Sequential testing in high stakes OSCE: a stratified cross-validation approach

Giovanni Mancuso[1], Stephanie Strachan[1], Steven Capey[2]

Corresponding author: Mr Giovanni Mancuso g.mancuso1983@gmail.com
Institution: 1. King’s College London, 2. Swansea University
Categories: Assessment, Research in Health Professions Education

Received: 13/05/2019
Published: 13/06/2019

Abstract

Sequential testing has been employed in clinical assessments to support student progression decisions by strategically targeting assessment resources towards borderline students. In this context resampling techniques have been utilised in the attempt to determine the appropriate blueprint number of stations to include in the screening phase of a sequential exam. However, statistical overfitting undermines the generalizability of examination psychometric properties and the uneven distribution (imbalance) of borderline vs. non-borderline students may cause resampling methods to produce biased results. Both phenomena may mislead educational practitioners when redesigning sequential assessments.

We demonstrate how to mitigate against the problems of overfitting and imbalanced cohorts whilst finding the optimal 9 screening stations out of an 18-station OSCE. To prevent overfitting our statistical model was developed on one set of data (train and test) and then validated on a different dataset (validation) with imbalance accounted for by operating a stratified sampling scheme.

The outcomes demonstrate the importance of validation: in the development phase, the accuracy was initially 91% (train) but the actual predictive accuracy when mitigating against overfitting and imbalance was 86% (test). Similarly, when we validated the model on completely new data – with a comparable assessment – the predictive accuracy was 83% (validation).

Keywords: Sequential Objective Structured Clinical Examination - OSCE; blueprint development; progression decisions; sensitivity; specificity; overfitting.

Introduction

Objective Structured Clinical Examinations (OSCEs) are considered the current ‘gold standard’ for the assessment of clinical and practical competences in undergraduate medicine (Norman, 2002, Harden, 2016). Despite their
popularity, OSCEs are the most resource intensive components of undergraduate assessment and pose a series of problems related to their costs, resources and implementation (Reznick et al., 1993, Cusimano et al., 1994). For these reasons, the concept of sequential design for OSCEs has emerged. Sequential OSCEs can achieve robust and reliable progression decisions economically, by targeting assessment resource at the borderline students, and progressing competent students on a reduced assessment blueprint, thereby providing a cost-effective alternative to standard OSCE formats while still producing robust pass/fail decisions (Cookson et al., 2011, Pell et al., 2013). The most commonly used sequential design is where all students are required to sit for an initial screening test; the screening test comprises a diverse diet of stations representative of the whole OSCE, and can therefore be much smaller than a traditional OSCE. Students that achieve a high standard on the screening test are exempted from further testing; students who do not achieve the set standard, or are considered to have only reached a borderline performance, are required to complete the extended OSCE and the final pass/fail decision is based on their performance on all the examination components combined – both the screening test and the extended exam.

When designing an OSCE assessment faculty need to consider, amongst a variety of aspects, the total number of stations to include. However, deciding the appropriate variety and number of stations remains a complex issue.

Part of the problem lies in the trade-off between the exam length and its psychometric characteristics. Longer exams are more reliable, but does the gain in reliability really justify the associated increment in costs (Wainer and Feinberg 2015)? Also, longer exams, in principle, allow testing of a broader range of elements of medical practice with a consequent gain in examination content validity (Kane, 2006, Clauser, Margolis and Swanson, 2008).

Another consideration in choosing the appropriate number of stations in a screening test concerns the balance between (a) test classification errors (are borderline and non-borderline students correctly classified) and (b) the necessity to minimize the cost of delivery of the exam. Several studies have investigated the best methodology to determine the appropriate number of stations to include in a screening OSCE (Colliver et al., 1991, Colliver, Vu and Barrows, 1992, Cass et al., 1997, Currie, Selvaraj and Cleland, 2015, Smee et al., 2003, Hejri et al., 2016), however currently no consensus has been reached (Muijtjens et al., 2000, Muijtjens, Van Luijk and Van Der Vleuten, 2006). Receiver operator characteristic (ROC) curves are the most prevalent statistical methodology utilised to model the predictive ability of the initial screening test to achieve appropriate certainty of the borderline decision (Regehr and Colliver, 2003).

The recent advancements of computing power together with the availability of historical data sets constitute an emerging territory that is proving to be extremely fruitful in educational settings (Muijtjens et al., 2003, Kotsiantis, 2012, Homer et al., 2016). In the context of sequential OSCE, Currie, Selvaraj and Cleland (2015) recently suggested a new resampling procedure to establish the minimum number of stations to include in the initial screening exam. By randomly sampling a variable number of stations, the authors were able to estimate the number of stations needed in order to achieve a minimum level of desired accuracy; but as the authors state in the discussion section: "The results presented are only simulations and fail to guarantee what could actually might happen in real life…” (Currie, Selvaraj and Cleland, 2015).

In our investigation we expanded this line of research to directly address the problem of the predictive accuracy of a theoretical smaller screening OSCE. In other words, once the number of stations has been established through resampling, how closely will this short exam forecast future outcomes? Stated differently, how precisely will a shorter screening OSCE predict exam outcomes employed on different cohorts of students? Is it possible to gain such insight from resampled data?

In order to answer these questions we focused our analysis on a theoretical 9-station screening assessment in an 18-station examination. The number of stations chosen in this study was motivated by previous findings where a
reduction from 15 to 8 stations was found to represent an appropriate and robust approach for optimal exam length and accuracy (Currie, Selvaraj and Cleland, 2015).

