PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form (**http://bmjopen.bmj.com/site/about/resources/checklist.pdf**) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

ARTICLE DETAILS

TITLE (PROVISIONAL)	External validation and clinical usefulness of three commonly used cardiovascular risk prediction scores in an Emirati population: a retrospective longitudinal cohort study
AUTHORS	Al-Shamsi, Saif; Govender, Romona; King, Jeffrey

VERSION 1 – REVIEW

REVIEWER	Jianchao Quan University of Hong Kong
REVIEW RETURNED	06-Jul-2020

GENERAL COMMENTS	The number of patients 1,048 patients for validation is small and the recruitment from a single outpatient clinic is unlikely to be representative of general population. How did the authors account for (competing risks) deaths from recruitment in 2008 to the follow-up date in 2019? Could the authors test different risk thresholds to classify high-risk for improved performance. Several parts of the TRIPOD statement (e.g. methods) appear unavailable when they could be stated in the manuscript. Line 98 How was “history of CVD” assessed – e.g. manual review of patient medical records, ICD diagnosis codes, patient recall? Line 106 Suggest the authors list out the variables in each respective model for the reader’s ease of comprehension. Line 120 “derivation cohort” – unsure why this is a derivation cohort given no model development. It seems only validation was conducted. Line 129 How was missing data were assessed, and be “determined” to be missing at random? Multiple imputation is recommended rather than single imputation - this should be quick to do given the small sample size. Line 136 Medians (25th percentile, 75th percentile). Suggest replacing with “quantiles/percentiles”. Table 1&2 –It would be useful to include the event rate by person-year. Table 3 – Could include an overall C-statistic row. Useful to present 95% CI. Calibration – unclear how the Figure 2 plots are generated. If it is individual-level calibration an individual may have a predicted risk of 25% but the observed frequency for an individual is 0 or 100%.

	Is this calibration plot grouped? Given the small sample size of ~500, it is may be beneficial to group the predicted risks by deciles of predicted risk. Figure 2B (women) PCE (black) line - not clear why the observed risk falls to 0% at higher predicted risks unlike the other risk models in men and women.

REVIEWER	Nadim Mahmud Hospital of the University of Pennsylvania United States of America
REVIEW RETURNED	06-Jul-2020

GENERAL COMMENTS	In this very interesting study, Al-Shamsi et al. evaluate the performance of cardiovascular disease (CVD) risk prediction scores in an Emirati population. They highlight that while the recommended CVD prediction models have been derived and validated primarily in Caucasian and African American populations, they have not been previously tested in an Arab cohort. The authors found that the Framingham and Pooled Cohort Risk Equation (PCE) models overestimated CVD risk in both men and women in higher ranges of predicted risk, and the models underestimated risk in women in the lower range of predicted risk. The discrimination of the scores in men was C-statistic = ~0.7 and in women C-statistic = ~0.75. Overall, this is a well-conducted study with generally appropriate statistical methods and valid conclusions. However, I do have a number of recommendations and critiques that I hope may strengthen this work prior to publication. In particular, aspects of the methods regarding cohort selection and outcomes adjudication need to be described in further detail. Major: - I am admittedly not familiar with the UAE health system, but it seems incredible that only four patients screened in 2008 were lost to follow-up through 2019. Can the authors provide further details as to how frequently patients were assessed in a clinical setting, and how maximum follow-up was defined? - In applying the screening criteria, the authors say that patients with baseline CVD were excluded, including various component diagnoses. However, can the authors detail how this assessment was made? Was this based on administrative coding data, manual chart review, laboratory or imaging data, etc.? This needs to be clearly stated to provide a better appraisal of possible bias. Indeed, in the Framingham cohort, >50% of the cohort was taking lipid lowering agents at the time of cohort entry, and >50% were on antihypertensives. In the PCE cohort, these were each >60%, and >50% had diabetes. Could patients who truly had prior established CVD have entered into this cohort? The authors should consider and discuss the possibility of selection bias and exposure misclassification bias. - Similarly, that authors should detail the adjudication of cardiovascular endpoints. How were these assessed for each model? Administrative codes, manual adjudication? How were death events reviewed to assess for CVD or non-CVD causes? Do all of these patients receive medical care exclusively at this one hospital, or could they have experienced a CVD event in a different health system, thus leading to under-ascertainment of CVD outcomes? These issues should be addressed in the methods and limitations, as applicable. - In the limitations section, the authors state that the “retrospective nature of our study presents an inherent limitation.” This comes

	across as very vague. Can the authors expand on this to be more concrete? For example, a discussion of possible misclassification of exposures and outcomes, and the possible implications of this. - The impact of this work would be significantly strengthened by concomitant derivation and validation of a refitted/recalibrated ASCVD score for use in the Arab population. This may be a goal of the authors in a subsequent manuscript if not felt to be within the scope of the current study.

