Developing Screening Services for Colorectal Cancer on Android Smartphones

Hui-Ching Wu, PhD Candidate,^{1,*} Chiao-Jung Chang,²
Chun-Che Lin, MD,^{3,4} Ming-Chang Tsai, MD,^{3-5,*}
Che-Chia Chang,⁶ and Ming-Hseng Tseng, PhD⁶

Schools of ¹Medical Sociology and Social Work, ³Medicine, and
⁶Medical Informatics, Chung-Shan Medical University, Taiwan,
Republic of China.

²Computer Center and ⁵Institute of Medicine, Chung-Shan
Medical University, Taiwan, Republic of China.

⁴Division of Gastroenterology and Hepatology, Department of
Internal Medicine, Chung-Shan Medical University Hospital,
Taiwan, Republic of China.

*H.-C.W. and M.-C.T. contributed equally to this study and share first
authorship. H.-C.W. was mainly in charge of system development, and
M.-C.T. was responsible for the clinical study.

Abstract

Introduction: Colorectal cancer (CRC) is an important health problem in Western countries and also in Asia. It is the third leading cause of cancer deaths in both men and women in Taiwan. According to the well-known adenoma-to-carcinoma sequence, the majority of CRC develops from colorectal adenomatous polyps. This concept provides the rationale for screening and prevention of CRC. Removal of colorectal adenoma could reduce the mortality and incidence of CRC. Mobile phones are now playing an ever more crucial role in people's daily lives. The latest generation of smartphones is increasingly viewed as hand-held computers rather than as phones, because of their powerful on-board computing capability, capacious memories, large screens, and open operating systems that encourage development of applications (apps). **Subjects and Methods:** If we can detect the potential CRC patients early and offer them appropriate treatments and services, this would not only promote the quality of life, but also reduce the possible serious complications and medical costs. In this study, an intelligent CRC screening app on Android™ (Google™, Mountain View, CA) smartphones has been developed based on a data mining approach using decision tree algorithms. For comparison, the stepwise backward multivariate logistic regression model and the fecal occult blood test were also used. **Results:** Compared with the stepwise backward multivariate logistic regression model and the fecal occult blood test, the proposed app system not only provides an easy and efficient way to quickly detect high-risk groups of potential CRC patients, but also brings more information about CRC to customer-oriented services. **Conclusions:** We developed and implemented an app system on Android platforms for ubiquitous healthcare services for CRC screening. It can assist people in achieving early screening, di-

agnosis, and treatment purposes, prevent the occurrence of complications, and thus reach the goal of preventive medicine.

Key words: colorectal cancer, mobile health telehealth, data mining, health monitoring

Introduction

More than half of Americans 25–29 years of age now live in households with mobile phones but no traditional landline telephone according to a December 2010 report on phone use by the U.S. National Center for Health Statistics at the Centers for Disease Control and Prevention.¹ The same study also found that the younger children are, the more likely they are to live in homes that only have wireless phones, suggesting that younger parents are becoming increasingly reliant on mobile phones even as they adjust from being single to a more settled family lifestyle.²

Although the mobile phone has been widely used for several decades, smartphones are a more recent advance. The latest generation of smartphones is increasingly viewed as hand-held computers rather than as phones, because of their powerful on-board computing capability, capacious memories, large screens, and open operating systems that encourage development of applications (apps).³

According to a recent video report by Mobile Future,⁴ there has been a massive increase in the numbers of consumer smartphone apps downloaded over the past 2 years, with figures going up from 300 million apps downloaded in 2009 to 5 billion in 2010. Kailas et al.⁵ stated that there are already in excess of 7,000 documented cases of smartphone health apps. The potential for the creation of simple and easy-to-download apps for smartphones has created a vibrant emerging healthcare industry.

Android™ (Google™, Mountain View, CA) has a large community of developers writing apps that extend the functionality of the devices. Developers write primarily in a customized version of Java.⁶ Apps can be downloaded from third-party sites or through online stores such as Google Play (formerly Android Market), the app store run by Google. In October 2011, there were more than 500,000 apps available for Android,⁷ and the estimated number of apps downloaded from the Android Market as of December 2011 exceeded 10 billion.⁸ The operating system itself is installed on 130 million total devices.⁹ The Android platform has also grown to become a favorite among mobile developers.

The goal of this study was to design an intelligent app on Android smartphone for colorectal cancer (CRC) screening. The app would provide three models with different parameter complexity of the diagnosis of colorectal adenomatous polyp based on some prediction rules mined from a clinical dataset. It will reduce overhead costs and examination duration for CRC candidates or patients. In addition, this application also delivers some useful information

about medical service referral on customer-oriented services when a user is diagnosed as a high-risk patient. To the best of our knowledge, there has been no previous app for CRC screening on a smartphone platform.

Background

It is clear that the potential for mobile communication to transform healthcare and clinical intervention in the community is tremendous. Significant economic benefits have also been reported where mobile communication is used in the provision of remote healthcare advice and telemedicine.¹⁰ Extensive reviews^{11–16} of the use of mobile phones and hand-held computing devices in health and clinical practice can be found, notably in the collection and collation of data for healthcare research, and as used in support of medical and healthcare education and clinical practice in the community.

