International Journal of Learning, Teaching and Educational Research
Vol. 21, No. 11, pp. 188-207, November 2022
https://doi.org/10.26803/ijlter.21.11.11
Received Aug 17, 2022; Revised Nov 17, 2022; Accepted Nov 23, 2022

Alternative Digital Credentials: UAE's First Adopters' Assessment and Evaluation Part (2)

El-Farra Samar
Higher Colleges of Technology
Emirates Medical Association-Radiographers Society of Emirates,
United Arab Emirates (UAE)

Abstract. Despite the wealth of research on full credentials assessments, standardized approaches are still scarce. This is even more threatening to the acceptance of higher education alternative digital credentials. To address this threat, validated and transparent assessments and evaluation processes are of paramount importance. This study is a continuum to our previous review on the pedagogical program analysis, design, development, and implementation. This paper reviews, assesses and evaluates the alternative digital credential offering case study. We review the development and administration of seven requirements and assessment tools used to evaluate students' performance and use Kirkpatrick's model to evaluate the effectiveness of the alternative credential offered. The predominantly clinical-based assessment tools and assessment decision criteria are reviewed in detail in this paper, allowing educators to leverage the outcome of this work.

Results: The reviewed alternative digital credential case study in the human thorax and extremities from medical imaging has achieved Kirkpatrick level three, as evident in results, particularly from clinical assessments and clinical site viva-voce. When introducing a new competency-based assessment, professional standards can be used as a reference point to develop Behavioral Marker System rubrics. The Ebel method in calculating the cut score, which reflects expert judgment, should be considered when developing competency-based rubrics. Standardization of at least the top common technical and NTS is possible when researchers consider international collaboration by publishing comprehensive methodologies, frameworks, and results. This paper is unique as we are unaware of any publication on alternative digital credentials combining medical imaging and technical and non-technical skills within entrustable professional task assessment, verification, and program evaluation.

Keywords: alternative digital credentials; clinical-based assessment; non-technical skills; medical imaging; entrustable professional tasks

©Author
This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (CC BY-NC-ND 4.0).

1. Introduction

The contemporary healthcare Non-Technical Skills NTS training programs have mainly originated from aviation programs adopted since the 1970s after high-profile plane crashes. Thus, clinical aspects are not well addressed in the curricular components of those programs (Thomas, 2018). This paper is a continuum review of the Alternative Digital Credential (ADC), offering two distinguishable Entrusted Professional Tasks (EPTs) program analysis, design, development, and implementation (El-Farra, 2022). The findings and discussions in this paper have a heavy context dependency on part one reviewed by El-Farra (2022).

1.1. Importance of the Study

This paper reviews student and program evaluation methodology and framework to achieve two EPTs evaluation as the final step of the ADDIE instructional model by Gagne and Briggs (1974). The EPTs assume that a learner would be able to practice effective and safe communication and collaboration while demonstrating empathy in performing extremities and thorax Medical Imaging (MI) accurately and independently (El-Farra, 2022).

1.2. Study strategies and instruments

Seven assessment tools and requirements were developed and administered to evaluate the students' performance and the effectiveness of the five pedagogical collective interventions previously published. The overall objective of the evaluation is to verify the awarded ADC, which combines technical and NTS within the EPT scope to ensure that the ADC recipient can be trusted to perform clinically without direct supervision. Student assessments described herewith are grade/decision-bearing assessments. So, all formative non-graded assessments like discussions, feedback, and debriefing were part of the training phases before the clinical experience.

Competency/Clinical Based Assessment (CBA) addresses the witnessed paradigm shift in how educators and policymakers judge the effectiveness of educational programs in regard to how learners 'perform' and 'do' (Gruppen et al., 2012). Within a CBA context, summative assessments were carried out at the clinical site, including eight clinical assessments and a final viva-voce. We maintained the working hours with staggered shifts to prepare the students for their clinical assessments without impacting their contact hours load. Still, we doubled the patient training quota compared to the parent course requirements. However, summative judgments should not be made in isolation. Multiple observers of various patient encounters throughout training are needed to provide a collective judgment (Schuwirth & Vleuten, 2019). Also, evaluating NTS by only one faculty or clinical supervisor is not enough to correctly assess NTS. In fact, multi-sourcing information is essential for more accurate findings (Thomas, 2018). Further, assessment approaches, such as reflective practice, portfolios, and multisource feedback, help evaluate technical and NTS directly related to patient safety (Goldman & Wong, 2020). Consequently, we used the curricular course assessment results to triangulate our findings using the newly developed assessment tools and requirements. Seven different added and/or altered assessment tools or requirements were introduced to the ADC. The

added requirements were training patient quota, reflective journal requirements, and an aggregated e-portfolio. While the assessments introduced were eight additional clinical assessments on actual patients, Jefferson's Score of Empathy (JSE) (Hojat, 2016), viva-voce, and patient satisfaction and critical incident reports reviews.

In a systematic review, tools used to assess NTS were reported to be often locally developed to specific educational interventions, without reference to conceptual frameworks. Consequently, those tools are rarely validated, limiting their dissemination and replication (Gordon et al., 2019). Moreover, literature reports a gap in educators' CBA literacy and preparedness (Gallardo, 2020) in line with our qualitative, exploratory, descriptive, and contextual research design. We applied Guba's model of trustworthiness, which suggests dependability, credibility, transferability, and confirmability as criteria suitable for qualitative studies (Shenton, 2004). As such, we utilized triangulation, peer and industrial review, and a thorough description of the setting and the research methods in developing the assessment tools, calculating the cut-scores, and rolling out the assessments. Finally, we maintained a trail detailing the process of creating the assessment tools, cut scores, and interpretations. Some of that evidence will be in the form of the lifelong e-portfolio, and some are published in this paper and other interlinked papers.

