

1. 何を評価するのか？

臨床医のコンピテンシー，プロフェッショナル・アイデンティティ^{1,2)}という概念が浸透するに伴い，状況に応じて統合して発揮する医師の実践的能力や専門職としての日常の行動を評価，可否判定することが教育・研修プログラムの修了認定や資格認定に求められるようになった。学習者評価には評価する能力が観察可能な行動として明確に定義されて，達成すべきレベルとともに明示されていることが必須である。そのためには，例えば，

「倫理観を持つ」→「当該学習・研修機関の倫理規範に基づいた行動をする」

「生涯，学習する」→「(現在の研修において)自己評価により課題を発見し，自主的に情報収集して判断し必要な技能の修得を行う」

等，具体的な行動として記述することが第一歩である。

CanMEDS2015 competency framework³⁾はカナダにおける研修プログラムの基盤であり，統合して発揮される8つの役割roleの定義と評価可能な具体的行動が示されている。日本でも医師に求められる能力・資質の概要作成が始まったが，学習者評価に利用できる明確な記述とそれを活用する制度の整備はこれからであり，各教育機関・プログラムに基準の作成が求められている。

2. アウトカム基盤型教育による学習者評価の変化

アウトカム基盤型教育の導入により，学習者評価も変化した。プログラム修了時の「期待する学

習成果」を保証するために，異なる背景を持つ学習者が同じ経験をして，理解や行動の変化は多様であることを踏まえ，学習者評価が必須となる。また，人材育成を目的とする医学教育では，可否，成績判定等の総括的評価よりも，形成的評価，個々の学習者に応じた支援と評価後の再教育の重要性が強調されている。

学習者の行動の改善，向上に効果的な形成的評価とは，何がどの程度修得され何が不足しているかの豊富な情報が，タイミングよく提供され，学習者の振り返りを促して能動的，自立的な学習を促すものである。そのためには数字や標語による判定結果に加え，修得の程度や進捗状況を示す評価（たとえばマイルストーン評価）(図1)⁴⁾を活用し，行動が観察された状況を加味した改善のための具体的なフィードバックを行う。

3. 学習者評価の質の検証

(1) 妥当性とは何か

妥当性は学習者評価の検討において最も優先すべき事項である。学習者評価の計画から実施，評価結果の利用，評価後の影響を含めて検討することが妥当性の検証⁵⁾であり，評価形式や評価ツールだけで評価の良し悪しは議論できない。ここでの妥当性は尺度開発における妥当性の検証とは異なる事項である。尺度としての妥当性が検証された評価ツールであっても，適切に利用されなければ妥当性は極めて低くなる。したがって「技能と態度をOSCEで評価しているので妥当である」「mini-CEXを用いている臨床実習の評価は妥当性が高い」では，評価の妥当性を説明していることにはならない。

妥当性の重要な要素に，「一般化」と「評価の影響」が挙げられる。

^{*1} Learner Assessment

^{*2} Masami Tagawa 鹿児島大学大学院医歯学総合研究科医歯学教育開発センター

一般化に着目したKaneの妥当性 I ~ IV⁷⁻⁹⁾に従って、A大学のOSCEを分析する。

1) 観察から点数へ：評価項目の妥当性

胸部診察ステーションでは、胸部診察の50余りの学習項目と医療安全等関連領域から選択された10数項目で一つの胸部課題が構成され、胸部ステーションの総合点が算出される。「呼吸音を聴診する」が評価項目の一つに選択された場合、受験生には前胸部、背部、あるいはそのどちらかの呼吸音を聴診する指示が出され、2名の評価者は受験生のパフォーマンスを観察して評価基準に従って項目の評点をつける。項目の受験生への説明文章、評価表のスケール、評価基準、評価者によるばらつき、評価項目の前後の流れ、受験生の時間配分等の結果として、ある学生の1項目の評価点が記録される。「呼吸音を聴診する」能力のある学生が高い点となり、能力が満たない学生は0点あるいは低い点ならば、項目の評価には妥当性がある。