To illustrate how predictive accuracy can be correctly estimated we have identified two areas that should be carefully considered in order to obtain unbiased results in simulation settings. Hence, we used the analysis of 9-stations screening exam as a proof of concept, to demonstrate the paramount importance of these two areas and the associated risks in neglecting them. The two issues are presented in the following:

1. The generalisability of the predictions based on statistical models. It is well known that statistical models have the tendency to overfit the data (capturing noise together with signal (Hastie, Tibshirani and Friedman, 2001, Pitt and Myung, 2002). This problem is associated with overly optimistic model forecasts of future performances and can mislead practitioners and compel them to take sub-optimal actions. It is important to stress that overfitting is pervasive and will affect any outcome measure chosen to evaluate the robustness of the classification (accuracy, sensitivity/specificity or composite metrics). We therefore adopted two measures to overcome this prediction bias. First, a cross-validation approach was employed (Browne, 2000, Stone, 1974) where the original data were split into train and test set and the statistical model was developed using only the train data. Subsequently, only the best-fitting models were assessed on the unseen observations contained in the test dataset. Effectively the test set constitutes "new" data that can be employed to objectively and impartially evaluate the predictive power of the model. In an effort to further strengthen the conclusions obtained, a separate data set (validation set) from a different academic year was employed to independently assess the predictions of our model.

2. The class imbalance problem, that is to say the uneven class distribution of borderline vs. non-borderline students. Typically borderline students represent only a small section of the entire cohort and therefore constitute the minority class. Estimates of the prevalence of borderline students in a screening assessment vary greatly across studies and institutions ranging from 4-7% (Currie, Selvaraj and Cleland, 2015) to 9-41% (Colliver, Vu and Barrows, 1992, Rothman et al., 1997, Hejri et al., 2016). In our experience the prevalence of borderline students in a screening assessment was estimated between 11-25% among the final year examinees. In such cases, if a random sampling scheme is adopted, there will be samples without borderline students, especially if the prevalence is very low. In a classification exercise, where the goal is to evaluate how a new decision-making test (i.e. 9-station screening) agrees with the current standard (18-station screening), the lack of the examples from one class can severely bias the outcomes (Forman and Scholz, 2010). In order to overcome this problem, we opted for a stratified sampling scheme whereby the split between train and test set was random but the overall prevalence of borderline students was kept constant in the two sets. Indeed the stratified sampling guarantees that, irrespectively from the random split, there will always be enough examples from the minority class to develop a robust statistic (Kohavi, 1995).

The above two points were addressed in our simulation scheme (i.e. stratified cross-validation) where the stochastic nature of the simulations covered different possible scenarios and naturally evaluated the random variability of the process.

Whilst the predictive power of a test is a key component in assessing the generalisation of that test’s conclusions. In undergraduate medical school OSCE examinations content validity is achieved with a diverse blueprint designed to test important areas of medical practice as well as align the exam content with internal requirements (e.g. school curricula) and national standards of care; construct and concurrent validity can be measured through extrapolation, relating the exam outcomes to the expertise and proficiency of the examinees in work-place environments (Kane, 2006, Clauser, Margolis and Swanson, 2008).
However, in the context of station-selection-by-resampling aspects such as blueprint diversity, validity and extrapolation have previously been neglected with outcomes evaluated solely in terms of sensitivity and specificity (Currie, Selvaraj and Cleland, 2015). This constitutes an important limitation because resampling methodologies rely heavily on chance and it is therefore imperative that extra checks are put in place to guarantee that the blueprint targets and exam content are attained. This lapse may erroneously suggest a restricted applicability or practical utility and discourage the employment of the methodology. Resampling methods are, on the contrary, very robust and should be flexibly adapted to inform blueprint selection. In this work we advocate the need to evaluate if the resampling methodology contributed to the development of an effective blueprint safeguarding both the content validity as well as the ability of the exam to discern between borderline and non-borderline students. In this study, we found that the best short version of the exam preserved the content targets of the longer examination. Moreover, we unequivocally found that the blueprint diversity was necessary to obtain the highest levels of predictive accuracy. Despite our results, the question of what to do in cases in which the most predictive resampled exam is not representative of the content of the original exam remains open. Since the blueprint reflects internal as well as national standards, how much of the content can be sacrificed is a policy question and the answers can only be found in the rules and regulations of individual institutions. In this study we suggest that by focusing on subscales – rather than single items – it should be possible to (1) diagnose a departure from blueprint targets and (2) perhaps more importantly, to weigh the depth of the exam against its associated predictive capacity. In other words, the joint analysis of the two dimensions, i.e. the predictive power and the range and diversity of the exam, will allow practitioners to achieve an optimal blueprint.

Finally, in the effort to investigate how resampling techniques perform in a different scenario we designed a simple simulation exercise using the framework of Item Response Theory (See Supplementary File 1). This simplified study is designed to reassure that the methodology is robust, applicable outside the field of medical education and of interest for a wider audience.

**Methods**

Ethical approval has been granted by the Biomedical Sciences, Medicine, Dentistry and Natural & Mathematical Sciences (BDM) Research Ethics Panel for this retrospective observational study.

**OSCE overview**

To illustrate how our stratified cross-validation resampling scheme operated we included only those academic data bases that were sufficiently comparable. Of all the historical data available three final year MBBS student OSCE results (academic years 2011-12 to 2013-14) were selected for statistical analysis. These cohorts provided the most homogeneous data sets, based on the similarity of blueprint and station construction. For each academic year cohort the OSCE were of the same sequential design with an initial 18 station screening assessment. The station content was blueprinted against both the GMC Tomorrow’s Doctor (2009) outcomes and the King’s College London curriculum. The OSCE comprised stations belonging to 5 broad clinical domains: history taking, examination, management, communication and practical skills; each domain was not equally represented in the exam (3, 5, 6, 2 and 2 stations respectively).

The logistics of the exam were also comparable across these academic years: individual stations lasted for seven and a half minutes with a gap of one minute between consecutive stations. The OSCE screening test took place each year over a number of days due to the large cohorts and the stations vary on each of these days whilst retaining a near identical blueprint.

