Development and Validation of a Risk Score to Predict the First Hip Fracture in the Oldest Old: A Retrospective Cohort Study

Ming-Tuen LAM[4,*], Chor-Wing SING[1,*], PhD, Gloria HY LI[1], PhD, Annie WC KUNG[2], MD, Kathryn CB TAN[2], MD, Ching-Lung CHEUNG[1,2,3], PhD

*Contributed equally

[1]Department of Pharmacology and Pharmacy, [2]Department of Medicine, [3]Centre for Genomic Sciences, [4]Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong, China.

Correspondence and reprint requests:

Ching-Lung Cheung, PhD

Department of Pharmacology and Pharmacy

The University of Hong Kong

Pokfulam, Hong Kong, China

Email: lung1212@hku.hk

Tel: +852-2831-5085 Fax: +852-2816-2095

ABSTRACT

Background: To evaluate whether the common risk factors and risk scores (FRAX, QFracture and Garvan) can predict hip fracture in the oldest old (defined as age≥80), and to develop an oldest-old specific 10-year hip fracture prediction risk algorithm.

Methods: Subjects aged≥80 without history of hip fracture were studied. For the derivation cohort (N=251, mean age=83), participants were enrolled with a median follow-up time of 8.9 years. For the validation cohort (N=599, mean age=85), outpatients were enrolled with a median follow-up of 2.6 years. A 5-factor risk score (the HKOS score) for incident hip fracture was derived and validated, and its predictive accuracy was evaluated and compared with other risk scores.

Results: In the derivation cohort, the C-statistics were 0.65, 0.61, 0.65, 0.76 and 0.78 for FRAX with bone mineral density (BMD), FRAX without BMD, QFracture, Garvan, and the HKOS score, respectively. The category-less net reclassification index and integrated discrimination improvement of the HKOS score showed a better reclassification of hip fracture than FRAX and QFracture (all P<0.001) but not Garvan, while Garvan, but not HKOS score, showed a significant over-estimation in fracture risk (Hosmer-Lemeshow test p-value<0.001). In the validation cohort, the HKOS score had a C-statistic of 0.81 and a considerable agreement between expected and observed fracture risk in calibration.

Conclusion: The HKOS score can predict 10-year incident hip fracture among the oldest old in Hong Kong. The score may be useful in identifying the oldest old patients at risk of hip fracture in both community-dwelling and hospital settings.

Keywords: HKOS score, osteoporosis, BMD, CVA, falls

INTRODUCTION

Hip fracture is a leading cause of morbidity and mortality in the elderly. With an ageing world population, hip fracture will remain a huge public health problem. Our recent projection of hip fractures in major Asian countries using the most updated incidence rates and projected population sizes showed that the number of hip fractures will rise from 1.12 million in 2018 to 2.56 million in 2050, equivalent to a 2.28-fold increase (1). The direct cost associated with hip fracture is going to rise from 9.5 billion USD to 15 billion USD (1). In particular, the oldest old (defined as people aged 80 or above), who are among the most vulnerable to fracture (2), is the fastest growing age-group (3). According to a study performed by the World Health Organization, the proportion of countries with average life expectancy exceeding 80 was 6.6% in 2005, and it increased to 15.4% in 2015 (4). Prevention of hip fracture in the oldest old is therefore a key public health issue.

Many risk factors of hip fracture have been identified, yet most of them were based on epidemiological studies with a mean age of around 50 (5). Therefore, these findings may not be applicable to the oldest old, who have distinct features in terms of physiological functioning, comorbidity patterns and lifestyles (6, 7) while these are relatively rare in people aged below 80. Therefore, it remains largely unknown whether

conventional risk factors of hip fracture are still valid for the oldest old. We hypothesized that the commonly used hip fracture prediction scores have limited predictive ability in the oldest old population. To evaluate whether these conventional risk factors can predict hip fracture in the oldest old and to derive an oldest-old specific 10-year hip fracture prediction risk score, we conducted a retrospective study among participants of the Hong Kong Osteoporosis Study (HKOS) aged 80 or above who had no history of hip fractures, with subsequent validation in an independent cohort comprising of patients at high risk of osteoporotic fracture using robust statistical analyses.

MATERIALS AND METHODS

Participants

HKOS is a prospective cohort study of osteoporosis in Hong Kong since 1995. Details of HKOS have been described elsewhere (8). Ethical approval was obtained from the Institutional Review Board, HKU/HA HKW, HKSAR, China. In brief, the cohort participants were community-dwelling Southern Chinese men and women of Han descent recruited from public road shows and health fairs held in various districts of Hong Kong from 1995 to 2010. A total of 9,449 participants aged between 12 and 105 were recruited. Among all these participants, 387 were aged 80 or above and had no

history of hip fracture. After excluding those with missing data, 251 remained in the final analysis and they had at least 10 years of follow up (Figure 1). Among the excluded participants (N=136), most of them did not have their bone mineral density (BMD) measured, either because they were too frail, chair-bound, or refused to have dual-energy X-ray absorptiometry (DXA) scan. The risk score developed from the derivation cohort was validated in an independent cohort.