聴診器を胸部に当てれば呼吸音の聴取にかかわらず高い得点となるならば、この項目の点が示すのは「呼吸音の聴診部位に聴診器をあてて見せる能力」であり、「呼吸音を聴診する能力」の評価ではない。

2) 観察した点数から該当領域の点数へ：ステーション課題の妥当性

次の検討は胸部診察の能力を表すステーションの得点である。胸部ステーション課題の合計点、すなわち10数項目の総合点が、この課題で評価する胸部診察の能力、すなわち視診、聴診、触診、打診による胸郭、胸部皮膚、血管、心臓、呼吸の診察を患者に配慮して行う能力を表していると一般化できるならば、課題の評価は妥当性がある。妥当性は、100%の得点（満点）を構成する評価項目の選択、組み合わせ、総合点に対する比重の根拠ある計画と実際の評価データ解析で示される。

現在共用試験OSCEでは胸部診察には複数の課題が用いられており、大学によって、また同じ大学でも年度によって異なる課題で評価している。どの課題でも胸部診察の能力を一般化できる判定

が可能ならば（これは全課題の平均点が同じという意味ではない）、大学間、年度間の成績の比較が可能になり、学生能力を大学、年度を越えて保証できる。

3) 領域の点数から評価対象とした領域全体へ：OSCEによる能力判定の妥当性

胸部だけではなく全ステーション課題の妥当性を検証した次に、6つのステーションの総合点で判定するOSCEの妥当性を検討する。評価の目的である「基本的医療面接技能、身体診察、基本的臨床手技を安全に実施する」能力にOSCEの成績を一般化できるかを検討する。適切なブループリント（課題の組み合わせと評価項目、配点で、評価すべき能力を評価していることを示す仕様書）が存在し、実際に用いられた6課題の構成（組み合わせ）、各課題とその評価項目がブループリントに従っていれば、内容妥当性の根拠になる。ステーション数（課題数）が少ない、あるいは偏っていればconstruct underrepresentation、評価領域外の課題や評価者によるばらつきはconstruct-irrelevant varianceと評価される¹⁰⁾。このような内容妥当性が脅かされた状態は課題を増やすことなく、評価者数を増やしても改善しない。運営、評価データから算出する信頼性も合わせ、妥当性を判断する。

4) 評価対象領域から修得すべき能力への一般化：OSCEの成績、可否の利用に関する妥当性

診療参加型臨床実習で患者を受け持ち、医行為を実施するために必要な臨床能力の有無をOSCEだけで、しかも6ステーションで判断することは、実際には難しい。OSCEに関わる教員も受験する学生も、評価すべき能力のほんの一部を、実際の診療とは異なる設定で評価していることを理解している。不合格になった学生が十分な指導を受け再評価の機会を得られることと、評価しない技能を学生が十分に修得していないことも予想して、合格した学生に臨床実習での医行為を慎重に許可しているのであれば、評価結果の利用に大きな問題は生じないであろう。しかし、患者に「この学生は共用試験に合格しているので患者診療に

必要な臨床能力がありますので、実習への協力をお願いします」と説明するならば、妥当性は十分に検証され、保障されていることが前提である。

共用試験OSCEを実施している意義は極めて大きいですが、妥当性の点からは課題があり、それを踏まえて評価結果を活用し、改善を続けることが各医学部に求められる社会的責任と言える。

(3) 評価の影響consequences

能力を的確に評価している時、学習者はその能力を修得する学習を行い、指導者は望ましい指導を行うと期待される。OSCEの胸部ステーションで、決められたセリフを述べ、聴診器を胸部に当