Assessing NTS within professions is a topic in urgent need of research (Gordon et al., 2019). Reportedly, there is no gold standard for NTS setting standards to be assessed (Yune et al., 2018). Yet, predetermined technical and NTS criteria are helpful to indicate whether a student is competent (Thomas, 2018). Three assessment rubrics were developed for the student assessments. A Dedicated Behavioral Marker System (DBMS) rubric was designed to make EPT reliable judgments. The DBMS taxonomy was equivalent to the Professional Practice Standards modified list (PPS-modified) taxonomy used for training students in the lab and class (El-Farra, 2022). A Structured Viva-voce (S.Viva) conducted by the head of the clinical department as a lead panelist was performed using a rubric co-designed with the clinical experts. For both tools, calculating the cut-score was achieved by applying the Ebel method (De Champlain, 2019). During cut score identification for the clinical and S.Viva assessments, a panel of clinical experts provided difficulty estimates against content relevance per assessment item. This is essential because cut-scores should reflect expert judgment as to what constitutes competence, supported by several sources of evidence (De Champlain, 2019). The cut-score was then calculated by adding the cross-products of the difficulty and relevance decisions. The cut-score for the clinical was 80% and 84% for the S.Viva. Finally, the reflective journal entries rubric was developed based on Gibbs' reflective cycle of reflection (Gibbs, 1988). Suitable for formative requirements, an estimated 60% cut-score was used for the rubric to assess the reflective journal accounts.

To measure the response of patients, we deployed a test survey. We found that the patients were giving highly subjective encouraging surveys all of which reached 100% satisfaction. We retrospectively reviewed all patient complaints,

clinical peer dissatisfaction, and safety incident reports for any possible involvement of the ADC students as a reverse confirmation of satisfaction.

To document all relevant results supporting the value of the awarded ADC, a structured e-portfolio was used as an electronic space requirement to aggregate the evidence of achievement stipulated in the ADC standards for job applications. The digital artifacts linked with the students' digital record e-portfolio are the ADC standards document, clinical assessments results, JSE, and S.Viva results. The technology infrastructure is still in the sandbox phase, and the technology journey's subsequent publications are en route.

Finally, we conducted an ADC effectiveness evaluation using the Kirkpatrick method: reaction, learning, behavior, and results (Kirkpatrick & Kirkpatrick, 2016). The ADC program described as such has reached the third level of effectiveness. Further longitudinal research is required to provide evidence of achieving all four levels of ADC effectiveness.

2. ADC Student Completion Requirements

The ADC requirements were divided into three main categories based on the deviations from the parent curriculum, as follows:

- Unchanged course requirements such as the 60% pass score of the didactic and clinical courses and the clinical training hours. Student results were used for triangulation.
- Replaced or waived as an ADC requirement due to lack of relevancy or compatibility, and those Objective Structured Clinical Examination (OSCE), written image critiques, reflection journals, and clinical site supervisor assessments.
- Added seven and/or altered assessment tools or requirements:

1. Patient training quota doubled from 34 to 70 per EPT.
2. Eight clinical assessments, four per EPT, using a dedicated rubric in addition to the existing technical rubric.
3. Empathy score using Jefferson's Scale of Empathy (JSE) test (Hojat, 2016)
4. A final S.Viva conducted by a panel led by clinical experts using a dedicated rubric.
5. Two reflection journals using a dedicated rubric.
6. Patient satisfaction and critical incident report reviews for possible students' direct responsibility or significant involvement.
7. E-portfolio.

Table 1 represents a summary of the ADC completion requirements and the types of adjustment as compared to the parent curriculum.

Table 1: ADC Completion Requirements and the Types of Adjustments Compared to the Parent Curriculum

Category	Component	Parent curriculum components	ADC curricular adjustments
Unchanged requirement	Clinical hours	160 hours	160 hours.
Unchanged requirement used for triangulation	Passing didactic and clinical courses	As part of the full credential graduation requirements	As part of the ADC completion requirements
Waived but used for triangulation	Clinical supervisor assessment	Four assessments with a rubric	Student performance results were used for triangulation.
Waived but used for triangulation	Image critique	Two image critiques with grading criteria	Waived but used for triangulation
Waived but used for triangulation	OSCE	Two assessments with a heavily technical rubric	Waived but used for triangulation
Waived and replaced	Reflection journal	Four entries with no rubric	Replaced
Added as a replacement	Reflection journal with dedicated rubric	None	Two different Non-Technical Skills NTS- focused entries
Added	Patient quota	34 per EPA	70 per EPA
Added and altered	Clinical assessment	One assessment with a heavily technical rubric	Eight assessments using the DBMS and technical rubric
Added	S. Viva	None	By industry panel using a rubric
Added	Empathy score	None	JSE
Added	Patient satisfaction and critical incident reports reviews	None	Incident reports and patient-compliant reports, reviews of exclusion

3. Student Evaluation Framework

Methods for NTS assessment can be categorized into holistic judgment, standardized assessments, performance/competency assessments, and portfolio assessments. Each demonstrates strengths and weaknesses, so using a variety of those assessments is more valuable (Curtis, 2004). Furthermore, CBA is the most appropriate NTS assessment approach (Thomas, 2018). The proposed framework has seven different added and/or altered assessment tools or requirements that form a mix of all four categories described by Curtis (2004) with a comprehensive holistic judgment assumption. Between academics and clinical supervisors, a variety of assessors facilitated a holistic judgment. A faculty and a clinical instructor from our institution were directly involved in the assessment, in addition to a pool of clinical supervisors at the clinical site. The ADC students were all assigned to one hospital on a non-rotational schedule, so the same pool of clinical supervisors and faculty would get to know students'

attributes through frequent clinical or lab observations. At the same time, consistency of judgments within panels of assessors has been demonstrated through the collective holistic judgment via utilizing rubric or checklists to formulate the bases of EPT performance judgment. Additionally, two NTS-focused journals were also used to add a self-reflecting layer. Reflective practices are essential for patient safety and quality improvement within the healthcare context (Goldman & Wong ,2020). Multisource judgments are valuable when examining patient safety and quality improvement skills (Goldman & Wong, 2020). Therefore, we introduced an approach we debated to complete a 360-evaluation, suitable for small cohorts with challenged statistical inferences. In this assessment, we propose that no direct or indirect student involvement in patient and staff complaints and/or incident reports provides indirect proof of satisfaction and safe conduct. We argue that this evaluation tool is relevant to high-risk industries and fits under the CBA methodology.

Because the ADC was within an undergraduate parent curriculum, we did not introduce any additional standardized assessments. Although the parent's curriculum standardized assessments do not have segregated NTS items per se, we opted to use standardized assessment results for triangulation to correlate with the EPT individual performance. Because of the lack of comparability in the NTS assessment, a discrepancy trigger was considered relevant if a student's score was below a C grade in any of the curricular ADC courses. Although none of the ADC students scored less than a C grade, further research is required to establish better correlation triggering parameters.

