TITLE PAGE

Machine Learning and Prediction of All-Cause Mortality among Chinese Older Adults

Xurui Jin [1,8*] Yiyang Sun [1], Tinglong Zhu [1], Yu Leng [1], Shuyi Guan[1], Kehan Zhang [1], Kangshuo Li [8,9], Zhangming Niu [8,9], Chenkai Wu [1], Yi Zeng [4,5], Yao Yao [4], Lijing L. Yan[1,3,6,7*]

1. Global Health Research Center, Duke Kunshan University, Kunshan, Jiangsu, China
2. School of Population and Global Health, University of Melbourne, Melbourne, Australia
3. Duke Global Health Institute, Duke University, Durham, NC, United States of America
4. Center for Healthy Aging and Development Studies, National School of Development, Peking University, Beijing, China
5. Center for the Study of Aging and Human Development and Geriatrics Division, Medical School of Duke University, Durham, North Carolina, USA United States of America
6. School of Global Health and Development, Peking University, Beijing, China
7. School of Health Sciences, Wuhan University, Wuhan, Hubei, China
8. MindRank AI, Hangzhou, China
9. Department of statistics, Columbia University GSAS, NY, US

* Corresponding authors:

Professor Lijing L. Yan.
Global Health Research Center,
Duke Kunshan University,
Jiangsu Province, China.
E-mail: lijing.yan@duke.edu

Dr. Xurui Jin
Global Health Research Center,
Duke Kunshan University,
Jiangsu Province, China.
E-mail: xurui.jin@dukekunshan.edu

Abstract

Background and aim: Mortality risk stratification was vital for targeted intervention. This study aimed at building the prediction model of all-cause mortality among Chinese dwelling elderly with different methods including regression models and machine learning models and to compare the performance of machine learning models with regression model on predicting mortality. Additionally, this study also aimed at ranking the predictors of mortality within different models and comparing the predictive value of different groups of predictors using the model with best performance.

Method: We used data from the 2008 and 2011 waves of Chinese Longitudinal Healthy Longevity Survey (CLHLS). The follow-up of CLHLS was conducted every 2-3 years till 2018. The analysis sample included 2,448 participants. We used totally 117 predictors to build the prediction model for survival including 61 questionnaire, 41 biomarker and 15 genetics predictors. Four models were built (XG-Boost, random survival forest [RSF], Cox regression with all variables and Cox-backward). We used C-index and integrated Brier score to evaluate the performance of those models.

Results: The XG-Boost model and RSF model shows slightly better predictive performance than Cox models and Cox-backward models based on the C-index and integrated Brier score in predicting surviving. Age, activity of daily living and Mini-Mental State Examination score were identified as the top 3 predictors in the XG-Boost and RSF models. Biomarker and questionnaire predictors have a similar predictive value, while genetic predictors have no addictive predictive value when combined with questionnaire or biomarker predictors.

Conclusion: In this work, it is shown that machine learning techniques can be a useful tool for both prediction and its performance sightly outperformed the regression model in predicting survival.
Keywords: Machine learning, regression, mortality prediction, older adults

Introduction

Currently, aging is a global phenomenon, and in particular, the proportion of older individuals in low and middle-income countries (LMIC) is growing in an accelerated pace due to the increasing average life span. Among LMICs, China shows the fastest rate of aging, and the related health care cost has been growing rapidly. According to the UN statistics database, China's dependency ratio for retirees could rise as high as 44% by 2050 which would then be the highest around the world [1]. As the population is aging quickly, it is important to have adequate identification and risk stratification for individuals with reduced life expectancy. Adequate risk stratification of mortality can lead to a more precise intervention and more targeted care. It also has important clinical relevance as an accurate mortality risk assessment tool might improve the accuracy of the prognostic assumptions which in turn can influence clinical decisions [2, 3].

Most of the current estimation methods available for mortality prediction are based on a single independent factor including blood pressure [4], body weight [5], walking speed [6], self-reported health [7] and frailty [8]. There are some indices consisting of several predictors for short-term mortality, but they have mainly been developed for and assessed in older individuals or in high-risk populations [3]. To be more specific, a large number of clinical conditions were assessed their relevance in the prediction of 1-year mortality and they were summarized into the Charlson Comorbidity Index (CCI) [9], which has been validated in large populations and widely used to predict mortality among hospitalized patients. However, since more data, including biomarkers and genetic assessments, became available, a systematic approach has become a trend in medical and public health studies. However, it is still unclear whether only using several or only a selected group of predictors for prediction has limitations in providing accurate mortality risk stratification. Additionally, an inaccurate mortality risk stratification may lead to an untargeted mortality prevention strategy [2].

Following the fast-paced development of biomedical technology, more and more high dimensional data became available in clinical and public health studies [10]. These data include genetics, metabolomics, and proteomics. Additionally, the number of predictors in epidemiological studies were also increasing. Normally, hundreds of predictors are available for developing mortality prediction models in epidemiological studies or electronic medical record data. The core of traditional statistical methods is hypothesis testing and this process is user-driven which indicated that the researcher would have to specify dependent variables, regression family and link and interaction type. Therefore, user intervention may influence the results of those models. Another methodological concern was that the traditional statistical techniques such as regression models are often limited by the correlation between variables, nonlinearity of variables, and the possibility of overfitting [11, 12]. The newly developed machine learning method has provided the possibility

the address those challenges. With machine learning techniques, the hypothesis is that the associations between various predictors and the outcome were a pattern [13]. Thus, it is a hypothesis-free method. Machine learning models can analyze all predictor variables in a way that prevents overlooking potentially important predictor variables even if it was unexpected. Accordingly, it is reasonable that machine learning method could be an effective method in identifying the best predictors of outcomes from a large number of predictors.

Most of the prediction models are based on regression model and the researcher need to manually input the predefined interactions, such as the interaction between vitamin D and albumin on mortality I have demonstrated [14]. Missing those complex interactions in the regression model may result in an inaccurate prediction of outcomes. In the prediction model developed by machine learning methods, the model can automatically identify those interactive relationship from the data and it is unnecessary to specify interactions [11, 12]. However, it is still widely debated whether the performance of machine learning model was better than that of regression model and recently it was a very hot issue to demonstrate the usefulness of these methods.