In the area of mobile diagnostics using mobile phone technologies, there have been several recent developments. For example, Kaplan¹⁷ has highlighted the successful use of mobile phones to support telemedicine and remote healthcare in developing nations. Martinez et al.¹⁸ used camera phones and paper-based microfluidic devices in real-time off-site medical diagnosis. Jin et al.¹⁹ have developed a system to record electrocardiograms via a cellphone. Tan and Masek²⁰ have developed a system to interface with Doppler devices for fetal ultrasound assessment. Black et al.²¹ have created a low-cost pulse oximeter attached to a cellphone to try to distinguish pneumonia from other febrile illnesses. Chen et al.²² have developed an intelligent heart sound diagnostics capability on a cellphone based on the use of the audio channel to determine heart rate and heart rate variability.

CRC is an important health problem in Western countries and also in Asia. CRC is nowadays a leading cause of cancer death throughout the world. It is also the third leading cause of cancer death in both sexes in Taiwan.²³ In 2009, the age-adjusted CRC incidence in the United States was 61.2 cases per 100,000 population among men and 44.8 per 100,000 population among women.²³ According to the well-known adenoma-to-carcinoma sequence, the majority of CRC develops from colorectal adenomatous polyps through a series of genetic and epigenetic alterations.²⁴ Adenomatous polyps may take at least 10 years to undergo malignant transformation. In addition, early cancer usually needs several years to progress to the invasive stage. This concept provides the rationale for screening and prevention of CRC. The prevalence of adenoma among the average-risk population varied in previous literature reports. A recent meta-analysis found that the prevalence of adenoma and CRC was 30.2% and 0.3%, respectively.²⁵ A very recent study in Austria found that the prevalence of adenoma and CRC was 19.7% and 1.1%, respectively. Male gender was significantly associated with a higher prevalence of adenoma (24.9% versus 14.8%) and CRC (1.5% versus 0.7%).²⁶ Among healthy asymptomatic patients older than 50 years of age, screening colonoscopy should detect adenoma in 25% of men and 15% of women. Accumulating evidence supports the proposal that early detection and removal of colorectal adenomas could reduce the mortality and incidence of CRC.²⁷ Therefore, CRC is a good

candidate for screening, and several modalities are recommended, including the fecal occult blood test (FOBT), flexible sigmoidoscopy, and colonoscopy.²⁸ The FOBT is the mostly widely used screening method for CRC. Current guidelines suggest an annual FOBT test in asymptomatic adults 50 years of age and older.²⁸ Many potential risk factors have been investigated previously, including age, sex, family history, smoking, high-fat diet, inflammatory bowel disease, diabetes mellitus, post-cholecystectomy, and overweight.^{29–32}

Data mining analysis is the process of selecting and exploring large amounts of data, without setting the hypothesis, in order to discover unknown patterns or relationships. In recent years, data mining technologies have matured and become more popular, providing a useful tool to analyze, classify, and predict the large amounts of data gathered from various fields. Early detection of medical problems is important to increase the chance of successful treatment. Such detection is often formulated as a binary classification problem. Various classification methods have been developed for the detection of a potential medical problem.³³

Subjects and Methods

DATA SOURCE

We prospectively enrolled patients who had undergone diagnostic colonoscopy in the endoscopy unit of a 1,312-bed academic urban tertiary-care referral center because of clinical symptoms or screening for colon lesion between August 2009 and August 2010.

A trained nurse interviewed these subjects for eligibility and invited them to participate on the examination day. Subjects were excluded if (1) they were unable to answer questions from the investigator, (2) data from the chart were incomplete, (3) there was a failure to achieve complete colonoscopy without a diagnosis, (4) there had been poor colon preparation (semisolid or solid stool at colonoscopy), or (5) there was colorectal carcinoma with an incomplete check of the total colon.

Before a colonoscopy, patient characteristics and medical history were obtained from charts and interview. In total, 20 variables made up the patient's demographic data (age, sex, body height, body weight, and body mass index), reasons for colonoscopy (bloody stool, abdominal pain, constipation, bowel habit change, anemia, tenesmus, positive FOBT, colon polyp history, family history of CRC, or elevation in carcinoembryonic antigen level), and patient's habits (smoking, drinking, betel nut chewing, or tea or coffee consumption).

Colonoscopy was performed by one of three experienced endoscopists (with a minimum experience of 2,000 colonoscopies) with a colonoscope device (model CF-H260AZI; Olympus Optical Co., Ltd., Tokyo, Japan). The scope was passed to the cecum. Complete colonoscopy examination was defined when the scope reached the cecum and was documented by photography. The detected lesions and procedural variables were recorded, including size, morphology, location, biopsy, or polypectomy. Removed polyps were examined by a pathologist who was unaware of clinical history or endoscopic finding. The pathology results were divided into adenomatous neoplasm (included adenoma, dysplasia, and adenocarcinoma) or not (included hyperplastic polyp or others).

