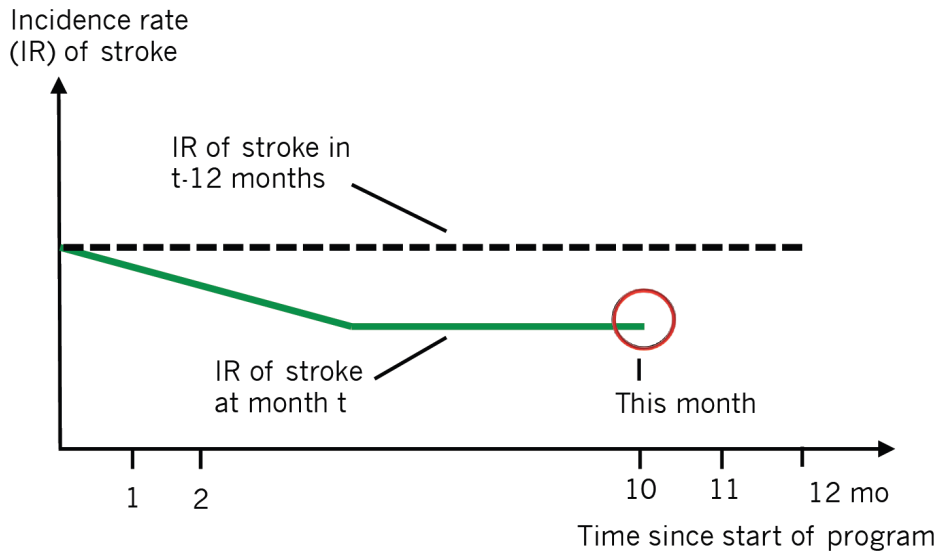
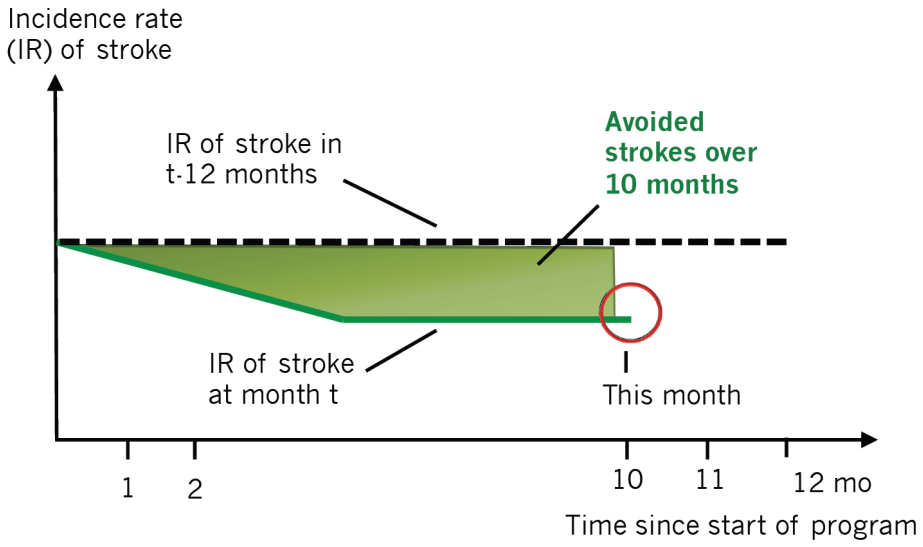
Supplementary Figure S1: Gain Sharing In Practice: Translating Evidence From Program Evaluation Into Financial Transactions

