Alternatives to the Randomized Controlled Trial in Implementation Science

Alexander S. Young, MD, MSHS
Greater Los Angeles VA HSR&D
VA VISN-22 MIRECC
UCLA Department of Psychiatry

Overview

- Randomized Controlled Trials (RCTs)
  - strengths, weaknesses, and barriers in Implementation Science
- Observational Methods
  - can they produce accurate results?
- What should you consider?
- Two examples
  - cohort study of the effect of insurance policy
  - instrumental variables analysis of effectiveness of treatment for depression

Poll Question #1

- What is your primary role in VA?
  - student, trainee, or fellow
  - clinician
  - researcher
  - manager or policy-maker
  - other

Poll Question #2

- Which best describes your research experience?
  - have not done research
  - have collaborated on research
  - have conducted research myself
  - have applied for research funding
  - have led a funded research grant

Why We Use RCTs

- Confounding factors can be associated with both exposure and outcomes
  - the association between alcohol and lung cancer is due to the confounding factor of cigarette smoking
- In RCTs, with a large sample size, confounding factors are equally present in exposed and unexposed groups
- RCTs are often blinded
  - knowledge of exposure can bias the evaluation of outcomes

What about RCTs in Implementation Research?

- Implementation usually occurs at the organizational or practice level
- Many sites are not interested in participating in implementation, research, plus random assignment
  - substantial site motivation is required
  - limiting research to highly motivated sites reduces generalizability of results
- How many variables affect intervention outcomes at the organizational level?
  - how many sites are needed to balance variables between intervention and control groups using randomization? 100? 500? 1,000?

What Observational Designs are Available?

- Many!
- Associations or correlations
- Regression analyses to control for measured confounders
  - propensity scoring
- Instrumental variables to control for <u>un</u>measured confounders
- Control group: none, historical, before-after, case-control, cohort

Results from RCTs vs. Observational Studies

- Concato and colleagues reviewed meta-analyses of RCTs and observational studies
  - 5 clinical areas, 99 articles
  - New England Journal of Medicine, 2000
- Cohort or case-control studies
  - excluded studies with historical controls
  - excluded clinical trials with non-random assignment to the intervention
- RCT & observational results were quite similar