VERSION 1 – AUTHOR RESPONSE

REVIEWER #1 Jianchao Quan, University of Hong Kong:

Thank you for your valuable comments on our manuscript. The following are our point-by-point responses.

1. The number of patients 1,048 patients for validation is small and the recruitment from a single outpatient clinic is unlikely to be representative of general population.

   Author's response: *Thank you for your comment. This is an important point raised by the reviewer. We agree that patients from one hospital may not correctly represent the whole population of the UAE. However, to clarify, Tawam Hospital is one of the largest publicly funded health facilities in the UAE, and its outpatient medical departments, which consists of both primary care clinics and specialty clinics, serve most of the UAE nationals of Al-Ain city. Furthermore, the ideal target population who would benefit the most from risk stratification are patients visiting outpatient clinics rather than the general population.*
   The following lines had been included in the limitation section addressing this point, Discussion, Strengths and limitations section page 16, lines 312−315:
   "Finally, our study population was from outpatient clinics of a single large medical center, and the results may not be applicable to the general Emirati population. However, patients visiting outpatient clinics would be the ideal target for risk stratification and subsequent preventive therapies."

   With regard to the sample size, it has been suggested that a minimum of 100 observed events are required to validate the predictive performance of models[1] which our study met.

   [1]Collins, G. S., Ogundimu, E. O., & Altman, D. G. (2015). Sample size considerations for the external validation of a multivariable prognostic model: A resampling study. Statistics in Medicine, 35(2), 214-226.

2. How did the authors account for (competing risks) deaths from recruitment in 2008 to the follow-up date in 2019?

   Author's response: *Thank you for your comment. We agree that non-CVD death is a potential competing risk event and if we were developing a novel risk prediction model, an alternative survival analysis method specifically designed for assessing competing risks data, such as the proportional subdistribution hazards model (i.e. Fine Gray model) would have been more appropriate. However, the aim of this study was to only validate existing prediction models in our population using a similar approach (i.e. outcomes) and statistical method to that of the original derivation studies and of subsequent validation studies in order to accurately compare our results (references # 9, 15, 16, 17, 18, 19, 20, 21 & 22).*

3. Could the authors test different risk thresholds to classify high-risk for improved performance.

   Author's response: *Thank you for your suggestion. Testing different high-risk thresholds for improved performance would be ideal after the recalibration of the models. However, the*

recalibration of the models is beyond the scope of this study. Furthermore, we believe that the more appropriate next step would be to develop a novel prediction tool to accurately estimate vascular risk by using local data that include younger age groups and emerging risk factors rather than recalibrating existing poor performing prediction models. This is being explored as a potential future project.

4. Several parts of the TRIPOD statement (e.g. methods) appear unavailable when they could be stated in the manuscript.

   Author's response: *Thank you for pointing this out. The parts to the TRIPOD statement have been updated that are relevant to model validation.*

5. Line 98 How was "history of CVD" assessed – e.g. manual review of patient medical records, ICD diagnosis codes, patient recall?

   Author's response: *Thank you for your comment. Baseline data which included a history of CVD were obtained by manual review of patients' medical records.*

   The following sentence has been edited for clarification. Methods section page 5, lines 95−96:
   "Baseline ambulatory electronic medical records of patients were manually extracted from April 1 to December 31, 2008."

6. Line 106 Suggest the authors list out the variables in each respective model for the reader's ease of comprehension.

   Author's response: *Thank you for your suggestion. The variables for each model have been listed in S1 Table (Supporting Information):*
   "S1 Table. List of variables used for each CVD risk prediction model"

7. Line 120 "derivation cohort" – unsure why this is a derivation cohort given no model development. It seems only validation was conducted.

   Author's response: *Thank you for your comment. The term 'derivation cohort' refers to the cohorts from the original studies. The following sentence has been edited for clarity. Methods section page 6, lines 121−122:*
   "During follow-up, the primary endpoints were defined for each model and assessed separately based on the original Framingham and PCE cohorts."

8. Line 129 How was missing data were assessed, and be "determined" to be missing at random? Multiple imputation is recommended rather than single imputation - this should be quick to do given the small sample size.

   Author's response: *Thank you for your comment. We agree that multiple imputation is recommended when there are large amounts of missing data. However, with small amounts of missing data (<10%) single imputation performs almost equally as well[2]. In our study, the variables with missing data included HbA1c (4.0% missing), BMI (<1% missing), total cholesterol (<1% missing), and HDL-C (<1% missing). As the amounts of missing data were small, it was decided to use the single imputation method.*
   With regard to how missing data was assessed, the validity of single imputation does not depend on whether the data are missing completely at random as it is in multiple imputation. This line has therefore been removed; we thank the reviewer for pointing this out.

