardness, fat; speed of eating; regularity of eating; time interval between dinner and sleep.
Risk behaviors	Alcohol drinking; smoking; passive smoking; stay up late; lack of physical activity; time spent on sleeping, sedentary work, heavy activities.
Family history	First degree family history of cancer, diabetes, hepatitis, tuberculosis, pancreatitis, hematological system diseases; urogenital infections of partner (s).
Digestive system symptoms and diseases	Tooth decay and-or lose; a toothache and-or gum inflammation; food reflux; swallowing difficulty; stomach discomfort; hepatalgia; reflux esophagitis; chronic gastritis; gastric polyps; gastroduodenal ulcer; helicobacter infections; gastric epithelial dysplasia; gastric intestinal metaplasia; stomach surgery; hepatitis; fatty liver; cirrhosis; cholecystitis or gallstones; pancreatitis; appendicitis; junction (straight) enteritis; intestinal polyp; schistosomiasis; constipation; blood and mucus in stool; hemorrhoids.
Respiratory system symptoms and diseases	Chest distress or breathing difficulties; chest pain; long-term asthma; chronic cough or sputum; long-term nasal blockage; long-term runny nose; chronic rhinitis or sinusitis; tuberculosis; asthma; pneumonia; chronic bronchitis; emphysema; bronchiectasis; silicosis; pneumoconiosis; chronic obstructive pulmonary disease.
Urinary system symptoms and diseases	Urinary frequency, urgency, micturition pain; urethral discharge; urine with blood; eyelids or lower extremity edema; chronic cystitis; urethritis; nephritis; urinary system lithiasis.
Reproductive system symptoms and diseases	Breast pain; breast mass; nipple discharge; repeated abnormal vaginal bleeding; leucorrhea abnormal repeatedly; abdominal mass; abdominal pain or straining feeling; dysmenorrhea; irregular menstruation; mastitis; breast hyperplasia; mammary duct expansion; galactoma; fibroadenoma of breast; gynecological inflammation; infertility; uterine fibroid; cervical cyst; ovarian cyst; prolonged urination; thin and weak urine; dribbling urine; urine overflow; nocturia; painful ejaculation; prostatic hyperplasia; prostatitis; age of first menstruation, marriage, sex, pregnancy, labor; days of menstruation before 40 years; age of menopause; times of marriage, induced abortion, spontaneous abortion, preterm births, live births; accumulated months of breastfeeding, separation from spouses.
Miscellaneous symptoms and diseases	Diabetes; hypertension; hyperlipidemia; rheumatoid arthritis; malaria; blood transfusion; yellow skin and sclera; flushing after alcohol drinking; insomnia and dreaminess; overweight; underweight; others.
Medical procedures	Use of herbs, aspirin, painkiller, clritin, vitamin B12, vitamin B3, chloramphenicol, glucocorticoid, progestin, androgen, estrogen, oral contraceptives, intrauterine device; tubal ligation; hysterectomy; vasectomy.
Exposure to environment hazards	Pesticide; lampblack; straw smoke; soot; dust; asbestos; formaldehyde; coal tar; livestock and pet; water source.
Mood and feelings	Sulking; introversion; impatience; irritability; impulsion; anxiety; tension; distress.
Negative life events	Loss of relatives; major injuries/diseases of relatives, self; major property damage; enmities with others; marital/love breakups/conflicts; stressful tasks prevailed life; natural disasters; law suits; financial hardship; threats of violence.
Anthro-biologic factors	Age; sex; education; blood pressure; glucose; height; weight.