To address the inherent cynicism in ADC offerings, three heavily clinical-based-oriented requirements and assessments were introduced. First, the patient training quota requirement was doubled to ensure student readiness for the subsequent eight clinical assessments on actual patients. For those assessments, we used the existing technical rubric and designed a DBMS rubric. Both tools facilitated standardized technical and NTS EPT-specific judgments. Because the tool was used for the first time, the outcomes correlated with the well-established JSE. The students had pre- and post-intervention JSE tests and the results were correlated with the empathy sections of the DBMS for major discrepancies. The S.Viva was the third CBA conducted by the head of the clinical department as the lead panelist. The structured setting and a special rubric were co-designed with the clinical experts.

Finally, an e-portfolio an electronic space requirement to aggregate the evidence of achievement stipulated in the ADC standards. The e-portfolio was structured to have all supporting evidence that may be used for job applications and the technology infrastructure was still in the sandbox phase. The e-portfolio has four sections: the ADC standards document, the clinical assessment, S.Viva results, and the JSE scores. The technological journey shall be published in the subsequent parts of this paper.

4. Clinical Requirements and Assessments

Authentic CBA through direct clinical observations is gaining popularity (Schuwirth & Vleuten, 2019). Consistent with clinical expectations, CBA reflects the complexity of the clinical environments and situations (Gallardo, 2020). This is relevant to the "entrust" component of the EPT concept in four ways. First, CBA focuses on performance and results rather than educational processes. Second, CBA criteria or performance standards are not determined by the performance of other learners but by the expert judgment of practitioners and educators in the field (Gruppen et al., 2012). Third, CBA is criterion-based performance judgment (Thomas, 2018; Yune et al., 2018). Predetermined technical and NTS criteria are utilized to identify whether a trainee is competent or not yet competent (Thomas, 2018). Lastly, the authenticity and interdisciplinarity features of CBA imply the involvement of work field experts in determining the essential evaluation indicators (Gallardo, 2020).

4.1. Supervised Patient Training Quota

Because the curriculum is accredited by regulatory bodies, we benchmarked the number of patients against the accreditation requirements to determine the suitable quota. Since we needed to distinguish the ADC from the associated clinical course, we estimated that doubling the training quota should provide students and future employers with enough confidence in the quality of training received under clinical site supervision. After discussions and reviews on the clinical site patient intake number, it was evident that students needed to work staggered shifts to meet the quota. Upon completing almost 60% of the quota by week nine of the semester, all seven students shared that they were ready for the first clinical assessment. This was considered an indication of a possibly lower required quota for future ADC offerings.

4.2 Clinical-Based Assessment

To determine the number of assessment encounters required, traditional estimates of reliability against feasibility should be observed. The purpose of the assessment has a direct impact on the encounters required; the more encounters, the more the width of the confidence interval declines, and the more the number of valid decisions increases (Norcini & Zaidi, 2019). Since the assessment aims to identify which students are not "entrusted" to perform the EPA, we did not use the standard error of measurement to refine making this decision. Instead, four clinical assessment encounters per EPA were estimated to be sufficient. The DBMS was designed to be used to rate direct observations of interactions with actual patients, while more lab training requirements were planned for students who had received scores below the cut-off.

4.2.1. Assessment tool

Assessing competencies in combination with NTS is a topic in urgent need of research (Gordon et al., 2019). In a literature review, 76 healthcare NTS measurement tools were identified with widely various methods of scoring. However, there is no golden standard tool for NTS measurement (Higham et al., 2019), with no available perfect way of setting standards (Yune et al., 2018). In a systematic literature review, the challenges of CBA rubrics were identified in integrating skills beyond technical and cognitive ones which account for the

interdisciplinarity of tasks and educators' CBA literacy (Gallardo, 2020). While CBA is based on observational judgments, the lack of homogenizing assessment exacerbates assessors' differences (Norcini & Zaidi, 2019). Further, expertise, experience, and firmness are substantial factors that affect CBA reliability (Yune et al., 2018).

Rubrics have evolved in the last three decades as a powerful tool to make judgments about students' learning. The importance of rubrics design for technical and NTS rubrics was reported in a recent systematic literature review (Gallardo, 2020). Despite the wealth of research, rubric design varies according to pedagogical and assessment needs, which continues to be challenging. Furthermore, few existing rubrics go beyond the cognitive and technical domains despite the urgent demand to capture NTS within complex situations (Gallardo, 2020, Velasco-Martínez & Hurtado, 2018). Furthermore, current CBA rubrics are relatively weak due to rubric design assessment literacy issues (Velasco-Martínez & Hurtado, 2018).

During the pilot analysis and design phases, the competencies were first identified and translated into two distinct EPT incorporating technical skills and NTS (El-Farra, 2022). Ideally, the same set of competencies used for training should be used to develop assessment methods (Gruppen et al., 2012; Thomas, 2018). Consequently, there was a need for a methodology to translate the PPS-modified checklist used for training (El-Farra, 2022) into an assessment rubric.

In thirty-three publications, checklists of observed interactions are the most frequently used assessment method (Cimatti, 2016). In addition to task-specific checklists, holistic and analytic rubrics are often used for CBA (Yune et al., 2018). Task-specific checklists are relatively objective; they capture the occurrence of behaviors that non-experts like simulated patients or peers can observe and evaluate. Also, scores based on checklists are strongly correlated with scores based on holistic rubrics and global rating scales. Further, task-specific checklist judgments focus on feedback to help students learn (Norcini & Zaidi, 2019). Finally, a unified checklist for training and assessment enhances faculty confidence in teaching and evaluating NTS (Duffy et al., 2004). Adversely, task-specific checklists limit the effects of the evaluator's expertise in evaluation. Also, there is scarce research with which to determine analytic rubrics holistically versus the efficacy (Yune et al., 2018).

While holistic rubrics, including global rating scales, underline an overall expert's judgment of a comprehensive complex tasks assessment, the analytic scoring process involves assigning points to individual performance. It adds the points to derive one or more dimension scores (Yune et al., 2018). For example, students with low scores in the cue "Modifies communication methods to account for patient diversity" from the PPS-modified checklist (El-Farra, 2022) can be trained separately on different MI-specific scenarios. Consequently, analytic rubrics are more reliable in checking the key content providing precise feedback per dimension. In congruence with Yune et al.'s (2018) findings that holistic and analytic rubrics are efficient tools for explaining task-specific checklist scores, we used the parent curriculum holistic rubrics with

behaviorally anchored scales to indirectly assess students by triangulating the individual performance outcomes. This way we would mitigate the limitations of task-specific checklist evaluation. Additionally, we debate that such triangulation is especially valuable when introducing a new assessment tool.

