Avoiding Randomization Failure in Program Evaluation, with Application to the Medicare Health Support Program

POPULATION HEALTH MANAGEMENT
Volume 14, Supplement 1, 2011

DOI: 10.1089/pop.2010.0074

Avoiding Randomization Failure in Program Evaluation, with Application to the Medicare Health Support Program

Gary King, PhD,[1] Richard Nielsen,[1] Carter Coberley, PhD,[2]
James E. Pope, MD,[2] and Aaron Wells, PhD[2]

Abstract

We highlight common problems in the application of random treatment assignment in large-scale program evaluation. Random assignment is the defining feature of modern experimental design, yet errors in design, implementation, and analysis often result in real-world applications not benefiting from its advantages. The errors discussed here cover the control of variability, levels of randomization, size of treatment arms, and power to detect causal effects, as well as the many problems that commonly lead to post-treatment bias. We illustrate these issues by identifying numerous serious errors in the Medicare Health Support evaluation and offering recommendations to improve the design and analysis of this and other large-scale randomized experiments. (*Population Health Management* 2011;14(suppl 1):S-11–S-22)

Introduction

Randomized experiments have revolutionized the field of program evaluation, as they offer the potential to generate validated information on what works, a path forward to create improvements, and demonstrated progress in a wide variety of programs in government, industry, and the nonprofit sector. However, many—perhaps even most—large-scale randomized experiments fail.[1] They fail because of unexpected interventions that disrupt the randomization, or its intended effects, when applied in the real world. These unexpected interventions can arise from myriad sources such as by research subjects, politicians, or others interested in affecting the assignment of people to treatment and control groups or affecting the results. They also fail when investigators do not anticipate these or other issues in the design and do not respond to them by choosing statistical methods that can correct the problems. The main advantage of experiments comes from random treatment assignment, which enables researchers to make inferences without modeling assumptions; however, failures of random treatment assignment also account for many of the problems in large-scale experiments. In this article, we describe several common problems with the design, implementation, and analysis of randomized experiments, and show how to avoid or fix them.

Throughout, we use as our example the Medicare Health Support (MHS) evaluation. A brief summary of the experiment appears in the following section. The Randomization Design Decisions section discusses ex ante choices in the assignment of subjects to treatment and control conditions. The Post-treatment Bias section considers what can go wrong in randomization and, because in the real world many problems arise, the Ex Post Adjustment section discusses how to fix a broken experiment by adjusting statistically for problems that may have arisen. The Recommendations section summarizes how to recover useful information from the MHS experimental data, despite the severe problems that occurred during its design and implementation.

The Medicare Health Support Program

The MHS program is used as an example to help illuminate the general methodological points we discuss here (for details, see https://www.cms.gov/CCIP/). MHS is an evaluation of chronic care (or disease) management programs as implemented by 8 private companies. MHS is a large and consequential experiment in its own right, the results of which will likely affect public policies with major fiscal, health, and social welfare implications. The cost of the evaluation, to either the government or the companies, totals nearly half a billion dollars. (This amount includes payments from the government to the companies, which legislation requires be returned if medical spending is not reduced by at least the cost of their services.)

We give a brief overview of the experiment here.[2] MHS was authorized by Section 721 of the Medicare Prescription

[1]Institute for Quantitative Social Science, Harvard University, Cambridge, Massachusetts.
[2]Health Research and Outcomes, Healthways, Inc. Franklin, Tennessee.

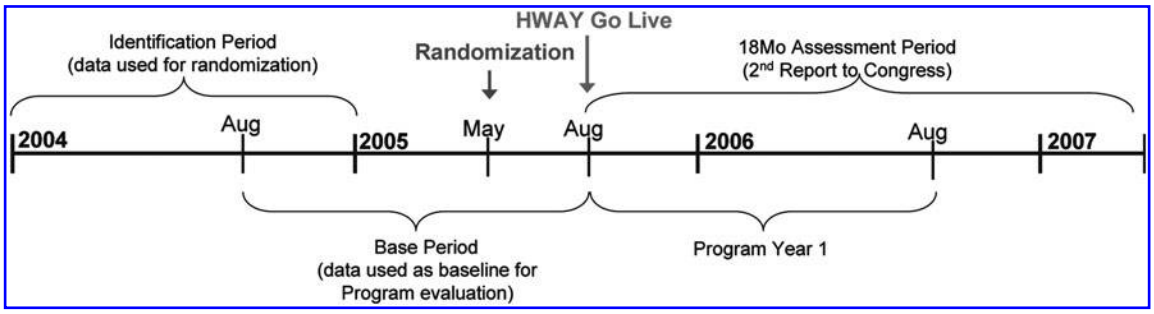

FIG. 1. Medicare Health Support program time line. Specific dates given are for Healthways; the other organizations had different time lines in similar or identical order.