In the academic year 2014-15 – as part of a new curriculum implementation and efforts to streamline various processes and operations – the OSCE changed, moving from 18 to 16 stations with. Only one – instead of two – of
both the prescribing and communication stations were employed in the exam. Also, the station length extended to 8 minutes of testing with a 2 minute gap between stations. Due to these differences the data set was not included in the development phase of the statistical model but instead employed as a validation data set to investigate the generalisability or predictive power of the proposed model.

**Data collection, screening and storage**

All the data employed here were produced in the context of summative year 5 OSCE examinations. The data, gathered during the exam, were subsequently digitalized with the commercially available NCS Opscan 4U optical reader and software ScanTool Plus. This operation was done after a calibration procedure of the instruments that insured that both the optical reader and the software produced valid data bases.

**Screening test outcomes**

The sequential testing is a short-term remediation framework that allows Medical Schools to achieve robust decisions about student's competency (Cookson et al., 2011, Pell et al., 2013). The sequential testing framework comprises two concatenated steps:

1. First the entire cohort of students sits for the screening examination. This is a short but representative version of the whole OSCE.
2. A decision about student's status (borderline vs. clear pass) is achieved by a combination of (a) minimum standard achievement (i.e. a pass-mark aggregated across stations) and (b) a minimum number of stations passed.

2.1 All the students that achieve a "clear pass" status are exempted from further examination.

2.2 The remaining students that do not achieve a "clear pass" status are invited for the extended component of the OSCE. This re-test comprises a diverse diet of OSCE stations and the final the final pass/fail decision is based on their performance on all the examination components combined.

In our institution the outcome of the screening test was established for each student as follows: in each station an examiner records the score for the student using a combination of item marking and global ratings. This practice is in line with previous studies that indicated that the objective assessment of the performance is guaranteed by the combination of (a) the standardization of the station tasks and (b) the joint employment of scoring checklists and global ratings (Swanson and van der Vleuten, 2013). On the one hand checklists ensure that procedural skills that require precision in execution are evaluated appropriately. This information supports the scoring of the global rating scales which, in turn, provide a more holistic judgment of the performance.

The student performance is then compared to the station pass mark (previously established with the Angoff method) to establish a station-level outcome (pass or fail). The outcomes of each station in the screening test are combined to derive an overall score attained by the student. In order to pass the OSCE at this screening stage there were two criteria that must be met; to reflect the uncertainty in the screening process and adopt a safe approach, the student needed to gain an overall score greater than the pass-mark plus the Standard Error of the Measurement (SEM) and to prevent compensation across the entire screening test, the student must pass a minimum number of stations (12 stations out of 18, equating to 67%). If both these criteria were met then the student was not required to sit for the remainder of the OSCE (the extended component) and is said to have passed at the screening test.

Mimicking these real standards, in our 9-station simulation of the exam we deemed a student achieved the status of "clear pass" if a minimum overall score plus SEM was attained and the minimum number of 6 stations was passed: six stations out of nine represent an identical threshold (67% of the total stations) to the threshold employed in the existing 18-station screening test.
Cohorts

The statistical modelling was developed on a total number of 1306 students (see Table 1) from the academic years 2011-12 to 2013-14 (n=431, 454 and 421 respectively). Following the screening tests between ~11% and ~25% of the students were classified as borderline and were invited for an extended assessment.

In the academic year 2014-15 (n=411) about 11% of the students were classified as borderline after the screening test.

Table 1: The prevalence (class-distribution) of borderline students in the screening test.

<table>
<thead>
<tr>
<th>Academic Year</th>
<th>Number of borderline students</th>
<th>Total number of students</th>
<th>% borderline</th>
</tr>
</thead>
<tbody>
<tr>
<td>2011-12</td>
<td>108</td>
<td>431</td>
<td>~25%</td>
</tr>
<tr>
<td>2012-13</td>
<td>76</td>
<td>454</td>
<td>~17%</td>
</tr>
<tr>
<td>2013-14</td>
<td>48</td>
<td>421</td>
<td>~11%</td>
</tr>
<tr>
<td>2014-15</td>
<td>45</td>
<td>411</td>
<td>~11%</td>
</tr>
</tbody>
</table>

Analysis

The statistical analysis consisted of two parts: the iterative search and the cross-validation. The iterative search was nested inside the cross-validation and repeated multiple times.

Iterative search

In order to determine the optimal set of 9-stations, for the screening OSCE from the total of 18 stations that characterized the complete test, an iterative search was employed. The diagram below depicts the iterative procedure: in an 18-station OSCE there are 48620 possible combinations of a 9-station screening. For instance, in iteration 1, the selected stations were station 1 to 9; in iteration 2 the selected stations were station 2 to 10 and so on until every possible permutation was exhausted (see Figure 1).

Figure 1: Overview of the iterative search employed to find the best-fitting set of 9 stations.
screening classification (Borderline vs. Clear Pass) mapped the classification decision of the 18-station screening by means of sensitivity and specificity. Historically, a number of different key indicators have been employed to evaluate the accuracy of the screening test including pass rate, passing error rate, accuracy and positive predictive value or composite scores to name a few (Colliver et al., 1991, Colliver, Vu and Barrows, 1992, Cass et al., 1997, Rothman et al., 1997, Hejri et al., 2016). Here we opted for the calculation of sensitivity and specificity. Sensitivity was defined as the number of students that passed the theoretical 9-station screening test over the total number of students that passed the real 18-station test. Similarly, specificity was defined as the number of students that were failed by the theoretical 9-station screening test over the total number of students that failed the real 18-station test. As a key indicator of the classification performance, we wanted to equally capture both sensitivity and specificity. However, as explained in the previous sections, the data set contains a strong imbalance in the class labels with only a minority of students (about 18% pooling across academic years) who actually failed the 18-station test. Therefore, we employed the balanced accuracy metric (average of sensitivity and specificity) to select the best-fitting 9-station screening. Balanced accuracy is proven to be a better choice than overall accuracy in case of unbalanced data (Velez et al., 2007).