Also relevant are the Behavioral Marker Systems (BMS) being at the forefront of the NTS assessment in high-risk work environments. Relevant to the cues described in the PPS-modified checklist are the standards of performance described as "good" versus "bad" behaviors stated in BMS (Thomas, 2018).
Based on the key characteristics of BMS, we transformed the PPS-modified checklist used for training with the below alignment constituents.

The critical characteristics of BMS and PPS-modified constituents align, as both are relevant for NTS related to safe and efficient operations. Also, both tools state observable behaviors as per a taxonomy. Furthermore, both tools have comparable taxonomies as follows:

1. The BMS term *category* is used to describe a domain of NTS, such as communication, which is the equivalent of the term *standard* used in the PPS-modified checklist.
2. The terminology *element* is equally used across both tools to describe a specific skill, such as assertiveness.
3. The BMS lexicon *indicators'* equivalent is *performance criteria* defined as evident actions.
4. *Cues* in the PPS-modified checklist that aid with clarification of the indicators of performance is the equivalent of the good/bad *behavioral marker*. As such, performing a cue when needed is "good" and not performing it is "bad". Table 2 outlines the equivalency in the taxonomic structure between the BMS and the PPS-modified checklist (Thomas, 2018).

Table 2: BMS and the PPS-modified Taxonomic Structure Equivalency

PPS-modified taxonomy (El-Farra, 2022)	BMS taxonomy (Thomas, 2018)
Standard The explicit professional activity requirements to be demonstrated in the clinical setting, such as behavior elements.	*Category* Used to describe a domain of NTS, such as communication.
Key element components/responsibilities within the standard, such as "Sound communication methods"	*Elements* Are used to describe a specific skill, such as "Assertiveness"
Indicators Performance criteria and evident actions to ensure the standards are being met, such as "Recognize and overcome communication barriers"	*Performance criteria* Evident actions, such as "Identify and communicate any alternative Diagnoses"
Cues Aid with clarification of the indicators of performance, such as "Provides aftercare instructions"	*Behavioral markers* Performance indicators, such as "Read-back is used to confirm information is received correctly"

4.2.2. Assessment Decision Criteria

Although grading with sufficient levels of granularity is desired to differentiate performance, the simplest NTS rating scales use a binary 'pass/fail' to produce greater levels of consistency between assessments (Thomas, 2018). Using the DBMS, each constellation of indicators was categorized per at least one of the targeted NTS and allocated a binary grading per indicator. Each indicator has one or more possible cues and is assigned to a value of one if achieved and zero if not. Finally, when an underlying assumption that a set of NTS are safety-related, a jeopardy 'pass/fail' NTS assessment can be justified (Thomas, 2018). Therefore, based on the foreseeable or actual safety implications, certain cues were tagged as a jeopardy 'Zero-tolerance' (e.g., subjecting a person to unnecessary radiation or failing to call for help if a patient deteriorates/collapses).

4.2.3. Cut-score calculations

The underlying assumption in CBE is that the translation of qualitative standards to a *cut-score* number would reflect competencies. Also, in competency-based methods, it is desirable that all learners achieve 'competence' after training. Consequently, criterion-referenced cut-scores are more relevant in healthcare education as they indicate that a candidate has mastered the EPT's underlying components during an assessment (Gruppen et al., 2012). Finally, the involvement of the clinical experts in determining essential evaluation indicators and scores is important (Gallardo, 2020).

We applied the Ebel method to determine whether the scores at or above a certain cut-score would indicate that the performance standard has been met. The Ebel method involves asking a panel of clinical experts to provide difficulty estimates per assessment item/cue categorized as easy, average, or difficult along with content relevance classified as essential, important, acceptable, and questionable. The cut-score is then calculated by adding the cross-products of the difficulty and relevance decisions. This approach correlates item relevance and difficulty, and the total score can be interpreted as an overall reflection of candidates' competencies in interrelated domains (De Champlain, 2019).

The two experts with more than 20 years of experience (one is with a mixed extensive academic/clinical and one is the head of the clinical department) were required to independently consider the relevance and degree of difficulty of the DBMS 58 cues and then estimate the proportion of questions that the minimally proficient student would correctly demonstrate in each cell. Disagreements between the two panelists were reconciled through inter-discussions. Finally, the panelists reported that nine of 58 cues were essential with 4,2,3 care categorized as easy, average, and difficult, respectively. Panelists agreed that 92% of the students should be able to achieve the easy essential cues, 85% should achieve the average essential cues and 71% should be able to achieve the difficult essential cues. Further, 49 of 58 items were classified as important items with 5,29,15 cues categorized as easy, average, and difficult respectively. Panelists agreed that 88 % of the students should be able to achieve the easy essential cues, 83% should achieve the average essential cues and 70% should be able to demonstrate competency in achieving the difficult essential cues. The resulting

cut-score was calculated as the sum of the relevance/ difficulty of cell cross-products. None of the cues were classified as acceptable or questionable which supports the expectation of the EPT "entrustable" component. Also, this supported the ASMIRT (2018) extraction processes described by El-Farra (2022). Table 3 depicts the two-dimensional Ebel grid relevance/ difficulty cells and the cut-score calculation sum of the relevance/ difficulty cell cross-products with examples.

Table 3: The DBMS Two-Dimensional Ebel Grid

Relevance	Easy	Average	Difficult
Essential	0.92 (4 cues)	0.85 (2 cues)	0.71 (3 cues)
Essential cue example	Exchanges and shares information with members of the interprofessional team	Involve other professionals as needed	Watches for non-verbal cues
Important	0.88 (5 cues)	0.83 (29 cues)	0.7 (15 cues)
Important cue example	Ensures patient identification policy has been adhered to	Ensure language style is situation appropriate	Adopts and adjusts communication style appropriately
Cut-score = 0.92(4)+ 0.85(2)+0.71(3)+0.88(5)+0.83(29)+0.7(15)= 46.48/58 (80%)			

4.3. Empathy Assessment

Cognizant of the shortfalls of our proposed DBMS in terms of maturity, we also used the JSE to establish a correlation that might support our work. Since Hojat and colleagues developed the JSE in 2001, it has been extensively employed as one of the most common psychometrically sound tools used to measure empathy in the context of health professions education and patient care (Hojat & Gonnella, 2017; Hojat et al., 2018). The associations between scores of the JSE and pertinent variables have been extensively reported in empirical research (Hojat et al., 2018). We deployed the (HPS-Version) for administration to all health professions students other than medical students. We ran the baseline and a post-intervention of the JSE assessment at the beginning of the first semester and towards the end of the second. The initial compressions between the post-intervention did not reveal significant discrepancies between the empathy constituents of the devised DBMS. However, to improve our statistical inferences and to establish retention of the attained empath, the after-intervention score will be analyzed as compared to all health science students (*n)*=1000 students. Future work will be published on a none randomized parallel arm-controlled trial.