Fig. 1. System architecture. 3G, third generation.

After being approved by the Institutional Review Board, the study was completely de-identified to all subjects, and all subjects signed written informed consent before participation. This study initially collected 343 subjects' records. Finally, 225 records were used in the following analysis by using data processing techniques to exclude missing data and outliers.

SYSTEM ARCHITECTURE

The system architecture (shown in Fig. 1) is proposed by the combination of five modules, including the Android smartphone app, the data analysis server, the medical information service servers, the physiological signals measurements, and the multiple physiological signals gateway.³⁴ The analysis server extracts the decision rules from a clinical dataset by using a data mining approach for app programming, and the users can directly retrieve the related medical treatments and services information from some Web servers by smartphones with Wi-Fi or third-generation wireless communication technology. The connection between the Android smartphone and the data analysis server is adopted in an offline manner. The new version of the intelligent app will be automatically updated to the user's smartphone if the decision rules are renewed from the data analysis server. The dashed box in Figure 1 represents the optional facilities of the proposed system.

In this prototype system, users can directly input their personal characteristics (such as sex, age, body weight, and body height) to the Android smartphone through the app interface. Users also can use the multiple physiological signals gateway to transmit the user's body weight value to the Android smartphone from the physiological measuring device by using Bluetooth® (Bluetooth SIG, Kirkland, WA) communication.

DECISION TREE ALGORITHMS

In order to develop a simple, reliable, and accurate app in this study, the detection rules were first extracted at the data analysis server from a clinical dataset by a data mining approach with decision tree algo-

rithms, and then these rules were confirmed by hepatogastroenterology medical experts.

Large amounts of data have been gathered routinely in the course of day-to-day management in medicine and the delivery of health services. For mining useful information and helpful knowledge from these large databases, an important research area has been developed called data mining.³⁵

Data mining is the process of extracting hidden knowledge from large volumes of raw data. It can also be defined as the process of extracting hidden predictive information from large databases. Data mining is more intuitive, allowing for increased insight beyond data warehousing. An implementation of data mining in an organization will serve as a guide to uncovering inherent trends and tendencies in historical information. It will also allow for statistical predictions, groupings, and classifications of data. Most hospitals collect, refine, and deduce massive medical data. Data mining techniques can be implemented rapidly on existing software and hardware platforms to enhance the value of existing information resources and can be integrated with new products and systems as they become part of the system. Some value results can be extracted by using data mining tools to deliver answers to many different types of predictive questions.³⁶ There are many tasks associated with data mining, such as classification, clustering, regression, and association rule mining.

Decision trees are powerful and popular methods for classification and prediction. In medical applications, both the accuracy of a prediction and the ability to explain the reason for a decision are equally crucial. The attractiveness of decision trees is due to the fact that, in contrast to neural networks and support vector machines, decision trees are able to generate understandable rules. Rules can readily be expressed so that humans can easily understand them, and they can be used directly in applications. Decision trees also can provide a clear indication of which fields are most important for prediction or classification.

The goal of decision tree learning is to create a model that predicts the value of a target variable based on several input variables. The decision tree is a classifier in the form of a tree structure, where each node is either a decision node or a leaf node. Each decision node specifies some test to be carried out on a single attribute-value, with one branch and subtree for each possible outcome of the test. Each leaf represents a value of the target variable given the values of the input variables represented by the path from the root to the leaf. A decision tree can be used to classify an example by starting at the root of the tree and moving through it until a leaf node, which provides the classification of the instance.

Most algorithms that have been developed for constructing decision trees usually work top-down by choosing a variable at each step that is the next best variable to use in splitting the set of items.³⁷ The best

Table 1. Descriptive Statistics of Normal and Colon Adenoma Cases

ATTRIBUTE	NORMAL (N=128 [56.89%])		COLON ADENOMA (N=97 [43.11%])		P VALUE
	MEAN ± SD	RANGE	MEAN ± SD	RANGE	
Age (years)	47.80 ± 14.70	21–80	59.95 ± 9.94	34–80	0.000 ^b
Height (cm)	162.62 ± 9.13	145–187	164.99 ± 6.75	150–178	0.026 ^a
Weight (kg)	60.04 ± 12.76	35–85	68.78 ± 10.05	51–93	0.000 ^b
BMI (kg/m ²)	22.50 ± 3.14	16–31	25.25 ± 3.15	20–35	0.000 ^b
Sex (male/female) ^c	54 (42.2%)/74 (57.8%)	—	72 (74.2%)/25 (25.8%)	—	0.001 ^b

^a $\alpha=0.05$, ^b $\alpha=0.01$.^cData are number (%).

BMI, body mass index; SD, standard deviation.

attribute is defined by how well the variable splits the set into homogeneous subsets that have the same value of the target variable. Different algorithms use different formulas for measuring the best attribute. A tree can be constructed by splitting the training set into subsets based on the best attribute. This process is repeated on each derived subset in a recursive manner called recursive partitioning. The recursion is completed when the subset at a node all has the same value of the target variable or when splitting no longer adds value to the predictions.