The stratified Monte Carlo cross-validation approach

The iterative search explored in the previous section was repeated 100 times with random samples. The procedure was based on three sequential steps (see Figure 2):

1. The entire data set is first randomly split into train (60% of the data) and test (40% of the data) sets (left and right column respectively of Figure 2). Stratification imposes that the percentage of borderline students (prevalence) is constant in the two sets to insure a comparable imbalance in the data.

2. The classification model is developed on the train set alone while the test set is kept aside. The best-fitting set of 9-station screening OSCE is then identified by means of the iterative search (Figure 2 overlapping panels). At each step of the iteration (iteration 1 to iteration 48620) a subset of 9 stations is drawn from the total number of available stations. This set of 9 stations is then considered the screening test and the results are calculated. Finally the results of this 9-station screening OSCE are then compared with the outcome of the actual 18-station screening exam to derive performance metrics including sensitivity, specificity and balanced accuracy.

3. In order to test the model’s predictive value, the set of 9 station associated with the highest balanced accuracy was subsequently drawn on new observations contained in the test set.

Figure 2: Overview of three repetitions (1, 2 and 100) the cross-validation process.
At the end of the simulation, the model's predictive performance was further validated on a completely new data set, i.e. the academic year 2014-15.

Taken together, the above steps address the pitfalls of simulation studies in sequential testing: the stratification is of paramount importance given the high imbalance in the data. Indeed it guarantees that, irrespective of the random split, there will always be enough examples from the minority class (in our case borderline students) to develop a
robust statistic (Kohavi, 1995, Forman and Scholz, 2010). The split into train and test sets insures that the model is validated on unseen observation and therefore avoids the pitfalls of over-fitting. And the stochastic nature of the repetitions insures that the results are not contingent of the data set at hand but provides an estimate of the uncertainty related to the predictive accuracy and the other summary metrics.

Results/Analysis

Train data set results

Table 2 shows the median value and the 5th-95th percentiles of sensitivity, specificity and balance accuracy across the 100 cross-validation repetitions.

Table 2: Train data set summary of sensitivity, specificity and balanced accuracy.

<table>
<thead>
<tr>
<th>Metrics</th>
<th>SEM cutoff</th>
<th>Median</th>
<th>5th-95th Percentile</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensitivity 1</td>
<td>0.87</td>
<td>0.75 0.95</td>
<td></td>
</tr>
<tr>
<td>Sensitivity 2</td>
<td>0.64</td>
<td>0.46 0.81</td>
<td></td>
</tr>
<tr>
<td>Sensitivity 3</td>
<td>0.32</td>
<td>0.17 0.51</td>
<td></td>
</tr>
<tr>
<td>Specificity 1</td>
<td>0.81</td>
<td>0.56 0.95</td>
<td></td>
</tr>
<tr>
<td>Specificity 2</td>
<td>0.95</td>
<td>0.81 0.99</td>
<td></td>
</tr>
<tr>
<td>Specificity 3</td>
<td>0.99</td>
<td>0.96 1.00</td>
<td></td>
</tr>
<tr>
<td>B. Accuracy 1</td>
<td>0.84</td>
<td>0.75 0.88</td>
<td></td>
</tr>
<tr>
<td>B. Accuracy 2</td>
<td>0.79</td>
<td>0.72 0.84</td>
<td></td>
</tr>
<tr>
<td>B. Accuracy 3</td>
<td>0.66</td>
<td>0.58 0.74</td>
<td></td>
</tr>
</tbody>
</table>

The effects of the strong imbalance can be easily seen in the table: when the cut-off SEM was set to be very harsh (i.e. 3) then the statistical model was able to correctly identify virtually all the failing students and the specificity became very high (about 0.99 [0.96-1.00]) but the sensitivity plummet dramatically (about 0.32 [0.17-0.51]). However, a more lenient SEM threshold (i.e. 1) doesn’t have the opposite effect on specificity (0.81 [0.56-0.95]) and sensitivity (0.87 [0.75-0.95]) as one would expect in perfectly balanced data sets.

To identify the characteristic of the "optimal" exam blueprint, in each simulation we choose the set of 9-station screening associated with the highest balanced accuracy. The table presented below (see Table 3) shows the highest values of balanced accuracy across the 100 cross-validations. The 1 SEM is associated with the highest balanced accuracy which reflects the best trade-off between specificity and sensitivity.

Table 3: Train data set summary of the highest balanced accuracy in the 100 cross-validation.

<table>
<thead>
<tr>
<th>Metrics</th>
<th>SEM cutoff</th>
<th>Median</th>
<th>5th-95th Percentile</th>
</tr>
</thead>
<tbody>
<tr>
<td>B. Accuracy</td>
<td>1</td>
<td>0.91</td>
<td>0.90 0.92</td>
</tr>
</tbody>
</table>
B. Accuracy 2 0.87 [0.86 0.89]
B. Accuracy 3 0.80 [0.79 0.82]

The reliability (Cronbach’s alpha) of the best station set was 0.52 (95% confidence interval: 0.45-0.57). Also, the predicted reliability of the extended examination (18 stations), as calculated with the Spearman–Brown prophecy formula, 0.69 (95% confidence interval: 0.62-0.73).

**Test data set results**

The analysis of the test set allows us to make predictions of what might happen when the best-fitting stations are applied on a different data set and to directly address the problem of over-fitting. As expressed by Pitt and Myung (2002, p.422) over-fitting can be quantified by “…assessing how well a model's fit to one data sample generalises to future samples generated by the same process”.