Drug, Improvement, and Modernization Act of 2003 (Pub. L. 108–173). It was administered by the Centers for Medicare and Medicaid Services (CMS), which then hired a financial reconciliation contractor (the Actuarial Research Corporation) and an independent evaluation contractor (RTI International) to perform the data analyses.

The design was initially tested on a group of 240,000 Medicare beneficiaries, and a second separate intervention cohort of 47,000 beneficiaries added a year later. Because our access to data was limited to parts of the Healthways data set, we focus on the original sample of the Healthways portion of the evaluation, which included approximately 30,000 beneficiaries.

Figure 1 gives a time line for the Healthways portion of the experiment. During the calendar year 2004, experimental subjects were identified. Each Medicare beneficiary identified for participation in the experiment had to meet the *inclusion criteria,* which are that they each (1) were eligible for benefits under Medicare Part A, enrolled under Part B, and not enrolled in a plan under Part C; (2) were diagnosed with heart failure and/or diabetes; and (3) had a Hierarchical Condition Category (HCC) risk score of 1.351 or greater, which is designed to select beneficiaries with at least 35% higher estimated payments than average. To participate, beneficiaries also must have not met any of the *exclusion criteria,* which included (1) Medicare was not their primary payer; (2) they were not eligible for Medicare Part A and Part B; (3) they were enrolled in a Medicare end-stage renal disease program, hospice, Medicare Advantage plan, or CMS-sponsored Medicare fee-for-service chronic care demonstration. These rules imply that a beneficiary who dies also will be excluded. These criteria were applied continuously through the experiment so that each beneficiary could be included in the experiment, removed, or included again depending on whether he or she met these eligibility criteria at any point.

On May 11, 2005, approximately 30,000 beneficiaries meeting all inclusion criteria and none of the exclusion criteria (during calendar year 2004) were randomized to receive treatment or control. Treatment involved the Healthways chronic care management program that included a complex and contingent set of telephonic and in-person patient contacts and advice. Control involved no contact with the beneficiaries.

The specific assignment procedure involved randomization within strata defined by the cross-classification of low, medium, and high HCC scores (above the threshold for inclusion), Medicaid eligibility, and heart failure diagnosis. For each control, 2 subjects were assigned to treatment. Spouses and others living in the same household eligible for participation in the experiment were always assigned to treatment or control together (based on the random assignment of the eligible person in the household with the earliest birthday in the calendar year).

On August 1, 2005, Healthways began to contact the approximately 20,000 beneficiaries assigned to the treated group. To be in the treatment (but not control) arm of the experiment, a beneficiary was required to consent verbally to be part of the experiment via a formal, scripted procedure. Those who refused consent or who could not be reached were not removed from the study; they counted toward treatment group medical costs but did not receive treatment. At any time during the experiment, beneficiaries could withdraw consent by informing Healthways or CMS, in which case they would not receive treatment. They could also opt back into the experiment at any time by giving consent and agreeing to be contacted. In all situations, medical costs accrued during periods of treatment and no treatment were counted toward the pool of treated group costs.

Investigators were not required to (and did not) obtain consent from those assigned to the control group; control group members thus could not opt out of participation in the experiment, but they would be removed and included at any time depending on the inclusion and exclusion criteria.

Financial evaluation was conducted in several ways. The most prominently discussed (and the method included in the contract) included a difference-in-difference analysis. This involved computing the medical costs of beneficiaries in the control group minus costs for the treated group at various time points up to a final determination 3 years following randomization, minus baseline differences between the treated and control groups. CMS also allowed an "actuarial adjustment" to be applied, which added a constant amount to the estimated treatment effect to adjust for what they described as a drift in the treated and control groups away from balance after randomization but before treatment began (CMS, unpublished data, January 9, 2008).

Randomization Design Decisions

Once data are collected, statistical inference typically involves a trade-off between bias and variance. Both require imagining the same procedures being applied to hypotheti-

cal repeated samples of different sets of subjects selected in the same ways in different runs of the same experiment. *Bias* refers to deviation in an estimate from the true causal effect on average across experiments, whereas *variance* summarizes differences in the estimate of the causal effect across experiments. Having an unbiased estimate with a very large variance is of little use because in the one experiment we actually run, our estimate may be far from the truth; similarly, having an estimate with a small variance is of no use if there is large bias. As much as possible, we wish to minimize bias and variance, but given a fixed set of already collected data, different analytic strategies usually lead to a trade-off.

Yet, when the investigator creates the data and controls the assignment of subjects to treated and control groups, bias and variance need not be in conflict, as it is often possible to reduce both by carefully designing how the data are to be collected in the first place. We thus discuss fundamental decisions involved in designing the procedure by which MHS beneficiaries are assigned to the treated and control groups. These decisions fall into 4 related categories involving *control* of variability, the *level* of randomization, the relative and absolute *size* of the experimental groups, and the *power* of the experiment to detect causal effects. Each will be discussed in turn.