Results from RCTs vs. Observational Studies
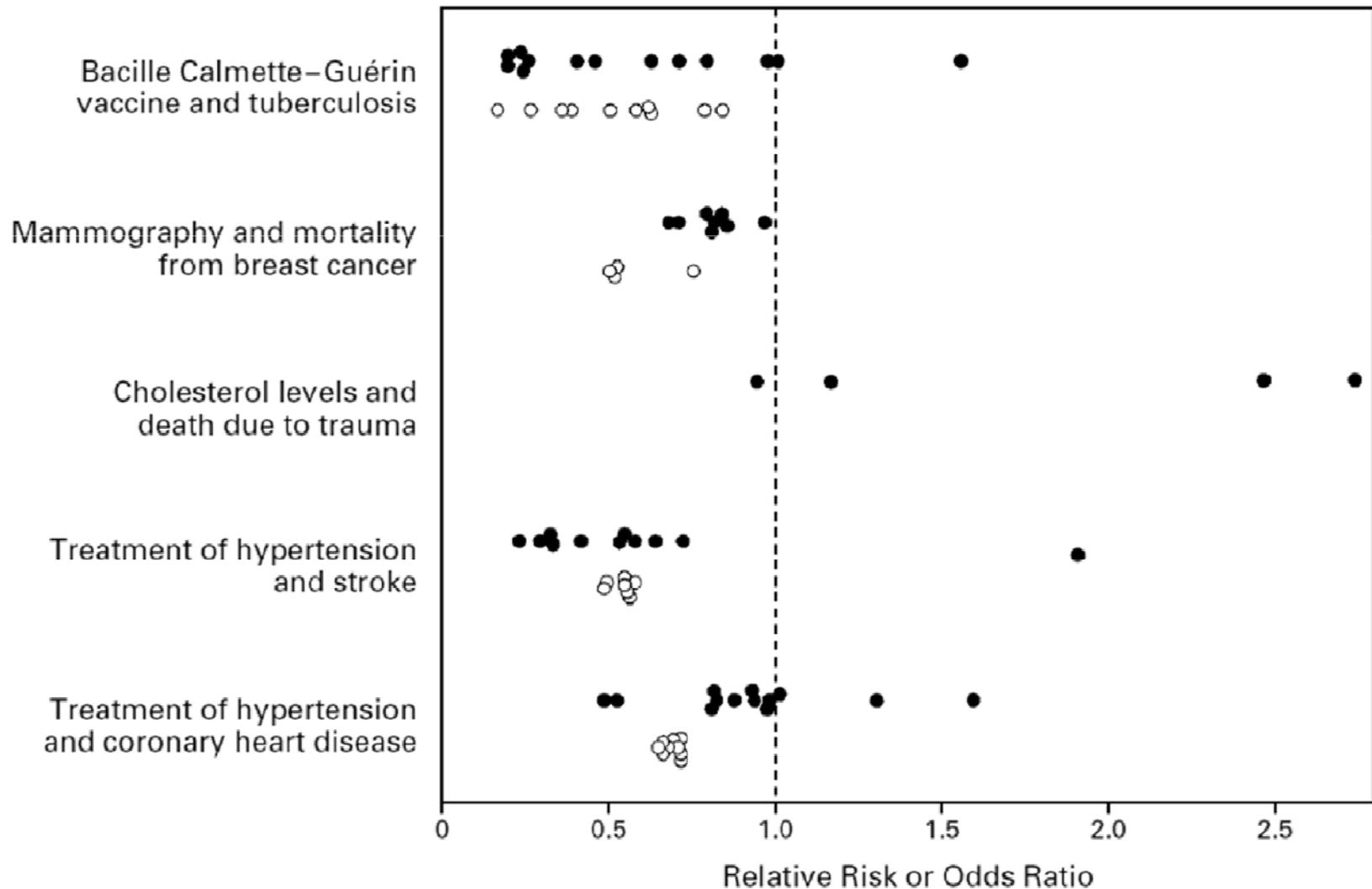
Concato, New England Journal of Medicine, 2000

Non-RCTs in Implementation Science:
Two Examples

- Cohort with matched control
  - evaluate parity of mental health insurance
  - policy implementation
  - 9 implementation plans, 9 control plans
- Instrumental variables
  - evaluate the effectiveness of depression treatment
  - quality improvement for depression in primary care
  - 938 patients

Cohort with Matched Control: Evaluation of Parity in the Federal Employees Health Benefits (FEHB) Program

Goldman et al, *New England Journal of Medicine*, 2006

FEHB Program

- Largest employer-sponsored health insurance program in the Nation
  - 8+ million beneficiaries
  - over $29 billion annually in health care benefits
- U.S. OPM administers FEHB Program
- Over 250 health plan choices

Mental Health & Substance Abuse: History of Benefits in FEHB

- Beginning in 1975
  - MH/SA coverage began to erode with ongoing diminution of benefits
- Copays, deductibles, limits
- From 1980 to 1997
  - coverage became much more limited than for other medical treatments
  - the share of total claims accounted for by MH/SA claims declined from 7.8% to 1.9%
  - interest in restoring parity, but at what cost?
- Similar in private sector insurance market

Parity Policy

- 1999: President Clinton directed OPM to institute a policy of parity improving MH/SA coverage within FEHB
  - MH/SA insurance benefit design equal to benefits for general medical services
  - equal copays, deductibles, limits
  - parity to begin January 2001
- Included funding for research evaluation
  - played major role supporting recent enactment of national MH/SA parity

Research Questions

- Did FEHB plans comply with the parity policy?

- How did the FEHB parity policy affect MH/SA benefit design and management?

- How did parity affect access to MH/SA care?