**Table 4: Summary of the highest balanced accuracy in the 100 cross-validation repetitions of the test set.**

<table>
<thead>
<tr>
<th>Set</th>
<th>Metrics</th>
<th>Median</th>
<th>5th-95th Percentile</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test</td>
<td>Sensitivity</td>
<td>0.84</td>
<td>[0.79 0.89]</td>
</tr>
<tr>
<td>Test</td>
<td>Specificity</td>
<td>0.88</td>
<td>[0.78 0.95]</td>
</tr>
<tr>
<td>Test</td>
<td>B. Accuracy</td>
<td>0.86</td>
<td>[0.83 0.89]</td>
</tr>
</tbody>
</table>

As expected (see Table 4), the overall fitness of the best-fitting 9-station screening models on the test set (0.86 [0.83-0.89]) was lower that the equivalent on the train set (0.91 [0.90-0.92]). Again, this reflects the overly-optimistic results of the train set (overfitting). This result highlights the perils of overfitting and suggests that a more conservative estimate – as the one obtained on the test set – is more likely to be accurate in real settings.

The robustness of the cross-validation process is also demonstrated by the reliability calculated on the test data. As can be seen the obtained reliability was remarkably similar (0.52 [0.45-0.56]) to the reliability obtained on the train data (0.52 [0.45-0.57]). Since the selection of the best 9-station screening in the train data was based on the balanced accuracy and not on the reliability itself, no overfitting is observed in the estimate of the reliability.

**Validation data set results**

The academic year 2014-15 wasn’t employed in the development of the statistical model as it contained a different number of stations (16 instead of 18). However, the station content of the exam remained unchanged so the data was set aside and employed as validation set to further investigate the generalisability or predictive power of statistical model.

As can be seen from Table 5 the balanced accuracy (0.83 [0.79-0.91]) is smaller than the corresponding accuracy in the train set (0.91 [0.90-0.92]). Again, this reflects a realistic estimation of the predictive validity of our model on a "fresh" data set. The balanced accuracy of the validation set is also smaller than the corresponding value in the test set (0.86 [0.83-0.89]). Although this result may seems odd, it becomes clear when we consider that – as explained above – not all the stations tested in the academic years 2011-12 to 2013-14 featured in the academic year 2014-15. Thus, effectively the predictive results were calculated on 7 or 8 stations instead of 9 stations (about 94% of the cross-validation repetitions were based on less than 9 stations).

In fact, the smaller number of stations available in the validation set also determined a decrease in the reliability
measure (from 0.52 to 0.42).

Table 5: Metrics summary in the 100 cross-validation repetitions of the validation set.

<table>
<thead>
<tr>
<th>Set</th>
<th>Metrics</th>
<th>Median</th>
<th>5th-95th Percentile</th>
</tr>
</thead>
<tbody>
<tr>
<td>Validation</td>
<td>Sensitivity</td>
<td>0.87</td>
<td>[0.68 0.94]</td>
</tr>
<tr>
<td>Validation</td>
<td>Specificity</td>
<td>0.84</td>
<td>[0.68 0.98]</td>
</tr>
<tr>
<td>Validation</td>
<td>B. Accuracy</td>
<td>0.83</td>
<td>[0.79 0.91]</td>
</tr>
<tr>
<td>Validation</td>
<td>Reliability</td>
<td>0.42</td>
<td>[0.33 0.49]</td>
</tr>
</tbody>
</table>

Domain Analysis of the results

Since the resampling method proposed here is completely agnostic about the exam content, it is imperative that the prospective optimal 9-station blueprint is evaluated against significant deviations from the original content. In the effort to assess this point we focus the analysis on the best-fitting stations clustered according to their domain membership – i.e. important areas of medical practice tested in the original exam. Table 5 shows the domain frequency in the optimal 9-station screening as established through the cross-validation procedure. Each domain frequency (third column) is compared with the relevant chance level (i.e. the frequency of the domain if a random sampling was employed – second column). The communication domain was observed between 10 and 14% of the cases. Such frequency doesn't deviate from what we would expect from a random sampling (2/18 ≈ 11%). This means that the amount of communication stations in the screening OSCE is already at an optimal level. In contrast, the estimated proportion of the examination domains is lower (21% ± 2.7%) than the actual proportion (5/18 ≈ 28%). This means that probably the examination stations could be reduced accordingly without any significant loss of classification accuracy.

The results lead us to propose a summary blueprint for the 9-station screening OSCE where the domain frequency replicates the estimated frequency obtained in the best-fitting exam (Table 6).

Table 6: OSCE blueprint breakdown by domain content: actual count and estimated optimal frequency (95% CI).

<table>
<thead>
<tr>
<th>Domain</th>
<th>Actual count</th>
<th>Estimated optimal frequency</th>
<th>Proposed 9-Station screening OSCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Communication</td>
<td>2/18 (11.1%)</td>
<td>11.5% (9.5-13.6%)</td>
<td>1</td>
</tr>
<tr>
<td>Examination</td>
<td>5/18 (27.7%)</td>
<td>20.6% (18-23.2%)</td>
<td>1</td>
</tr>
<tr>
<td>History Taking</td>
<td>3/18 (16.7%)</td>
<td>17.5% (15.1-20%)</td>
<td>2</td>
</tr>
<tr>
<td>Management</td>
<td>6/18 (33.3%)</td>
<td>40.7% (37.6-43.9%)</td>
<td>4</td>
</tr>
<tr>
<td>Skills</td>
<td>2/18 (11.1%)</td>
<td>9.6% (7.7-11.5%)</td>
<td>1</td>
</tr>
</tbody>
</table>

The prospective optimal 9-station blueprint raises the question of content validity: to what extent the reduction from 18 to 9 stations impacted the assessment of the exam content? From the cross validation exercise we already know that the best classification performance is achieved by a maximally diversified blueprint where all the domains of the medical practice are represented proportionally to their initial contribution in the 18-station exam. Here we provide evidence that such diversity in domain representation is not a consequence of random sampling, but it is actually needed to achieve the best classification accuracy in the 9-station exam. To support the idea that the 9-station exam
is really representative of its initial 18-station version, we looked at the numerosity of the domains that featured in each of the best 100 folds of training set. We hypothesized that, if our selection method capitalizes only on the most discriminatory stations, ignoring altogether the blueprint content and diversity, then the best 9-station exam could have been equally achieved by a 2, 3 or 4 domains exam. In contrast if the 9-station exam well represents the full blueprint of the initial version, then we may expect that the best classification performance is achieved by sampling broadly from all the 5 domains.