Control

The purpose of statistical control is to reduce bias by adjusting for potential confounding variables. For example, if the treated group is comprised of healthier beneficiaries than the control group, then health is a confounder that can bias our causal inferences—which, in this instance, could lead one to conclude that the treatment has an effect even if it does not; the reverse bias can also occur. If a researcher could measure all possible confounders, it may be possible to produce an unbiased causal estimate through statistical adjustment. However, this strategy is not always feasible. The remarkable contribution of random assignment is that it is known to be unrelated (on average across experiments) to all possible confounders, known and unknown, and so unbiasedness is guaranteed without adjusting for any other variable. In other words, properly executed random assignment eliminates all potential confounders, on average.

Unfortunately, the widespread adoption of this powerful bias-reduction procedure seems to have led to complacency with respect to variance reduction. In fact, different methods of random assignment, even if unbiased, have very different variance profiles. These choices can therefore have dramatic effects on the size of an experiment's confidence intervals around causal effects for a given sample size, or enable one to spend less on the experiment (ie, include fewer beneficiaries) to yield the same degree of confidence in the outcome.

We discuss 3 randomization designs. First is *complete randomization*, which involves a separate randomization for each subject, such as flipping a coin to decide on the treatment assignment. A better alternative procedure is that used in the MHS study, which is *randomized blocks.* Although both procedures are unbiased, randomized blocks have lower variance than complete randomization. That is, whereas complete randomization equalizes the treated and control groups on average across experiments, blocking guarantees that in the one experiment we actually run (and across the hypothetical repeats of this experiment) there is zero variability in the causal effect because of variation in the blocking variables. In the MHS study, the treated and control group data are exactly balanced with respect to the 3 variables blocked on and their cross-classifications. This means that chance differences between the 2 groups cannot exist with respect to these variables and the variance is accordingly reduced.

A final randomization procedure is *matched pairs,* whereby similar subjects are paired together using all available measured pretreatment variables and then a coin is flipped for each pair, assigning either the first to treatment and the second to control or the reverse. When exact matches are unavailable for the pairs, the closest available matches are used.[3] The matched pairs design is the logical extension of the randomized blocks procedure to blocking on all available pretreatment variables. In the same way that blocking reduces variance relative to complete randomization, matched pairs reduces variance relative to either procedure. In fact, for all observations in a stratum with the same value on each of the blocking variables, randomized blocks is equivalent to complete randomization and, as such, some variation is clearly left to chance when it could be controlled exactly. In contrast, with matched pairs, everything known and measured is matched on and thus controlled ex ante; only that which is unknown is left to be controlled by randomization.[4,5]

Thus, when feasible, one should block on all potentially important variables. Statistically, one is uniformly better off by making the blocks as small as possible so that all known pretreatment variability is controlled as well as possible. The result is that matched pairs is preferable in most situations.[6] The exceptions to this rule are not statistical but administrative. For example, in drug trials for serious but rare diseases, eligible patients may become available only sporadically, and medical personnel cannot wait until a good match is found to treat one patient. In other situations, collecting information on important pretreatment variables may be too expensive or infeasible. In still others, a strong case can be made that some measured pretreatment variables are unrelated to the outcome variable (eg, hair color), and so blocking may matter little and be more effort than its worth in terms of variance reduction.

What about MHS? The pretreatment variables that are most important to block on are those likely to be predictive of the outcome variable, which is the cost of a beneficiary's medical care. As indicated in the previous section, this experiment blocked on measures related to health, chronic conditions, access to health care, and income level, all of which may be predictive to some degree of future medical spending. However, the data collected included more information that could (and probably should) have been used to block on, such as prior baseline medical spending, age, sex, race, end-stage renal disease diagnoses, nonorganic mental psychosis diagnoses, and likelihood of death.

To illustrate the unnecessary variability (and hence statistical inefficiency) introduced into the analysis by the failure to block on available covariates, Figure 2 re-randomizes the treatment assignment 1000 times and plots a histogram of the difference in means between treated and control groups for each of 6 variables. This is accomplished in 2