The Microsoft® (Redmond, WA) Decision Trees (MSDT)^{38,39} algorithm was developed in the Microsoft SQL server 2005 for classification and regression tasks. The MSDT algorithm mainly uses an interesting score or Shannon's entropy in tree splitting and has support for analyzing discrete and continuous valued attributes. However, like the Chi-Square Automatic Interaction Detector algorithm, it does not have any pruning step but presents a parameter named "complexity penalty" for controlling tree growth. Both multiway and binary-splitting capabilities are provided in MSDT. Furthermore, MSDT has also automatic feature selection and dimension reduction, which are not available in other data mining packages. Bozkir and Sezer³⁹ reported that MSDT gives a better prediction than the Chi-Square Automatic Interaction Detector and Classification and Regression Tree algorithms because multisplitting decision tree approaches are more suitable than binary-splitting techniques and because multiway splitting decision tree models are clearly more explainable. Therefore, the MSDT algorithm is used in the present study.

A decision tree graph can be converted into a number of If–

Then rules. In this study, support factor, confidence factor, and odds factor are also provided to measure the performance of each rule. The support factor measures the coverage of a rule, which is the ratio of the number of instances covered by the rule to the total number of instances. The confidence factor measures the accuracy of a rule. The "odds" of an event is the fraction of the probability that the event will happen to the probability that it will not happen. For a rule "If X then Y" and a training set of N instances, the support factor and confidence factor are defined as

$$\text{Support} = \frac{\text{number of instances with both } X \text{ and } Y}{N} \quad (1)$$

$$\text{Confidence} = \frac{\text{number of instances with both } X \text{ and } Y}{\text{number of instances with } X} \quad (2)$$

$$\text{Odds} = \frac{\text{number of instances with both } X \text{ and } Y}{\text{number of instances with both } X \text{ and not } Y} \quad (3)$$

Results

This study used the Eclipse Java development tools to develop the intelligent Android app using a total of five decision rules extracted from the MSDT algorithm for prediction of high-risk groups of patients with colorectal adenomatous polyp.

DESCRIPTIVE STATISTICS

Quantitative data were summarized and are presented as mean ± standard deviation values. Continuous variables were compared

Table 2. Descriptive Statistics of Five Noninvasive Features

ATTRIBUTE	ALL PATIENTS	MODEL BUILDING	INTERNAL VALIDATION	P VALUE
Number of patients (%)	225 (100%)	150 (67%)	75 (33%)	
Age (years)	53.040 ± 14.184	54.127 ± 13.446	50.867 ± 15.420	0.121
Height (m)	1.636 ± 0.083	1.634 ± 0.083	1.641 ± 0.082	0.561
Weight (kg)	63.804 ± 12.427	64.383 ± 12.719	62.645 ± 11.819	0.312
BMI (kg/m ²)	23.683 ± 3.421	23.943 ± 3.428	23.164 ± 3.370	0.106
Sex (male/female)	126 (56.0%)/99 (44.0%)	89 (59.3%)/61 (40.7%)	37 (49.3%)/38 (50.7%)	0.154

Data are mean ± standard deviation values or number (%) as indicated.

BMI, body mass index.

Fig. 2. Decision tree of the simple model. F, female; M, male.

using Student's *t* test for normal data and abnormal data. Categorical variables were compared using Chi-square χ^2 test for different patient groups. A value of $p < 0.05$ indicates statistical significance.

There are 20 attributes in the original dataset. Five noninvasive features were extracted by feature selection, and their definition and basic descriptive statistics are shown in Table 1. There are 97 patients (43.1%) with colon adenoma (abnormal) and the remaining 128 patients (56.9%) without colon adenoma (normal). The prevalence rate of colon adenoma in participants aged 50–79 years is 36.89% in this study. From the result of the significance test between the normal and colon adenoma groups in Table 1, one can find that four noninvasive features (age, weight, body mass index, and sex) pass the significance test with $p < 0.01$. In addition, the feature of height passes the significance test with $p < 0.05$.

The records of all patients used in this study were divided into two datasets: the training dataset and the validation dataset. There were no significant differences in the clinical backgrounds between these two groups as shown in Table 2. The training dataset is used for learning the colon adenoma pattern and then generating the decision rules, and the validation dataset, which was not used to develop the model, was used to validate the results. The training set contained 150 instances (66.7% of the total records) with respect to the ratio with and without colon cancers that were randomly selected. The remaining 75 (33.3% of the total records) were retained as the validation dataset. In validation instances, the numbers of patients' records classified as "normal" and "abnormal" were 45 and 32, respectively.

PREDICTION MODELS

Based on the different complexity levels of parameters, three decision tree models were built in this study. The first model is just to input two parameters (sex and age) into the simple model, which has an overall accuracy of 75.1%. We denoted this as the "simple model" in this study. Figure 2 shows the decision tree graph of this simple

model. The second model is the intermediate model: two users' features (body weight and age) gave an overall accuracy of 85.8%. The corresponding decision tree of the intermediate model is shown in Figure 3. The third model used four parameters (age, sex, body weight and body height) in the model, which had an overall accuracy of 88.0% as demonstrated in Figure 4. We considered this as the "best model" in the study.