As we predicted (see Table 7), the iterations with the best classification performance were composed, by far (76%), by a diverse blueprint with all 5 domains represented. This was followed by 23% of the cross validation folds being composed of 4 domains. Only 1% of the best cross-validation was associated with a 3-domain exam (and none with a 2-domain exam). These results deviate markedly from what can be expected from a random resampling, where domain frequencies would be about 48, 45, 7 and 0.1% respectively.

Table 7: Possible domain distribution in 9-station OSCE blueprint.

<table>
<thead>
<tr>
<th>Domain Numerosity</th>
<th>Possible Combinations</th>
<th>Frequency</th>
<th>Observed</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Best</td>
<td>Worst</td>
</tr>
<tr>
<td>2</td>
<td>56</td>
<td>~0.11%</td>
<td>/</td>
</tr>
<tr>
<td>3</td>
<td>3405</td>
<td>~7%</td>
<td>1%</td>
</tr>
<tr>
<td>4</td>
<td>21842</td>
<td>~44.9%</td>
<td>23%</td>
</tr>
<tr>
<td>5</td>
<td>23317</td>
<td>~47.95%</td>
<td>76%</td>
</tr>
</tbody>
</table>

Crucially, we repeated this analysis, taking the worst performing folds of the cross validation. In agreement with the previous results we discovered that the poorest classification performance is achieved by sampling poorly from the 5 domains: only 2% of the worst-folds were obtained by a fully diversified blueprint whereas the vast majority (98%) of the worst exams was composed by 4 or less domains.

Finally, in order to strengthen this idea that a diversified blueprint is also more predictive of the exam outcomes, we studied how the classification accuracy relates to the number of domains contained in different versions of the 9-station exam by means of linear regression. If the cross validation method presented here captures the depth and breadth of the 18-station exam, then we should expect a positive association between number of domains tested and classification accuracy. A linear model was employed to test this hypothesis and the results confirmed that the classification accuracy (transformed into z-scores) grows lightly but significantly with the increase in number of domains tested (slope = 0.014, p<2e-16).

Taken together, these results indicate that the cross validation sampling strategy is not only leveraging the most discriminative or hard stations, but is really investigating the underlying skills and abilities tested by the full exam.

Discussion

The constant pressure on healthcare systems not only demands evidence to support students’ progression decisions but it also requires that the evidence is worth the scarce resources. In the context of undergraduate medicine qualifications, the concept of sequential OSCE testing, offers a tool to strategically allocate assessment resources towards borderline students allowing robust progression decisions to be made for this group of students (Cookson et al., 2011, Pell et al., 2013). Whilst the sequential OSCE approach has proved to be robust and reliable, how
individual institutions should decide which, and how many stations to include in the screening part remains an unsolved problem.

In recent years the meeting of computational approaches and the availability of historical data revitalised this research field. Currie, Selvaraj and Cleland, (2015) demonstrated how resampling strategies can be employed to select a theoretical smaller exam from a larger screening OSCE.

Despite their flexibility and robustness, resampling methods present their challenges. Our research identified two major limitations of the computational approach suggested by Currie, Selvaraj and Cleland, (2015) and proposes statistical strategies to overcome these problems. The first issue is related to the application of classification strategies with imbalanced data that is, when borderline students represent only a minor fraction of the entire cohort. Second, we examined the effects of statistical overfitting on the predictions of future examinations. Here we have demonstrated how, classification models – if not properly validated – lead to inflated estimates of key indicators such as specificity, sensitivity or overall accuracy. Both problems constitute serious threats to the interpretation of statistical analysis for decision-making: because they are associated with overly optimistic forecasts, they may compel institutions to take actions not supported by the evidence.

In order to address these points we choose the specific task of selecting a theoretical 9-station screening OSCE from a full set of 18-station screening and we illustrated how resampling studies should be designed to alleviate the above-mentioned problems.

Finally, once established an unbiased way to calculate outcome indicators (such as balanced accuracy) our analysis proceeded with the investigation of the effects of resampling on the validity of the blueprint of the exam. Particular emphasis was given on the idea that a highly-predictive short assessment should still be weighed against a diversified blueprint.

Our results, obtained in the context of a simulation analysis of three cohorts of final year students, closely match previous findings. When the screening OSCE was halved (from 18 to 9 stations), the accuracy of the best-fitting model was 0.91 (0.90-0.92). This observation suggests that resampling studies provide robust and consistent results and strengthen the idea that simulations can be flexibly adapted to serve individual circumstances. Furthermore, our results, not only reinforce previous findings, but also highlight the importance of the imbalance in the class distribution (the prevalence of borderline students was about 18% in our sample). When the pass-mark was set to be very high (+3 SEM) – the exam was very harsh as the model captured virtually all the failing students and consequently the specificity reached 0.99 (0.96-1.0). However, this result should be interpreted in light of the very low prevalence of borderline students (similar specificity estimates were documented by Currie, Selvaraj and Cleland, (2015) in which the prevalence of borderline students was between 4 and 7%). In fact, at the same time the sensitivity plummets to 0.32 (0.17-0.51).