Table 2. Sample Items Included in Rapid and Detailed Risk Assessment Questionnaire

Part A Rapid Risk Assessment					
Have you ever been diagnosed with any of following digestive diseases?					
<input type="checkbox"/> Tooth loss or cavities	<input type="checkbox"/> Duodenal ulcer	<input type="checkbox"/> Fatty liver	<input type="checkbox"/> Intestinal polyps		
<input type="checkbox"/> Chronic gastritis	<input type="checkbox"/> Hepatitis	<input type="checkbox"/> Cholecystitis /Cholelithiasis	<input type="checkbox"/> Pancreatitis		
<input type="checkbox"/> Chronic gastritis ulcer	<input type="checkbox"/> Cirrhosis	<input type="checkbox"/> Chronic appendicitis /enteritis	<input type="checkbox"/> Hemorrhoids		
<input type="checkbox"/> Helicobacter pyloriInfection					
Please offer the following information (applies only to females)?					
<input type="checkbox"/> Age of first menstruation	<input type="checkbox"/> Age of first sexual activity	<input type="checkbox"/> Age of first pregnancy	<input type="checkbox"/> Age of first parturition		
<input type="checkbox"/> Times of marriage	<input type="checkbox"/> Times of abortions	<input type="checkbox"/> Accumulative years of taking contraceptives	<input type="checkbox"/> Accumulative months of breast feeding		
<input type="checkbox"/> Times of parturition	<input type="checkbox"/> Age of menopause	<input type="checkbox"/> Age of first marriage	<input type="checkbox"/> Times of premature birth		
<input type="checkbox"/> Days of menstruation per time					
Do you have the following dietary habit (s)?					
<input type="checkbox"/> Eating too full	<input type="checkbox"/> Eating cold food	<input type="checkbox"/> Eating within 1 hour before sleep	<input type="checkbox"/> Drinking coffee		
<input type="checkbox"/> Eating fast	<input type="checkbox"/> Eating at irregular time		<input type="checkbox"/> Drinking alcohol		
<input type="checkbox"/> Eating hot food	<input type="checkbox"/> Drinking tea				
Which of following describes you?					
<input type="checkbox"/> Smoker	<input type="checkbox"/> Sedentary person	<input type="checkbox"/> Night owl			
Part B Detailed Risk Assessment					
Tooth loss/cavities (If checked in Rapid Risk Assessment)					
--How old were you when you first found tooth cavity?					
<input type="checkbox"/> 10 and less	<input type="checkbox"/> 1-1-20	<input type="checkbox"/> 21-30	<input type="checkbox"/> 31-40	<input type="checkbox"/> 41-50	<input type="checkbox"/> 51 and older
--How many cavity teeth have you had in total?					
<input type="checkbox"/> 1-2	<input type="checkbox"/> 3-4	<input type="checkbox"/> 5-6	<input type="checkbox"/> 7-8	<input type="checkbox"/> 9-10	<input type="checkbox"/> 11 and more
--How many teeth have you lost?					
<input type="checkbox"/> 1-2	<input type="checkbox"/> 3-4	<input type="checkbox"/> 5-6	<input type="checkbox"/> 7-8	<input type="checkbox"/> 9-10	<input type="checkbox"/> 11 and more
Drinking alcohol (If checked in Rapid Risk Assessment)					
--How many years in your life can you be described as frequent drinker?					
<input type="checkbox"/> 1-5	<input type="checkbox"/> 6-10	<input type="checkbox"/> 11-15	<input type="checkbox"/> 16-20	<input type="checkbox"/> 21-25	<input type="checkbox"/> 26 and more
--How many times did you drink a month?					
<input type="checkbox"/> 1-6	<input type="checkbox"/> 7-12	<input type="checkbox"/> 13-18	<input type="checkbox"/> 19-24	<input type="checkbox"/> 25-30	<input type="checkbox"/> 31 and more
--How much alcohol did you drink per time?					
<input type="checkbox"/> 1-10 g	<input type="checkbox"/> 11-30g	<input type="checkbox"/> 31-50g	<input type="checkbox"/> 51-70g	<input type="checkbox"/> 71g and more	
--How many times did you over drink a month?					
<input type="checkbox"/> 1-6	<input type="checkbox"/> 7-12	<input type="checkbox"/> 13-18	<input type="checkbox"/> 19-24	<input type="checkbox"/> 25-30	<input type="checkbox"/> 31 and more