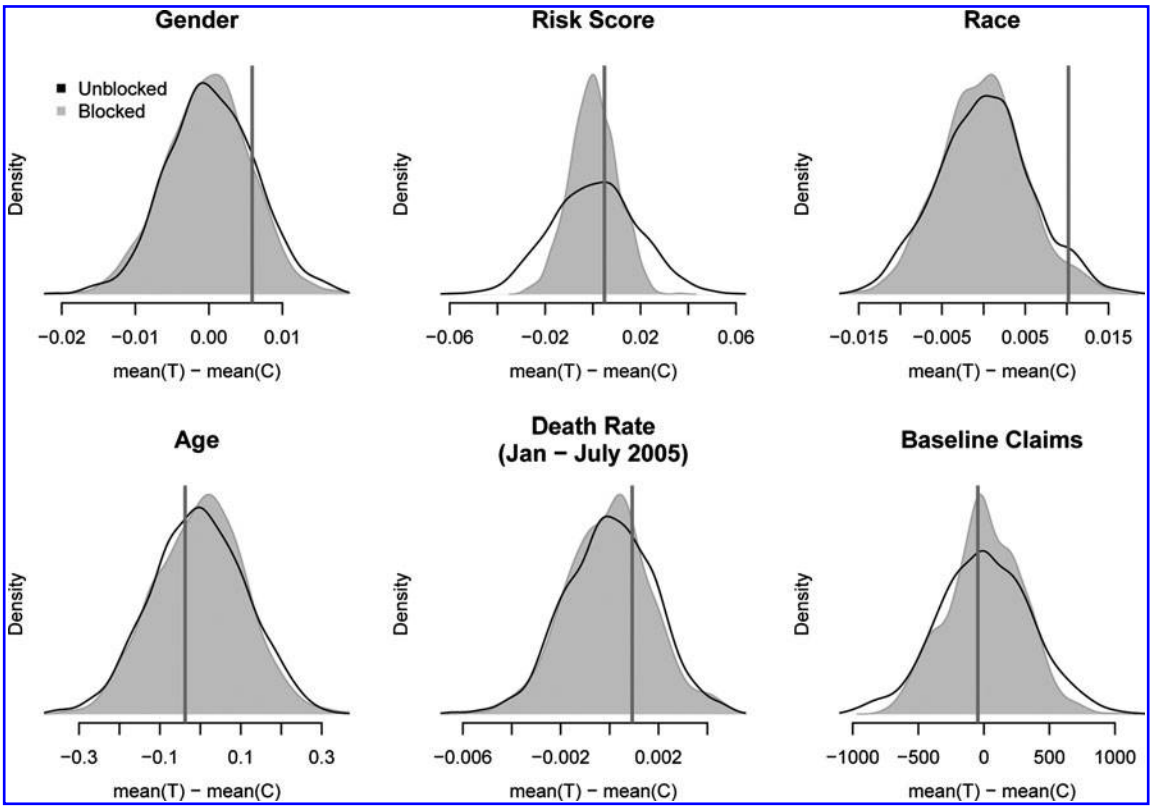

FIG. 2. Randomization distributions, for the actual blocked experiment (*gray*) and under complete randomization (*black*), with a *vertical line* representing the one actual randomization assignment. Note that the horizontal range, and thus the meaning of the distance from the vertical line from the zero point, of each variable differs.

ways, once for the actual manner in which the experiment was conducted including blocking (in gray) and once under the assumption of complete randomization (ie, without blocking [in a black line]).

For example, in the upper middle panel, we can see that the gray distribution resulting from partially blocking on the HCC risk score is narrower than what would have occurred from complete (or "unblocked") randomization. The narrowed variance that results illustrates the power of blocking on this important covariate, but it also shows that by blocking only on the coarsened 3-category version of the HCC, much variability was unnecessarily left to random chance. If instead the experiment had used matched pairs, the distribution would be a spike (or nearly a spike) over zero.

In the bottom right panel of Figure 2, we see a different result for baseline claims. This variable was not blocked on, but the experiment nevertheless produced a slightly smaller variance for it than complete randomization would. The reason for this pattern is that baseline claims are partially related to the variables that were blocked on, and so the blocking that was done proxied to a small extent for this key variable and controlled some of the variabilty.

In contrast, other variables commonly used in health care outcomes research were not blocked on, including sex, race, and age. Because these variables were also apparently not correlated with the variables blocked on, the actual distribution and complete randomization versions had approximately the same variance. These variances are substantial. For example, the age histogram in the bottom left panel has a range of ± 0.3 of a year, which is ± 3.6 months. Because mortality is approximately exponential in age, this is a substantively important difference for the elderly. (Fig. 2 studied only main effects of 6 variables. It excludes all other measured variables and the interactions among them, all of which may be worth additional study.)

In fact, although implementation began in August of 2005 (Fig. 1), randomization occurred in May of 2005 based on data through only the end of 2004. In the 7 months that elapsed between the data and implementation, many in this older and relatively sick population died. As such, this experiment would also have benefited from randomizing at the time of implementation, or equivalently blocking on death and, of course, only choosing those alive at the time of implementation. The variation in the bottom center panel of Figure 2 summarizes the consequences of inefficiency and bias resulting from the 7-month delay. The bias is portrayed in the figure by the deviation of the vertical line from the zero point. This experimental inadequacy also interacts adversely with the biased method of constructing the outcome variable, which we discuss in the next section.

