International Journal of Technology Assessment in Health Care, 31:4 (2015), 214–222.
© Cambridge University Press 2015
doi:10.1017/S0266462315000410

DECISION-MAKING ALIGNED WITH RAPID-CYCLE EVALUATION IN HEALTH CARE

Sebastian Schneeweiss
Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School
schneeweiss@post.harvard.edu

William H. Shrank
CVS Health, Woonsocket, Rhode Island

Michael Ruhl
Aetion Inc.

Malcolm Maclure
Department of Anesthesiology, Pharmacology and Therapeutics, University of British Columbia

Background: Availability of real-time electronic healthcare data provides new opportunities for rapid-cycle evaluation (RCE) of health technologies, including healthcare delivery and payment programs. We aim to align decision-making processes with stages of RCE to optimize the usefulness and impact of rapid results. Rational decisions about program adoption depend on program effect size in relation to externalities, including implementation cost, sustainability, and likelihood of broad adoption.
Methods: Drawing on case studies and experience from drug safety monitoring, we examine how decision makers have used scientific evidence on complex interventions in the past. We clarify how RCE alters the nature of policy decisions; develop the RAPID framework for synchronizing decision-maker activities with stages of RCE; and provide guidelines on evidence thresholds for incremental decision-making.
Results: In contrast to traditional evaluations, RCE provides early evidence on effectiveness and facilitates a stepped approach to decision making in expectation of future regularly updated evidence. RCE allows for identification of trends in adjusted effect size. It supports adapting a program in midstream in response to interim findings, or adapting the evaluation strategy to identify true improvements earlier. The 5-step RAPID approach that utilizes the cumulating evidence of program effectiveness over time could increase policy-makers' confidence in expediting decisions.
Conclusions: RCE enables a step-wise approach to HTA decision-making, based on gradually emerging evidence, reducing delays in decision-making processes after traditional one-time evaluations.

Keywords: Rapid-cycle evaluation, Decision making, Healthcare data analytics

SETTING

In a rapidly transforming healthcare environment, policy makers and providers are increasingly challenged to make timely evidence-based decisions on complex evolving issues such as payment policies, care management improvements, or delivery system implementations. Rapid cycle evaluation (RCE) has been proposed as a strategy for the Centre for Medicare and Medicaid Innovations (CMMI) to accelerate health system innovations (1;2), but the details of how to align the stages of health system decision making with the stages of emerging evidence have yet to be clarified (Figure 1a). RCE is increasingly feasible as methods have emerged for collecting, and managing massive amounts of data from multiple sources in "real-time" (3–5), which allows for sequentially repeated and increasingly automated analyses of the effectiveness of interventions and products each time the data refresh (6–11).

The traditional approach of health technology assessment (HTA)—to base initial decisions on pre-approval data and then wait several years until effectiveness testing is complete before evaluating results and making policy decisions—is not consistent with the pressures on policy makers and providers for more timely decisions. Initial HTA is about *efficacy* and often involves clinically irrelevant comparison groups (e.g., placebo or an intervention that is not an available alternative), unrealistic assumptions about routine clinical care (e.g., close to 100 percent adherence or optimal program implementation), and selected populations that do not reflect the populations that will be exposed to the new technology. HTA of *"real-world" effectiveness* requires ongoing updates of evidence as use of the technology evolves and spreads to different populations and contexts.

Using one-time HTA methods to evaluate programs or policies can be fruitful but, if programs evolve over time and across contexts, slow evaluations are often obsolescent before they are finalized. With the emergence of methods for real-time monitoring of very large cohorts of patients, the window of observation now can be so shortened that the evaluation results become available while still relevant to the current program decision makers. Then rigorous HTA methods can be exploited rapidly as long as the knowledge transfer to decision makers is efficient.

Historically, knowledge transfer from researchers to decision makers has been viewed as a final stage of a study. But low impacts of such a sequential approach led to calls for "ongoing knowledge exchange" between researchers and decision makers, preferably with researchers embedded within the

William H. Shrank is formerly of the Center for Medicare & Medicaid Innovation, Centers of Medicare and Medicaid Services.

https://doi.org/10.1017/S0266462315000410 Published online by Cambridge University Press