Risk Factor Selection

Various risk assessment tools are currently available to evaluate one's probability of fracture in 10 years, and the most widely used ones are FRAX, QFracture and Garvan, which are all online available. FRAX was developed from population-based cohorts from Europe, North America, Asia and Australia, that comprised of a total of 60,000 men and women (9). FRAX score is calculated from a country-specific algorithm, which incorporates 12 readily available clinical risk factors of fracture and can be used with or without BMD. QFracture was developed from a UK cohort study using data from over 2 million men and women. QFracture included at least 24 risk factors in its calculation but BMD is not used in the model (10). Garvan was developed using data from an Australian cohort of approximately 2,500 men and women aged 60 years or above. Unlike the other two scores, Garvan uses only five risk factors in the calculation

(11). Comparison of these risk scores has been discussed elsewhere (12) and the clinical risk factors used in each score are shown in Supplementary Table 1. In our study, the risk factors used in FRAX, QFracture and Garvan were analysed.

BMD and Other Measurements

BMD at the femoral neck was measured by DXA (Hologic QDR 4500 plus). Daily calibration of the equipment was performed and the in vivo precision of DXA measurement at the femoral neck was 1.5%. All other variables in the derivation cohort were self-reported in a structured questionnaire with the help of a trained nurse or a research assistant. Details have been described elsewhere (13). For the validation cohort, data on the history of falls, history of non-hip fractures, and cerebrovascular accident (CVA) were obtained from electronic medical records that have been validated (14, 15).

Incident Hip Fracture

Data of the hip fracture incidence were obtained from the database of the Hong Kong Hospital Authority. The majority of the Hong Kong population attend the public outpatient clinics and hospitals managed by the Hong Kong Hospital Authority (16). In Hong Kong, 97.8% femoral fracture patients were admitted to hospitals under the Hong

Kong Hospital Authority (13). Hip fracture was identified using the ICD9 code 820.XX. We previously validated the records of hip fracture coding with a positive predictive value of 100% (14).

Statistical Analyses

We developed a 10-year sex-specific risk prediction score of hip fracture using logistic regression. The final prediction model was built using the forward stepwise selection procedure based on p-value, effect size, and clinical importance of the variables. Survival time was calculated from the baseline date to the date of diagnosis of hip fracture, death, or end of study (10 years of follow up time for the derivation cohort). Bias-corrected accelerated 95% CI and p-value were estimated by 500 bootstrap resamples. Due to the high risk of mortality in the oldest old, competing risk of death may affect the validity of the findings. As a sensitivity analysis, competing risk regression was used.

A predicted probability (the HKOS score) of hip fracture for the oldest old was generated. The beta-estimates from the final logistic regression model were used as the weights of the risk factors and combined with the intercept to generate the HKOS score in the derivation cohort. On the other hand, calculations of FRAX

(https://www.sheffield.ac.uk/FRAX/tool.aspx?country=20), QFracture (https://qfracture.org/) and Garvan (https://www.garvan.org.au/promotions/bone-fracture-risk/calculator/) risk scores were done using their online calculators. These fracture prediction scores provided prediction for "any osteoporotic fracture" and "hip fracture", and the risk score of hip fracture was recorded for the current analysis. Harrell's C-statistics were then calculated to evaluate the discriminative ability of the HKOS score, FRAX with BMD, FRAX without BMD, QFracture and Garvan, and the statistical differences between C-statistics were compared using the Z-score test. The improvement in discrimination for incident hip fracture was evaluated by using category-less net reclassification index (NRI) and integrated discrimination improvement (IDI). Category-less NRI was used because it is considered the most objective and versatile measure of improvement in risk prediction (17). The R package "Hmisc" was used for the analysis. The performance of the HKOS score was further evaluated by calibration using the Hosmer-Lemeshow statistics. We divided the cohort into 5 groups based on the quintiles of predicted fracture risk, and provided calibration plots of predicted versus observed fracture risks.

Model Validation

The performance of the HKOS score was validated in an independent validation cohort.

The validation cohort was a retrospective cohort from the Osteoporosis Centre in Queen Mary Hospital, which consisted of patients at high risk of osteoporotic fracture being referred to have a DXA scan and management in the Osteoporosis Centre. All patient identities were anonymized, and electronic medical records were linked to patients' details using a unique anonymized reference key. We identified patients aged 80 or above who had femoral neck BMD measured (N=1,085). After excluding patients with history of hip fracture (N=465) or with missing data (N=26), 599 participants remained in the final analysis (Figure 1). A comparison of basic demographics and risk factors in the final prediction model between the derivation and validation cohorts is shown in Supplementary Table 2. The Harrell's C-statistics and area under ROC curve (AUC) were determined. Z-score (for Harrell's C-statistics) and deLong tests (for AUC) were used to compare the C-statistics of the HKOS score with other existing fracture prediction risk scores. A corrected HKOS score based on follow-up time was derived for calibration. All statistical analyses were conducted using SPSS Version 22.0 software (IBM, Inc., Chicago, IL, USA) and R (R Foundation for Statistical Computing, Vienna, Austria; https://www.r-project.org/).

RESULTS

Participants in the Derivation Cohort

After a median follow-up time of 8.9 years (1,801.4 person-years in total), 40 participants had an incident hip fracture, with an overall incidence rate of 22.2 per 1,000 person-years. Those with incident hip fracture were more likely to be female, with lower body weight, lower BMD at femoral neck, history of falls over last 12 months, history of non-hip fracture, CVA, higher FRAX score, higher QFracture score, and higher Garvan score (Table 1 and Supplementary Table 3).