4.4. Structured Viva-Voce

Traditional viva-voce assessments have a long history in medical assessment, yet students feel that they lack standardization and objectivity (Jefferies et al., 2011; Shenwai & Patil, 2013). These shortcomings are better addressed by structuring the event. Faculty members and students favor S.Viva as its structure reduces the bias and makes viva-voce a fair assessment tool (Shenwai & Patil, 2013). Furthermore, S.Viva assessments entail many advantages such as assessing NTS like problem-solving and recognition of safe practice. They also provide an in-

depth assessment of knowledge and flexibility to tailor the questions asked to the needs of each individual candidate (Davis & Karunathilake, 2005). Moreover, the overall and inter-rater reliabilities achieved in S.Viva exceeds those of traditional viva-voce (Ganji, 2017).

In the S.Viva clinical cases, questions, criteria, possible answers, and marking schemes are all predetermined (Jefferies et al., 2011). Before the S.Viva session, each student was given a fixed, proctored 30 minutes to review two pre-selected cases with known diagnostic outcomes (a case per EPT). To improve the reliability, each student was provided with the checklist used for grading and was allowed to write comments that could be referred to during the S.Viva sessions by two examiners of different backgrounds, as described by Schuwirth and Vleuten (2019).

4.4.1. Assessment Tool and Decision Criteria

The objective of the S.Viva was to test the knowledge of students to determine the diagnostic quality of an image produced as the product of any of the EPTs. This decision is critical to provide the "entrust" component of concluding a patient's encounter safely and successfully. Image critique standards used were those which affect the diagnostic value of the produced medical image (Lampignano & Kendrick, 2017).

The criteria were tabulated in a checklist that included 15 items that describe medicolegal information, positioning, collimation, centering, technical factors, image artifacts and anatomy identification. Seventy percent of the cases were selected by experts at the clinical site with questionable diagnostic value, fifty percent of which had to be repeated due to suboptimal positioning and/or procedure. The predetermined answers were based on the actual outcome documented in the radiologist report.

We applied a binary 'met/unmet' rating scale to improve the consistency. We also pre-identified jeopardy assessment items as those which would result in the need to unnecessarily repeat the examination and subject a patient to unjustifiable hazards of radiation. Only one student responded that an ankle x-ray should be repeated, but that image was reported as acceptable, and repetition was not warranted. A single make-up attempt was allowed, but the score of the second attempt was capped by the cut-score for fairness to other students.

4.4.2. Cut-Score Calculations

Applying the same Ebel methodology, the panelists who used the rubric for the S.Viva were involved in determining the cut-score for the tool. They agreed that all items in the checklist are essential to the EPT scopes, with five items categorized as easy, six as average and four as difficult. The calculated cut-score was 84%. Table 4 depicts the two-dimensional Ebel grid relevance/ difficulty and the cut-score calculation sum of the relevance/ difficulty cross-products with examples.

Table 4: The S.Viva Two-Dimensional Ebel Grid

Relevance	Easy	Average	Difficult
Essential	0.95 (5 items)	0.85 (6 items)	0.70 (4 items)
Cue example	Student accurately critiques the image in terms of medico-legal requirements.	Student states if positioning is accurate using a minimum of two evaluation criteria.	Student states if the exam must be repeated due to positioning error. Minimum of one criterion.
Cut-score = 0.95(5) + 0.85(6) +0.70(4) = 12.65/15 (84%)			

4.5. Patient Satisfaction and Critical Safety Incident Report Reviews

Reviewing patient complaints, reports, and malpractice and patient surveys following an encounter effectively assesses interpersonal and communication skills (Duffy et al., 2004). Due to the limitations imposed by the small cohort to complete a 360 evaluation, we reviewed all patient and clinical peer dissatisfaction reports in addition to safety incident reports during the fifteen and subsequent five weeks past the semester for any possible involvement of the ADC students as a reverse confirmation of satisfaction. None of the reports showed any direct or indirect involvement of the ADC students in any safety compromises or dissatisfaction encounters during their training and assessment periods.

4.6. Reflection Journal

Reflective activities align well with NTS development as they show that achieving safe and efficient performance requires both technical and NTS (Thomas, 2018). The parent curriculum requires four entries that prompt a student to submit an account of encounters to state the challenges and success, with no rubric for marking. Those were excluded from the assessment and replaced by two structured reflection journals. The rubric was designed with a focus on NTS following the Gibbs' reflective cycle, one of the most famous cyclical models of reflection. The cycle is a six-stage step by step journey that guides the student to explore an experience through context description, feelings, evaluation, analysis, conclusion, and action plan (Gibbs, 1988). Due to the foreseen language barriers and to foster the reflective capacity in the students, journals were used as a formative assessment. The rubric used for grading has five criteria and ten guiding questions, each of which was assigned a mark. The maximum possible grade was 15 and the cut-off score was 60% (9/15). The first journal draft was due mid-semester and the second was due by the end. One-on-one discussions were offered to students who scored zero in any of the six stages, to ensure their comprehension of the concept of reflection.

5. E-Portfolio Structure and Content

Collecting data about NTS is more practical and accessible using e-portfolios (Cimatti, 2016). E-portfolios are useful for supporting and assessing clinical practice learning. Depending on their purpose, portfolios differ in scope, structure, and content (Driessen & Tartwijk, 2019). The primary purpose of the ADC e-portfolio is to aggregate show-case evidence of attainment to address

uncertainties around the ADC concept as a higher education paradigm shift. Upon discussions with the clinical sites, we balanced student privacy, credential authenticity, and relevancy of the evidence. The e-portfolio we used was systematically structured as a transparent and practical evidence reservoir suitable for potential employees to have confidence in the issued ADC. The sandbox environment used for issuing the ADC and the e-portfolio components was our institutional property. Similar to the full credentials, the ADC and the students' e-portfolios will be digitally available through Blockchain. Full autonomy is granted for a student to share or not share the ADC e-portfolio component(s) with prospective employers. Each student e-portfolio has four components: the ADC standards document, the clinical assessments results, the second Jefferson's scale of empathy score and the S.Viva results.