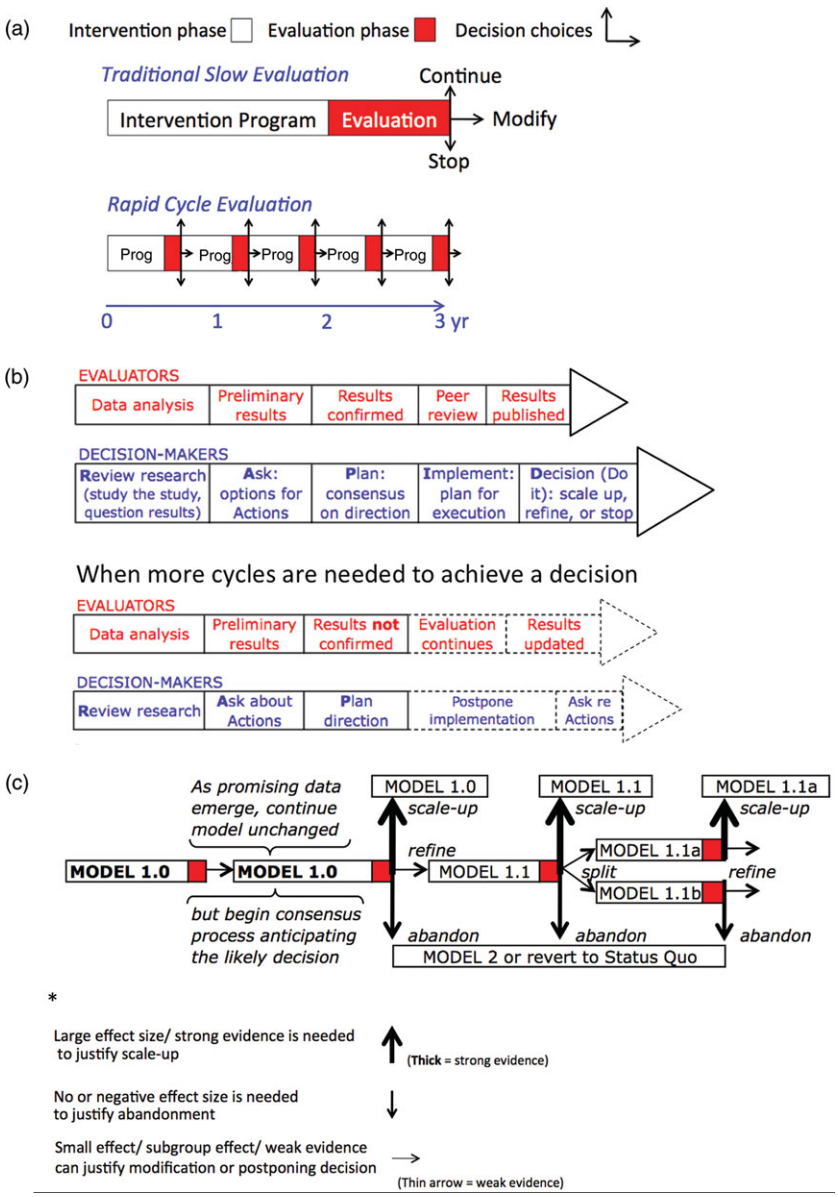


Figure 1. Rapid-cycle evaluation and HTA decision making.

decision-makers' organizations (12). As the pace of data management and analysis has accelerated to previously unimaginable speeds, the rate-limiting steps are shifting from the research process to the translation and decision processes. When windows of observation and translation are greatly shortened, the time of decision making is seen no longer as a *point*, but an *interval*. To shorten the overall time from the start of assessment to completion of the decision, we need better design of the interface between researchers and decision makers, using interim analyses in the early phases of the decision-making process.

Enacting a new style of evaluation is not without risks. Drawing *interim* conclusions more rapidly and frequently will raise concerns about too hastily making the *final* policy decisions (13). We must not sacrifice scientific rigor of the evidence for the *final* decision. However, the opportunity for frequent updates of accumulating high-validity evidence on program impacts allows the use of interim findings that are not yet statistically precise to influence the timing and speed of the early decision-making process.

In other words, RCE gives decision makers opportunities to commence processes of consensus building while data continue to accrue, when there is a suggestive trend in outcomes that points toward need for a program change. If subsequent data updates confirm the trend, the decision makers are ahead of the game (Figure 1b). On the other hand, if subsequent data updates negate the initial trend, the consensus process can be put on hold and the decision deferred until rigorous evidence of the program's impact is obtained.

HTA and program decisions typically involve multi-attribute analyses and consensus processes from multiple perspectives. Instead of starting this process at the very end of an evaluation, the process can be expedited and made more

https://doi.org/10.1017/S0266462315000410 Published online by Cambridge University Press

inclusive by starting the decision making process while RCE findings are adding evidence over time.

Therefore, to get the most out of RCE, we must understand the nature of decisions over time, and align the timing of consensus-building processes with the timing of evidence updates. Building on the "Plan-Do-Study-Act" cycle from continuous quality improvement, we assume here we are midway through the first cycle: the planning of a demonstration is complete, and the doing and studying are underway. This article focuses on aligning decisions as we move from studying to acting to a second cycle of planning and doing. We envision an Implementation Planning stage for scaling up before the final decision in a cyclic five-step process, called R.A.P.I.D. (**R**eview the Research, **A**sk about Actions, **P**lanned Direction, **I**nitiating Implementation, **D**ecision to Do It; see Table 1 and the appendix).

Having aligned the *stages* of decision-makers' roles with the evaluators' roles (Figure 1b), important questions arise about the optimal methods at each stage for linking HTA decision-makers' thinking with evaluators' thinking and evidence. In this article, we present a framework and examples to facilitate that linkage.

NATURE OF DECISIONS

A key aspect of RCE is that it permits the decision-making process and evaluation process to overlap so that decision makers can improve the alignment of evaluation to decisions and that evaluators can improve the alignment of decisions to the evidence.