Association and Model Development

Results of the logistic regression models are shown in Table 2. Female sex, lower weight, lower BMD at femoral neck, history of falls over last 12 months, history of non-hip fracture and CVA were significantly associated with increased risk of hip fracture in the crude model. Using forward stepwise regression, the final prediction model comprised of femoral neck BMD (OR 0.47; 95% CI 0.32-28.01), history of non-hip fracture (OR 2.44; 95% CI 1.01-5.87), and CVA (OR 5.94; 95% CI 1.26-28.01) (Table 2). Well established risk factors, such as age, sex, and weight, were not significantly associated with hip fracture and not included in the final model (Supplementary Table 4). The bias-corrected p-values were generally similar to the conventional logistic regression analysis. We selected variables with p-value < 0.1 in the final model to estimate the HKOS score. These included femoral neck BMD, history

of non-hip fracture, CVA, height, and history of falls over last 12 months. Given that sex-specific association was observed, we used the beta estimates and intercept of the sex-specific model to calculate the HKOS score (Supplementary Table 5).

Comparison of the C-statistic of the HKOS score (femoral neck BMD + height + history of falls + history of non-hip fractures + CVA) with that of FRAX with BMD, FRAX without BMD, QFracture and Garvan is shown in Table 3. The HKOS score (C-statistic: 0.78; 95% CI: 0.72-0.84) had the highest yet comparable C-statistic with other prediction scores, followed by Garvan, QFracture, FRAX with BMD, and FRAX without BMD with C-statistics of 0.76 (95% CI: 0.70-0.81), 0.65 (95% CI: 0.58-0.73), 0.65 (95% CI: 0.58-0.72) and 0.61 (95% CI: 0.53-0.69), respectively. Using the Z-score test, the HKOS score was more concordant with the outcome when comparing with other fracture prediction scores (all $P<0.05$; Table 3), except that similar C-statistics were observed between HKOS score and Garvan (Table 3). Similar results were observed using AUC and deLong test (Supplementary Table 6).

Table 4 shows the category-less NRI and IDI of the HKOS score when compared with FRAX with BMD, FRAX without BMD, QFracture and Garvan respectively. The category-less NRI and IDI showed that the HKOS score had an improved

reclassification of incident hip fracture when compared with FRAX with BMD (NRI: 0.672; IDI: 0.109), FRAX without BMD (NRI: 0.913; IDI: 0.153), and QFracture (NRI: 0.470; IDI: 0.142) (all P<0.001). No significant difference in performance was observed between the HKOS score and Garvan.

Model Calibration

The calibration plots demonstrated considerable agreement between the predicted and observed fracture risks of the HKOS score in the derivation cohort (Hosmer-Lemeshow test p-value: 0.638; Figure 2a). Although Garvan showed a similar NRI and IDI to the HKOS score, it significantly over-estimated the risk of fracture (Hosmer-Lemeshow test p-value: <0.001), as shown in the calibration plots (Supplementary Figure 1).

Association and Model Validation in the Validation Cohort

After a median follow-up time of 2.6 years (2,1965.9 person-years in total), 29 participants had an incident hip fracture, with an overall incidence rate of 13.2 per 1,000 person-years. Baseline characteristics of the validation cohort have been shown in Supplementary Table 7.

In the independent validation cohort, BMD at femoral neck (OR: 0.59; 95% CI: 0.39-

0.89), history of falls over last 12 months (HR: 3.54; 95% CI: 1.44-8.69), and CVA (HR: 3.89; 95% CI: 1.45-10.42) were also significantly associated with the first hip fracture in the oldest old (Supplementary Table 8). In addition, using competing risk regression as a sensitivity analysis, similar results were observed (Supplementary Table 9).

Using the intercept and beta-estimates from the derivation cohort, the HKOS score was calculated in the validation cohort. The Harrell's C-statistic of the HKOS score in the validation cohort (Harrell's C-statistic: 0.81; 95% CI: 0.75-0.87) was even higher than that in the derivation cohort. Moreover, the calibration plots showed considerable agreement between predicted and observed fracture risks in the validation cohort (Hosmer-Lemeshow test p-value: 0.528; Figure 2b).

DISCUSSION

Summary of Key Findings

This study evaluated the conventional risk factors for prediction of hip fracture in the oldest old and the discriminative performances of the risk scores for incident hip fracture. It was found that the HKOS score comprising of BMD at femoral neck, height, history of falls over last 12 months, history of non-hip fractures, and CVA might be a

good predictor of 10-year hip fracture risk in the oldest old, and the finding was subsequently validated in an independent validation cohort in a separate setting. Other well-established risk factors such as age, sex, and weight were not significantly associated with hip fracture in the oldest old. The commonly used hip fracture prediction scores showed limited predictive ability of hip fracture in the oldest old.

Risk factors of Hip Fracture Among the Oldest Old

In the development of the HKOS score, five risk factors were included in the final model. Among the five risk factors, BMD, height, history of fall, and history of fracture are well-established risk factors of hip fracture (18). Yet, CVA is comparatively less recognized as a risk factor of fracture, which is only used for calculating QFracture. Stroke survivors are known to have a higher risk of fracture, owing to impairments in balance, sensory and motor functions, and hence predisposing them to falls (19). This also explains why the effect of history of falls, which was significantly associated with incident hip fracture in the crude model, was attenuated after adjustment for CVA. Moreover, we previously also showed that warfarin, a drug commonly used for stroke prevention, was associated with increased risk of fracture (20). Reduction in BMD in stroke survivors also contributed to their higher fracture rates (21, 22).