Figure 2. Sample Web Pages of Risk Assessment Questionnaire (Top) and Scoring (Bottom) Browsed Via A Smart Phone

Model building and validation

Model production, validation and optimization proceeds in the following steps. Initial step centers on descriptive summaries intended to examine patterns of the various variables and check for normality of the continuous variables. Necessary transformations are tried and selected, if necessary, to induce approximate normality. The next step focuses on building combined score or index (for predicting overall risk of all the leading cancers) and specific models (for each of the leading cancers). This step stresses exploring various approaches to maximize the potential of alternative models including the Harvard Cancer Index, the tentative Score in use by eCROPS-CA and models using rapid assessment variables only and those incorporating both rapid and detailed assessment variables. The third step evaluates the performance of each of the alternative models generated and calculates the concordance statistics and the positive and negative predictive values. The final step decides upon optimal models and variable sets for future use and cutoff value (s) for selecting priority individuals from rapid assessment into detailed assessment and from detailed assessment into focused interventions or follow up.

The modeling adopts a stage-wise approach in reaching two-stage models. The first stage produces rapid assessment models using the rapid risk assessment variables and all the case (N=2600) and control data (N=72,000). The second stage builds detailed assessment models using the detailed risk assessment variables and a subset (rather than the whole set) of the case and control data. This subset is determined by a cutoff score of rapid assessment. In order to maximize the potential for choice, a series of cutoff values (say the 10th, 20th, 30th, 40th, 50th, 60th, 70th, 80th, and 90th percentile of rapid assessment scores) will be tested. Anticipated methods for building both the rapid and detail assessment models include empirical analysis, consensus group process, logistic regression, proportional hazards models, log incidence, neuro-network analysis, decision theory, and

even combinations of these.

Selection of optimal models strives to reach a balanced decision upon: a) the highest predictive value, sensitivity and specificity of the model; and b) the highest percentage of individuals being filtrated by the rapid risk assessment so as to reduce the detailed assessment workload to the minimum. One potential roadmap toward this end reads: a) selecting, among all the potential rapid assessment models, a limited number (say 5) of best performers in terms of concordance statistics (or ROC curves); b) calculating a rapid assessment score for each of the cases and controls using each of the best performer models selected; c) setting a series of cutoff values for each of the selected rapid assessment models; d) selecting eligible subsets of cases and controls into detailed risk assessment modeling using each of the cutoff values set; e) exploring various detailed assessment models using each of the subsets; f) evaluating the performance of all the detailed assessment models built and deciding on a few best performers using concordance statistics and calibration via bootstrapping.

Ethics

The study protocol had been reviewed and approved by the Biomedical Ethics Committee of Anhui Medical University. Participation of farmers and village doctors are voluntary and written informed consent is sought from all participants.

Discussion

As the stated by the study aim, the assessment tool this study tries to develop stresses several important features. Different from prediction models for single specific cancers, our intended tool produces not only a combined score predicting the overall risk for developing any of the leading cancers, but also a whole set of specific scores for estimating the risk of each of the cancers. Such a "mixed" tool may be useful at individual as well aggregate levels by various means, e.g., identifying individuals at elevated risk; improving clinical decision-making; planning intervention trials, estimating the cost of population cancer burden and designing population prevention strategies (Freedman et al., 2005). Of these, one point worth particular noting is that interventions guided by overall risk score tackle critical paths leading to multiple cancers simultaneously. This strategy may prove to be more cost-effective than that focusing on a single cancer. Most cancers share similar causes. Smoking, for instance, is not only linked with lung cancer, but also colorectal (Cross et al., 2014), gastric (Zhong et al., 2014), and breast cancer (Ilic et al., 2014). Therefore, smoking cessation prevents all these cancers at the same time. Targeting at multiple cancers may also benefit from "economies of scale" (Trogon et al., 2014). Taking the example of a typical village included in our eCROPS-CA project, given the trial design and cutoff scores per se, the number of high risk farmers needing personalized intervention is estimated as some 120. If the village doctor (s) were requested to deliver intervention against only one type of the cancers, the service volume is reduced to about 10 and thus the unit cost for training, supervision etc. will increase substantially.