- How did the parity policy affect cost of MH/SA care to the beneficiary and insurance companies?

Research Design

- Prospective, cohort, matched control
- 9 FEHB plans selected on the basis of
  - 500 or more enrollees
  - geographic location
  - plan type and structure
  - size of enrollee population
  - plan's interest in evaluation
- Matched 9 comparison plans to account for secular trends
  - matched on location and type of plan
- Difference-in-differences analysis

Results

Research Question	Findings
Did all FEHB plans comply with the parity policy?	All plans complied.
How did parity affect MH/SA benefit design and management?	Most plans enhanced benefits and entered into managed care carve-outs.
How did the parity policy affect access to MH/SA care?	
How did parity affect cost of MH/SA care to the insurance companies and beneficiary?	

MH/SA Use and Spending: Difference-in-Difference by Plan

Plan	Difference-in-Difference in probability of MH/SA use from pre- to post-parity		Difference-in-Difference in spending on MH/SA use from pre- to post-parity	
	Estimate	Significance	Estimate	Significance
FFS-NAT	-0.12%	NS	-$68.97	p≤0.05
FFS-MA1	-0.96%	p≤0.05	-$42.13	NS
FFS-MA2	0.78%	p≤0.05	$27.11	NS
FFS-NE1	0.23%	NS	-$5.50	NS
FFS-NE2	-0.38%	NS	-$119.26	p≤0.05
FFS-W	-0.24%	NS	-$22.60	NS
FFS-S	-0.35%	NS	-$201.99	p≤0.05
HMO-W1	0.32%	NS	$32.96	NS
HMO-NE	-2.73%	p≤0.05	-$77.82	p≤0.05

Results

Research Question	Findings
Did all FEHB plans comply with the parity policy?	All plans complied.
How did parity affect MH/SA benefit design and management?	Most plans enhanced benefits and entered into managed care carve-outs.
How did the parity policy affect access to MH/SA care?	Access increased on par with secular trends.
How did parity affect cost of MH/SA care to the insurance companies and beneficiary?	

MH/SA Use and Spending: Difference-in-Differences by Plan

Plan	Difference-in-differences: probability of MH/SA use from pre- to post-parity		Difference-in-differences: estimate of MH/SA spending per user from pre- to post-parity	
	Estimate	Significance	Estimate	Significance
FFS-NAT	-0.12%	NS	-$68.97	p≤0.05
FFS-MA1	-0.96%	p≤0.05	-$42.13	NS
FFS-MA2	0.78%	p≤0.05	$27.11	NS
FFS-NE1	0.23%	NS	-$5.50	NS
FFS-NE2	-0.38%	NS	-$119.26	p≤0.05
FFS-W	-0.24%	NS	-$22.60	NS
FFS-S	-0.35%	NS	-$201.99	p≤0.05
HMO-W1	0.32%	NS	$32.96	NS
HMO-NE	-2.73%	p≤0.05	-$77.82	p≤0.05

MH/SA Out of Pocket Spending: Difference-in-Difference by Plan

Plan	Out-of-pocket spending measure	
	Estimate	Significance
FFS-NAT	$4.48	p≤0.05
FFS-MA1	-$15.43	p≤0.05
FFS-MA2	-$13.82	p≤0.05
FFS-NE1	-$8.78	NS
FFS-NE2	-$48.12	p≤0.05
FFS-W	-$49.80	p≤0.05
FFS-S	-$87.06	p≤0.05
HMO-W1	$25.16	p≤0.05
HMO-NE	$23.40	p≤0.05

Results

| Research Question	Findings
Did all FEHB plans comply with the parity policy?	All plans complied.
How did parity affect MH/SA benefit design and management?	Most plans enhanced benefits and entered into managed care carve-outs.
How did the parity policy affect access to MH/SA care?	Access increased on par with secular trends.
How did parity affect cost of MH/SA care to the insurance companies and beneficiary?	Total costs increased in line with secular trends. In most plans, out-of-pocket costs declined.