Our framework for aligning decision making with RCE begins with the crux of the decision-making process. Typically policy makers face a threefold decision about direction: (i) scaling up a successful program; (ii) continuing a partially successful program with modifications while continuing to monitor its effectiveness; or (iii) terminating an unsuccessful program.

Figure 1c shows these three main directions to take when evidence from a program evaluation is presented: (i) to scale up, (ii) to refine, or (iii) to abandon. A fourth option—to continue the program unchanged and postpone choosing among the three paths—often involves a consensus process, in which case it can be considered a decision in itself.

Scaling Up and Continued Funding (Option 1)

RCE enables earlier discovery and spread of superior technologies or successful healthcare improvement programs. RCE of technologies or programs is *analogous* but not identical to periodic monitoring of randomized clinical trials of drugs or devices, so the trials can be stopped early or their designs adapted when a meaningful improvement in outcomes has been achieved earlier than predicted (superiority) or it is unlikely that any difference between therapies will be shown (futility) (14).

Because so much theory and methodology have been developed on early stopping and adaptation of clinical trials, it is tempting to assume that the same principles and methods can be transplanted to RCE of complex programs. However, unlike a drug, a program is expected to undergo structural change over time and its implementation involves people and systems whose attitudes toward change will be influenced by interim results. Frequent feedback of interim findings will support change management through cycles of improvement.

Therefore, unlike drug trials, RCE of complex programs should not have rigid stopping rules that are triggered by reaching threshold p-values or statistical tests. In contrast to the U.S. legislation for drug approval that requires only a binary decision on whether an effect of any size is observable at $p < .05$, the drug *marketing* question (Is X sufficiently better than Y that the company should scale up production?) or the drug *coverage* question (Is X sufficiently cost-effective within the context of a risk-sharing Managed Entry Agreement?), is different from the scientific question addressed by the drug trial.

Similarly, the adoption and expansion of complex healthcare improvement programs are not driven by demonstrating the existence of a minimal impact that is statistically significant at a given p-value threshold. It is instead primarily driven by the size of effects on key outcome parameters and the uncertainty of the effect estimates, which will then be put in context with externalities for decision making, including the cost of implementation, the chance of sustainability, the likelihood of broad adoption by all parts of the system, and so on. In other words, the decision makers' question is not primarily "Is the program working" but "Is it working well enough to be worth the trouble of change?" Therefore, effect size and continuous improvement are key characteristics for rational program choice.

An example of RCE is FDA's Sentinel program which focuses on identifying potential drug safety issues with newly marketed medications as drug use spreads to populations of users that were not included in the original drug trials. Sentinel uses a distributed system of eighteen longitudinal healthcare databases with an analysis system that allows pooling adjusted effect estimates. For newly marketed medications, preprogrammed analysis modules with extensive confounder adjustment procedures can be re-run once data refresh. The decisions that FDA makes with Sentinel data are different in their specifics from those of program policy decision makers but both pay a price for false positive and false negative decisions. Supplementary Table 1 summarizes differences and similarities between RCE of prescription drug safety by FDA's Sentinel program and RCE of complex programs by a program like CMS' Center for Medicare & Medicaid Innovation (CMMI).

The decision to scale up a particular healthcare program could be a massively costly decision if it is made prematurely on the basis of misleading data. It is also risky because scaling up will involve spread to clinicians who are less enthusiastic about the change and might exhibit the opposite of a "Hawthorne

https://doi.org/10.1017/S0266462315000410 Published online by Cambridge University Press

Table 1. The Five-Step R.A.P.I.D. Pyesrocess

Review the Research: During the data analysis, decision makers can pose questions about the analyses to the evaluators. For example, the evaluators might learn that the decision makers are determined to think of the problem in terms of percentage changes rather than absolute differences, so they modify their analytic methods so the results are more understandable for decision makers. Decision makers have some power to influence the evaluation, but at this stage the evaluator has maximum influence.
Ask about Actions: As the decision maker begins to grasp the emerging findings from the evaluation, and translate them into possible actions, the evaluators guide them in more directly linking the findings to the proposed actions. "No, you really cannot conclude that," says the evaluator. "But what we now know is X. So, if you want to take Action A, we should present our findings in the following manner." As the pressure to make decisions increases, the decision-maker's power increases and the evaluator begins to feel a reduction in power as the decision maker is compelled to create a simpler interpretation of results than the evaluator is comfortable with.
Planned Direction: Decision makers now face the daunting options of scaling up a promising program or abandoning a program that people have invested a lot of hope and energy into, or steering a middle course – making refinements to it and continuing to evaluate it. Many value judgments enter at this point that evaluators might have little understanding of. The power of the decision maker to make a decision based on their interpretation of the findings now exceeds the evaluator's power to influence the decisions.
Initiating Implementation: Decision makers might act as if the final decision has been made to implement scaling up. But grappling with the logistics of implementation takes time and has an impact on how people think about the problem and the findings of the evaluation. It is an opportunity for the evaluator to continue having an influence on the spin and translation of the findings into policy, although the evaluator might be dismayed at how much his or her influence on the interpretation has dwindled by this stage. Still, the evaluator can prevent decision makers from making major errors in implementation, because the evaluator has been thinking about the problem in much greater detail than the decision maker for many months. Also, if another cycle of RCE results comes in that casts doubt on the decision-maker's implementation plan, the final decision to go ahead can still be postponed.
Decision to Do It: There comes a point of no return in Implementation. It is not always obvious when that will be. One day the planner might discover the final decision to scale up has been made. "I thought we agreed that we would have one more look at the latest findings." "Well, circumstances have changed. Stakeholders have run out of patience."

effect," purposefully reducing the impact of an intervention. The potential for nonrepresentativeness of the pilot demonstration means that skeptics of the model will want the ongoing RCE to continue as the program is scaled up so their concerns can be either refuted or validated in time to avoid wasteful scale-up.