   [2]Shrive, F. M., Stuart, H., Quan, H. & Ghali, W. A. (2006). Dealing with missing data in a multiquestion depression scale: a comparison of imputation methods. BMC Medical Research Methodology 6 57.

9. Line 136 Medians (25th percentile, 75th percentile). Suggest replacing with "quantiles/percentiles".

Author's response: *Thank you for your suggestion. The following line has been edited. Methods section page 7, line 140:*
"Medians (quantiles/percentiles) are used…"

10. Table 1&2 –It would be useful to include the event rate by person-year.

Author's response: *Thank you for your comment. The incidence rate per 1000 person-years has been included in Tables 1 & 2.*

11. Table 3 – Could include an overall C-statistic row. Useful to present 95% CI.

Author's response: *Thank you for your comment. Table 3 has been updated to include the overall C-statistic and 95% CIs.*

12. Calibration – unclear how the Figure 2 plots are generated. If it is individual-level calibration an individual may have a predicted risk of 25% but the observed frequency for an individual is 0 or 100%. Is this calibration plot grouped? Given the small sample size of ~500, it is may be beneficial to group the predicted risks by deciles of predicted risk. Figure 2B (women) PCE (black) line - not clear why the observed risk falls to 0% at higher predicted risks unlike the other risk models in men and women.

Author's response: *Thank you for your comment. Plotting the calibration curve by deciles is suitable when validating cross-sectional data, however, with survival data, a comparison using the Kaplan-Meier curves (as was used in this study) provides a more suitable assessment of the models' calibration.*
The likely reason for the observation that the observed risk falls to 0% at higher predicted risks in the calibration curve of the PCE model in women (Figure 2B) is the lower incidence rate of outcome events seen among women in the PCE cohort compared to the other risk models in men and women. The incidence rate of events (per 1000 person-years) in men and women in the Framingham cohort were approximately 25 and 19, respectively (Table 1), while in the PCE cohort the incidence rate of events (per 1000 person-years) in men and women were approximately 20 and 10, respectively (Table 2).

REVIEWER #2 Nadim Mahmud, Hospital of the University of Pennsylvania, United States of America:

Thank you for your valuable comments on our manuscript. The following are our point-by-point responses.

1. In this very interesting study, Al-Shamsi et al. evaluate the performance of cardiovascular disease (CVD) risk prediction scores in an Emirati population. They highlight that while the recommended CVD prediction models have been derived and validated primarily in Caucasian and African American populations, they have not been previously tested in an Arab cohort. The authors found that the Framingham and Pooled Cohort Risk Equation (PCE) models overestimated CVD risk in both men and women in higher ranges of predicted risk, and the models underestimated risk in women in the lower range of predicted risk. The discrimination of the scores in men was C-statistic = ~0.7 and in women C-statistic = ~0.75. Overall, this is a well-conducted study with generally appropriate statistical methods and valid conclusions. However, I do have a number of recommendations and critiques that I hope may strengthen this work prior to publication. In particular, aspects of the methods regarding cohort selection and outcomes adjudication need to be described in further detail.

Author's response: *Thank you for your comment. Further detail below and in the manuscript has been provided to describe cohort selection and outcomes adjudication.*

Major:

2. - I am admittedly not familiar with the UAE health system, but it seems incredible that only four patients screened in 2008 were lost to follow-up through 2019. Can the authors provide further details as to how frequently patients were assessed in a clinical setting, and how maximum follow-up was defined?

Author's response: *Thank you for your comment. To clarify, Tawam Hospital is one of the largest publicly funded health facilities in the UAE, and its outpatient medical departments, which consists of both primary care clinics and specialty clinics, serve most of the UAE nationals of Al-Ain city. Furthermore, Tawam Hospital is the only publicly funded hospital in Al-Ain city that provides medical care exclusively to UAE nationals therefore the rates of patient's lost-to-follow-up are relatively low.*
In this retrospective study, patients' charts were reviewed from the baseline clinic visit in 2008 annually, until December 31, 2019. Patients were defined as lost to follow up if their last clinic visit was prior to 12 months from the baseline visit and they were not known to have a cardiovascular outcome.

The following sections have been modified/included to address the points raised above,
Methods section page 5, lines 93−97:
"This 10-year retrospective cohort study was conducted at outpatient clinics of Tawam Hospital, a large, government-subsidized tertiary care hospital in Al-Ain, UAE, which provides medical care exclusively for UAE nationals. … Follow-up data were reviewed annually until December 31, 2019."

Methods section page 5, lines 102−103:
"…four patients who were lost to follow-up, defined as last clinic visit prior to 12 months from the baseline visit and were not known to have a cardiovascular outcome."