Two, four, and five decision rules can be generated from Figures 2–4 for the simple model, intermediate model, and best model, respectively. After three domain experts' review and verification, these decision rules were used to develop the intelligent CRC screening app.

USER SCENARIO

For a sample female user 39 years old, the three models would give the following scenarios:

1. *The simple model.* Two parameters (age and sex) were input into the simple model of the proposed app. The prediction result showed that she was a non-colon polyp subject.
2. *The intermediate model.* If she input age and body weight (64 kg) into the intermediate model, the prediction result was that she may be suffering from colon polyp with confidence = 85%, odds = 5.6, and support = 20%.
3. *The best model.* If she further input her age, sex, body weight (64 kg), and body height (150 cm) together into the best model (shown in Fig. 5), the prediction result tells her that she may be suffering from colon polyp with confidence = 100%, odds = ∞, and support = 4.9% (shown in Fig. 6). The app system suggests medical service referrals as shown in Figure 7.

The concept of customer service orientation is that every customer should receive perfect service. In order to perform better customer service, a service provider must first understand what each customer

Fig. 3. Decision tree of the intermediate model. BW, body weight.

Fig. 4. Decision tree of the best model. BMI, body mass index; BW, body weight; F, female; M, male.

Fig. 5. Best model interface 1.

thinks and what he or she wants. When a provider is screening high-risk groups of CRC potential patients, the developed app provides a series of Web link services as shown in Figure 8. The system can link to the hospital network registration system to allow users to directly seek further medical care. Users can also link directly to the associated CRC health education information to obtain more medical knowledge related to CRC. Or, the user can watch or join patient-related associations or a patient's Facebook Web site.

Discussion

Logistic regression (LR) is one of the traditional most commonly used statistical methods to identify the risk factors of colorectal adenomatous polyp in previous studies.^{40–42} It could handle both categorical and continuous independent variables and calculate the disease probability. Decision tree analysis is another exploratory technique of data mining and predictive modeling.^{36–39} Using this method, the study results represent a series of rules and identify priorities of these variables by a tree structure.^{36–39} It could generate understandable rules and provide a clear indication of which fields are most important for prediction or classification. Recently, decision tree analysis was used in defining prognostic factors in various diseases such as diabetes mellitus,⁴³ colorectal carcinoma,^{44,45} prostate cancer,⁴⁶ liver failure,⁴⁷ and chronic hepatitis C treatment.⁴⁸ To our knowledge, decision tree analysis has not been used in colorectal adenomatous polyp prediction.

Fig. 6. Best model interface 2.

692 TELEMEDICINE and e-HEALTH AUGUST 2014

Fig. 7. Medical service referrals.

In this study, several proven risk factors of colon polyps were not selected for decision tree analysis, including family history and smoking. Demographic data do not show a statistically significant association between these two risk factors and colon adenomatous polyp in this article. Therefore, family history and smoking were not selected as variables for decision tree analysis.

To compare the performances between our proposed three prediction models and the stepwise backward multivariate LR model, the performance results of these four models are presented in Table 3. The number of features is two, two, four, and four in the simple model, intermediate model, best model, and LR model, respectively. The number of decision rules is two, four, five, and one in the simple model, intermediate model, best model, and LR model, respectively. Based on the results in Table 3, the best model has the highest sensitivity, accuracy, and the area under the receiver operator characteristic curve and the lowest false-positive rate. The accuracy of the best model has a 10.76% higher predictive value than that of the LR model. The false-negative rate of the best model was 7.2% lower than that of the LR model. The false-positive rate of the best model was 13.3% lower than that of the LR model. In addition, the false-negative and false-positive rates have a significant effect in the medical prediction system. Performance analysis results demonstrated that the intermediate and best models had a better performance than the LR model.

At present, the FOBT is used as a first-line CRC screening tool in many countries.^{28,49} Annual or biennial screening with FOBT (guaiac-based or

immunochemical) could reduce both CRC and CRC-related mortality.⁵⁰ The sensitivity of guaiac-based FOBT is 25–38% for CRC and 16–31% for advanced adenoma.⁵¹ Qualitative immunochemical FOBT provided higher sensitivity for CRC than guaiac-based FOBT. Identification of colorectal adenomas by a quantitative immunochemical FOBT depends on adenoma characteristics, development threshold used, and number of tests performed. There were 11 patients with positive immunochemical FOBT, and 5 subjects were diagnosed with colorectal adenomatous polyp by colonoscopy in this study. The corresponding accuracy is 45.5%. By comparing with the performance result in Table 3, we can see that our proposed application system for colon adenoma screening in this patient group is better than immunochemical FOBT.