Refinement (Option 2)

On an intermediate path between scaling up and abandonment are decisions to refine the program model (Models 1.1 and 1.2 in Figure 1c). These might involve changes throughout the model (applying to all patients, all services, all products, and all settings) or changes to only part of the model. There are three major classes of decision after dividing the model into parts based on differences in success: (i) scaling up the successful part and abandoning the unsuccessful part, (ii) scaling up the successful part and refining the remaining part, and (iii) abandoning an unsuccessful part and refining the remaining part.

This underscores the importance of understanding subgroup performance, especially locations of 'positive deviants'—the rare very creative clinicians and administrators—who discover better ways of implementing the policy. The locations (subgroups) where performance has been poorer are usually very keen to know what worked elsewhere.

Additional cycles of RCE are needed to evaluate the refined models, but the evaluation need not start again at square one if the refinements are relatively small. If the changes are so large that evaluation must start all over again and prior evaluation results cannot be combined with future evaluation results, then it is clearer to refer to it as a new model (Model 2 in Figure 1c).

Abandonment (Option 3)

RCE enables pilot policies and programs to "fail fast," reducing waste and potential harm. This is particularly relevant when patients or clinicians are enthusiastic about a policy or program but the first available evidence is clearly disappointing. It is more difficult to terminate a failing program that patients and clinicians have come to expect, than to just say "no, sorry" after an early pilot study demonstrates failure.

Again, there is a difference between demonstrating failure of a drug with fixed structure and demonstrating failure of a policy or program that can be modified. While failure of a pilot will often be followed by a return to the status quo, in today's climate of unsustainable healthcare growth, there will be pressure instead to try alternative innovations (Model 2 in Figure 1c). In both cases, continuing RCE of the pilot sites might produce additional evidence for decision makers. RCE not only enables earlier detection of complete failures, it also is likely to reduce their number because proponents of the pilot of a new funding model will react to disappointing preliminary results by increasing their efforts and creative solutions to their implementation problems. In other words, proper use of RCE helps struggling models evolve rather than die.

Moreover, judgments are blurred by the inseparability of costs from the decisions. For example, to continue a payment policy or care delivery system that costs more without yet

https://doi.org/10.1017/S0266462315000410 Published online by Cambridge University Press

producing more benefit, or that has poorer results but costs less, might be rational if the policy or system is evolving in a direction expected to produce better outcomes in the future.

Postpone Decisions and Continuing Evaluation (Option 4)

Because of the above-mentioned additional complexities of evaluating real-world healthcare programs, decisions about the fate of a funding model will be made by committees whose members often disagree. As evidence accrues through periodic updates by RCE, the balance of opinion in the committee gradually shifts from "wait and see" to "time to decide." The option to postpone a decision on the program and continue evaluation can involve strong differences of opinion. Some members might believe that "No decision" is a decision to disagree or a decision to be indecisive. However, RCE makes such differences of opinion more manageable because proponents and opponents can be persuaded to wait for another cycle of evidence, hoping that the next update from the RCE will favor their standpoint. Generally, the speed and assertiveness of decision making is inversely proportional to the observed effect size. One of the most basic decisions for information overloaded policymakers and clinicians is "Should I spend time and energy on this issue?" Frequent interim results help answer this question. Preliminary worrisome findings will lead to a "Yes, I should."

Like the first three decisions, postponement is also subject to the consequences of false decision making: (i) Incorrectly continuing or scaling up a program that does not improve quality and/or increases costs is wasting resources and prohibits real progress by underfunding successful programs. (ii) Incorrectly discontinuing a program is a lost opportunity for saving cost and/or improving care.

Whether the decision is to scale up, abandon or refine a model, the opportunity and need for RCE is likely to continue. It might not be long before the number of evaluations that could be done thoroughly has grown to be more than even a well-resourced evaluation organization could handle. Then the nature of the decision making begins to include the trade-off between waiting a little longer to receive the results of an evaluation or moving it to top priority (15).

CASES: EXAMPLES OF DECISIONS IN RELATION TO EVIDENCE

If a positive impact of a program is clear and desirable, and the results generalizable to a broader population, the policy decision should be to scale up the program. An example of such a decision is the prospective payment system, in which Medicare pays hospitals a prospectively determined amount for the hospital costs of a hospitalization. This program was predated by demonstration projects in the early 1980s, where prospective payments for hospitalizations were tested (16). Substantial reductions in Medicare spending were seen, without consequent reductions in quality or adverse clinical outcomes, so the program was scaled nationally; it is widely believed to have led to massive savings for the federal government in the short term.

