10 プログラム評価，カリキュラム評価の考え方[*1]

大西　弘高[*2]

1．はじめに

医学教育の領域において，何か新たな授業を採り入れたとき，以前の授業を改善したときなどに，そのプログラムやカリキュラムの評価が必要になると感じた方は多いだろう．実際，年次大会での発表，学会誌での論文の中にも，プログラム評価に関連した内容は非常に多い．しかし，医学教育領域においてプログラム評価がどうあるべきかについて包括的に論じられた機会は非常に少ない．

本稿では，新しいカリキュラムの実施，あるいはカリキュラムの改革を行ったときに，それをどう評価すればよいかというようなレベルの話題を論じながら，プログラム評価，カリキュラム評価の話題について概説したい．

2．用語の定義

プログラム評価という用語における「プログラム」を，安田は「特定の社会的・教育的目標を達成するために，人が中心となって介入やアクションを行う事業」と定義している[1]．一方，従来「カリキュラム」は，教育プログラム，教育課程とほぼ同じ意味を指すとされてきた．しかし，成人学習理論が広く受け入れられるようになった後，教育課程は「履修」を規定するものではあっても「習得」や「修得」について踏み込んだ議論をしていなかったため，教育者中心の教育観における用語とみなされる傾向が強まった．学習者中心の教育観に立脚すると，カリキュラムは学習者が学び取ったすべてのものを含むため，潜在的カリキュラムの影響なども考慮する必要が生まれた[2]．

こういった点を考慮すると，「カリキュラム評価」においては，学習者の学び取ったもの，あるいはそれによって生じる社会的影響などを含めて包括的に評価する必要があることが見えてくる．では，「プログラム評価」は，“プログラムにおける計画”の通りに事が進んだかどうかを評価するだけでよいのだろうか．この点については，カリキュラムに関する議論と同様，以前はプログラムを策定する最初の段階である「プログラム目標」がどの程度達成できたかを重視していた．しかし，近年ではプログラムが住民や社会にどのような影響を及ぼし，価値を加えたかという部分を重視する傾向が強まっている．

英語では，プログラム評価やカリキュラム評価の「評価」は，いずれもevaluationの語を当てはめる．この語は，value，すなわち「価値」を推し量ることを言い表している．単にプログラム目標の達成度を一元的に測定，査定するといった観点には留まらないことを理解していただけるだろう．

日本語で単に「評価」と言うと，学習者の到達度や改善点に対する評価（assessment，学習者評価）のことを指すことも多い．学習者評価を寄せ集めたデータもプログラムやカリキュラムの評価の一部をなすが，かなり次元や対象規模が異なることに注意が必要である．

また，大学評価や病院機能評価など，認証評価（accreditation）と呼ばれる領域に関する内容も時に混同されることがある．Webster dictionaryにおいては，教育機関に関連したaccreditationの語義として，「当該教育機関の課程修了者が更に上級の教育機関に入学してよいか，専門職としての就労を行えるか，に関する要件を満たしていることを認定すること（筆者訳）」とされている．

---

*1 The concept of program and curriculum evaluation

*2 Hirotaka Onishi　東京大学医学教育国際協力研究センター

認証評価は，それぞれのプログラムの価値を判断するよりは同種のプログラムや教育組織全体の質を管理するための手法であり，今回の議論には含めないことにする．

より一般的な用語として，「調査」というものもあり，調査と評価の違いも議論になりうる．Davidsonは，評価＝事実特定（Factual identification）＋価値判断（Value determination）であり，調査は「事実特定」の部分のみを目的にしているものとして区別している[3]．

3．プログラム評価の実例

例として，2004年春から開始された新医師臨床研修制度に関する評価を挙げてみたい．福井らは，平成17～19年に実施された厚生労働科学研究の「新医師臨床研修制度の評価に関する調査研究」にて，臨床研修体制・プログラム・処遇に対する満足度，臨床研修終了後の進路，基本的臨床能力（知識，技術，態度に関する99項目）の修得状況，それに症例経験数（82の症状・病態，4種類の医療記録）などを明確にする評価を行った[4]．そして，「研修医の満足度で見る限り，大学病院での研修の質の向上が大きく，旧制度下で見られたような臨床研修病院との差は縮まりつつある．基本的臨床能力や症例経験数についても，新制度下での満足すべき達成度が維持されていて，臨床研修病院と大学病院との間でもほとんど差はない」と結論づけた．

一方で，2008年9月より文部科学省，厚生労働省が開催した「臨床研修制度のあり方等に関する検討会」においても議論が重ねられ，2009年2月に「臨床研修制度等に関する意見のとりまとめ」が出された[5]．ここでは，医学部教育改革との連携不足，大学病院若手医師の減少，研修医を含めた若手医師の都市部集中といった点も指摘され，(1)研修プログラムの弾力化，(2)募集定員や受入病院のあり方の見直し，(3)関連する制度等の見直し，といった対策の方向性が示された．この検討会は，より包括的なプログラム評価，しかも研修制度改革の方向性を決定する「総括的プログラム評価」の色を帯びていると言えるだろう．