We adopted two measures to alleviate the bias in this study. First, we opted for a stratified sampling scheme to insure that the same proportion of borderline students was present in every data set (Kohavi, 1995, Forman and Scholz 2010). This step is particularly important if the students are sampled at random from their respective cohorts: if, there are no borderline students in a given pool, how a model can probabilistically learn how to classify the observations into borderline vs. non-borderline students? Second, as a key indicator, we employed the balanced accuracy (averaging sensitivity and specificity). The reason for this choice are rooted in the statistical properties of balanced accuracy that supersedes those of the overall accuracy as they are less sensitive to uneven distribution of class memberships (Velez et al., 2007). The phenomenon of class imbalance has been overlooked in the literature and we suggest future simulation studies should try to mitigate the imbalance as demonstrated here to avoid biased results (Davis and Goadrich 2006, He and Garcia 2009, Saito and Rehmsmeier 2015).
In terms of cut-off scores it is interesting to note some similarities/differences with the published literature. Notably, despite other studies have employed different outcome measures, our results are consistent with previous research which essentially reported an acceptable cut-off at +0.5 SEM (Colliver, Vu and Barrows, 1992). However, practices vary greatly across institutions and geographies: other authors have found that a thresholds of +2 SEM is more appropriate as it virtually eliminates false positive – a feature that, in licensing exams, reassures external stakeholders (Pell et al., 2013). This was also replicated in our experience: a harsh threshold (2 or 3 SEM corresponding to 95.4% and 99.7% confidence respectively) was able to capture the majority of the borderline students. However, in our opinion this result may reflect the very small percentage of borderline students as explained above.

Our findings uncovered and explored another problem neglected by previous research: the problem of overfitting. Overfitting is associated with poor predictive performance and indicates that models capture random noise as well as the underlying signal (Hastie, Tibshirani and Friedman 2001, Pitt and Myung 2002). To illustrate the unwanted consequences of overfitting consider the balanced accuracy obtained in the training set (0.91 (0.9-0.92)); if our institution decided that a minimum satisfactory standard was a classification accuracy of 0.9 then we would have been inclined to accept the set of 9 stations that generated that level of accuracy. However, when the same set of 9-stations was tested on new data (from the test set), the balanced accuracy decreased on average of 5 points dropping below the defined threshold (from 0.91 to 0.86). In this example, a balanced accuracy of 0.86 reflects a more parsimonious estimate of "real" capacity of our test to discern borderline from non-borderline students and should therefore be considered as more representative of future applications of the test. The extent of the predictive ability of our models (and the robustness of our conclusions) is further demonstrated by the results on a different set of data (validation set). Indeed, when our models were applied to a slightly different – but comparable – version of the screening OSCE, a balanced accuracy of 0.83 (0.79-0.91) was achieved. The further 3 point decrement is probably due to the smaller number of stations contained in the validation set (from the original 18 stations the exam moved to 16 stations, therefore the balanced accuracy was calculated on 7 or 8 stations instead of 9 as in the previous results).The problem of overfitting is not related to the choice of metrics employed in this study and similar reductions of the classification performance would have been observed if we had employed other metrics such as positive or negative predictive values or even composite scores. In fact any metric, if used for station selection, will adapt to unique features of the data at hand and will generate an inflated estimate. This argument can be easily seen when we considered the opposite scenario that is to say, when we consider the effects of validation on metrics that were not employed in the station selection process. This is demonstrated with the reliability which remained essentially the same in both train and test sets. Instead, the drop of reliability in the validation set is associated with the decrease of the number of stations (since the full screening was 16-stations long and not 18, 94% of the cross-validation samples contained less than 9 stations).

Taken together, our results (1) represent a realistic estimate of the predictive value of the theoretical 9-station screening and (2) suggest that, if not properly validated, OSCE models may "overstate" their predictive value and (3) therefore we may inadvertently compel education providers to take suboptimal actions when designing and implementing future assessments.

Once the optimal set of station has been established through sampling procedures, the question of content validity remains open. This is because the resampling methodology – as it was formulated here – is completely blind to the blueprint targets and iterates through all possible stations’ combinations. We, therefore, advocate the need to investigate content representation in the putative best short exam. This is a necessary step to guarantee the adherence to the policies that regulate exam content. The strategy pursued here – that can easily be applied to different testing circumstances – is to focus on the range and diversity of exam sub-scales represented in the short test. By looking at
exam sub-scales, test developers can diagnose a departure from the original scope of the exam. In OSCE examinations, different stations can be clustered into "domains" – i.e. fundamental elements of medical practice – such as history taking, communication and management. In the present study we asked if our approach could reveal important patterns at the domain level. In order to answer this question we looked at the best-fitting 9-station models identified with the iterative search and then replicated in the cross-validation phase. The proportion of domains contained in the optimal 9-stations sets was then compared to the proportion expected by a random sampling and the discrepancy between the two revealed where the exam could be improved.

In the present data, the frequency of communication (11.5% [10-14%]), history taking (17.5% [15-20%]) and skills (9.6% [8-12%]) domains don't deviate from what we would expect from random sampling (11.1%, 16.7% and 11.1% respectively). This means that the frequencies of communication, history taking and skills stations are already at an optimal level in the screening exam. In contrast, the estimated proportion of the examination domains is lower (20.6% [18.0-23.2%]) than the actual proportion (28%). Or the observed frequency of management is higher (40.7 [37.6-43.9%]) than the baseline rate (33.3%). These deviations indicate that the stations frequency can be adjusted accordingly: (1) examination stations could be reduced, (2) management stations can be increased (3) and these changes will help to replicate – up to the level of precision defined by the balanced accuracy – the outcomes of the 18-station screening.