In other cases, the results of a program are not definitive but the value appears likely, and the program is scaled. An example of such a result in the commercial insurance business is the MI-FREEE trial (17). This randomized controlled trial, conducted by Aetna, assessed the effect of eliminating copayments for essential medications for patients who experience a myocardial infarction. While no statistically significant ($p < .05$) benefit was seen in the primary outcome—a composite cardiovascular outcome, a significant improvement was seen in the secondary outcome—a reduction in subsequent vascular events. The study was powered on clinical events, and not costs. A trend of cost reduction was seen, but it was not statistically significant. While the results appeared to be mixed, they were sufficiently compelling to Aetna to scale the program to their fully insured book of business. This is a typical example where multiple outcomes of a complex intervention were evaluated and based on a mix of effect sizes and a decision to scale up was made. Part of the consideration was likely that some of the "nonsignificant" findings would have become "statistically significant" if the pilot would have run longer or the sample size was larger. The *p*-values may be helpful quantifications of chance despite their misinterpretations but their dichotomized reports, for example $< .05$, are least helpful when assessing the added value of medical technologies and interventions (18).

Certainly, not all programs tested are successful, and there is similar pressure to identify failing programs and terminate them promptly to save precious resources. One example is a Medicare demonstration project evaluating the effect of expanded chiropractic services on total cost of care. The demonstration found no evidence of reduced spending, and suggested increased overall spending, leading Medicare to abandon that program for the status quo (19). Similarly, Medicare supported a national randomized controlled trial to test the efficacy of a clinical intervention—lung volume reduction surgery, on clinical outcomes to determine their coverage decision of the program. The trial, although statistically significant, convinced the clinical community to abandon the surgery (20). More examples of decisions are provided in Supplementary Material 1.

EVIDENCE THRESHOLDS FOR DECISION MAKING IN RCE

Setting Evidence Thresholds

Evidence thresholds are used in the first two steps of the R.A.P.I.D. decision-making process. The reason for trying to define thresholds in advance is to increase the objectivity of the Review of Research, to reduce the temptation to lower thresholds when results are tending toward decision-makers' hopes (or to raise the bar against results that conflict with decision-makers' beliefs.) Another approach is to keep the decision

https://doi.org/10.1017/S0266462315000410 Published online by Cambridge University Press

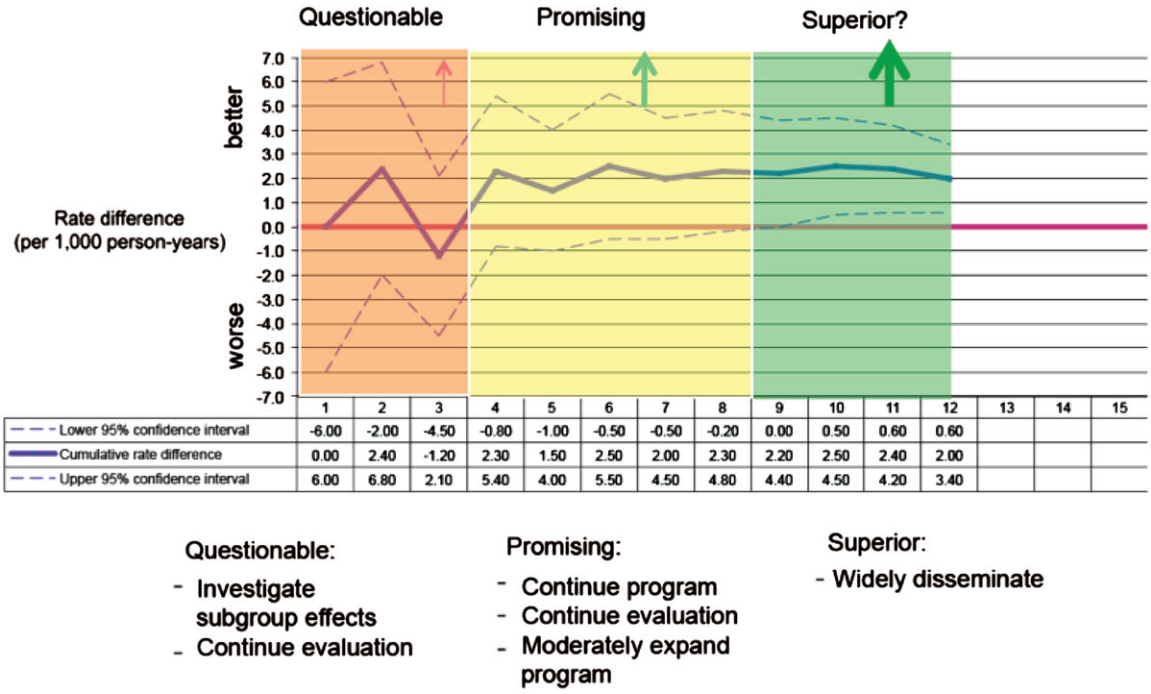


Figure 2. "Strength" of evidence is due to the effect size of the impact and increasing narrowness of the confidence intervals.

makers out of the room while the analysis is done, but total disconnection between decision makers and evaluators at this stage is likely to cause misunderstandings. Some degree of engagement of decision makers in the analysis is desirable. Thresholds defined in terms of statistical significance or non-significance reduce complex decisions to a binary aggregation and are conducive to disengagement by evaluators and decision makers, whereas thinking about the absolute size of the impact over a period time is more conducive to better decision making.