これらを比較すると，前者の結果は，後者の議論に一定の役割を果たしているものの，プログラム全体の価値判断という観点では，後者の議論の方がはるかに政策決定に大きな役割を果たしていることが分かる．あるいは，前者は「評価に関する調査研究」と称しているだけあって，調査や研究であっても，評価そのものではないと考えるのがより妥当かもしれない．

4．プログラム評価にまつわる様々な課題

新医師臨床研修制度での例を比較し，プログラム評価を議論する上で，**表1**のような課題があることが示唆される．以下，それぞれに関して議論を進めていく．

1） 目的

医療者教育プログラム評価のレベルに関して

表1　プログラム評価における論点

項　目	論　点
目　的	• 評価のレベル：Kirkpatrickの4段階 • 4種類の評価目的 • 研究と評価の違い • プログラムの進行と評価の変化
モデル	• プログラム理論 • ロジックモデル • 評価デザイン
情　報	• 量的データと質的データ • 内部評価（自己評価）と外部評価（他者評価）の比較 • ステークホルダー分析

は，Kirkpatrickの4段階がよく知られている（**図1**）[6]．レベルは上がる方が，より社会的な観点からインパクトが大きいが，妥当性の高い評価はより難しくなる．レベル1は学習者の満足度を自記式に測定することが一般的に行われる．これは，学習者の満足感が，レベル2の達成を導きやすくするからである．レベル2は，タキソノミー（taxonomy：学習目標分類）の考え方と似ている[7]．レベル3の行動変容に至るには，知識やスキルが身に付くだけでなく，態度や認識が変容し，日常的な行動を変えようという気持ちを伴うことが求められるため，これらの要因がレベル2に分けて記載されていると考えるのが妥当である．レベル4の一つ，患者の利益に関しては，医療者の行動変容の次のステップとして最も重要なアウトカムである．一方で，組織内で業務を行う多くの医療専門職の行動が変容し，組織のあり方自体が変化すると，それは「組織変革」と呼ばれる状況である．この変革はトップダウンで行われることもあるが，キーパーソンの主導で，あるいはボトムアップで行われることもあり，持続発展性（sustainability）にも関わる重要なアウトカムであると言える．

ロッシらは，プログラム評価の目的を，①プログラムの改良，②説明責任（accountability），③知識生成（knowledge generation），④裏の目的，の4種類に分けて述べている[8]．①は形成的評価（formative evaluation）やプロセス評価とも呼ばれ，改善点をいち早くフィードバックすることが重視される．②は総括的評価（summative evaluation）とも呼ばれ，信頼性の高い評価が必要となる．プログラムの効果などを定量的に示し，所期の目的を達成したかどうかが問われる．③は理論に基づいて計画されたプログラムが効果的かどうかを評価することで，介入対象や介入内容，プログラム自体について理解が深まり，新たな知識や知見が得られることで，研究的な側面の必要性が高い．④は広報活動，既に決定された内容の裏付けといった本来的でない評価であり，学問的見地からは避けたいものと言えよう．

評価は研究の形で行われることもあるが，それが一般像というわけではない．プログラム評価自体は当該プログラムのことだけが分かればよいが，研究の目的はプログラムの範囲に留まらない真実を明らかにし，新たな知見を生むことだからである．研究は「真実（truth）の追究」，評価は「価値（value）の追求」とも言われている[9]．上記の4段階，4種類の評価を研究という視点から見ると，できればKirkpatrickのレベルは3や4など高いレベルのものを含めて評価し，①，②，③の評価をプログラムの規模や実施時期との関連から適宜選択して実施すれば，一般化可能な結論を含めた研究の視点が含まれやすい印象がある．

プログラム評価は，プログラムの進行に伴って**表2**のような形で変化していく．1〜6のプロセスは繰り返し改善サイクルを描くように実施されてもよい．1や6においては，プログラムの良し悪しだけでなく，費用対効果に優れたプログラムかどうかも問題になっている．教育現場では時にこの側面を入れずに議論してしまうことがあり注