Both history of falls and history of non-hip fractures were significantly associated with incident hip fracture in the crude model, but the associations became statistically insignificant in the full model. We further investigated these two factors in the derivation and validation cohort. The results showed that these risk factors were both significantly associated with incident hip fracture in the crude model. However, history of falls became statistically insignificant after mutual adjustment in the derivation cohort, whereas history of non-hip fracture became statistically insignificant in the validation cohort (Supplementary Table 10). These findings suggest that both history of falls and history of non-hip fractures are dependent on each other and captured common information.

Age is not a significant predictor of incident hip fracture in the oldest old. Age is included in all hip fracture prediction scores, whereas our previous study also showed that age is a significant predictor of hip fracture in the HKOS participants with a mean age of 63.4 (23). However, it seems that it is no longer useful in predicting hip fracture in the oldest old, which is in agreement with a recent study of nursing home residents (mean age of 83.9) showing that age contributed modestly towards the risk of hip fracture (24), compared with other community-based studies. Indeed, the null association of age with clinical outcomes (such as mortality in critically-ill patients (25)

and blood pressure decline) have been well reported in the oldest old, in which overall health status (26) or biological age (13), instead of chronological age, may be a more important factor in predicting clinical outcomes.

Comparison with Existing Risk Prediction Scores

Garvan had a similar performance in predicting fracture risk to the newly-derived HKOS score. Same as the HKOS score, the Garvan score is calculated from five risk factors, in which three of them (history of falls, history of fractures, and BMD) were also incorporated in the HKOS score, which may explain the similar performance in reclassification of fracture risk. Regarding the calibration plot, the Garvan score was inferior to that of the HKOS score, in which Garvan significantly over-estimated the risk. The development of Garvan score was based on a different population (Australian), thus the estimates derived from Garvan are not applicable in the oldest old Chinese population. Although the Harrell's C-statistic of the HKOS score was the highest among various prediction scores, its 95% confidence interval overlapped with that of Garvan, QFracture, and FRAX with BMD. This suggested HKOS score has similar discriminative performance when compared with the above prediction scores. On the other hand, the HKOS may have better risk reclassification than FRAX and QFracture in predicting incident hip fracture. This again highlighted the importance of the current

study in deriving an oldest-old specific fracture risk prediction score. The FRAX algorithm was originally developed from 11 Caucasian and 1 Japanese prospective cohort studies (27), and the mean age in the majority of these studies was 50-65 years (5). Our previous study showed that ethnic-specific clinical risk factors may be more important than FRAX in predicting fracture (28). The difference in age-group and ethnicity of FRAX's development cohorts may therefore explain its relatively low accuracy of FRAX in predicting fracture in the oldest old in Hong Kong. Similarly, the mean age of the cohorts deriving QFracture and Garvan was 50 (10) and 70 (11), respectively. This may explain why Garvan had a better predictive ability than FRAX and QFracture in the oldest old (data not shown). On the other hand, although the calculation of QFracture includes CVA, its ability in predicting hip fracture in the oldest old was modest in the current study. This could be because QFracture was developed from the UK population, and it consists of 21 clinical variables in addition to age, sex, ethnicity, and BMI. Among these variables, only three were significantly associated with hip fracture in the current study. Inclusion of additional predictors which are not associated with hip fracture in the calculation of QFracture may indeed introduce noise to the prediction.

Implications for Clinical Practice

This study has several clinical significances. To the best of our knowledge, this is the first study to evaluate the performances of various risk prediction scores in predicting hip fracture in the oldest old, and to derive a 10-year risk score for predicting the first hip fracture in the oldest old with validation. Notably, the settings of the derivation and validation cohorts were very much different. The derivation cohort was a cohort study in the community setting, whereas the validation cohort was a retrospective cohort study in the hospital setting. Those participants in the validation cohort had osteoporosis or were at high risk of fracture being referred to DXA scan and management by physicians. Thus, it was shown that the incidence of hip fracture was lower in the validation cohort (13.2 and 22.24 per 1,000 person-years in validation and derivation cohorts, respectively). Together with the considerable calibration data, these findings suggest that the HKOS score may be useful in predicting hip fracture in both community-dwelling and hospital settings.

Strengths and Limitations

This study has several other strengths. The HKOS is a prospective cohort study of osteoporosis, and hence many osteoporosis-related risk factors are included in the study (8). The findings of our study were independently validated in a validation cohort which comprised of a group of high-risk patients from hospitals, suggesting the

generalizability of the findings. Secondly, the detection of hip fracture was validated with high accuracy (14). Thirdly, this study had a long follow up time (10 years) for the oldest old, which would be difficult to achieve in other places where life expectancies are shorter. Fourthly, since external validation by other cohorts is expected to be difficult in the oldest old with a long follow up time, bootstrapping was first performed as internal validation. Similar results observed in 500 bootstrap resamples, together with subsequent validation in the validation cohort, suggest that the findings are robust. Fifthly, due to competing risk of death, competing risk regression was used to account for the effects of risk of death on risk of hip fracture, and similar results were observed.