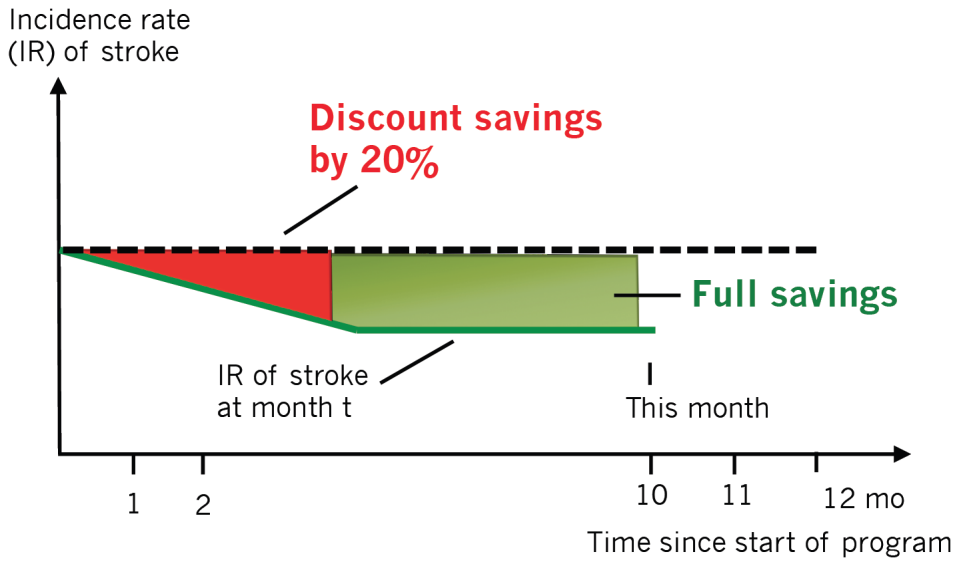
S1a) Comparison of the observed incidence rate of strokes versus the expected based on historical data with monthly data updates.



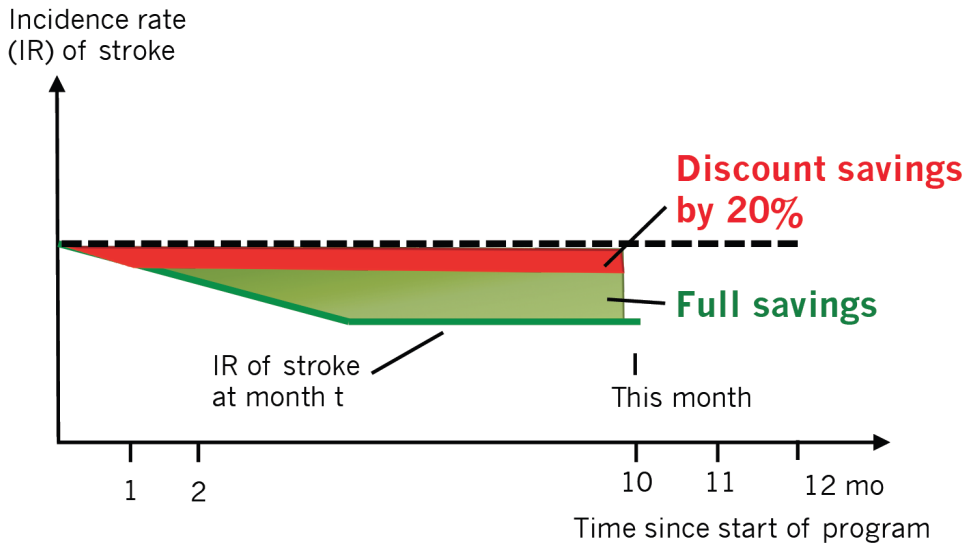
S1b) Computation of the total number of strokes prevented up to month 10.



S1c) Calculation of savings accounting for uncertainty in the statistical inference until sufficient certainty is achieved (20% discount was picked as an example not guidance).



S1d) Calculation of savings accounting for uncertainty in the statistical analysis (20% discount was picked as an example not guidance).



Supplementary Material 1

Nine Examples of Decisions in Relation to the Evidence

The following selected examples help illustrate how decisions in healthcare are made based on empirical evidence. This is not a collection of RCE examples but any type of evidence generation was considered with a focus on understanding the decision making process not the evidence generation process.

1) 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 demonstrations project in the early 1980s, such as Kaiser Permanente's in Oregon,[1] where prospective payments for hospitalizations were tested. 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.