6. Program Evaluation

We refer to the Kirkpatrick method to evaluate the ADC training program's effectiveness. The Kirkpatrick model is a widely recognized tool for evaluating and analyzing the results of educational and training programs. It consists of four levels: level 1 *Reaction*; level 2 *Learning;* level 3 *Behavior*; and level 4 *Results* (Kirkpatrick & Kirkpatrick, 2016).

In the first week of the semester, nine eligible students were offered an awareness session with the faculty and employer. Seven of them have voluntarily enrolled based on individually expressed written interest (El-Farra, 2022). The reaction level of the program evaluation model refers to the degree to which those seven students find the training favorable and relevant to their future jobs (Kirkpatrick & Kirkpatrick, 2016). All students had the option to withdraw from the ADC without penalty at any stage. Yet, none of the students have expressed any intention to drop out at any stage and all seven students have successfully completed the ADC. Therefore, we assume that the offering has fully achieved the first level of the Kirkpatrick model. *Learning* is the second level of the model, and it is about describing the degree to which participants acquire the intended outcomes including knowledge, skills, attitude, confidence, and commitment. At the same time, the *behavior* level is about the effects beyond what learners attain in terms of course marks and measuring the observable behavior at work (Kirkpatrick & Kirkpatrick, 2016). The progression offering of the ADC was designed to optimize and assess learning transfer and subsequent employee performance in two ways. First, the deployment of CBA as an integral assessment directly involves clinical demonstration of technical and NTS as well as the direct industrial involvement in developing the assessment tools and assessing the students. Both approaches provide evidence to reflect the individual competency in performing the EPT stated. Therefore, achieving the Kirkpatrick levels 2 and 3, learning, and behavior, respectively, was evident in the assessments results, particularly in the clinical assessments and the S.Viva.

However, we cannot hold the debate that the ADC program has achieved the *results* level. The fourth and final level encompasses the broader *results* and consequences of the learning and the gains for the organizations and stakeholders. We argue that achieving tangible results to align with the fourth

Kirkpatrick level could be established through three possible avenues. First, a follow-up survey on the acceptance of the ADC amongst the employers, a periodical assessment if those students had a better job or volunteering offers, and a full-fledged rollout of the ADC framework proposed and /or national alignment with the same.

7. Trustworthiness Analysis

This work is a qualitative, exploratory, descriptive, and contextual research design. To enhance the value of this research we aligned with Guba's model of trustworthiness. This model lists dependability, credibility, transferability, and confirmability as criteria that enhance qualitative research trustworthiness (Shenton, 2004), with overarching transparency in publishing an elaborate description of our work to enhance all four trustworthiness criteria. Triangulating data collection and inferences was used across the process as a valuable tool for formative guidance and to extend credibility and dependability to our work. We used formative and summative triangulation approaches. The formative approach was to triangulate student performance variation across the different assessment tools. Those checking points were mainly utilized to inform the individualized training and mentoring plans. Table 5 plots the triangulation points used to establish individual performance correlations across the different requirements and assessments. At the same time, the summative triangulation was intended to provide an overall program triangulation. Both ADC and the clinical course have a 'pass/fail' derived from percentage grading schema. So initially, deviation in performance was considered if a student passed the ADC but did not pass the parent courses or vice-versa. This is because the cut-score of all course assessments is 60%, compared with variable cut-scores described in this paper. We also compared the aggregated percentage of letter grades. Table 6 depicts individual ADC and parent curriculum clinical course performance differences in letter grade. In addition to transparency and triangulation, the ADC curriculum and assessment had substantial input from peers and clinical experts. The extensive involvement of various experts improves the credibility and confirmability of our work. External involvement, contribution, and validation roles were stated in the above sections and are reviewed in Table 7.

Table 5: The Formative Triangulation Points Matrix Across the Different Requirements and Assessments

	Passing courses	Clinical supervisor assessment	Image critique	OSCE	Clinical assessment technical skills rubric
Reflection journal with dedicated rubric	Yes				Yes
Clinical assessment using the DBMS	Yes	Yes		Yes	Yes
S.Viva	Yes		Yes		
Patient satisfaction and critical incident report reviews	Yes	Yes			Yes

Table 6: Individual ADC and Parent Curriculum Clinical Course Student Performance Differences

Student ID/ Final Grade	1MI	2MI	3MI	4MI	5MI	6MI	7MI
ADC	A	A	A	A	A-	A	A
Course	A	A	A	A	B+	A-	B+

Table 7: External Peer and Clinical Experts' Involvement, Contribution, and Validation Roles

Assessment	Contributor(s)	Role
Clinical hours	Accreditation requirement	-
Passing didactic and clinical courses	Two different faculty and clinical supervisors	Deliver the three courses and assess and grade students per parent curriculum requirements.
Clinical supervisor assessment	Clinical supervisors	Assess and grade students' clinical performance as per parent curriculum requirements
Image critique	Faculty member	Assess and grade students' submissions as per parent curriculum requirements
OSCE	Faculty member and lab technician	Assess and grade students' clinical performance within the lab setting as per parent curriculum requirements
Reflection journal	Faculty member	Assess and grade students' submissions as per parent curriculum requirements
Reflection journal with rubric	ADC Faculty	Rubric design Train the students on the requirements Grade entries for formative feedback
Patient quota	ADC Faculty and clinical site head of department	Estimate the number of supervised training patient quotas required for the EPT to be attainable based on the patient intake numbers and competency expectations.
Clinical assessment ADC faculty	ADC Faculty	Rubric design Train the students on the requirements Orient evaluators to use the rubric Student result interpretation and performance triangulation
	Clinical site expert and ADC Faculty	Cut score determination
	Clinical instructor	Conduct clinical assessment
S. Viva	ADC Faculty	Rubric design Panelists in the S.Viva sessions Result interpretation and student performance triangulation
	Clinical site expert and head of department	Cut-score determination Case image selection Lead the S.Viva panel to make the final

		student performance decision
Empathy score	Thomas Jefferson University	Online test
Patient satisfaction and critical incident report reviews	ADC Faculty and clinical site head of department	Review and report

8. Ethical Considerations

In addition to the ethical considerations followed during the pre-assessment phases falling under the institutional Research Ethics and Integrity Committee (REIC2-113) permission (El-Farra, 2022), additional assessment-specific ethical considerations were adhered to. To ensure non-maleficence, all ADC-related assessments and requirements did not have any input or influence on the grade-bearing assessment that contributed to the Grade Point Average (GPA) calculations. Further, participants' rights to privacy, confidentiality, and anonymity were protected by not sharing the assessment results with the rest of the MI faculty. Additionally, the sandbox e-portfolio permissions are designed with a security layer that allows the earner of the ADC to share with employers of their own choice.