Nevertheless, there are several limitations. First, with only 40 events in the derivation cohort, the power was limited to the extent that we could only identify risk factors with large effect size. It is possible that other risk factors with a smaller effect size may also predict hip fracture, especially those categorical variables that are less common. Thus, it is possible that the prediction performance of FRAX and QFracture may be improved in a larger sample size. Similarly, there were only 29 hip fracture events in the validation cohort. Thus, the HKOS score may behave very differently in other cohorts, and future validation study in other population is warranted. Secondly, there are no data for some of the variables used in the calculation of QFracture and assumptions had to

be made. For example, we do not have data regarding a history of epilepsy or anticonvulsant use. However, given that the annual incidence rate of epilepsy in Asian elderly is only around 1.6 per 1 000, it is reasonable for us to assume that there are no patients with a history of epilepsy in our study group of 251 participants. There were data on COPD but not asthma, therefore we could not evaluate the predictive ability of asthma. In addition, our study does not have data on bone microarchitecture (29), trabecular bone score (30), diet (31), and cognitive impairments (32), which may be important for hip fracture prediction. Furthermore, there was quite a large proportion of missing data in the derivation cohort, mostly because those participants were too frail or chair-bound to have their BMD measured, or refused to have DXA measurement. However, multiple imputation was not performed. This is because frailty itself is associated with an increased risk for fracture (33), so these participants would already have a higher fracture risk than the rest of the derivation cohort, potentially rendering the results of imputation invalid. This also means that the HKOS score may not be applicable in predicting fracture risk in the oldest old who are already too frail. Yet this would not undermine the clinical utility of this score because frail individuals would generally be under special care to minimize fracture risk even without the aid of a clinical risk score. On the other hand, since both the derivation and validation cohorts were ethnically Chinese, the extent to which the HKOS score can be applied to other

ethnic groups remains unknown. Thus, future validation studies in other ethnic groups are needed. Besides, it is unclear if other advanced statistical modeling, such as machine learning approach, can derive a risk score with higher accuracy. However, a recent systematic review suggests that machine learning approach is not necessarily better than the conventional logistic regression approach (34). Finally, like other observational studies, there may be residual confounding effects.

Conclusions

In conclusion, the HKOS score derived from the five predictors can predict 10-year incident hip fracture in the oldest old. The good performance of HKOS score was replicable in terms of C-statistics and calibration in the independent validation cohort. These findings suggest that HKOS score may be useful in identifying oldest old patients at risk of hip fracture in both community-dwelling and hospital settings. Our study calls for more research not only to further validate these findings, but also to focus on research on hip fracture in the oldest old, which is associated with extremely high mortality and treatment cost and reduction in quality of life.

Contributors:

MTL undertook literature review, interpretation, and drafting of the paper; CWS

contributed to data collection, analysis, interpretation, and drafting of the paper; GHYL contributed to data collection and analysis. AWCK contributed to study plan, data collection, interpretation of the results; KCBT contributed to interpretation of the results; CLC initiated the study, undertook literature review, data extraction, data manipulation, and primary data analysis, and wrote the first draft of the paper.

Declaration of interests:

We declare no competing interests.

References

1. Cheung CL, Ang SB, Chadha M, Chow ESL, Chung YS, Hew FL, Jaisamrarn U, Ng H, Takeuchi Y, Wu CH, Xia W, Yu J, Fujiwara S. An updated hip fracture projection in Asia: the AFOS study. *Osteoporosis and Sarcopenia (Accepted)* 2018.
2. Cauley JA, Cawthon PM, Peters KE, Cummings SR, Ensrud KE, Bauer DC, Taylor BC, Shikany JM, Hoffman AR, Lane NE, Kado DM, Stefanick ML, Orwoll ES, Osteoporotic Fractures in Men Study Research G. Risk Factors for Hip Fracture in Older Men: The Osteoporotic Fractures in Men Study (MrOS). *J Bone Miner Res* 2016; 31: 1810-1819.
3. Zeng Y, Feng Q, Hesketh T, Christensen K, Vaupel JW. Survival, disabilities in activities of daily living, and physical and cognitive functioning among the oldest-old in China: a cohort study. *Lancet* 2017; 389: 1619-1629.
4. World Health Organization. World Health Statistics: Life expectancy and Healthy life expectancy Data by country. http://apps.who.int/gho/data/view.main.SDG2016LEXv?lang=en. Accessed March 1, 2018.
5. Kanis JA, Johnell O, Oden A, Johansson H, De Laet C, Eisman JA, Fujiwara S, Kroger H, McCloskey EV, Mellstrom D, Melton LJ, Pols H, Reeve J, Silman

A, Tenenhouse A. Smoking and fracture risk: a meta-analysis. *Osteoporos Int* 2005; 16: 155-162.