2) Another example is the Medicare Recovery Audit Contractor program. The program was tested as a demonstration that started in 2005 in which contractors were assigned to assess the medical necessity of hospitalizations. Hospital stays that were not deemed necessary, were not reimbursed at inpatient rates, leading to a substantial increase in rates of observation status utilization. While unpopular with hospitals, the program led to massive Medicare savings without directly impacting medical care, and was scaled nationally in 2010.[2,3]

3) 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.[4] 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 ($p<0.05$) 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 "non-significant" findings would have become "statistically significant" if the pilot would have run longer or the sample size was larger.

4) 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. 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. Although the trial concluded that lung volume reduction

surgery was statistically significantly better regarding some health outcomes the magnitude of the findings and benefit-risk considerations convinced the clinical community to abandon the surgery.[5]

5) Sometimes, programs are abandoned for a different model altogether. An example is the Medicare Electronic Medical Record demonstration, which used a randomized controlled trial to test the effect of physician incentive payments on electronic medical record use and overall cost and quality of care. A national program was subsequently passed by the legislature (HITECH) that applied a different payment incentive to providers, and rendered the trial irrelevant; the trial was terminated early without producing usable results.[6]

6) Frequently, the available evidence from a test is not conclusive, and that information is used to adjust or modify the program and continue it, allowing for re-evaluation. –An example is the High-Cost Beneficiaries Demonstration which provided per beneficiary per month fees to primary care providers to better manage complex Medicare beneficiaries.[7]
The results were promising, but not overwhelming for Medicare beneficiaries, and substantial learning was available from participants who assessed their own successes and failures. This program was then modified to become the Comprehensive Primary Care initiative,[8] a multi-payer approach to providing primary care providers with larger per beneficiary per month payments to reward improved quality of care and efficiency for the entire patient populations they serve.

7) Alternatively, policymakers may choose to split up results, and scale part of a program while abandoning other parts. An example of this is the Medicare plus Choice demonstration, a test of the ability of health plans and provider organizations to manage the health of populations with risk-adjusted capitated payments for the beneficiaries they serve. The demonstration found health plans to be more credible partners than provider groups, and the Medicare Advantage program was scaled nationally in partnership with health plans alone.[9,10]

8) Similarly, policymakers may choose to scale part of a program and alter another part for continued study. One recent example is the Physician Group Practice demonstration, which tested the effect of a shared savings payment model for health systems and found mixed results.[11] When programmatic decisions were made about expansion, only interim analyses were available. Those early results informed the development of several shared savings approaches that were variations on the Physicians group Practice. The Medicare Shared Savings Program[12] a national program for Accountable Care Organizations, was implemented, and new demonstration programs were implemented to test the effect of greater risk sharing on care (the Pioneer ACO program) and the possibility of providing capital to rural providers who choose to start their own ACO (Advance Payment ACO program).

9) Another potential response for policymakers is to abandon unsuccessful portions of a program and to modify more promising parts. One recent example is the Medicare Care Coordination Demonstration.[13] The demonstration awarded to 15 applicants the opportunity to test care coordination approaches to improve care and reduce costs for Medicare beneficiaries with serious chronic diseases. Despite great enthusiasm for the program, only 1 awardee, Health Quality Partners, had promising results at the end of the demonstration period.[14] All others were terminated, but HQP was continued for 12 years, and tasked with expanding to different regions and demonstrating sustainability.

Supplementary Material 2

Different Quantities Of Effect Size/ Evidence Are Needed For Different Types Of Decisions – Illustrations Using The FDA Sentinel And CMMI Programs

Different quantities of effect size/ evidence are needed for different types of decisions because the cost of being wrong differs for scaling up versus abandonment versus continuing with refinements. There are many examples of healthcare programs scaling up without adequate evaluation, that then become almost impossible to evaluate after scaling up because they are regarded as standard care that cannot ethically be withheld from patients.[15] The opportunity cost of a false positive (in terms of lost opportunity for more rigorous evidence of program failure) can be very high; as costs of ineffective programs keep on multiplying, such expenditures cannot be used for truly effective programs.

In contrast, the lost opportunity for savings due to a false negative is not such a loss as proponents are likely to come up with an improved version of the model and test it. Ultimately the cost of a false negative might be a delay in achieving gains, whereas the cost of a false positive might be permanent losses.