Instrumental Variables: The Effectiveness of Primary Care Depression Treatment on Patients' Clinical Status and Employment

Selection Bias

◆ Unmeasured variable associated with both treatment and outcome
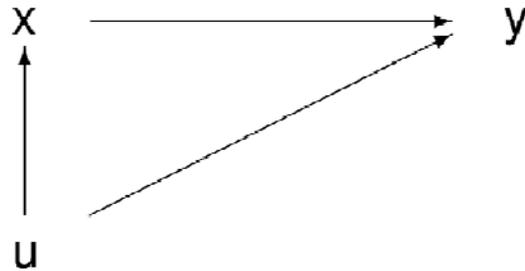
◆ Example: people who choose depression treatment have more severe depression

◆ Result: depression treatment is correlated with more severe depression

  – so, depression treatment makes depression worse?

How to Evaluate the Effectiveness of Depression Treatment?

◆ Clinical trials problem
  - depression treatment is well established and available
  - cannot ethically deny access to treatment
  - who would enroll in a randomized trial? people who are treatment-refractory, mildly ill, or doing it for the money
  - response rates in trials getting smaller and smaller

◆ Implementation science problem
  - quality improvement trial
  - organizations agree to participate
  - patients & providers choose whether or not to use treatment
  - the effectiveness of treatment cannot be directly estimated due to selection bias

Instrumental Variables

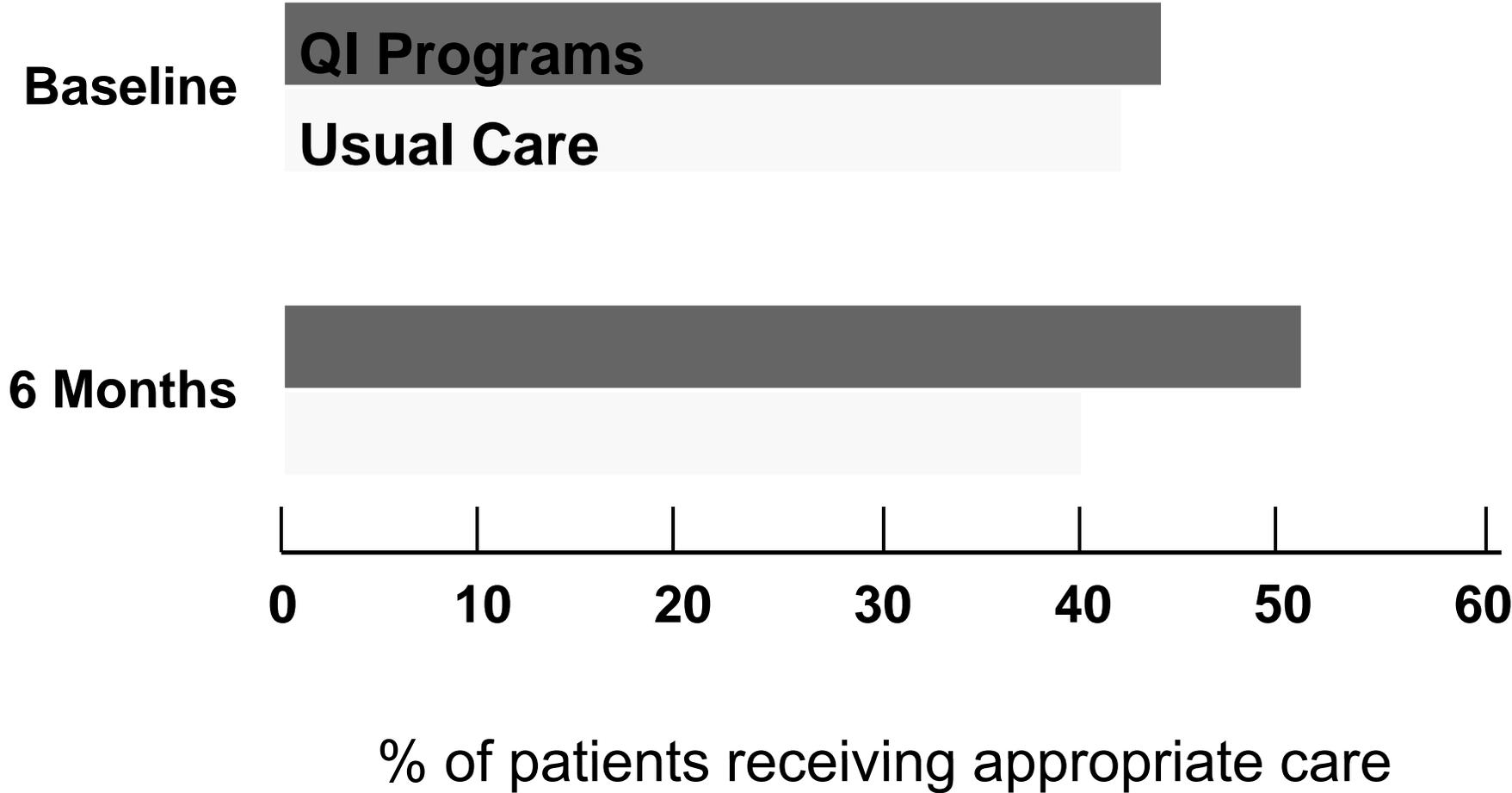◆ Estimate impact of treatment, given unmeasured selection bias
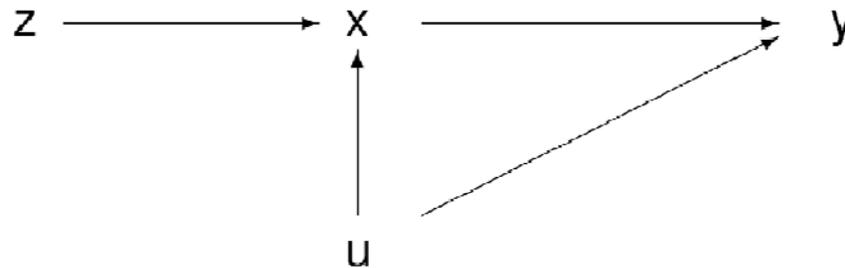
◆ Requires instrumental variable (z): affects the key independent variable, but only impacts the outcome through the key independent variable.