9. Limitations and Future Work

The originally intended scope of the ADC is not limited to undergraduate students with relatively high GPA scores. De facto, targeting upskilling and reskilling for the existing workforce is a significant drive. However, ethical and feasibility considerations have restrained this study to be only conducted for a relatively small number of undergraduate students with high GPA scores. Further study is required to determine if the results are generalizable to other cohorts like adult learners and those with non-GPA-dependency eligibility. Also, certain limitations arise due to the dependency on CBA to assess the students. CBA implementation requires careful feasibility and validity tradeoffs. Factors like suboptimal evaluator training could compromise the validity of the assessment (Gallardo, 2020). Our findings are based on optimized feasibility and validity tradeoffs because of the relatively small number of evaluators trained on using the assessment tools. Further, clinical training and assessment opportunities were feasibly optimized due to the small numbers. However, generalizability and variations in workplace contexts will contribute to differences in skills development (Gruppen et al., 2012).

Future research should be focused on the digital journey, the stakeholder's acceptance studies, and longitudinal studies to examine NTS changes in ADC graduates. Finally, there is a need to set broad yet distinguishable areas of competence that would collectively constitute a general descriptive framework for a profession (Englander et al., 2013). At the same time, standards of MI practice might differ from country to country, common technical and NTS form most of the MI professional expectations. One of the objectives of the International Academic Network (IAN) of the International Society of Radiographers and Radiological Technologists (ISRRT) is to contribute to the technical and NTS patient-centered care and patient safety educational materials

and research (ISRRT, 2022). Future work is required in collaboration with IAN to establish a generalized MI competency framework that incorporates interrelated and purposeful competencies like radiation and contrast injection safety communication. Such a framework would constitute the backbone of a more homogenous pedogeological and assessment method.

10. Conclusion

Our initial assessment and evaluation findings support the assumption of the need for various approaches, primarily CBA tools. Leveraging the local or international professional standards as a reference point to develop rubrics within a BMS framework is a practical and reliable starting point in any newly introduced CBA tool. The difficulty of a question item or cue of performance is not the only factor that should be considered while designing an assessment. Addressing the dependency between assessment items' relevancy and domains is equally important. Thus, the Ebel method in calculating the cut-score should be considered when developing CBA tools and BMSs, particularly when clinical expertise is required for the tool development and grading to enhance the ADC prospects and trustworthiness. Also, cut-scores can be viewed as a translation of the standards which reflect expert judgment as to what constitutes competence and should be supported by evidence. However, careful planning and local considerations to balance privacy and transparency should be followed to decide which digital footprint artifacts should be available to the stakeholders. Finally, there is a need for transparency in assessment research. Standardization of at least the top common technical and NTS is possible when researchers consider international collaboration by publishing comprehensive methodologies, frameworks, and results.

11. References

Australian Society of Medical Imaging and Radiation Therapy (ASMIRT). (2018). *Professional Practice Standards*. https://www.asmirt.org/asmirt_core/wp-content/uploads/371.pdf

Cimatti, B. (2016). Definition, development, assessment of soft skills and their role for the quality of organizations and enterprises. *International Journal for quality research, 10*(1), 97–130. https://doi.org/10.18421/ijqr10.01–05

Curtis, D. (2004). The assessment of generic skills. In J. Gibb (Ed.), *Generic Skills in Vocational Education and Training: Research Readings* (pp. 136-156). Centre for Vocational Education Research Ltd. https://files.eric.ed.gov/fulltext/ED493988.pdf

Davis, M. H., & Karunathilake, I. (2005). The place of the oral examination in today's assessment systems. *Medical teacher, 27*(4), 294-297. http://doi.org/10.1080/01421590500126437

De Champlain, A. (2019). Standard Setting Methods in Medical Education: High-stakes Assessment. In T. Swanwick, K., Forrest & B.C. O'Brien (Eds.). *Understanding Medical Education: Evidence, Theory, and Practice* (3rd ed., pp. 347–360). The Association for the Study of Medical Education (ASME). John Wiley & Sons Ltd. https://doi.org/10.1002/9781119373780.ch24

Driessen, E., & Tartwijk,J. (2019). Portfolios in personal and professional development. In T. Swanwick, K., Forrest & B.C. O'Brien (Eds.). *Understanding Medical Education: Evidence, Theory, and Practice* (3rd ed., pp. 225–161). The Association for the Study

of Medical Education (ASME). John Wiley & Sons Ltd. https://doi.org/10.1002/9781119373780.ch18

Duffy, D., Gordon, H., Whelan, G., Cole-Kelly, K., & Frankel, R. (2004). Assessing competence in communication and interpersonal skills: the Kalamazoo II report. *Academic medicine*, *79*(6), 495–507. https://journals.lww.com/academicmedicine/fulltext/2004/06000/assessing_competence_in_communication_and.2.aspx

El-Farra S.A. (2022). Alternative Digital Credentials: UAE's First Adopters' Design, Development, and Implementation Part (1). *International Journal of Learning, Teaching and Educational Research. 10*(21), *pp. 64-87.* https://doi.org/10.26803/ijlter.21.10.4

El-Farra, S., Mohaidat, M., Aldajah, S., & Alshamsi, A. (2022) Alternative Digital Credentials—UAE's First Adopters' Quality Assurance Model and Case Study. In K. Cheng, B. Koul , T. Wang , & X. Yu. (Eds.). *Artificial Intelligence in Education: Emerging Technologies, Models and Applications. Lecture Notes on Data Engineering and Communications Technologies*, 104, (pp. 339–359). Springer. https://doi.org/10.1007/978-981-16-7527-0_25