6. Covinsky KE, Lindquist K, Dunlop DD, Yelin E. Pain, functional limitations, and aging. *J Am Geriatr Soc* 2009; 57: 1556-1561.
7. Kydd A, Fleming A. What doctors need to know: Prescribing or not for the oldest old. *Maturitas* 2016; 90: 9-16.
8. Cheung CL, Tan KCB, Kung AWC. Cohort Profile: The Hong Kong Osteoporosis Study and the follow-up study. *International journal of epidemiology* 2018; 47: 397-398f.
9. Kanis JA, McCloskey EV, Johansson H, Oden A, Strom O, Borgstrom F. Development and use of FRAX in osteoporosis. *Osteoporos Int* 2010; 21 Suppl 2: S407-413.
10. Hippisley-Cox J, Coupland C. Derivation and validation of updated QFracture algorithm to predict risk of osteoporotic fracture in primary care in the United Kingdom: prospective open cohort study. *BMJ* 2012; 344: e3427.
11. Nguyen ND, Pongchaiyakul C, Center JR, Eisman JA, Nguyen TV. Identification of high-risk individuals for hip fracture: a 14-year prospective study. *J Bone Miner Res* 2005; 20: 1921-1928.
12. Kanis JA, Harvey NC, Johansson H, Oden A, McCloskey EV, Leslie WD. Overview of Fracture Prediction Tools. *J Clin Densitom* 2017; 20: 444-450.
13. Cheung CL, Nguyen US, Au E, Tan KC, Kung AW. Association of handgrip strength with chronic diseases and multimorbidity: a cross-sectional study. *Age (Dordr)* 2013; 35: 929-941.
14. Sing CW, Woo YC, Lee ACH, Lam JKY, Chu JKP, Wong ICK, Cheung CL. Validity of major osteoporotic fracture diagnosis codes in the Clinical Data Analysis and Reporting System in Hong Kong. *Pharmacoepidemiology and drug safety* 2017; 26: 973-976.
15. Sing CW, Wong AY, Kiel DP, Cheung EY, Lam JK, Cheung TT, Chan EW, Kung AW, Wong IC, Cheung CL. Association of Alendronate and Risk of Cardiovascular Events in Patients With Hip Fracture. *J Bone Miner Res* 2018; 33: 1422-1434.
16. Cheung CL, Tan KC, Bow CH, Soong CS, Loong CH, Kung AW. Low handgrip strength is a predictor of osteoporotic fractures: cross-sectional and prospective evidence from the Hong Kong Osteoporosis Study. *Age (Dordr)* 2012; 34: 1239-1248.
17. Pencina MJ, D'Agostino RB, Sr., Steyerberg EW. Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. *Statistics in medicine* 2011; 30: 11-21.

18. Majumdar SR, Leslie WD, Lix LM, Morin SN, Johansson H, Oden A, McCloskey EV, Kanis JA. Longer Duration of Diabetes Strongly Impacts Fracture Risk Assessment: The Manitoba BMD Cohort. *The Journal of clinical endocrinology and metabolism* 2016; 101: 4489-4496.
19. Benzinger P, Rapp K, Konig HH, Bleibler F, Globas C, Beyersmann J, Jaensch A, Becker C, Buchele G. Risk of osteoporotic fractures following stroke in older persons. *Osteoporos Int* 2015; 26: 1341-1349.
20. Lau WC, Chan EW, Cheung CL, Sing CW, Man KK, Lip GY, Siu CW, Lam JK, Lee AC, Wong IC. Association Between Dabigatran vs Warfarin and Risk of Osteoporotic Fractures Among Patients With Nonvalvular Atrial Fibrillation. *JAMA* 2017; 317: 1151-1158.
21. Jorgensen L, Jacobsen BK, Wilsgaard T, Magnus JH. Walking after stroke: does it matter? Changes in bone mineral density within the first 12 months after stroke. A longitudinal study. *Osteoporos Int* 2000; 11: 381-387.
22. Ramnemark A, Nyberg L, Lorentzon R, Olsson T, Gustafson Y. Hemiosteoporosis after severe stroke, independent of changes in body composition and weight. *Stroke* 1999; 30: 755-760.
23. Kung AW, Lee KK, Ho AY, Tang G, Luk KD. Ten-year risk of osteoporotic fractures in postmenopausal Chinese women according to clinical risk factors and BMD T-scores: a prospective study. *J Bone Miner Res* 2007; 22: 1080-1087.
24. Berry SD, Zullo AR, Lee Y, Mor V, McConeghy KW, Banerjee G, D'Agostino RB, Sr., Daiello L, Dosa D, Kiel DP. Fracture Risk Assessment in Long-term Care (FRAiL): Development and Validation of a Prediction Model. *J Gerontol A Biol Sci Med Sci* 2018; 73: 763-769.
25. Ford PN, Thomas I, Cook TM, Whitley E, Peden CJ. Determinants of outcome in critically ill octogenarians after surgery: an observational study. *Br J Anaesth* 2007; 99: 824-829.
26. Weidung B, Toots A, Nordstrom P, Carlberg B, Gustafson Y. Systolic blood pressure decline in very old individuals is explained by deteriorating health: Longitudinal changes from Umea85+/GERDA. *Medicine (Baltimore)* 2017; 96: e9161.
27. Kanis JA, Johnell O, Oden A, Johansson H, McCloskey E. FRAX and the assessment of fracture probability in men and women from the UK. *Osteoporos Int* 2008; 19: 385-397.
28. Cheung EY, Bow CH, Cheung CL, Soong C, Yeung S, Loong C, Kung A. Discriminative value of FRAX for fracture prediction in a cohort of Chinese postmenopausal women. *Osteoporos Int* 2012; 23: 871-878.
29. Boutroy S, Khosla S, Sornay-Rendu E, Zanchetta MB, McMahon DJ, Zhang CA,

Chapurlat RD, Zanchetta J, Stein EM, Bogado C, Majumdar S, Burghardt AJ, Shane E. Microarchitecture and Peripheral BMD are Impaired in Postmenopausal White Women With Fracture Independently of Total Hip T-Score: An International Multicenter Study. *J Bone Miner Res* 2016; 31: 1158-1166.