てることが評価されると知れば、受験生は述べるべきセリフをひたすら暗記し、聴診器をいかに素早く所定の場所に触れるかを練習する。視診で診断し、聴診して音を聞き分けた学生が手技に時間がかかり、診断はせずに試験対策の手技を実施した受験生が高得点を得るのであれば、学生と指導者に視診、聴診をしない学習を誘導する。ましてOSCEは合格をしなければ留年となり、OSCEの成績提出がマッチングで受験する病院から求められるのであれば、なおさらである。学生の行動が評価方法から生じるものである限り、学生を批判することはできない。

医学部の臨床実習後OSCEも実施方法次第で、

表1 試験開発の12ステップ (Downing SM. 引用¹¹⁾ TABLE1.1を日本語訳、一部引用本文に基づき改編¹²⁾)

Step		業務の例
1	全体の計画	試験開発全体の体系的ガイダンス： 測定する構成概念：望まれている点数の解釈；単一あるいは複数の出題形式；妥当性の根拠となる重要な資料；明確な目標；望ましい試験の影響；心理測定モデル；日程；守秘；質の管理
2	評価内容の定義	領域／分野からのサンプリング計画；評価目的に合致した多彩な方法；内容に関連した妥当性の基礎資料；構成概念の記述
3	試験の設計書	運用上の内容の定義；領域からのサンプリングに関する体系的で弁明できる妥当性の根拠を示す枠組み；集団における相対評価か能力を基準とした評価か；望ましい問題の特性
4	問題（評価項目）の作成	効果的な問題の開発；形式；根拠に基づく原理を厳守した妥当性の根拠；問題作成者と検閲者のトレーニング；効果的な問題の校正；問題の不備による構成概念に無関係な分散
5	試験のデザインと組立	試験形式のデザインと作成；設計された試験形式のための問題の選択；計画されたブループリントによるサンプリング；評価対象とはならない試用問題の検討
6	試験の作成	製本；印刷あるいはコンピューター試験の完成；守秘に関する事項；質管理に関する妥当性の事項
7	試験の実施	標準化に関する妥当性の事項；障害者法への対応；試験監督；漏洩；時間設定
8	採点	妥当性；質管理；正解の妥当性確認；項目分析
9	合格点	弁明できる合格点の設定；相対的あるいは絶対的；基準点の妥当性の事項；複数の試験実施時の基準の相似性；得点尺度の恒常性の維持（複数の試験結果の同一化、関連付け）
10	結果の報告	妥当性；正確さ；質管理；適時か；意義；誤用；改善点；再受験
11	問題の蓄積	守秘／漏洩；有用性；融通性；効果的な問題蓄積の原則
12	試験の技術的報告	妥当性の根拠の体系的ですべてにわたる詳細な記述；12ステップの構成；勧告

学生は臨床実習に積極的に臨むかもしれないし、病棟ではなくシミュレーションセンターでの実習に熱心になるかもしれない。血圧測定が出ないとわかれば、臨床実習での受け持ち患者の血圧を計らなくなる。診断が評価され治療は評価対象外であれば、診断を学習し、治療への関心が薄れることが予想される。

実践とは異なる人為的設定で行うOSCEでは、課題をこなすための試験技能が必ず必要になるが、これは本来学習すべき能力ではない。実践に近い種々の状況で真正性の高い診察、診断の能力を得点に反映させる計画と、本来評価すべき能力を識別できているかの調査、分析も行う。試験手順や人為的設定に対する学生の戸惑い、不安の解消も、学生が本来の能力を試験で発揮するために重要である。評価対象の能力、課題の構成、合格基準、使用する機材の取り扱い方法、時間配分など、受験上必要な情報はすべて事前に周知し、試験では診療の遂行に集中できる環境を整える。一方、ブループリント（どのような疾患を出題し、どのように採点するのか）は学生には開示する必要はない。

試験により学習成果が明らかになり、教育改善が進むことも期待される。OSCEは多くの資源を用いる試験であるが、好ましい影響が明らかになれば試験への理解も得られ、さらに教育への良い波及効果も期待できる。学生の行動を予測し、観察して試験対策に走らないようにする方法を考え、またあらかじめ意図していなかった試験の影響も調査して妥当性を検討し、対応することが教育者の課題である。