Recently, a systematic review that included 71 from 927 searched studies suggested that some common-used machine learning methods were classification trees, random forests, artificial neural networks, and support vector machines [12]. It was found that the difference in logit between logistics regression and machine learning was 0.00 (95% CI: −0.18, 0.18). Moreover, a meta-analysis published in Nov. 2020 suggested that machine learning models provide better discrimination in mortality prediction after cardiac surgery [11]. However, among those studies included in those two reviews, the number of predictors and the sample has a very large variation. Therefore, more studies are warranted to compare those two methods in different cohorts and different populations.

The overarching goal of this study is to provide mortality risk stratification tools for the Chinese older adults. Specifically, there are three major research objectives and their sub-objectives of this study:

1. Using the data from CLHLS with maximal number of predictors to build the survival prediction model for Chinese older adults using different methods including regression models (Cox and Cox-backward) and machine learning models (random survival forest [RSF] and XG-boost) and to compare the performances of machine learning models with regression models in predicting survival.
2. To rank the predictors of mortality within different survival models.
3. To compare the predictive values of different groups of predictors using the best performing model

Method

Study population

The present study uses data from the Chinese Longitudinal Healthy Longevity Survey (CLHLS). CLHLS is a longitudinal study since 1998 with follow-up surveys every 2-3 years till 2018. The CLHLS surveys were conducted in randomly selected counties and cities in China, which accounted for half of the counties and cities in 23 out of 31 provinces covering over 85% of China's population. Based on gender and place of residence (ie, living in the same street, village, city or county) for a given centenarian, randomly selected octogenarians and nonagenarians were also sampled. This matched recruitment procedure resulted in an oversampling of the oldest old and older men. In the CLHLS, a weight of age-sex urban/rural residence in the sample with the distribution of the total population in the sampled 22 provinces was employed to reflect the unique sampling design. More details of this survey have been published elsewhere [15]. In this study, we derived data from the CLHLS biomarker study. The CLHLS biomarker study was conducted in eight longevity areas [16]. This study covers similar domains that CLHLS has investigated and shared the sampling strategy as CLHLS. Participants with age less than 65 years old (n=51) and with missing values in higher than 30% of the predictors were excluded (n=33) and my analyses included 2,448 elderly aged 65 years or over who had both phenotypic, biomarker, and genotypic measurements.

The Ethics approval of CLHLS biomarker study was obtained from the Research Ethics Committees of Peking University and Duke University. All participants or their legal representatives signed written consent forms in the baseline and follow-up surveys.

Predictors and imputation techniques

After excluding those variables with missing value higher than 30%, 117 predictors were included in this study including demographic variables, lifestyle (smoking, drinking, diet and physical activities), health indicators (cognitive function, activity of daily living and leisure activities), comorbidities, biomarkers and genetic information.

The questionnaire data were collected through in-home interviews by trained interviewers who are local staff members from the county-level network system of the National Bureau of Statistics of China. All interviewers have received 12+ years of schooling, and most have earned a college degree. Each interviewer was accompanied by a local doctor, a nurse, or a medical college student so that some health check-ups could be performed. In the physical examination, body weight and height were measured by trained medical staff using a standardized protocol. Totally 61 questionnaire predictors were included in the analysis.

The CLHLS collected blood and urine sample since 2008. During the investigations, blood samples were centrifuged within 1 hour after collection and heparin anticoagulant blood samples were centrifuged at 3000 rpm for 10 min at 18°C–25°C. Then blood and urine samples are immediately stored at −80°C in the local

Center of Disease Control and Prevention (CDC). And then the sample were transported at −20°C with transport cases provided by CCDC by specially assigned persons to designated testing units. A variety of blood and urine biomarker were measured and all laboratory analyses are conducted by the central clinical laboratory at Capital Medical University in Beijing. The protocol of biomarker measurement was published elsewhere. In this study, totally 41 biomarker predictors were included. [16]

The genotyping was performed by a customized chip targeting about 287,898 candidate SNPs associated with longevity, chronic disease or health indicators based on multiple studies. There were no familial/kinship relations among the participants within and across different waves. Beijing Genomics Institute (BGI) performed the genotyping, and the BGI genotyping quality control procedures of the genetics study have been published elsewhere. [17] In this study, totally 15 SNPs were selected by previous meta-Genome wide association studies (GWAS) focusing on longevity and our previous GWAS study using the same data [17-20]. All the predictors are listed in Table 1.

Overall, there are 4.7% missing data. We use the *missForest* algorithm to impute those missing values [21]. This method is a nonparametric imputation method that builds a random forest model for each variable and it was demonstrated that this method outperformed many imputation methods especially in data settings where complex interactions and non-linear relations are existing.

Data on mortality

Vital status and date of death were collected from officially issued death certificates when available or otherwise from the next-of-kin or local residential committees who were familiar with the decedents. Duration of follow-up was calculated by the time interval between the first interview date and date at death. Survivors at the last wave (2018) were censored at the time of the last survey.

Statistical analysis

Totally four models for predicting survival were built: 1. The RSF model, 2. the XG-boost model, 4. the Cox model with all predictors, 4. the COX-backward model. For the analysis using data from CLHLS, I build the XG-boost model to predict the two years' mortality.

In order to train and validate the models and optimize the machine learning models, I used 10-fold internal cross validation. The training data was randomly divided into ten folds, and each time, nine folds of data were included in the training model and the rest one fold of data was used as the testing test. This process was repeated ten times for all combinations of folds. I used the grid search to determine the hyper-parameters of machine learning model.

To evaluate the performance of those models, I use Harrell's concordance index (C-index) and integrated brier score (IBS) for the survival model. Higher C-index and lower IBS/BS indicate better discrimination and calibration performance.

Cox proportional hazard regression models and Cox-backward models: In this survival analysis, the focus is on the period till the occurrence of mortality. The Cox proportional hazards model is usually used to estimate the hazard ratio of the interested factors on the outcome. I build two Cox models. One is the Cox with all predictor and the other is the Cox model with a backward elimination model (Cox-backward). The Cox-backward model was widely used for variable selection. I used R (version 3.6.1) to build the Cox model.