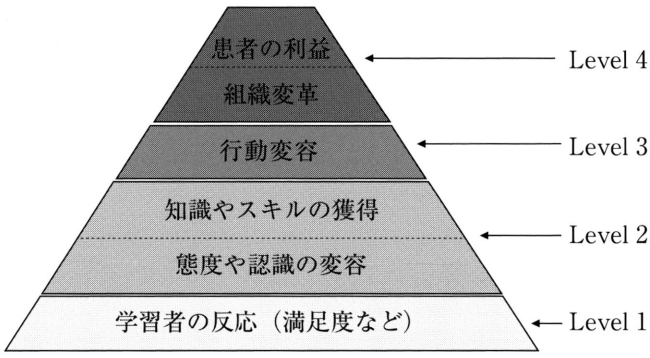

図1　Kirkpatrickの4段階

表2　プログラムの進行と評価プロセスの関係

	プログラムの進行	評価プロセス
1.	プログラムやその改革に対するニーズ調査	ニーズアセスメント
2.	プログラム自体やその改革の計画	資料の取りまとめと評価可能性アセスメント（evaluability assessment）
3.	プログラムの実施	形成的評価
4.	プログラムの効果のモニタリング	形成的評価＋総括的評価
5.	プログラム実施後のアウトカムのアセスメント	総括的評価
6.	プログラムの効率の改善	経済効率の評価

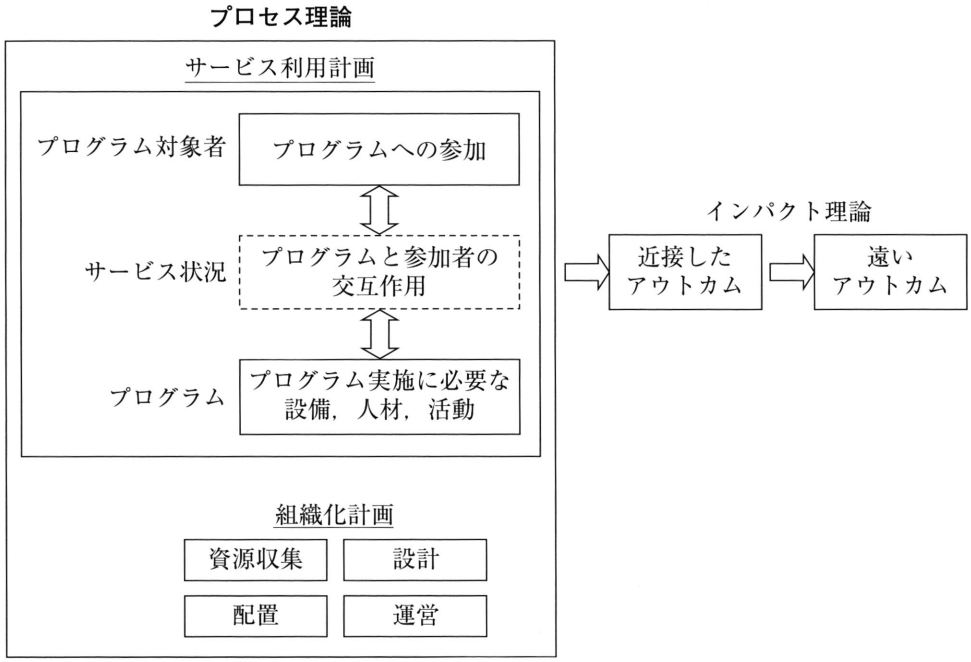

図2　プログラム理論の概念図

意が必要だろう．

2)　モデル

表2の2では評価可能性アセスメントという耳慣れない用語が出ている．評価の実施には多くの資源が必要だが，当該プログラムが“評価される”段階になければ，評価が資源を浪費してしまうこともあるため，プログラムの計画時点で評価可能かどうかをアセスメントしておく必要がある．例えば，「プログラムを何となく実施し，何となく上手くいった」という場合，本当に上手くいったのか，なぜ上手くいったのかの評価が困難なこともありうる．こういった事態を回避するために，プログラム理論を明確にした上で進める必要がある．

図2はプログラム理論の概念図である[10]．大きくプロセス理論とインパクト理論からなる．アウだけでなく組織化計画も必要であり，この観点で組織の運営・管理を行う者の力量が大きく物を言う面がある．

プログラム理論の一つとして近年重視されることが多いのは，ロジックモデルである．実施は資源の投入から始まるが，それがどのようにアウトカムやインパクトにつながっていくかを順序立てて図示するものと言える．**図3**にはロジックモデ

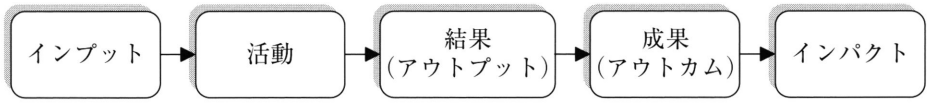

図3　基本ロジックモデル

ルの基本構造を描いた[11]．ロジックモデルをプログラム実施前に描いておくことで，たとえ計画通りにプログラムが進行しなかったとしても，そのインプットとアウトプット，アウトカム，インパクトの関係は見えやすくなるだろう．