4. 医学教育における評価の計画と運営

(1) 評価計画～評価の検証・報告

試験開発の12のステップ（表1）は学習者評価計画、運営の基本である^{11,12)}。評価する能力の明示と試験の設計書（Specification, ブループリント）に従った問題、課題、評価項目を作成し、適切な運営で評価、判定が行われれば評価結果の妥当性の根拠となる。

(2) 能力の絶対評価：Criterionで評価する

医学教育における学習者評価の目的は、医療者としての能力と次の学習・研修段階に進むための能力の保証である。医療者の資格認定は、年度を越えて一定の能力を保証する合否判定でなければ市民を守ることはできない¹³⁾。

対象グループ内での相対的評価、例えば、正規分布を想定した得点分布のどの位置にいるかで判断する場合（例：-2.0SD未満を不合格とする）、集団の順位で合否を決定する場合（例：上位100名までを合格とする）は、集団によって、また試験実施の度に合格者の能力は変化する。このような集団準拠norm-referenced、また「一律に60点を合格点とする」という絶対評価では能力の質保証は不十分であり、医学教育では基準準拠criterion-referencedで判断すべきとされている。観察した行動をルーブリックに従って判定したり、多肢選択問題ではアンゴフ法により問題の難易度を考慮した合格基準点を設定するなど基準準拠を念頭に置いた尺度と判定方法で、合格者は常に求められたレベル以上の能力を有していると判断できる。

よく知られているMini-CEXもP-MEXもオリジナルの尺度には、能力の判定基準が示されていないため、運用上の評価基準を各教育機関で文書化するか、新たなルーブリックの作成が必要になる。北米で用いられているコンピテンシーのマイルストーンは国として制度化された研修中の評価のためのツールであり、基準準拠で行う評価判定の良い例である。

(3) 臨床現場での評価

(workplace-based assessment)

実践的能力と日常的な行動・態度の評価の必要性から、臨床実習、研修など、診療現場での評価が着目されている。修得する能力の多様性と測定、判定上の困難のため、能力評価に関して多くの工夫が重ねられてきた。コンピテンシーとマイルストーン、EPA (entrustable professional activities)¹⁴⁻¹⁶⁾も評価を適切に実施し、利用するための制度である。診療現場で種々の評価データを収集、ポートフォリオに集積し、データに基づく能力の判