XG-boost: In the survival analysis and the analysis of predicting two year's mortality, I built the extreme gradient boosting (XG-Boost), an ensemble machine learning method based on decision trees, to establish the prediction model for mortality with all the predictors. Gradient boosting is a machine learning model that involves combinations of prediction models into a strong model. The XG-Boost approximates the value of the loss function with the second-order Taylor series and reduces the probability of overfitting by regularization.

Some hyper-parameters parameters of XG-Boost model was defined as following: 1) number of trees: 400, learning rate: 0.005, 2) minimal loss to expand on a leaf node: 0; 3) maximum tree depth: 4, 4) subsample proportion: 1. Additionally, the XG-Boost model can provide the estimations of feature importance from the trained model. In this study, I used F scores in XG-Boost model to evaluate the feature importance which is the sum of Gini index among the corresponding splits in a tree and further averaged among all the trees. Python 3.7 was used to build this model.

RSF: In the survival analysis, I build the RSF model with all predictors. RSF designed for time-to-event data such as survival as a transformed RF. Some hyper-parameters parameters of RSF model were defined as following: 1) number of trees: 1000, number of random split points used to split a node: 5, 2) mtry value: 12; 3) node size: 50, 4) block size: 5. I applied variable importance (VIMP) to ranking the predictors. The absolute value of VIMP indicates the impact of this predictor on the overall performance. The positive VIMP value indicates the predictor improves predictive performance and negative value indicates that the predictor has negative effect on the performance of the prediction model. I used "randomForestSRC" and "cph" in R (version 3.6.1) to perform those analysis.

Results

Baseline characteristics

The median follow-up period was 3.6 years (range: 0.2-9.9 years). Totally 43.8% (n=1,071) of all participants died during the follow up. Table 1 presents the part of baseline characteristics by survival status at the last follow up. The mean age was 84.3 years (SD: 13.7). Participants who were survival are more likely to be younger, male, without impaired activity of daily living, married, with higher MMSE score, social activity score and psychological wellbeing score, with higher levels of album, 25-hydroxyvitamin D, red blood cells and hemoglobin (Ps<0.05).

Comparisons between models

Two Cox models were built: a) a model with all 117 predictors and a Cox model with backward selection. Furthermore, two machine learning models were built: a) a RSF model, b) XG-Boost model.

Table 3 presented the performance of the four models on test data. Regarding the IBS, the XG-Boost models have the lowest (IBS = 0.120) followed by the RSF (IBS = 0.128). Cox models have sightly higher IBS (COX with all variables: IBS = 0.131; Cox-backward: 0.129). In terms of C-index, the XG-Boost model has the highest C-index (C-index: 0.80), and the Cox models with all variables has slightly worse performance (Cox model with all variables: C-index: 0.77; Cox backward model: C-index: 0.78). C-index for Cox backward and RSF models are nearly the same (C-index: 0.78).

Figure 1 shows the average prediction Brier error over time for those four models (XG-Boost, RSF, Cox with all variables and Cox-backward). Only small differences can be observed between Cox and Cox-backward models. The XG-Boost model achieved better performance than the other models, but only sightly better than the RSF model.

Predictor ranking

Hazard ratios of the 5 most predictive variables for the Cox models with all predictors are shown in **Table 2** which was selected by the z-score values. The strongest predictor is chronological age. One-year increase in age increased a 5% higher mortality risk. The other most predictive variables are mean corpuscular hemoglobin concentration (One-unit increase: HR: 1.005, 95% CI: 1.002, 1.008), gender (male: 1.66, 95% CI: 1.23, 2.24), activity of daily living (impaired: HR: 1.40, 95% CI: 1.11, 1.77), urine microalbumin (one-unit increase: HR: 1.002, 95% CI: 1.000, 1.003). In the Cox backward model, after the backward selection, there are only five variables left which are urine microalbumin (HR: 1.003, 95% CI: 1.002, 1.004), MMSE score (HR: 0.98, 95% CI: 0.97, 0.99), age, (HR: 1.06, 95% CI: 1.05, 1.07), albumin (HR: 0.93, 95% CI: 0.90, 0.95), 25-hydroxyvitamin D (HR: 0.98, 95% CI: 0.97, 0.98). The predictor with highest predictive value was urine microalbumin.

Table 3 presented the top-10 predictors in the RSF and XG-Boost models. The importance of predictors is assessed by F score in the XG-boost models and VIMP in RSF. The strongest predictor is age in the two models. In those top predictors, age, MMSE score, activity of daily living, social activity score, 25-hydroxyvitamin D, albumin, marital status and hemoglobin were both identified in those two models. “Red blood cell” was identified as high predictive value in RSF model (VIMP: 0.0023) and “self-reported health” was identified by XG-Boost model (F score: 0.0018).

Comparisons between groups of predictors

The predictors were grouped as “questionnaire variables”, “biomarker variables” and “genetic variables” and I put different groups of predictors in the XG-Boost model which has the best predictive performance.

When solely add “questionnaire variables”, “biomarker variables” or “genetic variables” into the model, the C-index for each XG-Boost model was 0.74 (95% CI: 0.68, 0.80), 0.77 (0.71, 0.83) and 0.54 (95% CI: 0.48, 0.60) respectively. After adding the genetic variables into the XG-Boost model with questionnaire or biomarker variables, the predictive value did not improve and the C-indexes before and after adding the genetic variables was nearly the same.

Discussion

In this large, prospective cohort study, with the data form CLHLS biomarker cohort (N=2,448, 117 predictors), I did an extensive analysis of mortality prediction model which included over one hundred predictors with all-cause mortality followed up to ten years. I have ranked the predictive value of those predictors from questionnaire, biomarker and genetic assessment.

Several key messages can be concluded from this study. First, the performance of machine learning model was sightly better than the regression model in predicting survival. Second, predictors that can simply be obtained by interview without blood testing can predict all-cause mortality with a high predictive value among the older population. Age, activity of daily living and MMSE score were the strongest predictors. Third, generally, the predictive value of biomarker predictors was highest, and the predictive value of questionnaire predictors was comparable, however the predictive value of genetic predictors was low, and it did not increase the performance of prediction model when combined with other groups of predictors. In this report, I have presented only part of my findings, and a detailed version of my results are available in an open access database where the detailed predictive value for each variable in each model was available to generate new research hypotheses for other researchers.