評価デザインについては，以前はプログラム評価においても「ランダム化比較事前事後テストデザイン」などの実験デザインが重視されていた．しかし，プログラム評価の観点は，1969年のCampbellによる「政策やプログラムに関する決定は，社会状況を改善するための方法を検証する継続的な社会実験から引き出されるべき」といった行動主義なもの[12]から，1982年のCronbachによる「評価も科学的研究と同様の科学的手続はとるが，評価の目的は科学的研究から明確に区別されるべき．ステークホルダー（stakeholder：後に詳述）のニーズ充足を志向すべき」というようなもの[13]に変化してきた．佐々木は，このようなランダム化実験デザインへの忌避的な態度は，1980年代に拡がったと述べている[14]．

実験デザインと準実験デザインの主な違いは，ランダムに割り付けた比較群を設けるか否かである．例えば，前述した新臨床研修必修化というプログラムについては，ランダムに割り付けた比較群の配置は困難と言えよう．しかし，準実験デザインに関しても，より統計学的に意味のある解析が可能なものが開発されつつあることは知っておくべきかもしれない．

3)　情報

データが量的か質的かという点は，プログラム評価において本質的に重要とは言えない．しかし，評価を研究として実施する場合は，どちらかのデータのみを用い，それによって研究の結論の強さが変化することも多い．プログラム評価の文脈では，量的データは質問紙（アンケート）の形式で得られることが多いが，そもそも信頼性，妥当性に優れた質問紙が使われていることが必須条件となる．質的データを得る手段としては，観察，インタビュー，フォーカスグループ等が挙げられる．網羅的にデータが得られているか，無理のない結論が引き出されているかがポイントであろう．

内部評価（自己評価）と外部評価（他者評価）の比較は重要である．一般的に，内部情報には深みや厚みがあるが，バイアスがかかりやすいという特徴がある[15]．適宜，外部評価データと対照しつつ，妥当性を損ねないように注意する必要がある．

ステークホルダーは，日本語で利害関係者と訳されることもある．医療者教育プログラムにおいては，指導者，学習者に加え，教育施設の運営・管理に関わる人たち，学習者の親，患者や社会の人たちなど色んな人がステークホルダーとみなされる．様々なステークホルダーのニーズを十分汲んだ上で評価すること，必要に応じてステークホルダー自身を情報源として評価に採り入れること，も評価の要点となりうる．

■文　献

1）安田節之．プログラム評価の意義と展望：方法論の視点から．人事試験研究 2010；**214**：2-15.
2）大西弘高．新医学教育学入門，医学書院，東京，2005.
3）Davidson EJ. Evaluation methodology basics：the nuts and bolts of sounds evaluation. Sage Publications, Thousand Oaks, 2005.
4）福井次矢．厚生労働科学研究費補助金（医療安全・医療技術評価総合研究事業）平成19年度総括研究報告書「新医師臨床研修制度の評価に関する調査研究」.
5）臨床研修制度のあり方等に関する検討会（厚生労

働省）．臨床研修制度等に関する意見のとりまとめ．平成21年2月18日．

6）Kirkpatrick DL. Techniques for Evaluating Training Programs. In：Kirkpatrick DL, Evaluating Training Programs. American Society for Training and Development, Alexandria, 1975, p.1-17.

7）Kern DE, Thomas PA, Howard DMら著，小泉俊三監訳．医学教育プログラム開発：6段階アプローチによる学習と評価の一体化，篠原出版新社，東京，2003.

8）Rossi PH, Lipsey MW, Freeman HE著，大島巌，平岡公一，森俊夫ら監訳．プログラム評価の理論と方法：システマティックな対人サービス・政策評価の実践ガイド，日本評論社，東京，2005.

9）安田節之，渡辺直登．プログラム評価研究の方法，新曜社，東京，2008.

10）Rossi PH, Lipsey MW, Freeman HE. Evaluation：a systematic approach（6th ed.), Sage Publications, Thousand Oaks, 1999.

11）W. K. Kellogg Foundation. Using logic models to bring together planning, evaluation, & action：logic model development guide, W. K. Kellogg Foundation, Battle Creek, 2001.

12）Campbell DT. Reform as experiments. *American Psychologist* 1969；**24**：409-29.

13）Cronbach LJ. Designing evaluations of educational and social programs, Jossey-Bass, San Francisco, 1982.

14）佐々木亮．エビデンスに基づく開発援助評価：援助評価の歴史，ランダム化実験の起源，スクリヴェンとバナージェの考え方の比較．日本評価研究2010；**10**(1)：63-73.

15）Worthen BR, Sanders JR, Fitzpatrick JL. Program evaluation：alternative approaches and practical guidelines（2nd ed.), Longman, White Plains, 1997.