Refining Evidence Thresholds

Effect size thresholds are further used in the second stage as decision makers contemplate options for action. The translation of evidence into policy is not a simple linear process. Decision makers take into account aspects of context and are likely to propose options that are only indirectly addressed by the evidence. For example, "We observed a modest improvement with an incentive fee. What about scaling up with double (or half) the fee? Would we get twice (or half) the improvement?" The threshold for evidence concerning the impact of the incentive fee is still important for assessing the possible effect of doubling (or halving) the fee, but the discussion between the evaluator and the decision maker involves some negotiation about how far the results can be generalized to some of the options for action. Each new data point in a rapid-cycle evaluation as well as trends in effect estimates and precision over time combined with contextual knowledge (Bayesian priors) may lead to reconsidering evidence thresholds. Decision makers might ask for new evidence sources about the processes that are producing unexpected impacts or lack of impact.

Commonly the decision maker will be more ready than the evaluator to move forward with choosing among options for action. Their negotiation will often be about whether another cycle of RCE is needed, or do we have enough evidence now to begin planning scale up or abandonment of the model.

Minimum requirements often used to define success of a new program are: (i) improved effectiveness, (ii) no quality declines or cost increases.

Different quantities of effect size/ evidence are needed for different types of decisions, which is illustrated in Supplementary Material 2 using the FDA Sentinel Program and the CMS CMMI program as examples.

Figure 2 illustrates accumulating evidence categorized into phases, where the model is initially "questionable" with highly variable effect estimates, then "promising" with consistently elevated effect estimates, and finally possibly "superior" as the confidence limits around the elevated effect estimates increasingly narrow. Such a pattern might lead to a decision to scale up. Sequential value of information research can be helpful in deciding whether it is worth continuing to monitor or whether increasing precision can be assumed and a decision be made earlier (21).

Evidence Thresholds in Light of Context

The following lists selected issues in benefit/harm assessment of drugs during rapid cycle drug safety analyses of accruing

https://doi.org/10.1017/S0266462315000410 Published online by Cambridge University Press

data. They are issues whether we are doing slow traditional summative evaluation or RCE of healthcare programs. We list them here to remind ourselves that healthcare evaluation is influenced by various aspects of the context:

Aspects of context that influence benefit: (i) availability of an alternative treatment, (ii) comparative effectiveness of the new drug versus alternative drug, (iii) clinical value of benefits (lifesaving versus cosmetic), (iv) number of potential life years lost by the user population (cancer patients versus babies)

Aspects of context that influence harm: (i) frequency of adverse effects (benefit/harm analysis is based on absolute risks), (ii) seriousness of adverse events (death versus nose bleed).

In the setting of drug safety monitoring, many of these parameters are known or relatively well characterized. However in healthcare improvement programs many of the parameters listed above need to be evaluated simultaneously as no prior evidence on a new program may be available. Consequently, usually more than one parameter will be considered to determine success, for example, multiple quality parameters, cost parameters, administrative burden of implementation. If multiple parameters are considered, the decision maker needs to determine the relative value of each. Such weights will be factored with effect sizes of each outcome measure. It is in the nature of a multi-stakeholder environment that different committee members will have different weights. More often than not disagreements are due to diverse values on different outcomes, not due to disagreement about the direct evidence.

As stated earlier the validity of the effect size of several key outcome parameters and the weights of relative importance attributed to them (in terms of quality of life, quality-adjusted life-years etc.) are critical for decision makers. It is an advantage specific to a rapid-cycle evaluation system that precision is slightly less relevant because it can be extrapolated based on observed trends to some degree. Because no healthcare intervention/program is happening in a hypothesis-free space, neither *p*-values with corresponding adjustments for multiple testing that may suppress true benefits (22;23), nor shrinkage of effect estimates toward the null, are of value to decision makers, particularly when considering several endpoints about the effectiveness of complex healthcare interventions.

PUTTING THIS FRAMEWORK TO A TEST: OUTCOMES-BASED PRICING AGREEMENTS

The iterative evidence generation and decision making that RCE enables does not only provide the foundation for transforming the way healthcare is delivered, but also for changing the way it is paid for. RCE overcomes a critical technology barrier and, if transparently designed and fully shared, will be able to overcome the deep distrust between product manufacturers and payers that have hampered the transitioning from a fee-for-service model to outcomes-based payments. The essence of outcomes-based payment is that it better synchronizes the timing of payment for a drug and other medical interventions with the health outcomes and cost impact they create in the target population. The long duration of one-time impact evaluations, in conjunction with the difficulty to flexibility perform the necessary midcourse corrections to optimize the way drugs and other medical interventions are delivered, have made it almost impossible to implement outcomes-based payment systems in the past (24;25). RCE changes this situation fundamentally, by reducing the duration of typical evaluation cycles from 2–3 years to 3–6 months and enabling the early identification of effect size trends to inform the necessary midcourse corrections.