The disadvantages of multiple versus single cancer instruments originate mainly from data requirement and process. The scope of data needed to predict overall risk of multiple cancers is much broader than that to predict any single cancer. Calculation of the overall Harvard Cancer Risk Index involves 52 variables; while variables needed to generate scores for specific cancers covered by the Index ranged from 3 to 17 (Colditz et al., 2000). In our case, total items forming the overall instrument add up to 194; while those relating to specific cancers, only 13 to 46. So, overall risk models incur much heavier workload in collecting and processing data than that of specific models. The two-stage strategy adopted in our tool provides an effective solution to this issue. By setting a proper cutoff score and starting with rapid followed by detailed risk assessment, this workload can be reduced to a minimum. For example, if we set the cutoff point of rapid assessment score at the 70th percentile, then only 30% of the individuals enter detailed risk assessment. As mentioned earlier, the rapid risk assessment takes about 10 minutes and the detailed risk assessment, 20 minutes. Therefore, a two-stage assessment takes only about 16 minutes on average (i.e., 10 minutes for all individuals plus 20 minutes for 30% of the individuals). This saves 14 minutes per individual since one-stage complete assessment takes 30 minutes (=10+20). In addition, the web-based support system further facilitates this reduced work by means of automatic branching or skipping from item to item and instant calculation and presentation of resultant scores.

Given that our rapid and detailed assessment questionnaires contain all the variables included in the Harvard Cancer Index, this study enables comparing its performance with various models derived by us. Developed through a group consensus process in 2000, the Index aims to predict the relative risks of individuals, aged 40 and above, of developing the leading types of cancers that contribute to approximately 80% of cancer incidence in the US (Kim et al., 2004). The Index has only been tested for part of cancers in some American groups. Given the heterogeneity in the genetic, environmental, nutritional, and lifestyle factors, as well as precancerous illnesses across nations and ethnic groups and new evidences on the relationships between cancers and these factors, there is a clear need to compare and adapt the Index to reflect renewed evidences and suit different populations. The study also allows for comparisons between its resultant models with that in use by eCROPS-CA. Based mainly on meta-analysis, the eCROPS-CA scoring system also lacks population-based validation and adjustment.

Perhaps the greatest challenge relates to model building. The essence of risk modeling is to obtain accurate relative and attributable risk estimates for etiologic factors, e.g., demographics, reproductive history, smoking, dietary patterns and medications (Sun et al., 2013). This depends on a clear understanding of the nature of all the individual factors involved and interactions between them. Given the state of art of researches in this field, there runs a risk of being unable to produce models as good as expected, though this risk may be reduced to some extent by trying various methods and perspectives. Our intended model is

not inclusive; it covers 80% leading cancers in China for avoiding undue emphasis being placed on rare cancers that make little contribution to total cancer burden (Colditz et al., 2000). It incorporates only minimum easy and low-cost clinical and biologic markers (e.g., blood pressure, cholesterol, glucose) but relatively expensive ones (e.g., enzyme levels, histologic markers). This ensures affordability and sustainability yet may restrict the quality of the resultant model(s). Our modeling utilizes data from both "current" cases and cases identified via follow up surveys. Potential biases and differences between these data (David et al., 2014) merit careful consideration and proper correction.

Finally, some readers may raise the concern about anxieties and fears resulting from the risk assessment. According to Emmons and colleagues, part of the participants in their qualitative study of the Harvard Cancer Risk Index reported that the new information presented by the index was somewhat anxiety producing (Emmons et al., 1999). Some researchers, however, hold different view over this issue. They argue that change often requires some amount of anxiety as a precursor to action (Benight et al., 2004). Besides, the anxieties are tunable by appropriate presentation of the risk score (e.g., absolute vs. relative risk) and explanation of its meaning and contributing factors.