Englander, R., Frank, J. R., Carraccio, C., Sherbino, J., Ross, S., Snell, L., & ICBME Collaborators. (2017). Toward a shared language for competency–based medical education. *Medical teacher*, *39*(6), 582–587. https://doi.org/10.1080/0142159X.2017.1315066

Gallardo, K. (2020). Competency-based assessment and the use of performance-based evaluation rubrics in higher education: Challenges towards the next decade. *Problems of Education in the 21st Century*, *78*(1), 61-79. https://doi.org/10.33225/pec/20.78.61

Ganji, K. K. (2017). Evaluation of reliability in structured viva voce as a formative assessment of dental students. *Journal of dental education, 81*(5), 590-596. 88 https://doi.org/10.21815/JDE.016.017

Gibbs, G. (1988). *Learning by Doing: A guide to teaching and learning methods*. Further Education Unit. Oxford Polytechnic. https://thoughtsmostlyaboutlearning.files.wordpress.com/2015/12/learning-by-doing-graham-gibbs.pdf

Goldman, J., & Wong, B.M. (2020). Nothing soft about 'soft skills': core competencies in quality improvement and patient safety education and practice. *BMJ Quality & Safety, 29*(619–622). http://dx.doi.org/10.1136/bmjqs-2019-010512.

Gordon, M., Farnan, J., Grafton-Clarke, C., Ahmed, R., Gurbutt, D., McLachlan, J., & Daniel, M. (2019). Non-technical skills assessments in undergraduate medical education: a focused BEME systematic review: BEME Guide No. 54. *Medical teacher*, *41*(7), 732-745. https://doi.org/10.1080/0142159X.2018.1562166

Gruppen, L. D., Mangrulkar, R. S., & Kolars, J. C. (2012). The promise of competency-based education in the health professions for improving global health. *Human Resources for Health,* 10(1), 1-7. https://human-resources-health.biomedcentral.com/articles/10.1186/1478-4491-10-43

Higham, H., Greig, P. R., Rutherford, J., Vincent, L., Young, D., & Vincent, C. (2019). Observer-based tools for non-technical skills assessment in simulated and real clinical environments in healthcare: a systematic review. *BMJ Quality & Safety*, *28*(8), 672-686. https://ora.ox.ac.uk/objects/uuid:b9e628c5-9501-4ad4-9304-dcfd75d773b2/download_file?safe_filename=Higham_et_al_2019_Observer-based_tools_for_non-technical_skills_assessment.pdf&file_format=application%2Fpdf&type_of_work=Journal+article

Hojat, M. (2016). A definition and key features of empathy in patient care. *In Empathy in Health Professions Education and Patient Care* (pp. 71-81). Springer. https://doi.org/10.1007/978-3-319-27625-0_6

Hojat, M., DeSantis, J., Shannon, S. C., Mortensen, L. H., Speicher, M. R., Bragan, L., LaNoue, M. & Calabrese, L. H. (2018). The Jefferson Scale of Empathy: a nationwide study of measurement properties, underlying components, latent variable structure, and national norms in medical students. *Advances in Health Sciences Education*, *23*(5), 899-920. https://doi.org/10.1007/s10459-018-9839-9

Hojat, M., & Gonnella, J. S. (2017). What matters more about the Interpersonal Reactivity Index and the Jefferson Scale of Empathy? Their underlying constructs or their relationships with pertinent measures of clinical competence and patient outcomes?. *Academic Medicine*, *92*(6), 743-745. https://doi.org/10.1097/ACM.0000000000001424

International Society of Radiographers and Radiological Technologists (ISRRT). (2022). *Board of Management Motion International Academic Network (IAN) New membership Category.* https://www.isrrt.org/pdf/Item_6_new_membership_category_educational_institutes.pdf

Jefferies, A., Simmons, B., Ng, E., & Skidmore, M. (2011). Assessment of multiple physician competencies in postgraduate training: Utility of the structured oral examination. *Advances in Health Sciences Education*, *16*(5), 569-577. https://doi.org/10.1007/s10459-011-9275-6

Kirkpatrick, J. D., & Kirkpatrick, W. K. (2016). *Kirkpatrick's four levels of training evaluation.* Association for Talent Development.

Lampignano, J., & Kendrick, L. E. (2017). *Bontrager's textbook of radiographic positioning and related anatomy-E-book*. Elsevier Health Sciences.

Norcini, J., & Zaidi, Z. (2019). Workplace assessment. In T. Swanwick, K., Forrest & B.C. O'Brien (Eds.). *Understanding Medical Education: Evidence, Theory, and Practice* (3rd ed., pp. 319–334). The Association for the Study of Medical Education (ASME). John Wiley & Sons Ltd. https://doi.org/10.1002/9781119373780.ch22

Schuwirth, L.W.T. & van der Vleuten, C.P.M (2019). How to Design a Useful Test: The Principles of Assessment. In T. Swanwick, K., Forrest & B.C. O'Brien. (Eds.). *Understanding Medical Education: Evidence, Theory, and Practice* (3rd ed., pp. 277–290). The Association for the Study of Medical Education (ASME). John Wiley & Sons Ltd. https://doi.org/10.1002/9781119373780.ch20

Shenton, A. K. (2004). Strategies for ensuring trustworthiness in qualitative research projects. *Education for information*, *22*(2), 63-75. https://www.pm.lth.se/fileadmin/_migrated/content_uploads/Shenton_Trustworthiness.pdf

Shenwai, M. R., & Patil, K. (2013). Introduction of Structured Oral Examination as A Novel Assessment tool to First Year Medical Students in Physiology. *Journal of clinical and diagnostic research: JCDR*, 7(11), 2544–2547. https://doi.org/10.7860/JCDR/2013/7350.3606

Thomas, M. J. (2018). *Training and assessing non–technical skills: A practical guide*. CRC Press. https://doi.org/10.1201/9781315550336

Velasco-Martínez, L. C., & Hurtado, J. C. T. (2018). The use of rubrics in higher education and competences evaluation. *Profesorado, 22*(3), 183–208. https://doi.org/10.30827/profesorado.v22i3.7998

Yune, S. J., Lee, S. Y., Im, S. J., Kam, B. S., & Baek, S. Y. (2018). Holistic rubric vs. analytic rubric for measuring clinical performance levels in medical students. *BMC medical education, 18*(1), 1-6. https://doi.org/10.1186/s12909-018-1228-9