Outcome-Based Gain Sharing for Pharmaceuticals

One specific form of outcomes-based payment that is technically easy to implement with RCE is pharmaceutical gain sharing. The key principle underlying gain sharing is that a supplier of a drug and other medical products is not paid by the volume of drug/intervention at the time of delivery, but based on the incremental number of avoided cost-driving events in a target population versus a predefined baseline. Design and implementation of gain sharing is a programmatic effort and requires these key steps:

- Design of gain sharing: Supplier of a drug or other medical intervention agrees with health system/payer on.
  - How to measure outcomes, that is, the type of avoided events, to measure the benefit of the drug or medical product for health system/payer.
  - How to measure projected event rate baseline before the use of the drug or medical product in the target population (Supplementary Figure 1a).
  - How to measure financial or usage gains for the health system/payer by means of the avoided event, for example, calculation of the total cost-of-care reduction from the avoidance of a single event based on the actual cost structure of the health system/payer, or reduction in specialty care visits and shift to primary care.
  - How to estimate the incremental number of avoided events (compared with baseline) and total expected gain considering population size and risk adjustment.
  - How to design the specific gain sharing logic and payment model, that is, how exactly gains will be shared.
- Implementation of gain sharing, through iterative RCE cycles to
  - Establish early effect size trends at pilot scale (Supplementary Figure 1b), perform course corrections where possible (e.g., by focusing the drug or other medical intervention on the specific patient sub-group that benefits most), and scale up the (refined) program.
  - Make initial payments based on effect size thresholds and estimates as empirical evidence builds up and is captured in near real-time updated effect size trends (Supplementary Figure 1c).
  - Transition, over time, to payments based on precise effect size calculations in the target population (Supplementary Figure 1d).

Gain sharing is particularly powerful when embedded into a broader program for re-design of healthcare delivery in the target disease area and patient population. RCE provides the

https://doi.org/10.1017/S0266462315000410 Published online by Cambridge University Press

missing link between the two of them and is the key to achieving and sustaining affordable healthcare.

An additional test example on prospective payment for joint replacement can be found in Supplementary Material 3.

CONCLUSIONS

Rapid cycle evaluation provides unique advantages for decision making over traditional one-time evaluations. A step-wise approach to decision making makes optimal use of gradually emerging evidence. We suspect that RCE will be most successful in settings where the exposures and clinical outcomes of interest can be measured well in data sources that are continuously produced during the routine operation of a healthcare system. Those data are ideally integrated into an evidence generation environment that provides trusted analytics, which can display results to decision makers in a transparent way every time the source data refresh.

SUPPLEMENTARY MATERIAL

Supplementary Table 1
Supplementary Figure 1
Supplementary Materials 1–3
http://dx.doi.org/10.1017/S0266462315000410

CONFLICTS OF INTEREST

Dr. Schneeweiss is consultant to WHISCON, LLC and to Aetion, Inc., a software manufacturer of which he also owns equity. He is principal investigator of investigator-initiated grants to the Brigham and Women's Hospital from Novartis, Genentech, and Boehringer Ingelheim unrelated to the topic of this study. Dr. Shrank is an employee of CVS Health. Dr. Ruhl is an employee of Aetion, Inc, a software manufacturer of which he also owns equity. Dr. Maclure has no potential conflicts to report.