– geographic proximity to treatment
– assignment to quality improvement program

◆ Requires larger sample size

Partners in Care

◆ Six community-based organizations in 5 states
  – primary care clinics: 46 of 48 agree
  – 27,332 consecutive patients screened
  – 1,093 patients with depression enrolled
  – 938 patients completed 6-month follow-up survey

◆ Clinics randomized to quality improvement program or usual care
  – increase use of antidepressant medication and psychotherapy

◆ Patients assessed
  – treatment use, clinical status, employment

Quality Improvement Programs Increased Appropriate Treatment (psychotherapy or medication)

Baseline

QI Programs

Usual Care

6 Months

0   10   20   30   40   50   60

% of patients receiving appropriate care

Effectiveness of Treatment

◆ Appropriate antidepressant medication or psychotherapy

◆ Effect on depression, quality of life, employment

◆ Observational

◆ Instrumental variables
  – instrument: assignment to QI program
  – affected treatment use, but not outcome

Predicted Outcome at 6 Months By Receipt of Treatment

		(95% CI)	T-statistic*	P-value*
Global Mental HRQOL (MCS-12)				
No appropriate care	32.5	(26.2 38.9)		
Appropriate care	47.9	(39.7 56.2)	2.081	0.037
Percent with probable disorder				
No appropriate care	70.0%	(58.3% 81.7%)		
Appropriate care	23.6%	(9.5% 37.8%)	− 3.297	0.001
Percent employed				
No appropriate care	52.5%	(41.7% 63.4%)		
Appropriate care	72.2%	(65.3% 79.1%)	2.917	0.006

Alternatives to RCTs

- ◆ Weak designs
  - – no control group
  - – before - after
  - – historical controls
- ◆ Potentially strong designs
  - – restricted cohort
  - – case-control
- ◆ Must be alert for possible confounders
  - – carefully match implementation and control groups
  - – anticipate important confounders: policy shocks, differences in access to care, differences in patient SES and illness severity
- ◆ Must include robust qualitative methods characterizing the organizations and providers

Are RCTs the Gold Standard for Implementation Research?