Despite these differences, in a follow-up analysis, we provided three arguments to support this idea that the short test is indeed investigating the underlying skills and abilities tested by the 18-station exam. First, the best iterations of the cross validation are unequivocally (76%) associated with the full range of five domains and only in a small minority of instances (1%) are associated with 3 or less domains. Similarly, when we focused on the worst performing folds of the cross validation, we found that only 2% of them were associated with a fully diversified blueprint. Finally, in a regression analysis, we found that the classification accuracy performance correlates positively with the number of domains tested: as the number of domains increases, the ability of the test also improves in its capacity to accurately classify students. These results provide converging evidence in favour of the idea that the best short OSCE is not merely capitalizing on chance but parallels the long version in its testing characteristics.

The domain analysis provides a proof of concept and indeed it would be possible to repeat the investigation and look at other characteristics that could be part of a blueprint e.g. primary or secondary care setting, disease process type or body system. The main point of relating the outcomes of specific sub-scales with the outcomes of the exam as a whole is to provide a general strategy that can be employed to design a diverse and effective blueprint, in which the broad selection of elements is balanced against their effectiveness (predictive accuracy).

In spite of the focus on medical education, clinical exams and the need of exact classification mechanisms, the methods studied here can be applied to different settings. In the effort to illustrate this point we developed a simple simulation exercise using the IRT (See Supplementary File 1). The employment of IRT framework allowed us to strengthen and generalize the conclusions achieved in the OSCE results. First, we demonstrated that cross-validation can be successfully applied to different data sets. Second, we abandoned the dichotomy of binary decisions and directly monitored the latent abilities of the cohort. The results of this simulation reinforced the idea of the dangers of overfitting and the consequent needs for validation.

Conclusion

Healthcare systems are facing unprecedented challenges as they grapple with an epidemic of long-term conditions, inequality, co-morbidity and an aging population. In this scenario the field of medical education has become
increasingly valuable and clinical examinations have far-reaching implications for society. In the effort to develop cost-effective decision making for examination development a multitude of different methods has been proposed (Colliver et al., 1991, Colliver, Vu and Barrows, 1992, Cass et al., 1997, Currie, Selvaraj and Cleland, 2015, Smee et al., 2003, Hejri et al., 2016, Muijtjens et al., 2000, Regehr and Colliver, 2003, Muijtjens, Van Luijk and Van Der Vleuten, 2006). Computational methods entered this debate as they constitute powerful and flexible tools for those test developers interested in transitioning to sequential testing.

In spite of the large cohorts of students employed in this study, it has been demonstrated that cross-validation resampling schemes, work on small samples just as well (Hastie, Tibshirani and Friedman 2001). To illustrate this point, here we have replicated our main OSCE findings on a minimal simulated IRT data set (200 Subjects x 12 Items – See Supplementary File 1). These minimal requirements enable virtually any medical school to use resampling schemes without posing any constrain on the size of historical data bases.

Future research will establish how resampling schemes, such as the one proposed here, perform in comparison to more straightforward methods such as selecting stations based on their correlation with the adjusted total score, difficulty or discrimination indices (Smee et al., 2003, Hejri et al., 2016). Simpler methods may capitalize less on chance that results from trying every possible combination. However, considering the experience in other fields of statistics, it is possible to speculate that hybrid station-selection approaches, based both on targeted strategies and sampling methods, will constitute the best options for educational providers (eg. averaging model ensambles Domingos, 2012).

Despite that OSCE exams vary considerably between institutions and geographical regions our results are in agreement with previous reported findings suggesting the robustness of this methodology. Our results, together with the experience of other medical schools, indicate that fewer representative stations can provide good estimates of proficiency with the advantage of reducing the delivery costs. However, different standards of care operate across the world and interact with different economies to shape and define priorities in educational settings. It is therefore important that future research should explore the generalisability of these findings by utilising this approach with data from comparable assessments within other educational institutions. The definition of "optimal" or "best-fitting" may even depend on local priorities and fulfil different purposes (Jalili and Hejri, 2016). In this respect, our research provides a robust and flexible methodological approach that can be adapted to serve individual needs in designing and analysing resampling studies of sequential OSCE data.

**Take Home Messages**

- The availability of historical data together with the increase in the computational power have made resampling techniques appealing for the field of medical education.
- In the context of sequential OSCE, resampling techniques have been leveraged to inform blueprint construction.
- However overfitting undermines test construction by overemphasizing relationships and patterns in the data. Consequently, progression decisions become more error prone and therefore less defensible.
- Cross validation can be successfully employed to build trustworthy statistical models as well as valid OSCE blueprints.
- This methodology flexibly generalizes to different types of tests and context (eg. written examinations).
Notes On Contributors

**Giovanni Mancuso, PhD.** Previously psychometric analyst at King's College London and currently independent researcher.

**Stephanie Strachan, MD.** Critical Care Consultant, King's College Hospital NHS Trust and honorary Senior Lecturer, GKT School of Medical Education.

**Steven Capey, MD.** Assessment Director of MBBCh Programme, Swansea University Medical School.

Acknowledgements

None.

Bibliography/References

https://doi.org/10.1006/jmps.1999.1279

https://doi.org/10.1097%2F00001888-199710001-00009


https://doi.org/10.1097%2F00001888-199109000-00023

https://doi.org/10.1097%2F00001888-199209000-00010

https://doi.org/10.1111%2Fmjme.12923.2010.03926.x

https://doi.org/10.3109/0142159X.2015.1079309


Appendices

None.

Declarations

The author has declared that there are no conflicts of interest.

This has been published under Creative Commons "CC BY 4.0" (https://creativecommons.org/licenses/by-sa/4.0/)

Ethics Statement

Ethical approval has been granted by the Biomedical Sciences, Medicine, Dentistry and Natural & Mathematical Sciences (BDM) Research Ethics Panel for this retrospective observational study. Study number approval: KCL Ethics Ref: LRS-15/16-2378.

External Funding

This paper has not had any External Funding

AMEE MedEdPublish: rapid, post-publication, peer-reviewed papers on healthcare professions’ education. For more information please visit www.mededpublish.org or contact mededpublish@dundee.ac.uk.