Conclusions and Future Work

Smartphones are now widely used over the world and could accomplish the task of information transmission or data processing with a relatively low cost. Also, smartphones are convenient, flexible, and easy to use for most people. It is clear that the potential for mobile communication to transform healthcare and clinical intervention in the community is tremendous. Until now, the Android platform has a large community of developers writing apps that extend the functionality of the smartphones. The possibility of creating simple and easy-to-download apps for smartphones has created a vibrant emerging healthcare industry. In this sense, it is a good choice to realize efficient low-cost medical care via smartphones. It would be a

Fig. 8. More information services.

Table 3. Performances of Four Prediction Models

	DECISION TREE			LOGISTIC REGRESSION MODEL
	SIMPLE MODEL	INTERMEDIATE MODEL	BEST MODEL	
FPR (%)	18.8%	14.1%	7.8%	21.0%
FNR (%)	29.9%	14.4%	17.5%	24.7%
Sensitivity (%)	70.1%	85.6%	82.5%	75.3%
Specificity (%)	81.3%	85.9%	92.2%	78.9%
Accuracy (%)	76.4%	85.8%	88.0%	77.3%
AUC	0.846	0.920	0.937	0.853
Number of features	2	2	4	4
Number of rules	2	4	5	1

AUC, area under the receiver operator characteristic curve; FNR, false-negative rate; FPR, false-positive rate.

large accomplishment to use smartphones in medical care, and it will be a milestone to market smartphone-enabled healthcare apps.

Based on medical decision rules from decision tree analysis, an intelligent app on Android smartphones for colon adenoma detection has been developed in this article. The proposed app can provide an easy and efficient way to quickly screen high-risk groups of potential CRC patients. Based on the concept of customer service orientation, the system also provides some medical treatment referrals and Web links to aid users achieve early diagnosis and treatment purposes for sufferers from CRC.

Future directions for our work include the following two points. First, the sample size is small. After patients were divided into different risk groups, some groups contained only several patients. In our analysis, we compared the model building and validation groups. The diagnostic accuracy was similar. However, a larger sample size is needed in a future study. Second, the study subjects were diagnostic-based patients in this article. Rebuilding the model of colorectal adenomatous polyp prediction by another database from the average-risk population is needed in a future study.

LIMITATIONS

This study's subjects were diagnostic-based, not population-based, patients. The prevalence rate of colorectal adenomatous polyp in general populations 50–79 years of age is around 19.7–30.2% but was 36.89% in this study, higher than in the average-risk population. Because our study was conducted on patients referred by gastroenterology and hepatology clinics, a bias of performance characteristics will exist if our results are generalized to real populations.

Acknowledgments

This work was partly supported by the National Science Council, Taiwan, Republic of China, under the grant NSC 102-2218-E-040-003.

Disclosure Statement

No competing financial interests exist. This article is not an advertisement for Android smartphones.

REFERENCES

1. Boulos MNK, Wheeler S, Tavares C, Jones R. How smartphones are changing the face of mobile and participatory healthcare: An overview, with example from Ecaalyx. *Biomed Eng Online* **2011**;10:24.
2. Blumberg SJ, Luke JV. *Wireless substitution: Early release of estimates from the National Health Interview Survey*. Hyattsville, MD: National Center for Health Statistics, **2010**:1–14.
3. Wikipedia. Smartphones. Available at <http://en.wikipedia.org/wiki/Smartphone> (last accessed November 6, 2013).
4. Mobile Future. Mobile year in review 2010. Available at www.mobilefuture.org/content/pages/mobile_year_in_review_2010?year=2010 (last accessed November 6, 2013).
5. Kailas A, Chong CC, Watanabe F. From mobile phones to personal wellness dashboards. *IEEE Pulse* **2010**;7:57–63.
6. Shankland S. Google's Android parts ways with Java industry group. Available at http://news.cnet.com/8301-13580_3-9815495-39.html (last accessed November 6, 2013).
7. Epstein Z. Android Market surpasses 500,000 published apps. Available at www.bgr.com/2011/10/21/android-market-surpasses-500000-published-apps (last accessed November 6, 2013).
8. Bonnington C. Google's 10 billion Android app downloads: By the numbers. Available at www.wired.com/gadgetlab/2011/12/10-billion-apps-detailed (last accessed November 6, 2013).
9. Tsotsis A. Google Android now on 130M total devices, with 6B app downloads. Available at <http://techcrunch.com/2011/07/14/google-android-now-on-130-million-total-devices> (last accessed November 6, 2013).
10. Farmer A, Gibson O, Hayton P. A real-time, mobile phone-based telemedicine system to support young adults with type 1 diabetes. *Inform Prim Care* **2005**;13:171–177.
11. Lee HJ, Lee SH, Ha KS, Jang HC, Chung WY, Kim JY, Chang YS, Yoo DH. Ubiquitous healthcare service using Zigbee and mobile phone for elderly patients. *Int J Med Inform* **2009**;78:193–198.
12. Chen M, Sergio GV, Athanasios VV, Cao H, Victor CML. Body area networks: A survey. *MONET* **2011**;16:171–193.
13. Noel HC, Vogel DC, Erdos JJ, Cornwall D, Levin F. Home telehealth reduces healthcare costs. *Telemed J E Health* **2004**;10:170–183.
14. Free C, Phillips G, Felix L, Galli L, Patel V, Edwards P. The effectiveness of m-health technologies for improving health and health services: A systematic review protocol. *BMC Res Notes* **2010**;3:250.
15. Blaya JA, Fraser HS, Holt B. E-health technologies show promise in developing countries. *Health Aff (Millwood)* **2010**;29:244–251.
16. Terry M. Medical apps for smartphones. *Telemed J E Health* **2010**;16:17–22.
17. Kaplan W. Can the ubiquitous power of mobile phones be used to improve health outcomes in developing countries? *Global Health* **2006**;2:9.
18. Martinez AW, Phillips ST, Carillo E, Thomas SW 3rd, Sindi H, Whitesides GM. Simple telemedicine for developing regions: Camera phones and paper-based microfluidic devices for real-time, off-site diagnosis. *Anal Chem* **2008**;80:3699–3707.
19. Jin Z, Sun Y, Cheng AC. Predicting cardiovascular disease from real-time electrocardiographic monitoring: An adaptive machine learning approach on a cell phone. *Conf Proc IEEE Eng Med Biol Soc* **2009**;2009:6889–92.