The design of the MHS experiment meant that the actual balance was a random draw from the gray distributions, rather than, say, under matched pairs, which would fix them to exactly or nearly zero. The actual randomization in this experiment is represented by the vertical line in each panel. None are exactly zero, and the difference from zero indicates the clear potential for bias. The actual degree of bias induced by the vertical line in each graph not being zero is a function of this distance and the importance of each variable (or the conditional effect of a variable on the outcome). Note that the

scale of each variable in the figure, and thus the meaning of the difference between the vertical line and zero, differs.

By this measure, the most important predictor of medical spending this year is baseline medical spending, shown in the bottom right panel and which was not blocked on directly. The range for this variable in the difference in means is ± $1000 per month, which is enormous on the scale of the expected benefits of the experiment. The experimenters were lucky that the actual randomization chosen was near the center of this distribution, as the design allowed it to have been much farther with reasonably high probability. However, the actual difference in means in per beneficiary per month medical claims from perfect balance is still a considerable −45.75. The fact that the experiment was randomized means that, on average across experiments, there will be no bias because of these variables. But in the one experiment actually run, this discrepancy means that the estimated causal effect can be substantially farther from the truth solely because of the design decision not to block on spending. How much farther? The answer depends entirely on the predictive capacity of this variable on the outcome, after controlling for treatment assignment and the variables used to do the blocking.

One way to measure the predictive importance of this variable is to regress the outcome variable on baseline medical spending, controlling completely for the blocking variables and all their interactions. We do this within the treated group observations to control for treatment assignment. (We do not have access to the control units for this test.) When we run this analysis, MHS baseline claims has a coefficient of 0.221 and a t statistic of 31.247, which indicates an extremely strong pattern estimated with an unusually high degree of certainty. This result alone offers very strong evidence that baseline claims is a very important predictor of medical spending and should have been blocked on. This considerable predictive power could have easily been used to increase the power of the experiment massively at almost no additional cost to the government.

We illustrate this relationship in Figure 3, which gives 3 measures of the relationship between baseline (horizontally) and follow-up (vertically) medical spending totals. The 3 measures include a linear regression (straight solid line/straight dashed lines with confidence intervals), a semiparametric LOESS smooth line (curved line), and a nonparametric average of the outcome for each $500-width binning of baseline medical spending. All show approximately the same result: baseline medical spending is a tremendously important predictor of ultimate medical spending. In fact, because most of the observations occur toward the extreme left of the graph, we can see that the actual relationship between the 2 variables for most people is even steeper than the regression results indicate.

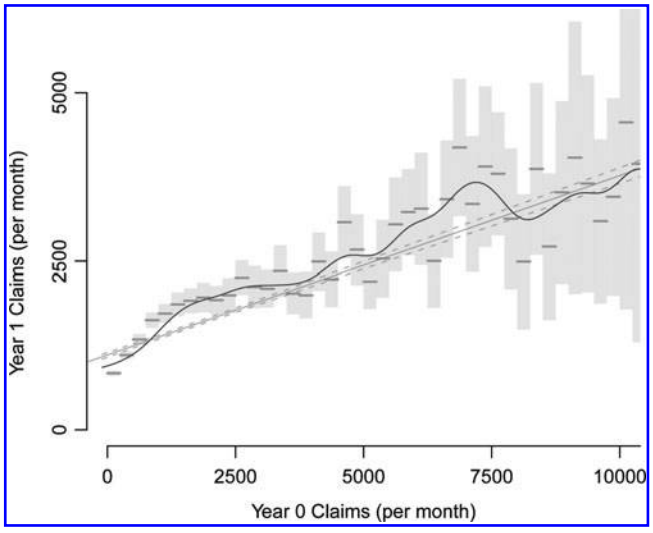

FIG. 3. The predictive capacity of prior medical spending on future medical spending.

Baseline medical spending and other variables were all available prior to the experiment but none were blocked on. This is unfortunate, because for the same cost and the same number of beneficiaries randomized, the experiment could have produced far more certain results with much narrower confidence intervals and results closer to the true answer with much higher probability. Or alternatively, it could have produced results with the same degree of uncertainty as reported at substantially lower cost to the government.

Level

A second issue is the level at which randomization is conducted. In MHS, individual beneficiaries were not randomized; instead, randomization was at the level of the household and all eligible beneficiaries living in the same household were assigned together. This is known as the difference between *unit randomization* and *cluster randomization*. Unit randomization usually has lower variance than cluster randomization. However, unit randomization is not always feasible. The MHS evaluation appears to be one such case, as giving services to one spouse but denying the other may lead to complaints, noncompliance, or lack of consent from both. And even when feasible, encouraging one spouse to take his or her medicines will plausibly have an effect on the behavior, and ultimately the medical spending, of the other spouse. This nonindependence of units then violates the "no statistical interference" condition of most statistical analysis procedures (this is the stable unit treatment value assumption[7]), and thus when present is a good reason to choose cluster randomization.