30. Schousboe JT, Vo T, Taylor BC, Cawthon PM, Schwartz AV, Bauer DC, Orwoll ES, Lane NE, Barrett-Connor E, Ensrud KE, Osteoporotic Fractures in Men Mr OSSRG. Prediction of Incident Major Osteoporotic and Hip Fractures by Trabecular Bone Score (TBS) and Prevalent Radiographic Vertebral Fracture in Older Men. *J Bone Miner Res* 2016; 31: 690-697.
31. Byberg L, Bellavia A, Larsson SC, Orsini N, Wolk A, Michaelsson K. Mediterranean Diet and Hip Fracture in Swedish Men and Women. *J Bone Miner Res* 2016; 31: 2098-2105.
32. Muir SW, Gopaul K, Montero Odasso MM. The role of cognitive impairment in fall risk among older adults: a systematic review and meta-analysis. *Age Ageing* 2012; 41: 299-308.
33. Tom SE, Adachi JD, Anderson FA, Jr., Boonen S, Chapurlat RD, Compston JE, Cooper C, Gehlbach SH, Greenspan SL, Hooven FH, Nieves JW, Pfeilschifter J, Roux C, Silverman S, Wyman A, LaCroix AZ, Investigators G. Frailty and fracture, disability, and falls: a multiple country study from the global longitudinal study of osteoporosis in women. *J Am Geriatr Soc* 2013; 61: 327-334.
34. 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.

Table 1. Baseline characteristics of the derivation cohort

	Without hip fracture (N = 211)	With hip fracture (N = 40)	p-value
Female	122 (57.8%)	31 (77.5%)	0.022
Weight (kg)	53.46 (10.04)	49.62 (9.99)	0.027
BMD at femoral neck (g/cm^2)	0.56 (0.13)	0.46 (0.09)	<0.001
History of falls over last 12 months	57 (27%)	18 (45%)	0.025
History of non-hip fracture	88 (41.7%)	30 (75%)	<0.001
Cerebrovascular accident	4 (1.9%)	4 (10%)	0.017
FRAX	10.42 (8.36)	15.27 (9.88)	0.002
QFracture	8.62 (6.44)	11.88 (6.12)	<0.001
Garvan	22.93 (24.26)	44.95 (26.59)	0.003

Note - Data are presented as mean ± SD for continuous variables, and frequency (percentage) for categorical variables.

Only variables showing significant differences are shown. Full table is provided as Supplementary Table 2.

Table 2. Association of conventional risk factors with incident hip fracture

Variables[a]	Crude model			Full model[b]			
	OR	95% CI	p-value	OR	95% CI	p-value	Bootstrap p-value[c]
Female	2.51	(1.14-5.54)	0.022				
Height (m)	0.09	(0.00-3.01)	0.176	64.16	(0.48-8500.29)	0.095	0.086
Weight (kg)	0.96	(0.93-1.00)	0.029				
BMD at femoral neck (per 0.1 g/cm^2)	0.50	(0.36-0.69)	<0.001	0.47	(0.32-28.01)	<0.001	0.002
History of falls over last 12 months	2.21	(1.11-4.42)	0.025	1.96	(0.90-4.25)	0.089	0.094
History of non-hip fracture	4.19	(1.95-9.02)	<0.001	2.44	(1.01-5.87)	0.047	0.064
Cerebrovascular accident	5.75	(1.38-24.04)	0.017	5.94	(1.26-28.01)	0.025	0.002
Cancer	3.24	(0.90-11.63)	0.072				

[a] Variables with P<0.1 are shown.

[b] Five predictors were selected based on forward stepwise regression.

[c] Internal validation of the multivariable logistic regression model by 500 bootstrap resamples.

Table 3. Harrell's C-statistics of the final prediction model, FRAX, QFracture and Garvan in the derivation cohort

	Harrell's C-statistics	Z-score test P-value*
Combined Model (HKOS score)	0.78 (0.72-0.84)	Ref
FRAX	0.65 (0.58-0.72)	<0.001
FRAX without BMD	0.61 (0.53-0.69)	<0.001
QFracture	0.65 (0.58-0.73)	0.006
Garvan	0.76 (0.70-0.81)	0.383

* Z-score test was used to compare the C-statistics of the combined model with other existing fracture prediction risk scores.

Table 4. Comparative performance metrics (category-less NRI and IDI) for predicting incident hip fracture in the derivation cohort

Statistics	Compared with	Estimates	p value
Category-Less NRI	FRAX with BMD	0.672 (0.379-0.965)	<0.001
	FRAX without BMD	0.913 (0.656-1.171)	<0.001
	QFracture	0.470 (0.228-0.713)	<0.001
	Garvan	0.066 (-0.188-0.319)	0.612
IDI	FRAX with BMD	0.109 (0.052-0.166)	<0.001
	FRAX without BMD	0.153 (0.088-0.219)	<0.001
	QFracture	0.142 (0.078-0.207)	<0.001
	Garvan	-0.045 (-0.122-0.031)	0.246

Figure 1. Flow diagram of the study.