A common result of RCE will be a mixture of positive and negative impacts. This is more likely to happen with multifaceted healthcare program evaluations than in single drug efficacy trials. Even if one wanted to set in advance the type 1 error and resulting tradeoff in type 2 error, getting agreement from proponents and skeptics would be unlikely. While clear thresholds are defined for decision-making on drug approval based on *efficacy* in a very narrow regulatory framework, for drug *safety* endpoints, such a framework does not exist. Safety decision-making is based on the preponderance of the entire evidence, mostly absolute effect sizes weighted by the clinical impact of harms and benefits.[16] One of the reasons for that is that decision-makers do not want to be locked into a single parameter when considering complex benefit-risk tradeoffs.

The contrast between rapid cycle evaluation and decision-making in FDA's Sentinel program and CMS's Center for Medicare and Medicaid Innovation (CMMI) is shown in **Table S1**. A key difference is that drugs evaluated in Sentinel have already passed several hurdles and are considered efficacious in at least some patients and not harmful in the short-term, otherwise they would not have been approved by FDA. Therefore, Sentinel is focused on accruing suggestive evidence of harm, not evidence of effectiveness. In contrast, a healthcare program, such as a new funding model, might not yet have passed even the first hurdle – showing some benefit– when RCE begins. Attention is focused initially on benefits but also cannot neglect evaluation of harm.

Table S1: Rapid-cycle evaluation of drug safety (Sentinel) vs. RCE of the effectiveness of delivery system innovations (CMMI) and resulting decision-making issues.

A) EVALUATION	FDA's Sentinel program	CMMI Rapid Cycle Evaluation
Intervention unit	Individual patients	Care delivery systems
Complexity of intervention	Simple: drug of interest vs. comparator	Complex, multifactorial interventions
Outcomes	Infrequent, unanticipated	More frequent, expected
Main data sources	Multiple data sources of similar structure	Single (CMS) data source?
Data lag time	Frequent asynchronous data refreshes	Near real-time claims data updates
Data availability	Medium lag time (6-12mo)	Short lag time (3 mo?)
Preferred design	Cohort study with control drug, self-controlled designs	Interrupted time trend, preferably with control group
B) DECISION MAKING		
Key decision after "alerting"	If a new drug might hurt patients, we need to do something about it (letters, REMs, withdrawal)	If a new delivery system produces superior outcomes or saves costs, it should be disseminated widely
False positive (FP)	Falsely conclude a drug may cause harm	Falsely conclude an intervention is effective
Reasons for FP	Confounding, multiple comparisons	Regression to mean, co-interventions *(i.e. confounding)*
Consequences of FP	Reduced use of a safe (and effective) drug	Dissemination of an ineffective program, waste
False negative (FN)	Falsely conclude a drug is safe	Falsely conclude an intervention is ineffective
Reasons for FN	Confounding, lack of precision	Contamination *(confounding)*, lack of precision

Supplementary Material 3

Putting This Framework To A Test: Prospective Payment Model For Joint Replacement

Setting: Prospective payment for joint replacement.

Hypothetical Model: Gain-sharing. Payment for success in avoiding readmissions for hospital-associated infections (and non-payment for readmissions.)

R. Review of Research: Decision-maker participation might result in the definition window for eligible readmissions to be changed in the analysis.

A. Ask about options for Action: Decision-maker asks whether there is enough evidence for scaling up now. Evaluator says yes for knee replacements but not for hip replacements. Decision-maker asks whether combined results for knee and hip replacement are 'statistically significant.' Evaluator argues that the size of the impact on readmissions after hip replacement is worrisome, suggesting the payment formula needs changing. The decision-maker argues that payment formulas must not be too complex or they result in too much gaming. The attractiveness of the formula for knee replacements should compensate from its problems with hip replacements. The evaluator asks whether that will result in more knee replacements and fewer hip replacements. The decision-maker returns to the question of whether another cycle of RCE is needed. The evaluator presses for it, arguing that it will increase the decision-makers options. The decision-maker then shifts to suggesting scaling up for knee replacements and modifying the formula for hip replacements, and extending the RCE for both streams. The evaluator agrees and offers to use the data to help justify a modified formula for hip replacements.