3. - In applying the screening criteria, the authors say that patients with baseline CVD were excluded, including various component diagnoses. However, can the authors detail how this assessment was made? Was this based on administrative coding data, manual chart review, laboratory or imaging data, etc.? This needs to be clearly stated to provide a better appraisal of possible bias. Indeed, in the Framingham cohort, >50% of the cohort was taking lipid lowering agents at the time of cohort entry, and >50% were on antihypertensives. In the PCE cohort, these were each >60%, and >50% had diabetes. Could patients who truly had prior established CVD have entered into this cohort? The authors should consider and discuss the possibility of selection bias and exposure misclassification bias.

Author's response: *Thank you for your comment. Baseline ambulatory data which included a history of CVD were obtained by manual review of patients' medical records and was based on diagnosis established by specialist physicians such as neurologists, neurosurgeons, cardiologists, cardiac surgeons, and vascular surgeons. As such, the probability of selection bias is low, however, we agree that retrospective studies are prone to such biases and is a possible limitation to consider.*

The following sentence has been edited for clarity. Methods section page 5, lines 95−96:
"Baseline ambulatory electronic medical records of patients were manually extracted from April 1 to December 31, 2008."

And page 5, line 99:
"A history of CVD was defined as a previous diagnosis established by specialist physicians of a…"

Strengths and limitations section page 16, lines 309−312:
"Second, the retrospective nature of our study presents an inherent limitation, possibly rendering it prone to selection bias. In addition, misclassification or under-ascertainment of cardiovascular endpoints may have led to the overestimation observed by the risk models."

4. - Similarly, that authors should detail the adjudication of cardiovascular endpoints. How were these assessed for each model? Administrative codes, manual adjudication? How were death events reviewed to assess for CVD or non-CVD causes? Do all of these patients receive medical care exclusively at this one hospital, or could they have experienced a CVD event in a different health system, thus leading to under-ascertainment of CVD outcomes? These issues should be addressed in the methods and limitations, as applicable.

Author's response: *Thank you for your comment. Adjudication of cardiovascular endpoints, including cardiovascular death, was conducted manually by reviewing the electronic medical records and death certificates of all study participants from their baseline visit in 2008 until December 31, 2019 (Methods section, page 5, lines 95−97 and page 6, lines 130−131).*
The following section has been added to better clarify how cardiovascular death was assessed, Methods section, page 6, lines 126−128:
"Coronary death or fatal stroke was determined from the manual review of electronic medical records or death certificates. In addition, sudden death outside the hospital was considered a coronary death unless documented otherwise."

As described in response to comment # 2, Tawam Hospital is the only publicly funded hospital in Al-Ain city that provides medical care exclusively to UAE nationals therefore the probability of *under-ascertainment of CVD outcomes is low, however, we agree, that retrospective studies are prone to misclassification bias.*
The following limitations section has been expanded for clarity, Strengths and limitations section 16, lines 309−312:
"Second, the retrospective nature of our study presents an inherent limitation, possibly rendering it prone to selection bias. In addition, misclassification or under-ascertainment of cardiovascular endpoints may have led to the overestimation observed by the risk models."

5. - In the limitations section, the authors state that the "retrospective nature of our study presents an inherent limitation." This comes across as very vague. Can the authors expand on this to be more concrete? For example, a discussion of possible misclassification of exposures and outcomes, and the possible implications of this.

Author's response: *Thank you for your comment. The following limitations section has been expanded for clarity, Strengths and limitations section 16, lines 309−312:*
"Second, the retrospective nature of our study presents an inherent limitation, possibly rendering it prone to selection bias. In addition, misclassification or under-ascertainment of cardiovascular endpoints may have led to the overestimation observed by the risk models."

6. - The impact of this work would be significantly strengthened by concomitant derivation and validation of a refitted/recalibrated ASCVD score for use in the Arab population. This may be a goal of the authors in a subsequent manuscript if not felt to be within the scope of the current study.

Author's response: *Thank you for your comments and suggestion. We believe that the more appropriate next step would be to develop a novel prediction tool to accurately estimate vascular risk by using local data that include younger age groups and emerging risk factors rather than recalibrating existing poor performing prediction models. This is being explored as a potential future project.*

VERSION 2 – REVIEW

REVIEWER	Jianchao Quan The University of Hong Kong
REVIEW RETURNED	19-Sep-2020

GENERAL COMMENTS	The authors have addressed my previous comments.

REVIEWER	Nadim Mahmud Hospital of the University of Pennsylvania United States of America
REVIEW RETURNED	23-Sep-2020

GENERAL COMMENTS	The authors have satisfactorily addressed my concerns.