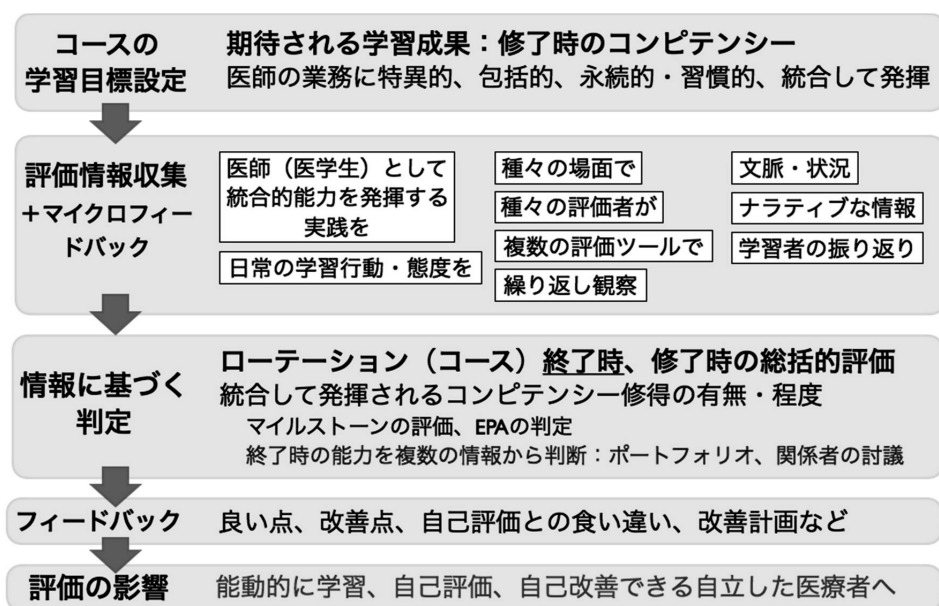


図3 臨床実習、研修での評価

断、修得のためのフィードバックを確実に実施していく（図3）。

・評価情報の収集

ある場面での特定患者に対するある診療活動を観察して判定できるのは、その時の診療活動遂行の良否、適否であり、その実施から判断できる評価対象者の能力である。評価データを左右するのは評価対象者の能力であるが、そこに場面、患者、患者と評価対象者との関係、評価対象者と評価者との関係をはじめとする種々の要因が加わり評価データは変化する。状況、症例特異性、評価者の誤差等が影響した評価データを多数収集して、評価対象者の真の能力を推定し、判定する。

臨床現場での評価は教育・研修プログラムで修得すべき能力の評価に直結し、ほぼすべてのコンピテンシーや診療活動に関する判定を行う。コンピテンシーは単一の評価方法、尺度、評価機会では判定できないことは明らかであり、計画的な評価が求められる⁷⁾。能力を判断するための評価データを収集する上で、以下を検討する。

①何を：コンピテンシーの各領域の具体的行動・

態度の項目、診療活動（疾患の診療、手技、場面）、患者など

②いつ：学習・研修期間のどの時点で、観察直後に、一定の観察期間の終了時に、随時など

③どこで：外来・入院・救急外来・手術室、カンファレンス・回診など

④誰が：指導医、研修医、同僚、医療専門職、患者・家族、学習者自身など

⑤どのように：観察評価、評価ツール、問題・質問、評価回数、ナラティブな記述など

コンピテンシーや特定の診療活動の実践を評価するために、どのような場面を組み合わせ、何回評価データを収集すれば、判断に十分な情報量となるかは難しい問題である。同じ評価者が同じ状況で同一の能力を判定すれば、少ないデータでも信頼性は高くなるが内容妥当性は低い。疾患、状況、患者が多様であるほど同じ数のデータでも信頼性は低くなるが、内容妥当性は高い。評価対象の能力に応じた評価が実施され、新たなデータを加えても判定結果に変更がなければ、十分なデータが得られていると判断しても良いとする現実的な方法も提案されている⁷⁾。

・情報に基づく判定

多数集積されたデータを用いて、コース（ローテーション）およびプログラム終了時の成績、可否を判定する。臨床実習、臨床研修終了時に修得しているべきコンピテンシーの各項目を基準準拠で判定して、すべて合格レベル以上であればプログラム修了を判定できる。繰り返し観察、評価した特定の診療活動の実践が、一人での実施が許可できる信頼できるレベル（entrustable）であるかも判定する。

コース（ローテーション）ごとに総括的評価をするのか、形成的評価のみに止めるのかは、各プログラムで決定する。コース横断的な能力、例えばコミュニケーション、倫理的対応、自己主導型学習はコースを超えて繰り返し学習して身につくものであるから、各コースで形成的評価とフィードバックで修得を支援し、プログラム終了時までに合格レベル達成を目指すことになるであろう。全ローテーション共通の尺度であるマイルストーンの評価は学習の進捗を明らかにできる優れた方法である。一方、内科、外科、小児科、精神科、救急等に特化した診療の実践や能力は、各コース（ローテーション）で成績判定をすることになる。

多彩な評価ツールで集められた多数のデータを用いた総合判定に絶対的な方法はない。責任者は、ポートフォリオに蓄積された評価データ（点数、記述）の評価時の状況を踏まえ、直接の指導者、観察者、必要に応じて学習者も加わり、判断することが望ましい。