20. Tan A, Masek M. TT fetal heart rate and activity monitoring via smart phones. In: *mHealth Summit 2009*, Washington, DC, **2009**. Available at <http://research.microsoft.com/en-us/events/mhealth2009/tan-masek-presentation.pdf> (last accessed November 6, 2013).
21. Black J, Sonenberg L, Scheepers R. Cell phone technology is key to better health in Africa. Available at http://research.microsoft.com/en-us/collaboration/focus/health/smartphone_clinical_diagnosis.aspx (last accessed November 6, 2013).
22. Chen T, Kuan K, Celi L, Clifford GD. *Intelligent heartsound diagnostics on a cellphone using a hands-free kit*. Menlo Park, CA: Association for the Advancement of Artificial Intelligence, **2009**.
23. Jemal A, Siegel R, Ward E, Hao Y, Xu J, Thun MJ. Cancer statistics, 2009. *CA Cancer J Clin* **2009**;59:225–249.
24. Vogelstein B, Fearon ER, Hamilton SR, Kern SE, Preisinger AC, Leppert M, Nakamura Y, White R, Smits AM, Bos JL. Genetic alterations during colorectal tumor development. *N Engl J Med* **1988**;319:525–532.
25. Heitman SJ, Ronksley PE, Hilsden RJ, Manns BJ, Rostom A, Hemmelgarn BR. Prevalence of adenomas and colorectal cancer in average risk individuals: A systematic review and meta-analysis. *Clin Gastroenterol Hepatol* **2009**;7:1272–1278.
26. Ferlitsch M, Reinhart K, Pramhas S, Wiener C, Gal O, Bannert C, Hassler M, Kozbial K, Dunkler D, Trauner M. Sex-specific prevalence of adenomas, advanced adenomas, and colorectal cancer in individuals undergoing screening colonoscopy. *JAMA* **2011**;306:1352–1358.
27. Shaikat A, Mongin SJ, Geisser MS, Lederle FA, Bond JH, Mandel JS, Church TR. Long-term mortality after screening for colorectal cancer. *N Engl J Med* **2013**;369:1106–1114.
28. Levin B, Lieberman DA, McFarland B, Andrews KS, Brooks D, Bond J, Dash C, Giardiello FM, Glick S, Johnson D, Johnson CD, Levin TR, Pickhardt PJ, Rex DK, Smith RA, Thorson A, Winawer SJ; American Cancer Society Colorectal Cancer Advisory Group; US Multi-Society Task Force; American College of Radiology Colon Cancer Committee. Screening and surveillance for the early detection of colorectal cancer and adenomatous polyps, 2008: A joint guideline from the American Cancer Society, the US Multi-Society Task Force on Colorectal Cancer, and the American College of Radiology. *Gastroenterology* **2008**;134:1570–1595.
29. Chak A, Post AB, Cooper GS. Clinical variables associated with colorectal cancer on colonoscopy: A prediction model. *Am J Gastroenterol* **1996**;91:2483–2488.
30. Michal S, Li L, Chen Z. Lifestyle and dietary risk factors for colorectal hyperplastic and adenomatous polyps. *Cancer Epidemiol Biomarkers Prev* **2012**;21:563.
31. Hu NC, Chen JD, Lin YM, Chang JY, Chen YH. Stepwise relationship between components of metabolic syndrome and risk of colorectal adenoma in a Taiwanese population receiving screening colonoscopy. *J Formos Med Assoc* **2011**;110:100–108.
32. Liu CS, Hsu HS, Li CI, Jan CI, Li TC, Lin WY, Lin T, Chen YC, Lee CC, Lin CC. Central obesity and atherogenic dyslipidemia in metabolic syndrome are associated with increased risk for colorectal adenoma in a Chinese population. *BMC Gastroenterol* **2010**;10:51.
33. Tseng MH, Liao HL. The genetic algorithm for breast tumor diagnosis—The case of DNA viruses. *Appl Soft Computing* **2009**;9:703–710.
34. Tseng MH, Hsu HC, Chang CC, Ting H, Wu HC, Tang PH. Development of an intelligent app for obstructive sleep apnea prediction on Android smartphone using data mining approach. *Ubiquitous Intelligence & Computing and 9th International Conference on Autonomic & Trusted Computing (UIC/ATC)*. Piscataway, NJ: IEEE, **2012**;774–779.