Acknowledgements

Development of the primitive protocol was supported by the Natural Science Foundation of China (grant number: 81172201). Refinement and Implementation of the protocol is lead and supported by Collaboration Center for Cancer Control of Anhui Medical University and Anhui Provincial Cancer Hospital.

References

- Bender R, Kuss O, Hildebrandt M, Gehrman U (2007). Estimating adjusted NNT measures in logistic regression analysis. *Stat Med*, **26**, 5586-95.
- Benight CC, Bandura A (2004). Social cognitive theory of posttraumatic recovery: the role of perceived self-efficacy. *Behav Res Ther*, **42**, 1129-48.
- Caplan L (2014). Delay in breast cancer: implications for stage at diagnosis and survival. *Front Public Health*, **2**, 87.
- Cross AJ, Boca S, Freedman ND, et al (2014). Metabolites of tobacco smoking and colorectal cancer risk. *Carcinogenesis*, **35**, 1516-22.
- Colditz GA, Atwood KA, Emmons K, et al (2000). Harvard report on cancer prevention volume 4: harvard cancer risk index. risk index working group, harvard center for cancer prevention. *Cancer Causes Control*, **11**, 477-88.
- David MC, Ware RS, Alati R, Dower J, Donald M (2014). Assessing bias in a prospective study of diabetes that implemented substitution sampling as a recruitment strategy. *J Clin Epidemiol*, **67**, 715-21.
- Eastham JA, May R, Robertson JL, Sartor O, Kattan MW (1999). Development of a nomogram that predicts the probability of a positive prostate biopsy in men with an abnormal digital rectal examination and a prostate-specific antigen between 0 and 4ng/ml. *Urology*, **54**, 703-13.
- Emmons KM, Koch-Weser S, Atwood K, et al. (1999). A