・フィードバック

プログラム終了時のコンピテンシー（教育・研修目標）の修得を促すために、実践直後の評価データに基づくマイクロフィードバック、実践と各種評価結果に対する学習者自身の振り返り、教育・研修目標の達成度を明らかにするマイルストーンとEPAの評価をポートフォリオに集積する。評価データは学習者自身の適切な自己評価に有効であり、評価に基づいて学習者が次の学習を計画し実施するための支援を行う。

まとめ

医療者のコンピテンシーの評価は容易ではなく、医学教育の専門的知識と的確な運営が求められる。臨床現場での評価とOSCEを含む評価それぞれの妥当性を高めることが教育、研修プログラムと資格認定制度による医療者の質保証に直結している。

■文 献

- 1) Irby DM, Cooke M, O'Brien BC. Call for reform of medical education by the Carnegie foundation for the advancement of teaching: 1910 and 2010. *Acad Med* 2010; **85**: 220-227.
- 2) Cruess RL, Cruess SR, Boundreau JD, Snell L, Steinert Y. Reframing medical education to support professional identity formation. *Acad Med* 2014; **89**: 1446-1451.
- 3) http://canmeds.royalcollege.ca/uploads/en/framework/CanMEDS%202015%20Framework_EN_Reduced.pdf (最終閲覧日: 2017.11.7)
- 4) <http://canmeds.royalcollege.ca/guide> (最終閲覧日: 2017.11.7)
- 5) Downing SM. Validity: on the meaningful interpretation of assessment data. *Med Educ* 2003; **37**: 830-837.
- 6) Messick S. Validity. In: Linn RL editor. Educational measurement 3rd ed., American Council on Education, Macmillan, New York, 1989, p.13-104.
- 7) Kane MT. Current concerns in validity theory. *J Educ Measure* 2001; **38**: 319-342.
- 8) Schuwirth LWT, van der Vleuten C PM. Programmatic assessment and Kane's validity perspective. *Med Educ* 2012; **46**: 38-48.
- 9) Cook DA, Brydges R, Ginsburg S, Hatala R. A contemporary approach to validity arguments: a practical guide to Kane's framework. *Med Educ* 2015; **49**: 560-575.
- 10) Downing SM, Haladyna TM. Validity threats: overcoming interference with proposed interpretations of assessment data. *Med Educ* 2004; **38**: 327-333.

- 11) Downing SM. Twelve steps for effective test development. In : Downing SM, Haladyna TM, editors. Handbook of test development. Lawrence Erlbaum Associates, Mahwah, New Jersey, 2006, p.3-25.
- 12) 田川まさみ, 西城卓也 : 医学教育における学習者の評価① 総論. 医学教育2013; **44**: 345-357.
- 13) Norcini JJ, Shea JA. The credibility and comparability of standards. *Appl Measurement Educ* 1997; **10**:39-59.
- 14) ten Cate O. Entrustability of professional activities and competency-based training. *Med Educ* 2005; **39**: 1176-1177.
- 15) ten Cate, O; Hart, D; Ankel, F; Busari, J; Englander, R; Glasgow, N; Holmboe, E; Iobst, W; Lovell, E; Snell, LS; Touchie, C; Van Melle, E; Wycliffe-Jones, K. Entrustment decision making in clinical training. *Acad Med* 2016; **91**: 191-198.
- 16) Englander, R; Flynn, T; Call, S; Carraccio, C; Cleary, L; Fulton, TB; Garrity, MJ; Lieberman, SA; Lindeman, B; Lypson, ML; Minter, RM; Rosenfield, J; Thomas, J; Wilson, MC; Aschenbrener, CA. Toward defining the foundation of the MD Degree: core entrustable professional activities for entering residency. *Acad Med* 2016; **91**: 1352-1358.
- 17) <http://canmeds.royalcollege.ca> (最終閲覧日 : 2017.11.7)