35. Han J, Kamber M. *Data mining: Concepts and techniques*, 2nd ed. San Francisco: Morgan Kaufmann, **2006**.
36. Using data mining for business intelligence. Available at www.aspfree.com/c/a/MS-SQL-Server/Using-Data-Mining-for-Business-Intelligence/ (last accessed November 6, 2013).
37. Rokach L, Maimon O. Top-down induction of decision trees classifiers—A survey. *IEEE Trans Syst Man Cybernet C* **2005**;35:476–487.
38. Tang Z, MacLennan J. *Data mining with Sql server 2005*. Hoboken, NJ: John Wiley & Sons, **2005**.
39. Bozkir AS, Sezer EA. *Predicting food demand in food courts by decision tree approaches*. *Procedia Computer Science* **2011**;3:759–763.
40. Lieberman DA, Prindiville S, Weiss DG, Willett W. Risk factors for advanced colonic neoplasia and hyperplastic polyps in asymptomatic individuals. *JAMA* **2003**;290:2959–2967.
41. Brazer SR, Pancotto FS, Long TT 3rd, Harrell FE Jr, Lee KL, Tyor MP, Pryor DB. Using ordinal logistic regression to estimate the likelihood of colorectal neoplasia. *J Clin Epidemiol* **1991**;44:1263–1270.
42. Imperiale TF, Wagner DR, Lin CY, Larkin GN, Rogge JD, Ransohoff DF. Risk of advanced proximal neoplasms in asymptomatic adults according to the distal colorectal findings. *N Engl J Med* **2000**;343:169–174.
43. Miyaki K, Takei I, Watanabe K, Nakashima H, Omae K. Novel statistical classification model of type 2 diabetes mellitus patients for tailor-made prevention using data mining algorithm. *J Epidemiol* **2002**;12:243–248.
44. Valera VA, Walter BA, Yokoyama N, Koyama Y, Iai T, Okamoto H. Prognostic groups in colorectal carcinoma patients based on tumor cell proliferation and classification and regression tree (CART) survival analysis. *Ann Surg Oncol* **2007**;14:34–40.
45. Zlobec I, Steele R, Nigam N, Compton CC. Clin, a predictive model of rectal tumor response to preoperative radiotherapy using classification and regression tree methods. *Cancer Res* **2005**;11:5440–5443.
46. Garzotto M, Beer TM, Hudson RG, Peters L, Hsieh YC, Barrera E. Improved detection of prostate cancer using classification and regression tree analysis. *J Clin Oncol* **2005**;23:4322–4329.
47. Nakayama N, Oketani M, Kawamura Y, Inao M, Nagoshi S, Fujiwara K, Tsubouchi H, Mochida SJ. Algorithm to determine the outcome of patients with acute liver failure: A data-mining analysis using decision trees. *Gastroenterology* **2012**;47:664–677.
48. Kurosaki M, Sakamoto N, Iwasaki M, Sakamoto M, Suzuki Y, Hiramatsu N, Sugauchi F, Yatsuhashi H, Izumi NJ. Pretreatment prediction of response to PEGinterferon plus ribavirin therapy in genotype 1 chronic hepatitis C using data mining analysis. *Gastroenterology* **2011**;46:401–409.
49. Sung JJ, Lau JY, Young GP, Sano Y, Chiu HM, Byeon JS, Yeoh KG, Goh KL, Sollano J, Rerknimitr R, Matsuda T, Wu KC, Ng S, Leung SY, Makharia G, Chong VH, Ho KY, Brooks D, Lieberman DA, Chan FK; Asia Pacific Working Group on Colorectal Cancer. Asia Pacific consensus recommendations for colorectal cancer screening. *Gut* **2008**;57:1166–1176.
50. Hewitson P, Glasziou P, Watson E, Towler B, Irwig L. Cochrane systematic review of colorectal cancer screening using the fecal occult blood test (hemoccult): An update. *Am J Gastroenterol* **2008**;103:1541–1549.
51. Whitlock EP, Lin JS, Liles E. Screening for colorectal cancer: A targeted, updated systematic review for the U.S. Preventive Services Task Force. *Ann Intern Med* **2008**;149:638–658.

Address correspondence to:
Ming-Hseng Tseng, PhD
School of Medical Informatics
Chung-Shan Medical University
No. 110, Sec. 1, J. K. N. Road
Taichung, Taiwan, Republic of China

E-mail: mht@csmu.edu.tw

Received: August 17, 2013

Revised: November 6, 2013

Accepted: November 6, 2013