- qualitative evaluation of the harvard cancer risk index. *J Health Commun*, **4**, 181-93.
- Feng R, Wang DB, Chai J, Cheng J, Li HP (2014). Total delay for treatment among cancer patients: a theory-guided survey in China. *Asian Pac J Cancer Prev*, **15**, 4339-47.
- Feng R, Li K, Cheng J, et al (2013). Toward integrated and sustainable prevention against diabetes in rural China: study rationale and protocol of eCROPS. *BMC Endocr Disord*, **13**, 28.
- Freedman AN, Seminara D, Gail MH, et al (2005). Cancer risk prediction models: a workshop on development, evaluation, and application. *J Natl Cancer Inst*, **97**, 715-23.
- Gao QY, Chen HM, Chen YX, et al (2013). Folic acid prevents the initial occurrence of sporadic colorectal adenoma in Chinese older than 50 years of age: a randomized clinical trial. *Cancer Prev Res*, **6**, 744-52.
- Gupta AK, Melton LJ 3rd, Petersen GM, et al (2005). Changing trends in the incidence, stage, survival, and screen-detection of colorectal cancer: a population-based study. *Clin Gastroenterol Hepatol*, **3**, 150-8.
- Grundy SM, Balady GJ, Criqui MH, et al (1998). Primary prevention of coronary heart disease: guidance from Framingham: a statement for healthcare professionals from the AHA task force on risk reduction. American Heart Association. *Circulation*, **97**, 1876-87.
- Gail MH, Brinton LA, Byar DP, et al (1989). Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *J Natl Cancer Inst*, **81**, 1879-86.
- Honein-Abouhaidar GN, Kastner M, Vuong V, et al (2014). Benefits and barriers to participation in colorectal cancer screening: a protocol for a systematic review and synthesis of qualitative studies. *BMJ Open*, **4**, 004508.
- Hady HR, Soldatow M, Lukaszewicz J, Luba M, et al (2013). Surgical treatment of malignant and benign colorectal neoplasms based on authors' clinical data. *Advances Clin Experimental Med*, **22**, 219-27.
- He J, Chen WQ (2012). Chinese cancer registry annual report 2012. military medical science press.
- Jemal A (2012). Global burden of cancer: opportunities for prevention. *Lancet*, **380**, 1797-99.
- Kim DJ, Rockhill B, Colditz GA. (2004). Validation of the Harvard Cancer Risk Index: a prediction tool for individual cancer risk. *J Clin Epidemiol*, **57**, 332-40.
- Lansdorp-Vogelaar I, Sharp L (2013). Cost-effectiveness of screening and treating *Helicobacter pylori* for gastric cancer prevention. *Best Pract Res Clin Gastroenterol*, **27**, 933-47.
- Levano W, Miller JW, Leonard B, et al (2014). Public education and targeted outreach to underserved women through the national breast and cervical cancer early detection program. *Cancer*, **16**, 2591-6.
- Ministry of health of the people's republic of China: China health statistic annual. peking union medical college press, 2009.
- Ozanne EM, Howe R, Omer Z, Esserman LJ (2014). Development of a personalized decision aid for breast cancer risk reduction and management. *BMC Med Inform Decis Mak*, **14**, 4.
- Popat K, McQueen K, Feeley TW (2013). The global burden of cancer. *Best Pract Res Clin Anaesthesiol*, **27**, 399-408.
- Patel D, Akporobaro A, Chinyanganya N, et al. (2012). Attitudes to participation in a lung cancer screening trial: a qualitative study. *Thorax*, **67**, 418-25.
- Ilic M, Vlajinac H, Marinkovic J. (2014). Cigarette smoking and breast cancer: a case-control study in Serbia. *Asian Pac J Cancer Prev*, **14**, 6643-7.
- Imperiale TF, Wagner DR, Lin CY, et al. (2003). Using risk for advanced proximal colonic neoplasia to tailor endoscopic screening for colorectal cancer. *Ann Intern Med*, **39**, 959-65.
- Seven M, Akyüz A, Robertson LB (2014). Interventional education methods for increasing women's participation in breast cancer screening program. *J Cancer Educ*, [Epub ahead of print].
- Sun CQ, Chang YB, Cui LL, et al (2013). A population-based case-control study on risk factors for gastric cardia cancer in rural areas of Linzhou. *Asian Pac J Cancer Prev*, **14**, 2897-901.
- Shi JF (2009). A cost-effectiveness analysis of China cervical cancer screening protocol for rural farmers. PhD thesis. China union Medical University.
- Spitz MR, Hong WK, Amos CI, et al (2007). A risk model for prediction of lung cancer. *J Natl Cancer Inst*, **99**, 715-26.
- Shimoyama T, Fukuda S, Tanaka M, Nakaji S, Munakata A (2000). Evaluation of the applicability of the gastric carcinoma risk index for intestinal type cancer in Japanese patients infected with *Helicobacter pylori*. *Virchows Arch*, **436**, 585-7.
- Tarraga Lopez PJ, Albero JS, Rodriguez-Montes JA (2014). Primary and secondary prevention of colorectal cancer. *Clin Med Insights Gastroenterol*, **7**, 33-46.
- Trogdon JG, Ekwueme DU, Subramanian S, Crouse W (2013). Economies of scale in federally-funded state-organized public health programs: results from the national breast and cervical cancer early detection programs. *Health Care Manag Sci*, [Epub ahead of print].
- Wang YD, Qu LY, Guan LZ, et al (2007). Colorectal cancer screening in Beijing community: effectiveness and problems. *Chinese J Primary Health Care*, **10**, 1586-8.
- Zhong C, Li KN, Bi JW, Wang BC. (2014). Sodium intake, salt taste and gastric cancer risk according to *Helicobacter pylori* infection, smoking, histological type and tumor site in China. *Asian Pac J Cancer Prev*, **13**, 2481-4.
- Zhai LF (2012). Effectiveness and implications of colorectal cancer screening in urban communities. *Chinese Primary Health Care*, **26**, 34-35.
- Zhang WK (1996). Control the spread of HIV/STI by incorporating prevention with treatment. *J Chin AIDS/STD Pre Cont*, **2**, 248-53.