REFERENCES

1. Gold M, Helms D, Guterman S. Identifying, monitoring, and assessing promising innovations: Using evaluation to support rapid-cycle change. *Issue Brief.* 2011;12:1-12.
2. Shrank W. The Center For Medicare And Medicaid Innovation's blueprint for rapid-cycle evaluation of new care and payment models. *Health Aff (Millwood).* 2013;32:807-812.
3. Hartzema AG. Utilizing Medicare claims data for real-time drug safety evaluations: Is it feasible? *Pharmacoepidemiol Drug Saf.* 2011;20:684-688.
4. Platt R, Carnahan RM, Brown JS, et al. The U.S. Food and Drug Administration's Mini-Sentinel program: Status and direction. *Pharmacoepidemiol Drug Saf.* 2012;21(Suppl 1):1-8.
5. Schneeweiss S. Learning from big health care data. *N Engl J Med.* 2014;370:2161-2163.
6. Schneeweiss S, Gagne JJ, Glynn RJ, Ruhl M, Rassen JA. Assessing the comparative effectiveness of newly marketed medications: Methodological challenges and implications for drug development. *Clin Pharmacol Ther.* 2011;90:777-790.
7. Rassen JA, Schneeweiss S. Using high-dimensional propensity scores to automate confounding control in a distributed medical product safety surveillance system. *Pharmacoepidemiol Drug Saf.* 2012;21(Suppl 1):41-49.
8. Gagne JJ, Rassen JA, Choudhry NK, et al. Near-real-time monitoring of new drugs: An application comparing prasugrel versus clopidogrel. *Drug Saf.* 2014;37:151-161.
9. Gagne JJ, Wang SV, Rassen JA, Schneeweiss S. A modular, prospective, semi-automated drug safety monitoring system for use in a distributed data environment. *Pharmacoepidemiol Drug Saf.* 30 2014.
10. Gagne JJ, Bykov K, Willke RJ, et al. Treatment dynamics of newly marketed drugs and implications for comparative effectiveness research. *Value Health.* 2013;16:1054-1062.
11. Franklin JM, Rassen JA, Bartels DB, Schneeweiss S. Prospective cohort studies of newly marketed medications: Using covariate data to inform the design of large-scale studies. *Epidemiology.* 2014;25:126-133.
12. Tetroe JM, Graham ID, Foy R, et al. Health research funding agencies' support and promotion of knowledge translation: An international study. *Milbank Q.* 2008;86:125-155.
13. Avorn J, Schneeweiss S. Managing drug-risk information–what to do with all those new numbers. *N Engl J Med.* 2009;361:647-649.
14. Chin R. *Adaptive and flexible clinical trials.* Boca Raton, FL: Chapman & Hall/CRC; 2012.
15. Fleurence RL, Meltzer DO. Toward a science of research prioritization? The use of value of information by multidisciplinary stakeholder groups. *Med Decis Making.* 2013;33:460-462.
16. Greenlick MR, Lamb SJ, Carpenter TM Jr, et al. Kaiser-Permanente's Medicare Plus Project: A successful Medicare prospective payment demonstration. *Health Care Financ Rev.* 1983;4:85-97.
17. Choudhry NK, Avorn J, Glynn RJ, et al. Full coverage for preventive medications after myocardial infarction. *N Engl J Med.* 2011;365:2088-2097.
18. Greenland S, Poole C. Living with statistics in observational research. *Epidemiology.* 2013;24:73-78.
19. Whedon JM, Goertz CM, Lurie JD, Stason WB. Beyond spinal manipulation: Should Medicare expand coverage for chiropractic services? A review and commentary on the challenges for policy makers. *J Chiropr Humanit.* 2013;20:9-18.
20. Ramsey SD, Sullivan SD. Evidence, economics, and emphysema: Medicare's long journey with lung volume reduction surgery. *Health Aff (Millwood).* 2005;24:55-66.
21. Patrick AR, Franklin JM, Weinstein MC, Glynn RJ, Schneeweiss S. Sequential value-of-information assessment for prospective drug safety monitoring using claims databases: The comparative safety of prasugrel v. clopidogrel. *Med Decis Making.* 2013;33:949-960.
22. Rothman KJ. No adjustments are needed for multiple comparisons. *Epidemiology.* 1990;1:43-46.
23. Rothman KJ. Six persistent research misconceptions. *J Gen Intern Med.* 2014;29:1060-1064.
24. Garrison LP Jr, Towse A, Briggs A, et al. Performance-based risk-sharing arrangements-good practices for design, implementation, and evaluation: Report of the ISPOR good practices for performance-based risk-sharing arrangements task force. *Value Health.* 2013;16:703-719.
25. Neumann PJ, Chambers JD, Simon F, Meckley LM. Risk-sharing arrangements that link payment for drugs to health outcomes are proving hard to implement. *Health Aff (Millwood).* 2011;30:2329-2337.
26. Shadish WR, Cook TD, Campbell DT. *Experimental and quasi-experimental designs for generalized causal inference.* Belmont, CA: Wadsworth Publishing; 2001.

https://doi.org/10.1017/S0266462315000410 Published online by Cambridge University Press

Appendix. Glossary of Selected Key Terms

Complex intervention: Interventions that require an interaction between multiple technologies and parties to succeed. Many healthcare interventions are complex. The evaluation of an implantable device depends on the device but also on the surgeon implanting the device. For a healthcare delivery system to change administrators, physicians, nurses and many other stakeholders need to collaborate in a coordinated way to make a change possible.

Rapid cycle evaluation (RCE): A type of evaluation that can be rapidly repeated as soon as data generated by the program are updated (refreshed). RCE is particularly efficient for evaluations that work with data that are electronically collected during the provision of routine care, including longitudinal claims data or electronic medical records. RCE must not sacrifice scientific validity for speed. RCE typically applies sequential cohort designs[6] or time trend analyses[26] for estimating the effect sizes of relevant program outcome parameters.

Effect size: The magnitude of an effect that a program or intervention has on the outcome of interest is called effect size. It can be measured as the increase in mean spending, the relative risk of a health outcome improvement, or the difference in rate of becoming non-adherent to name three examples. The effect size is not to be confused with estimation precision (see below) or p-values, which amalgamate effect size and precision.

Estimation precision: The precision of the intervention effect estimate is usually quantified as 95% confidence intervals or standard errors. It is based on the sample size (continuous data) and the number of outcomes (categorical data) precision of the effect estimation varies. As RCE goes on over time precision usually increases as the number of program participants increase and more outcomes are observed.

Bias and bias reduction: Biases are systematic errors in the effect size estimation leading to under or over-estimation of the true effect size. Elements to reduce bias include but are not limited to baseline randomization, multivariate confounding adjustment, accurate person-time accounting, standardized outcome assessment etc.

https://doi.org/10.1017/S0266462315000410 Published online by Cambridge University Press