In this study regression and machine learning models were applied for predicting all-cause mortality among Chinese community-dwellings within two cohorts. Generally, I found that machine learning models outperformed the regression model in predicting mortality in terms of C-index and IBS, but the difference was quite small. The results indicated that machine learning methods are well suited for meaningful risk prediction in large-scale epidemiological studies. Theoretically, compared with regression models, machine learning methods can avoid the problem of overfitting and non-convergence, and also considering the non-linearities. However, it is still controversial that whether machine learning could improve the accuracy of mortality risk stratification. In some study with relatively small sample size and limited number of predictors, the regression models have a comparable performance as the machine learning models. In one study including 1701 men with a follow up of 3 years and eight predictor available [22], it found that in predicting survival, machine learning models did not have a better performance as the regression models (C-static: Cox-regression: 0.78; survival tree: 0.71; binary tree: 0.66; logistics regression: 0.72) In another study including 603 patients from the hospital with ST elevation myocardial infarction, using 10-fold cross validation, the logistics regression

achieves the highest C-statistics of 0.82 and outperformed decision tree, naive Bayes classifier, artificial neural network and Bayesian network classifier . However, in some studies with larger simple size and larger numbers of predictor, machine learning models outperformed the regression model. For instance, a study including 6,520 patient with 66 predictors compared the performance of logistic regression model and different machine learning models on predicting the mortality in-hospital after elective cardiac surgery [23]. Four different machine learning models have been evaluated with the regression models: gradient boosting machine, random forest, support vector machine and naive bayes. The area under the ROC curve for the machine learning model (C-index = 0.80) was significantly higher than the logistic regression model (C-index = 0.742). Similar result was found in another study with a large size of simple and predictors, a study derived data from the Multi-Ethnic Study of Atherosclerosis (MESA) including 6814 participants aged 45 to 84 years [24]. Seven-hundred thirty-five variables from imaging and noninvasive tests, questionnaires, and biomarker panels were used to build the RSF and Cox model. In predicting all-cause mortality, the RSF model has a higher C-index (0.86) compared with that of the Cox regression models (AIC-Cox with forward selection: 0.78; LASSO-Cox: 0.80). To conclude, it is plausible that when the sample size was small and the number of predictors, the performance of machine learning model seems comparable with the regression model, while when the sample size was big enough and the number of predictors was larger enough, the performance of machine learning model will outperform the regression model. Combined with previous evidence, our analysis with the sample of over two thousand elderly with over one hundred predictors also partly demonstrated this hypothesis. However, further methodology studies were warranted to investigate in what situation the machine learning models would outperform the regression model. Additionally, the methodology development of the interpretative machine learning method is also needed to understand how those interactive relationships influent the predictive performance of the models. As we move into the age of precision medicine, understanding the use of phenotypic data and methods to analyze already acquired information is of paramount importance.

Previous studies were mainly focus on the impact of signal predictors and could not have a comparison among those predictors, while our analysis provided the information about the relative importance of each variable as predictor of all-cause mortality. Several previous studies on mortality have ranked the predictors from across domains. In a study using data from UK biobank, 655 predictors were evaluated on their predictive value of five-year mortality in nearly 500,000 adults aged over 50 years old [25]. Those predictors included blood biomarkers, disease histories, socio-demographics, early life health factors and family history, psychosocial factors, and healthy lifestyle. Among those predictors, self-reported health was the strongest predictors of all-cause mortality. In another study derived data from Heath and Retirement Study [26], totally 57 predictors of adverse socioeconomic and psychosocial experiences during childhood, socioeconomic

conditions, health behaviors, social connections, psychological characteristics, and adverse experiences during adulthood was evaluated. Smoking was the strongest predictor of mortality among those 13,611 American aged from 52 to 104 years old. Of note, in those two studies examined the predictive value of a comprehensive groups of predictors, age was regarded as covariate. In another study using machine learning methods the most powerful predictor for all-cause mortality was the chronological age. Additionally, in that study, some novel biomarkers were identified as tissue necrosis factor-α soluble receptor and interleukin-2 soluble receptor [24]. In my study, I have also identified some other predictors need to be addressed such as activity of daily living as a measurement of activity of daily living, blood pressure and MMSE score as a measurement of cognitive function which was previously underestimated in the mortality risk prediction models. This results were corresponding with the findings of UK biobank study which indicate that some general health indicator may be the most potent mortality predictor [25]. To note, the importance of cognitive health and blood pressure control were addressed in my study. In China, A large number of elderly people are expected to suffer varying degrees of cognitive impairment and hypertension in the future.[27] My findings along with evidence on the lack of adequate care for cognitive impairment and blood pressure control in China indicate that there is an urgent need for China's health care system and government to improve provision and quality of these services. My study also suggested that some biomarkers such as red blood cells, hemoglobin and urine microalbumin which was rarely studied also need to be addressed in the future mortality risk stratification studies. In terms of those finding, machine learning enabled the researcher to discover new relationships without prior without prior assumptions. Identifying effective mortality risk predictor may be of benefit for effective screening strategies and suggest specific targets for risk reduction. Additionally, although I have identified some predictors which seems not actionable on the personal level such as cognitive function and social and leisure activity. But this should not be interpreted as an impediment to improvement of health behaviors. Previous studies have demonstrated that healthier modifiable behaviors such as quit smoking, higher level of physical activities and healthier diet are beneficial and further reduce the risk of mortality.

Another finding of my analysis was that the predictive value of questionnaire and biomarker predictors was comparable, and the predictive value of genetic predictor was low. Additionally, adding genetic variables into the prediction model with biomarker or questionnaire predictors did not substantially improve its performance. Our results were corresponding with another study which also suggested that the predictive value of genetic predictors was limited. The study included 5,974 participants from the Rotterdam Study [28], followed for a median of 15.1 years, and it has demonstrated that specific genetic factors were independently associated with mortality, jointly they contributed little to mortality prediction (C-index = 0.56). Combined with those previous studies, it is possible that common SNPs only have very limited predictive power of mortality or longevity, when comparing with those traditional predictors. Considering its limited predictive

power, it may be not necessary to perform the genetic assessment when evaluating the mortality risk. Although my results suggested that it may be unnecessary or invalid to assess the individual's genetic background in mortality risk prediction, it is still need to be validated for some specific disease risk prediction such as the *APOE* gene in predicting the risk of dementia and *BCL2* gene in predicting breast cancer. Additionally, it may be possible that some epigenetic biomarker such as the methylation in genomes may improve the predictive performance of prediction models, however the cost of such measurement was much higher and the cost-effectiveness needs to be considered.