◆ Evidence reviews often exclude any study that is not an RCT

  – Cochrane reviews

◆ You want to evaluate implementation at 4 VA's: RCT? non-RCT?

  – does randomization make any difference?

◆ Consequences

  – do it randomly: appearance of internal validity, can do less qualitative study, journal reviewers accept the design

  – match intervention and control: generalizable, need a qualitative evaluation, does not fit the clinical research paradigm, journal reviewers may question the design

Are RCTs the Gold Standard for Effectiveness Research?

◆ Treatment guidelines often exclude or minimize the importance of non-RCTs
  - patients who enroll in RCTs can be atypical: highly motivated, little comorbidity
  - care is often atypical: patients are paid to be treated, high intensity of care, expert clinicians, strong treatment fidelity

◆ Consequences
  - treatments in guidelines may not be able to be implemented into routine practice
  - treatments in guidelines may not work in routine practice
  - guidelines do not include effective treatments, since the "definitive RCT" has not been done

Conclusions

◆ RCTs can be convincing with a large number of sites

- – results may not be relevant to sites that would not have participated

◆ High quality observational methods are available

- – cohort, case-control, instrumental variables, & propensity
- – inclusion criteria, outcome measurement

◆ Both benefit from robust qualitative methods

◆ Need to consider the implementation strategy, available data, and likely confounders

References

Brown AH, Cohen AN, Chinman MJ, Kessler C, Young AS: EQUIP: Implementing chronic care principles and applying formative evaluation methods to improve care for schizophrenia: QUERI Series. *Implement Sci.* 2008; 3: 9.

Concato J, Shah N, Horwitz RI: Randomized, controlled trials, observational studies, and the hierarchy of research designs. *The New England Journal of Medicine.* 2000; 342: 1887-92.

Goldman HH, Frank RG, Burnam MA, Huskamp HA, Ridgely MS, Normand SL, Young AS, Barry CL, Azzone V, Busch AB, Azrin ST, Moran G, Lichtenstein C, Blasinsky M: Behavioral health insurance parity for federal employees. *New England Journal of Medicine.* 2006; 354: 1378-86.

Schoenbaum M, Unutzer J, McCaffrey D, Duan N, Sherbourne C, Wells KB: The effects of primary care depression treatment on patients' clinical status and employment. *Health Services Research.* 2002; 37: 1145-58.

Wells KB, Sherbourne C, Schoenbaum M, Duan N, Meredith L, Unutzer J, Miranda J, Carney MF, Rubenstein LV: Impact of disseminating quality improvement programs for depression in managed primary care: a randomized controlled trial. *JAMA.* 2000; 283: 212-20.

Since 'sample size' is a challenge in implementation and improvement science, we suggest use of the research network, the Improvement Science Research Network. www.isrn.net

Can you elaborate on robust qualitative methods? Qualitative methods can be used to characterize the organizations being studied. These data can inform understanding of implementation success, both in terms of uptake and outcomes. It is possible to triangulate using qualitative data, to understand why and how implementation worked at particular sites. This is useful for strengthening and disseminating the intervention. To help readers understand how implementation can generalize beyond the study in question. With small sample sizes, this is particularly important, since success or failure in a small sample might be related in part to variation in the organizations, variation that cannot be controlled for with randomized assignment. To understand the methods themselves, I would refer you to a fellowship or professional training program in this area, or to work with someone that specializes this work.