P. Plan: The evaluator repeats the analysis and presents findings so they align with the direction of scaling up the successful part of the program and modifying the remaining part of the program. The decision-maker gets consensus from stakeholders on this strategy.

I. Implementation: The decision-maker develops a revised formula for hip replacement funding guided by the evaluator's advice based on the analysis of the pilot. Issues concerning the scale-up of the funding formula for knee replacement are addressed. The option of partial scale-up is chosen with further RCE to assess whether some assumptions about future trends in infection rates will need to be changed. The logistics of implementing these changes are negotiated with stakeholders.

D. Decision: The partial scale-up of the knee replacement funding formula is approved. The revised formula for hip replacement funding is delayed, pending the evaluation of the partial scale-up of the knee replacement formula.

Supplementary References

1. Greenlick MR, Lamb SJ, Carpenter TM, Jr., Fischer TS, Marks SD, Cooper WJ. Kaiser-Permanente's Medicare Plus Project: a successful Medicare prospective payment demonstration. *Health care financing review.* Summer 1983;4(4):85-97.

2. Nicoletti B. Coming to a theater near you: Recovery Audit Contract Initiative. *The Journal of medical practice management : MPM.* Jan-Feb 2008;23(4):252-253.

3. Rosenstein AH, O'Daniel M, White S, Taylor K. Medicare's value-based payment initiatives: impact on and implications for improving physician documentation and coding. *American journal of medical quality : the official journal of the American College of Medical Quality.* May-Jun 2009;24(3):250-258.

4. Choudhry NK, Avorn J, Glynn RJ, et al. Full coverage for preventive medications after myocardial infarction. *N Engl J Med.* Dec 1 2011;365(22):2088-2097.

5. 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. *Journal of chiropractic humanities.* Dec 2013;20(1):9-18.

6. Schroeder SD. Medicare demonstrations--electronic health records demonstration. *South Dakota medicine : the journal of the South Dakota State Medical Association.* Feb 2008;61(2):67.

7. Services CfMaM. Care Management for High-Cost Beneficiaries Demonstration. 2005; http://www.cms.gov/Medicare/Demonstration-Projects/DemoProjectsEvalRpts/Medicare-Demonstrations-Items/CMS1198967.html. Accessed March 28, 2015.

8. Peikes DN, Reid RJ, Day TJ, et al. Staffing patterns of primary care practices in the comprehensive primary care initiative. *Annals of family medicine.* Mar-Apr 2014;12(2):142-149.

9. Centers for M, Medicaid Services HHS. Medicare program; establishment of the Medicare advantage program. Final rule. *Federal register.* Jan 28 2005;70(18):4587-4741.

10. Dowd BE, Feldman R, Coulam R. The effect of health plan characteristics on Medicare+ Choice enrollment. *Health services research.* Feb 2003;38(1 Pt 1):113-135.

11. Pope G, Kautter J, Leung M, Trisolini M, Adamache W, Smith K. Financial and quality impacts of the Medicare physician group practice demonstration. *Medicare & medicaid research review.* 2014;4(3).

12. Centers for M, Medicaid Services HHS. Medicare program; Medicare Shared Savings Program: Accountable Care Organizations. Final rule. *Federal register.* Nov 2 2011;76(212):67802-67990.

13. Brown RS, Peikes D, Peterson G, Schore J, Razafindrakoto CM. Six features of Medicare coordinated care demonstration programs that cut hospital admissions of high-risk patients. *Health Aff (Millwood).* Jun 2012;31(6):1156-1166.

14. Coburn KD, Marcantonio S, Lazansky R, Keller M, Davis N. Effect of a community-based nursing intervention on mortality in chronically ill older adults: a randomized controlled trial. *PLoS Med.* 2012;9(7):e1001265.

15. Montini T, Graham ID. "Entrenched practices and other biases": unpacking the historical, economic, professional, and social resistance to de-implementation. *Implementation science : IS.* Dec 2015;10(1):211.

16. Psaty BM, Charo RA. FDA responds to institute of medicine drug safety recommendations--in part. *JAMA.* May 2 2007;297(17):1917-1920.