This is the first study where machine learning models are applied to data from Chinese dwelling elderly in predicting all-cause mortality where a comparison with the traditional Cox model was also performed. My study has some implications for at the individual, clinical or policy making level especially for those at the LMICs with a rapid pace of aging.

At the individual level, it can be applied to address the dwelling's self-awareness of the health. For example, taking more health behaviors such as quit smoking and alcohol drinking to prevent cognitive impairment and control the blood pressure. At the clinical and community level, the physician or health worker might use the model to identify individuals who were at high risk of mortality. To note, currently, frailty, as a measure consisting nearly one hundred variables, was gradually used in the real-world settings to perform the mortality risk stratification. However, it might be possible that my model which would be more convenient, as the development of medical informatics, could be applied to evaluate the mortality risk in different settings and may outperformed the frailty measurement. With such mortality risk stratification, targeted specific interventions or treatment can be implemented to those individuals and further improve the quality of healthcare. Finally, policy maker can use this information to allocate more medical or public health recourse to decrease the burden of specific risk factors.

My study has several limitations. Firstly, the sample size is limited of this study, but, it has included a wide range of predictors. It is difficult to find another cohort with such big numbers of predictors. Secondly, the accuracy of the predictor. I used many self-reported predictors and those predictors are always subject to misclassification bias. However, the data of CLHLS has a good overall quality and its reliability and validity were validated. Thirdly, the genetic information was selected from the GWAS studies, which may not fully represent the genetic risk of mortality, however, I extracted the genetic information according to the published GWAS study using the same data, which may reduce the under estimation of the genetic mortality risk. Fourth, some biomarker was not available in my data such as the Interleukin 6 or Tumour Necrosis Factor alpha which represented level of inflammation, however, in our data, there are other biomarker measured the inflammation level such as albumin and high-sensitivity C-reactive protein. Fifthly, I use the internal validation

method which is the training and test data sets were both drawn from the CLHLS study population. An external validation with data from other cohorts was needed.

Conclusions

In this population-based study, I built the prediction models with machine learning models and regression model and I found that machine learning model sightly outperformed the regression model in predicting all-cause mortality. Age, MMSE score and ADL were three of the most important predictors. Additionally, I have compared the predictive value of different groups of predictors and found that the predictive value of genetic predictors was much lower compared with those predictors from questionnaire and biomarker assessment. I provide a framework for big data applications to obtain meaningful risk prediction and generate data-driven hypotheses.

Reference

1. Cohen J. World population in 2050: assessing the projections. Conference Series ; [Proceedings]. 2001.
2. Bell SP, Saraf A. Risk stratification in very old adults: how to best gauge risk as the basis of management choices for patients aged over 80. Prog Cardiovasc Dis. 2014;57(2):197-203. Epub 2014/09/14. doi: 10.1016/j.pcad.2014.08.001. PubMed PMID: 25216619; PubMed Central PMCID: PMCPMC4174544.
3. Yourman LC, Lee SJ, Schonberg MA, Widera EW, Smith AK. Prognostic indices for older adults: a systematic review. JAMA. 2012;307(2):182-92. Epub 2012/01/12. doi: 10.1001/jama.2011.1966. PubMed PMID: 22235089; PubMed Central PMCID: PMCPMC3792853.
4. Ekbom T, Lindholm L, Oden A, Dahlof B, Hansson L, Schersten B, et al. Blood pressure does not predict mortality in the elderly. J Hypertens Suppl. 1988;6(4):S626-8. Epub 1988/12/01. doi: 10.1097/00004872-198812040-00196. PubMed PMID: 3241270.
5. Marti S, Munoz X, Rios J, Morell F, Ferrer J. Body weight and comorbidity predict mortality in COPD patients treated with oxygen therapy. Eur Respir J. 2006;27(4):689-96. Epub 2006/04/06. doi: 10.1183/09031936.06.00076405. PubMed PMID: 16585077.
6. Blain H, Carriere I, Sourial N, Berard C, Favier F, Colvez A, et al. Balance and walking speed predict subsequent 8-year mortality independently of current and intermediate events in well-functioning women aged 75 years and older. J Nutr Health Aging. 2010;14(7):595-600. Epub 2010/09/08. doi: 10.1007/s12603-010-0111-0. PubMed PMID: 20818476.
7. Miilunpalo S, Vuori I, Oja P, Pasanen M, Urponen H. Self-rated health status as a health measure: the predictive value of self-reported health status on the use of physician services and on mortality in the working-age population. J Clin Epidemiol. 1997;50(5):517-28. Epub 1997/05/01. doi: 10.1016/s0895-4356(97)00045-0. PubMed PMID: 9180644.
8. Stow D, Matthews FE, Barclay S, Iliffe S, Clegg A, De Biase S, et al. Evaluating frailty scores to predict mortality in older adults using data from population based electronic health records: case control study. Age Ageing. 2018;47(4):564-9. Epub 2018/03/17. doi: 10.1093/ageing/afy022. PubMed PMID: 29546362; PubMed Central PMCID: PMCPMC6014267.
9. Christensen S, Johansen MB, Christiansen CF, Jensen R, Lemeshow S. Comparison of Charlson comorbidity index with SAPS and APACHE scores for prediction of mortality following intensive care. Clin Epidemiol. 2011;3:203-11. Epub 2011/07/14. doi: 10.2147/CLEP.S20247. PubMed PMID: 21750629; PubMed Central PMCID: PMCPMC3130905.
10. Subramanian I, Verma S, Kumar S, Jere A, Anamika K. Multi-omics Data Integration, Interpretation, and Its Application. Bioinform Biol Insights. 2020;14:1177932219899051. Epub 2020/02/23. doi: 10.1177/1177932219899051. PubMed PMID: 32076369; PubMed Central PMCID: PMCPMC7003173.
11. Benedetto U, Dimagli A, Sinha S, Cocomello L, Gibbison B, Caputo M, et al. Machine learning improves mortality risk prediction after cardiac surgery: Systematic review and meta-analysis. J Thorac Cardiovasc Surg. 2020. Epub 2020/09/10. doi: 10.1016/j.jtcvs.2020.07.105. PubMed PMID: 32900480.
12. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol. 2019;110:12-22. Epub 2019/02/15. doi: 10.1016/j.jclinepi.2019.02.004. PubMed PMID: 30763612.
13. Rajkomar A, Dean J, Kohane I. Machine Learning in Medicine. N Engl J Med. 2019;380(14):1347-58. Epub 2019/04/04. doi: 10.1056/NEJMra1814259. PubMed PMID: 30943338.
14. Jin X, Xiong S, Ju SY, Zeng Y, Yan LL, Yao Y. Serum 25-Hydroxyvitamin D, Albumin, and Mortality Among Chinese Older Adults: A Population-based Longitudinal Study. J Clin Endocrinol Metab. 2020;105(8). Epub 2020/06/06. doi: 10.1210/clinem/dgaa349. PubMed PMID: 32502237.
15. Zeng Y, Feng Q, Gu D, Vaupel JW. Demographics, phenotypic health characteristics and genetic analysis of centenarians in China. Mech Ageing Dev. 2017;165(Pt B):86-97. Epub 2017/01/04. doi: 10.1016/j.mad.2016.12.010. PubMed PMID: 28040447; PubMed Central PMCID: PMCPMC5489367.