However, whatever level of randomization is chosen for an experiment, and whether the choice is made for statistical or administrative reasons, the resulting data must be analyzed using a statistical procedure designed for that level. The most common mistake is to randomize at the cluster level and analyze the data as if they were randomized at the unit level. The problem with this is that the number of independent pieces of information is the number of clusters but the computer program processing the data is incorrectly being told that the much larger number of units are all independent. Doing this can lead to significantly underestimated uncertainty estimates. As Cornfield wrote more than 3 decades ago, "Randomization by cluster accompanied by an analysis appropriate to randomization by individual is an exercise in self-deception … and should be avoided."[8] Unfortunately, this appears to be precisely the mistake made in the MHS analysis. The consequence is that the already wide confidence intervals reported in the MHS evaluation are actually somewhat larger than indicated. How much

larger? The answer depends on the number of households with clusters larger than 1 and the degree of dependence among those within households. The reported MHS confidence intervals would not be biased by this problem only if one were able to demonstrate that household members, such as spouses or sisters, have no effect on one another, which is an assumption contradicted by a considerable body of social science research. Full access to the complete data would be necessary to ascertain the correct confidence interval size, but clearly this information must be included in a proper analysis.

As a general rule, experimenters should randomize at the lowest level that is politically and administratively feasible, subject only to the constraint that interference is kept low. The statistical procedures selected must always be those appropriate to the level of randomization.

Size

The size-related design features of experiments involve the total sample size and how much of that sample is allocated to each treatment arm. Assuming that the expected variance in medical spending for any one person is exchangeable, conditional on the blocking covariates and treatment status, then the variance in the difference in means between randomly assigned treated and control groups equals the sum of the variance of each mean. The implication of this result is that the variance of the difference (ie, of the causal effect estimate) is mainly driven by the smaller of the 2 groups. (This is easy to see for a highly unbalanced experiment in which the sample size of one group is so large that the variance is essentially zero. In that situation, adding more observations to the larger group would have an imperceptible change in the variance of the difference, whereas adding the same number of observations to the other group would reduce the variance of the difference considerably.) Thus, the optimal allocation across groups is equal sizes. If exchangeability does not hold, then more complicated calculations may be necessary to produce more precise estimates.

In the MHS experiment, twice as many beneficiaries were placed in the treated group as the control group. This would only be efficient if the experimenters had reason to believe that the intervention, while attempting to reduce costs, would *double* the variance in costs. In fact, the variance might plausibly increase to some extent, because some beneficiaries might be made aware of or be encouraged to use medical services they would not have known about in the control group. Increased variance in some intervention groups is common[9]; the question is whether a 2-fold variance difference would have been expected, or at least observed after the fact. Given prior experience with chronic care management programs, a variance of twice the size seems highly implausible; if the underlying data collected were made available, it would be possible to directly estimate these quantities. The likelihood that a doubling of variance would occur is further reduced by the terms of the experiment, which, by prior agreement (between CMS and the MHS Organizations), caps total costs counted in the experiment. Thus, the 2-to-1 treatment-to-control size differential in the MHS study design may well have been a waste of resources that could have been better marshaled by equalizing the group sizes.

Power

In the specific sense of Neyman–Pearson hypothesis testing, power is the probability that a statistical test applied to data collected in a particular way will reject the null hypothesis of no causal effect, when in fact the true causal effect is zero. More generally, power refers to the level of uncertainty we can expect from a particular experimental design. In the MHS experiment, a key question is how much money a given health service intervention saves (eg, the causal effect of the Healthways intervention on medical spending); in this more general sense, "power" can refer to the width of the confidence interval around the quantity of interest.

In the Second Report to Congress on the MHS evaluation, McCall et al give a 95% confidence interval on the causal effect of $\pm$\$47 per month, which of course is an interval width of \$94.[10] This interval could then span a strong positive result to a negative one. And this difference is over and above the bias built into the experiment of $-$\$45.75 resulting from not blocking on baseline costs. To detect a savings of reasonable size in a way that is still clearly distinguishable from zero would require a much narrower confidence interval without bias; in other words, a substantially more powerful experimental design. This is a major issue, of course, since the point of the experiment was to detect this effect, but it appears that the design made detection highly unlikely whether the MHS Organizations were as effective as advertised.