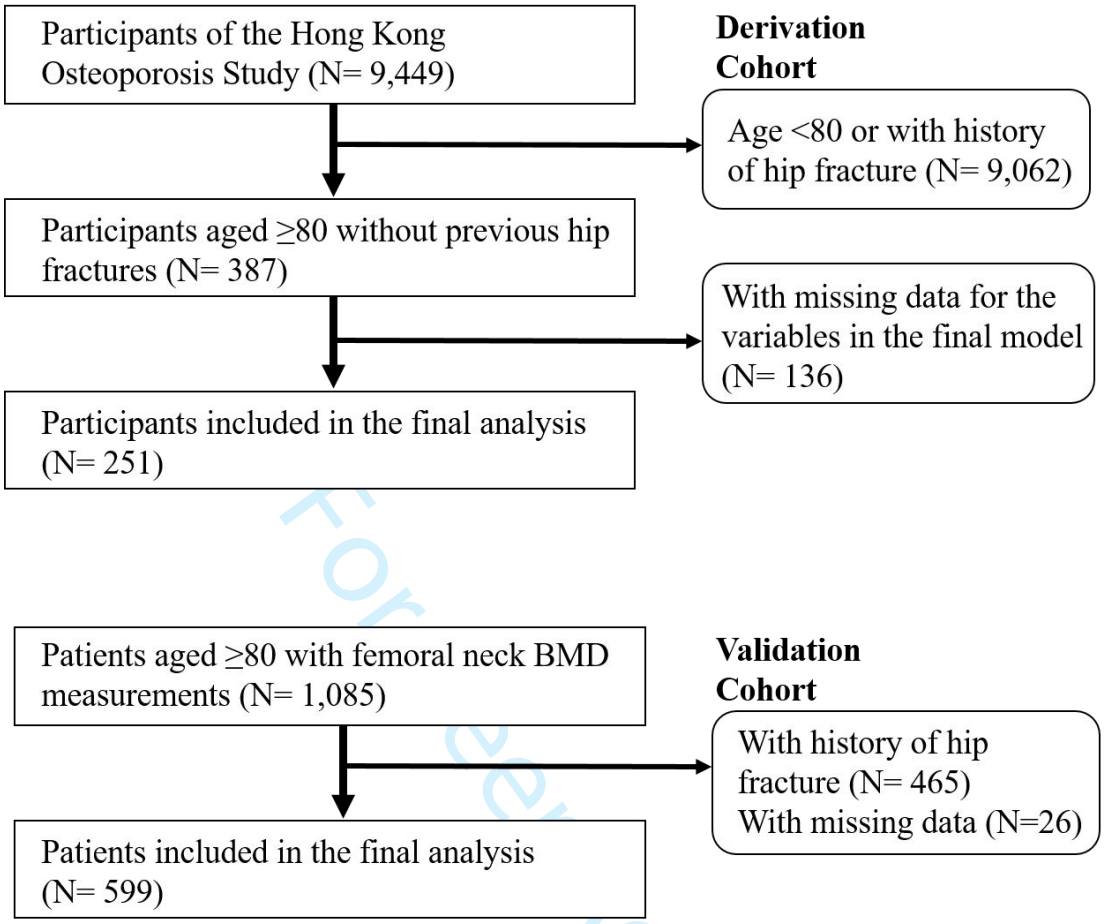
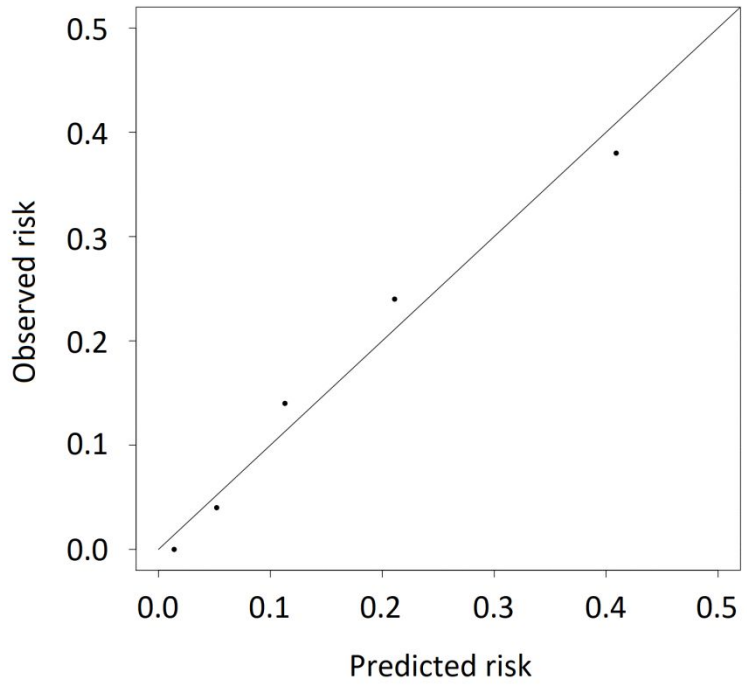


Figure 2. Calibration plots of the predicted vs. the observed HKOS score-derived hip fracture risk in the (a) derivation cohort, and (b) validation cohort.

(a) Derivation cohort



(b) Validation cohort