16. Lv Y, Mao C, Yin Z, Li F, Wu X, Shi X. Healthy Ageing and Biomarkers Cohort Study (HABCS): a cohort profile. BMJ Open. 2019;9(10):e026513. Epub 2019/10/12. doi: 10.1136/bmjopen-2018-026513. PubMed PMID: 31601581; PubMed Central PMCID: PMCPMC6797363.
17. Zeng Y, Nie C, Min J, Liu X, Li M, Chen H, et al. Novel loci and pathways significantly associated with longevity. Sci Rep. 2016;6:21243. Epub 2016/02/26. doi: 10.1038/srep21243. PubMed PMID: 26912274; PubMed Central PMCID: PMCPMC4766491.
18. Broer L, Buchman AS, Deelen J, Evans DS, Faul JD, Lunetta KL, et al. GWAS of longevity in CHARGE consortium confirms APOE and FOXO3 candidacy. The journals of gerontology Series A, Biological sciences and medical sciences. 2015;70(1):110-8. Epub 2014/09/10. doi: 10.1093/gerona/glu166. PubMed PMID: 25199915; PubMed Central PMCID: PMCPMC4296168.
19. Newman AB, Walter S, Lunetta KL, Garcia ME, Slagboom PE, Christensen K, et al. A meta-analysis of four genome-wide association studies of survival to age 90 years or older: the Cohorts for Heart and Aging Research in Genomic Epidemiology Consortium. The journals of gerontology Series A, Biological sciences and medical sciences. 2010;65(5):478-87. Epub 2010/03/23. doi: 10.1093/gerona/glq028. PubMed PMID: 20304771; PubMed Central PMCID: PMCPMC2854887.
20. Sebastiani P, Bae H, Sun FX, Andersen SL, Daw EW, Malovini A, et al. Meta-analysis of genetic variants associated with human exceptional longevity. Aging (Albany NY). 2013;5(9):653-61. Epub 2013/11/19. doi: 10.18632/aging.100594. PubMed PMID: 24244950; PubMed Central PMCID: PMCPMC3808698.
21. Stekhoven DJ, Buhlmann P. MissForest--non-parametric missing value imputation for mixed-type data. Bioinformatics. 2012;28(1):112-8. Epub 2011/11/01. doi: 10.1093/bioinformatics/btr597. PubMed PMID: 22039212.
22. Knuiman MW, Vu HT, Segal MR. An empirical comparison of multivariable methods for estimating risk of death from coronary heart disease. J Cardiovasc Risk. 1997;4(2):127-34. Epub 1997/04/01. PubMed PMID: 9304494.
23. Allyn J, Allou N, Augustin P, Philip I, Martinet O, Belghiti M, et al. A Comparison of a Machine Learning Model with EuroSCORE II in Predicting Mortality after Elective Cardiac Surgery: A Decision Curve Analysis. PLoS One. 2017;12(1):e0169772. Epub 2017/01/07. doi: 10.1371/journal.pone.0169772. PubMed PMID: 28060903; PubMed Central PMCID: PMCPMC5218502.
24. Ambale-Venkatesh B, Yang X, Wu CO, Liu K, Hundley WG, McClelland R, et al. Cardiovascular Event Prediction by Machine Learning: The Multi-Ethnic Study of Atherosclerosis. Circulation research. 2017;121(9):1092-101. Epub 2017/08/11. doi: 10.1161/CIRCRESAHA.117.311312. PubMed PMID: 28794054; PubMed Central PMCID: PMCPMC5640485.
25. Ganna A, Ingelsson E. 5 year mortality predictors in 498,103 UK Biobank participants: a prospective population-based study. Lancet. 2015;386(9993):533-40. Epub 2015/06/08. doi: 10.1016/S0140-6736(15)60175-1. PubMed PMID: 26049253.
26. Puterman E, Weiss J, Hives BA, Gemmill A, Karasek D, Mendes WB, et al. Predicting mortality from 57 economic, behavioral, social, and psychological factors. Proc Natl Acad Sci U S A. 2020;117(28):16273-82. Epub 2020/06/24. doi: 10.1073/pnas.1918455117. PubMed PMID: 32571904; PubMed Central PMCID: PMCPMC7369318.
27. Nations U. World Population Prospects: The 2017 Revision, Key Findings and Advance Tables. [Working Paper No. ESA/P/WP/248]. In press 2017.
28. Walter S, Mackenbach J, Voko Z, Lhachimi S, Ikram MA, Uitterlinden AG, et al. Genetic, physiological, and lifestyle predictors of mortality in the general population. Am J Public Health. 2012;102(4):e3-10. Epub 2012/03/09. doi: 10.2105/AJPH.2011.300596. PubMed PMID: 22397355; PubMed Central PMCID: PMCPMC3489349.