Some aspects of the uncertainty of statistical inferences cannot be computed until the final data available to the federal evaluators are made publicly available, but most can be computed as a direct result of design decisions, including those discussed in this section. The largest reduction in uncertainty would have come from a large-*n*, matched pair, unit-level randomization with no interference, and equal numbers of treatment and control units. This is not possible in all experiments but, fortunately, understanding the trade-offs makes many more choices possible. For example, we can increase the sample size to accommodate the fact that randomization needed to be at the household level, or we can collect and block on pretreatment variables that have a bigger effect on the outcome in order to avoid the costs of collecting more observations. But whatever point on the various trade-offs one chooses, we must choose a design that makes it possible to draw inferences about the quantities of interest with the level of uncertainty minimized.

Post-treatment Bias

To understand the paramount issue of post-treatment bias, consider a randomized experiment in which a drug is given to one group and a placebo to the other. A reasonable procedure for estimating the causal effect of the drug would be to compute the difference in the average life span of people in the 2 groups. However, suppose instead that the investigators had estimated the causal effect by taking the difference in means between the groups based only on those who had survived through the first 2 years of the experiment. The result would be what is known as *post-treatment bias*.[11] For example, if the drug kills 20% of the people who take it within the first 2 years but causes those who survive the first 2 years to live longer than expected, dropping those who

died early would (incorrectly) reverse the conclusions of the experiment.

To avoid post-treatment bias, researchers should follow a simple principle: Keep all subjects and do not alter measures of them based upon information available following randomized treatment assignment, other than measuring the outcome variable at the end of the evaluation period. (The only reasonable exception to this rule, in some circumstances, would be matching or another adjustment used in order to reduce variance in an inefficiently designed experiment. As the cost of doing so would be giving up the model-free benefits of randomization, this decision must be very carefully made.) Any adjustment that comes after treatment could be affected by that treatment assignment, and may therefore lead to post-treatment bias. Everything that is a consequence of the treatment gets attributed to the causal effect and so researchers must be careful to include only the effect of interest. We now discuss 4 key ways this principle is violated.

Compliance

The stated goal of the investigators of the MHS experiment was to "follow an intent-to-treat pre-randomization model."[2] Intent-to-treat designs focus solely on randomization to treatment as the causal factor, and leave the degree to which each randomly assigned subject complies with the experimental protocol as part of the causal effect. For example, if the intervention itself among those who comply has a positive effect, then the more subjects comply, the larger will be the intent-to-treat effect.

In another example, if a subject is assigned to take a particular drug, then the quantity of interest in an intent-to-treat design is the causal effect of the decision to assign the patient to take the drug. In an experiment, this assignment is the result of the randomization process. (In clinical practice, which the experiment approximates, the assignment is the result of the doctor advising the patient to take the drug.) This causal effect is distinct from (and usually smaller than) the effect of the drug itself among those who comply with the experiment and take the drug only when assigned to do so.

The distinction between intent-to-treat and complier causal effects is crucial in estimation because, from the perspective of an intent-to-treat design, the decision to comply with the treatment assignment by a research subject is post-treatment, and likely a consequence of the treatment, and so may lead to bias if used to adjust the sample or measures. For example, selecting for analysis only those patients who take the drug in the treatment group and do not take the drug in the control group leads to a classic example of post-treatment bias when attempting to estimate intent-to-treat causal effects. (Studying the complier causal effect—that is, the effect among those who would accept treatment if they were assigned to the treated group and would not take treatment if assigned to the control group—requires special statistical procedures).[6]

In the MHS experiment, because of the way the analysis was conducted, beneficiaries who received an intent-to-treat random assignment did not necessarily have the opportunity to receive treatment. The data analysis procedure included and excluded beneficiaries at different points in the study, including after randomization, based on the eligibility criteria. Some of the inclusion and exclusion of subjects was induced by the beneficiaries' own decisions (such as whether to move to a hospice), some were the result of uncontrollable events (such as death), and some were caused by administrative rules, such as exclusion because of alternative plan coverage. Either way, what was intended to be an intent-to-treat design was not. And because these inclusion and exclusion rules are at least in part a consequence of treatment assignment, such as if a beneficiary learns about a hospice from the MHS Organization, then post-treatment bias results.

Consent

In the United States and most other developed countries, human subjects must give explicit legal consent before participating in an experiment. This is usually dealt with by asking a large group of people for their consent and then randomizing the assignment of only those who agree to participate. Causal inferences can be made only with respect to those who give consent, but the decision to participate, occurring prior to treatment, cannot result in post-treatment bias.

In the MHS experiment, subjects were asked for consent only after treatment assignment and only in the treated group. Subjects were included in the experiment, and their costs were calculated, whether or not they consented, even though those who did not consent received no services. By not excluding patients who did not give consent, post-treatment bias resulting from this unusual consent procedure was avoided. This does not mean, however, that bias was avoided, a subject to which we now turn.