What would you say to a study design where there was one study site with an intervention and one control, and they randomized which got the intervention, and call that a RCT - is this design not publishable in any journal? Should they call it observational? I think the randomization in this instance does not actually accomplish the goals of randomization in RCTs. It will not balance covariates between the two sites. I think to call this an RCT is misleading, since it provides inappropriate reassurance. With an n of 2, one could be better finding a pair of matched sites. If the sites are well matched, then whether the choice is A or B, or a coin-flip for A or B seems not likely to make much difference.

What would you say to a study design where there was one study site with an intervention and one control, and they randomized which got the intervention, and call that a RCT - is this design not publishable in any journal? Should they call it observational? See response to the question immediately above

What about multiple baseline designs or permutation designs? There are very promising designs in this area. This is especially helpful if all sites are to receive the implementation. Then other sites can serve as controls. There was a very good recent QUERI Cyberseminar on a similar topic by C Hendricks Brown that I would refer you to. They focus on "Roll-Out Randomized Implementation Trials." Within this design class, I would caution you that studies with historical controls or a before-after design are thought to be easily biased or confounded.

Great presentation! I think the Concato NJM study may overstate the correlation between RCT and observational studies. I've seen a big literature on propensity scoring showing it can result in wrong-signed estimates compared to RCTs (using the same intervention and sample). Can you comment? Yes. It's important to consider what sort of observational study is being discussed. Observational can mean many things. The Concato study focused on cohort and case-control designs, with no historical controls. Propensity analyses, on the other hand, are fundamentally correlation studies, looking for associations, controlling for known confounders. The problem is that important confounders are often not known, or not well measured. In this instance, Propensity Scoring is not appropriate. I think some people do not appreciate that is as easily biased as correlations in this circumstance.

In the partners in care example, how is what you described different from an RCT randomizing at the level of the clinic? Partners in Care did assign to intervention or control at the clinic level. However, the outcome of interest was at the patient level. And depression treatment was received in both groups. Though it was more likely under the intervention. I would refer you to the Schoenbaum or Wells papers on this study.

Any comment on relative merit of regression discontinuity designs vs. RCTs? See response above

Can you suggest a design to be used for implementation of a QI project on a nursing unit with the goal(s) of the intervention to improve team performance and patient safety outcomes? I would suggest having a control group that is similar to the intervention group. So if there are temporal trends, policy changes, or unknown factors that affect outcomes over time, then this will be controlled for. So, a cohort design. Or, you could do this as a pilot study, with before-after or historical controls, knowing that this is a pilot and does not produce definitive outcome results, but can inform future work.

On the study of mh coverage in benefits plans-how did the researchers know about out of pocket costs if they could not be submitted to the insurance co? This is important since it could it could affect the outcome. Yes. Out-of-pocket costs here were measured based on co-pays, deductibles, and payment limits. So, if someone was hospitalized, how much was and was not covered. If there was no claim made at all for a service, then out-of-pocket costs would not have been detected. Though these plans all provided some mental health coverage, so one would expect claims for most services received.

Regarding the depression study, it was originally presented as an RCT of a QI approach, not as an instrumental variable study. Can you discuss why this study could be conceived of in either way, and whether there is a "right" way to consider this study? When this study was done, there was more interest in studying quality improvement than the effectiveness of services. How studies are characterized and justified, and the science itself, is often steered by what the priorities are of funding agencies.

I would like to do a study of implementation of a case manager for hospitalized patients with severe liver disease at 4 medical centers - what specifically would be the options for a cohort type study to look at this with respect to effect on re-hospitalization rates at 6 months? Or should it be patient level randomization at each site (more expensive though)? These are design choices for which pros and cons should be systematically considered and justified. There are advantages to patient-level randomization, however, future implementation is often at the clinic or organizational level, so the generalizability and future utility of patient-level results could be less. There are also challenges with cohort studies, finding good control groups, and minimizing confounders. You may also want to collaborate with an expert in study design or implementation research designs, as you consider your options.