Figures and tables

Table 1. Baseline Characteristics [a] of the CLHLS Participants by Survival Status

	Survival (N = 1377)	Death (N = 1071)	Total (N = 2448)	P -values
Age				<0.001
Mean (SD)	77.7 (12.6)	92.7 (9.8)	84.3 (13.7)	
Median (Q1, Q3)	77.0 (68.0, 86.0)	95.0 (86.0, 101.0)	85.0 (73.0, 97.0)	
Gender				<0.001
Female	570 (41.4)	663 (61.9%)	1233 (50.4%)	
Male	807 (58.6)	408 (38.1%)	1215 (49.6%)	
Marital status				<0.001
Married (spouse alive)	761 (55.3)	215 (20.1)	976 (39.9)	
Others [b]	616 (44.7)	856 (79.9)	1472 (60.1)	
Activity of daily living [c]				<0.001
Not impaired	1257 (91.3)	717 (66.9)	1974 (80.6)	
Impaired	120 (8.7)	354 (33.1)	474 (19.4)	
MMSE score [d]				<0.001
Mean (SD)	25.9 (6.24)	16.9 (11.2)	22.0 (9.8)	
Median (Q1, Q3)	28.0 (25.0, 29.0)	20.0 (5.0, 28.0)	27.0 (18.0, 29.0)	
Social activity score [e]				<0.001
Mean (SD)	13.0 (3.0)	10.5 (3.1)	11.9 (3.3)	
Median (Q1, Q3)	13.0 (11.0, 15.0)	10.0 (8.0, 12.0)	12.0 (9.0, 14.0)	
Psychological wellbeing score [f]				<0.001
Mean (SD)	25.4 (4.4)	23.8 (4.6)	24.7 (4.6)	
Median (Q1, Q3)	25.0 (23.0, 28.0)	23.0 (21.0, 26.0)	24.0 (22.0, 27.0)	
Album, g/L				<0.001
Mean (SD)	42.01 (3.6)	39.99 (3.7)	41.13 (3.8)	
Median (Q1, Q3)	42.0 (40.3, 43.8)	40.2 (38.4, 41.8)	41.2 (39.3, 43.0)	
25-hydroxyvitamin D, ng/mL				<0.001
Mean (SD)	46.9 (16.5)	38.1 (13.3)	43.0 (15.8)	
Median (Q1, Q3)	45.3 (37.8, 53.2)	37.9 (29.7, 44.6)	41.8 (33.8, 50.0)	
Red blood cells, million/mm^3				<0.001
Mean (SD)	4.9 (1.8)	4.6 (2.4)	4.8 (2.1)	
Median (Q1, Q3)	4.4 (4.0, 5.0)	4.1 (3.6, 4.7)	4.3 (3.8, 4.9)	
Hemoglobin, g/dL				<0.001
Mean (SD)	131.1 (22.7)	123.5 (35.6)	127.8 (29.3)	
Median (Q1, Q3)	131.0 (118.0, 145.0)	122.0 (109.0, 135.0)	127.3 (114.0, 141.0)	
TOMM40, rs2075650 genotype				0.45
AA	1150 (83.5)	904 (84.4)	2054 (83.9)	
AG	221 (16.0)		379 (15.5)	
GG	6 (0.4)	9 (0.8)	15 (0.6)	
FOXO3, rs10457180 genotype				0.66
AA	761 (55.3)	592 (55.3)	1353 (55.3)	
AG	517 (37.5)	406 (37.9)	923 (37.7)	

GG	99 (7.2)	73 (6.8)	172 (7.0)

Abbreviation: MMSE: Mini-Mental State Examination

[a] Numbers shown are N (%) unless otherwise noted.; [b] Other marital status includes separated, widowed, divorced, never married; [c] Activity of daily living: assessed by six self-reported questions: "Do you need assistance in bathing/dressing/toileting/transferring/eating/continence?". Impaired Activity of daily living was defined as if the participants answered 'Yes' for any of those questions; [d] MMSE score: The MMSE score includes 24 item, covering seven subscales including orientation (four points for time orientation and one point for place orientation); naming foods (naming as many kinds of food as possible in one minute, seven points); registration of three words (three points); attention and calculation (mentally subtracting three iteratively from twenty, five points); copy a figure (one point); recall (delayed recall of the three words mentioned above, three points); and language (two points for naming objectives, one point for repeating a sentence, and three points for listening and obeying). The MMSE score ranges from 0 to 30. Higher scores represent a better cognitive function; [e] Leisure activity score: eight activities are assessed: visiting neighbors, shopping, cooking, washing clothes, walking 1 km, lifting 5 kg, crouching and standing up three times, and taking public transportation. I scored each activity 1 for 'never', 2 for 'sometimes' 3 for 'almost every day'. The score ranged from 5 to 21 with a higher score indicating more leisure activities; [f] Psychological wellbeing score: Psychological wellbeing was measured by seven items. These seven items included four positive psychological wellbeing (tapping levels of optimism, conscientiousness, personal control and positive feeling about aging) and three negative aspects of psychological wellbeing (regarding neuroticism, loneliness and loss of self-worthy). The responses of these items used 5-point Likert scale ranging from "always" (5) to never (1) with negative items reversely recoded. The scores of the seven items were summed to obtain a total score with a possible range of 8-35 with a higher score represented a more positive level of psychological wellbeing.

Table 2. A List of the Predictors Used for Prediction

Questionnaire: basic demographics and lifestyle
Age, gender, ethnicity, co-residence, years of schooling, marital status, residence, occupation before retirement, smoking status, drinking status, physical activity, consumption of fruit, vegetable, meat, fish, egg, bean, salt-preserved vegetables, sugar, tea, garlic, height
Questionnaire: disease history and health indicators
Times of suffering from serious illness in past two years, suffering from hypertension, diabetes, heart disease, stroke, bronchitis, emphysema, pneumonia, asthma, tuberculosis, cataract, glaucoma, cancer, ulcer, Parkinson's disease, bedsore, arthritis, systolic and diastolic blood pressure, BMI, self-reported health, self-reported life satisfactory, activity of daily living, social and leisure activities, MMSE score, psychological well-being
Biomarkers
White blood cell count, red blood cell count, hemoglobin, erythrocyte hematocrit, erythrocyte mean corpuscular volume, erythrocyte mean corpuscular hemoglobin, erythrocyte mean corpuscular hemoglobin concentration, platelet count, plateletocrit, mean platelet volume, lymphocyte count, percentage of lymphocytes, platelet distribution width, white blood cell leukocytes, nitrates, urobilinogen, bilirubin, ketones, glucose, Vitamin C, high-density lipoprotein cholesterol, low-density lipoprotein cholesterol, urea acid, plasma creatine, urea nitrogen, triglyceride, total cholesterol, high sensitive c-reactive protein, malondialdehyde, superoxide dismutase activity, albumin, creatine, glycolated plasma protein, superoxide dismutase, 25-OH-D3, vitamin B12, urine microalbumin, urine creatinine, urine microalbumin/ urine creatinine ratio
Genetic
HLA-DQA1/DRB1: rs34831921, TOMM40: rs2075650, APOC1: rs4420638, NECTIN2: rs6857, APOE: rs769449, BEND4: rs1487614, EPHA6: rs10934524, ZFYVE28: rs57681851, ASIC2: rs7213812, FAM13A: rs2609284, LIMCH1: rs10007810, FOXO3: rs10457180, OR2W3: rs10888267, LINC02227: rs2149954, CSRNP3: rs6432832

Abbreviation: MMSE: Mini-Mental State Examination

Table 3: Integrated Brier Score (IBS) and C-index on the Test Data

CLHLS	**IBS**	**C-index**
Cox all variables	0.131	0.77
Cox backward	0.129	0.78
RSF	0.128	0.78
XG-boost	0.120	0.80

Table 4: Hazard ratios along with their 95% confidence intervals for the 12 most influential variables for the Cox models.

Variable name	Cox all variables, HR (95% CI)	Variable name	Cox backward, HR (95% CI)
One-year increase in age	1.05 (1.04, 1.06)	One-unit increase in urine microalbumin	1.003 (1.002, 1.004)
One-unit increase in mean corpuscular hemoglobin concentration	1.005 (1.002, 1.008)	One point increase in MMSE score	0.98 (0.97, 0.99)
Gender, male	1.66 (1.23, 2.24)	One-year increase in age	1.06 (1.05, 1.07)
Activity of daily living, impaired	1.40 (1.11, 1.77)	Albumin	0.93 (0.90, 0.95)
One-unit increase in urine microalbumin	1.002 (1.000, 1.003)	25-hydroxyvitamin D	0.98 (0.97, 0.98)

Variables are presented in decreasing order according to the absolute z-score values for the Cox model with all variable

.

Table 5: The 10 most prognostic factors for the XG-Boost and for the Random Survival Forest

Random Survival Forest		XG-Boost	
Variable	VIMP	Variable	F score
Age	0.0389	Age	0.068
MMSE score	0.0150	Activity of daily living	0.024
Activity of daily living	0.0137	MMSE score	0.023
Social activity score	0.0095	Social activity score	0.021
25-hydroxyvitamin D	0.0055	25-hydroxyvitamin D	0.0068
Albumin	0.0032	Marital status	0.0048
Marital status	0.0030	Psychological wellbeing score	0.0029
Red blood cells	0.0023	Self-reported health	0.0018

Abbreviations: VIMP: variable importance

Figure 1: Prediction Error Curves for all Models.

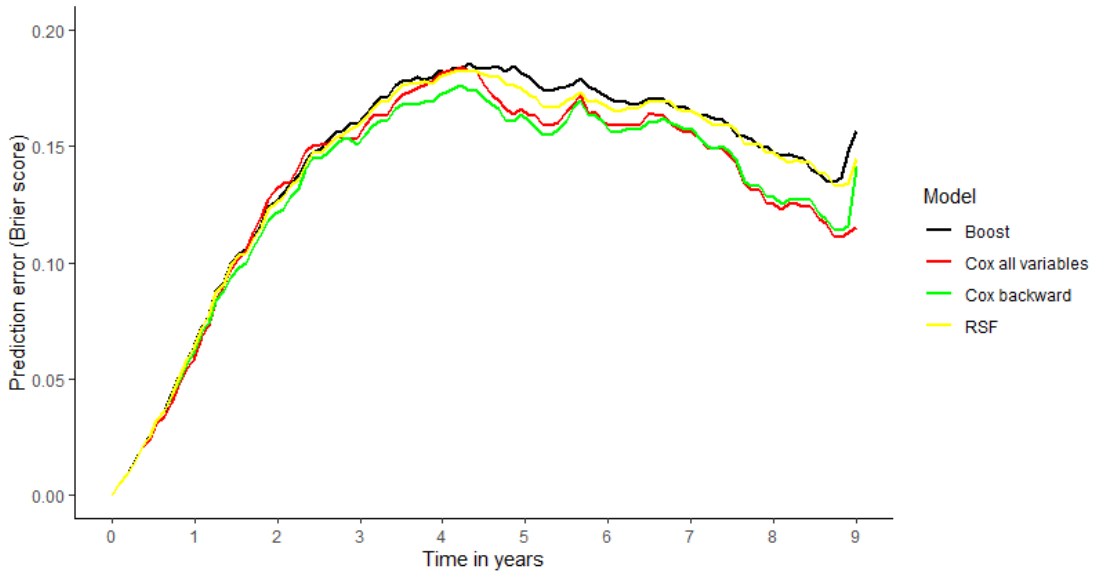
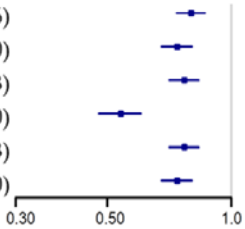


Figure 2: The C-index and 95% Confidence Interval of XG-Boost model with different groups of variables

	Number of variables	C-index (95% CI)
All variables	117	0.80 (0.74, 0.86)
Questionnaire variables	61	0.74 (0.68, 0.80)
Biomarker variables	41	0.77 (0.71, 0.83)
Genetic variables	15	0.54 (0.48, 0.60)
Biomarker + genetic variables	56	0.77 (0.71, 0.83)
Questionnaire + genetic variables	76	0.74 (0.68, 0.80)