Defining the treatment

The treatment in a randomized experiment includes whatever was assigned to the treated group and not the control group. This definition is sometimes not quite what it seems and requires careful attention in order to understand what the experiment actually measures. For example, in a drug trial, the treatment includes not only the molecular composition of the designated "active" ingredient, but also how that drug is packaged into a pill or other delivery device, how the doctor explains to the patient to take the drug, and even apparently irrelevant issues (eg, what the doctor is wearing, when the patient takes the drug).

In this light, a fundamental issue in the MHS experiment is that consent was obtained for some in the treated group and for no one in the control group, and so the consent process itself is in fact part of the definition of the causal effect. In other words, the (intent-to-treat) causal effect of this experiment includes *both* the intent to deliver MHS services *and* the process of attempting to obtain consent. This combined quantity can thus be estimated without post-treatment bias, but it is not the quantity of interest. This is a problem because the MHS organizations do not obtain consent in this way in the normal course of their business and so the combined causal effect cannot be considered relevant. Instead, the quantity of interest is only the intent to treat with MHS services. Estimating the wrong (ie, combined) quantity correctly is of course the same thing as estimating the quantity of interest with bias, which is the situation here.

In particular, the standard operating procedures of the Healthways business model has patients automatically opted in, without discussion, unless the patients raise the issue themselves and opt out. In these traditional programs, the initial telephone calls focus almost immediately on the nature of the beneficiary's chronic condition and actions they can take to change behavior and enable better self-care or doctor-patient interactions. In contrast, initial MHS calls were focused almost solely on providing members with a reason to want to opt in to the program, as well as delivering customized messages designed to engage and invite the member to be a part of the study. Requiring this opt-in process also delayed the onset of actual service delivery, which is a significant factor in a very sick population with a high mortality rate.

Moreover, the strong financial incentives provided to the chronic care management companies by CMS led to aggressive attempts to obtain consent. In practice, for Healthways, this meant designing customized messages, asking for consent on the welcome call, repeating calls if the beneficiary allowed it, recontacting the beneficiary after hospitalizations, and contacting primary care physicians and other providers to try to persuade the beneficiaries. From qualitative reports, difficulty in obtaining consent stemmed from reaching the beneficiary in the first place, overcoming objections from older people worried about fraud, and combating some inadvertently publicized misinformation about whether this was a genuinely authorized program. In contrast, the normal business practice outside of an experiment is to just start working with the beneficiary, with little or no discussion of consent, and with no immediate financial incentives involved.

To illustrate the size of this difference in the MHS data, ideally we would like to compare the estimated treatment effect of the chronic care management program separately from the treatment effect of the consent process. Of course, the randomization process makes this impossible without unverifiable assumptions, but we can see within the treated group how those who consent differ from those who do not consent. If the differences were zero, then we wouldn't have to worry about this source of bias.

For this purpose, the left panel of Figure 4 plots the standardized difference in means between those who consented and those who did not (among those in the treated group) for each of several variables. (Because beneficiaries were permitted to give or withdraw consent at any time, and even multiple times, during the experiment, this figure gives the last consent status for each person in our data set.) The vertical line marks the point of no difference, and each dot gives a difference in means; a 95% confidence interval is plotted as a horizontal line through each point. As can be seen, all but one or two of the differences are clearly distinguishable from zero. As such, we conclude that the source of bias in consenting only the treated group post-treatment is likely to be substantial.

Thus, the quantity of interest in the MHS experiment is the causal effect of the intent to assign the chronic care management program to an individual Medicare beneficiary. The purpose of the experiment is not to assess the effect on participation and medical spending of the consent process itself, although it and the quantity of interest are conflated. To estimate the quantity of interest with the data produced by the MHS experiment then involves modeling assumptions to separate the two out—some of the same assumptions that random treatment assignment is designed to avoid. The consequence of getting these assumptions wrong is post-treatment bias.

In addition, subjects who did not provide consent initially were allowed to opt in and opt out as often as they wished over time. This series of decisions are all post-treatment, all potentially a consequence of the chronic care management program, and thus another source of post-treatment bias.

Exclusion and inclusion criteria

Once subjects are randomly assigned to the treated or control group, they should remain in those groups and in the study until completion. Removing subjects after randomization risks post-treatment bias.

In MHS, the consent process removes no beneficiaries from the sample, but the inclusion and exclusion criteria are applied continually, leading to a constantly changing sample composition for both the treated and control groups. Unfortunately, each of these criteria are post-treatment and potentially a consequence of the treatment assignment. In addition, we can directly validate that the included group is not a representative sample of the entire group; we do this in the right panel of Figure 4, which gives the standardized difference in means between the eligible and ineligible (at the last time point for which they are observed in our data). Clearly, these inclusion and exclusion criteria were not remotely representative.

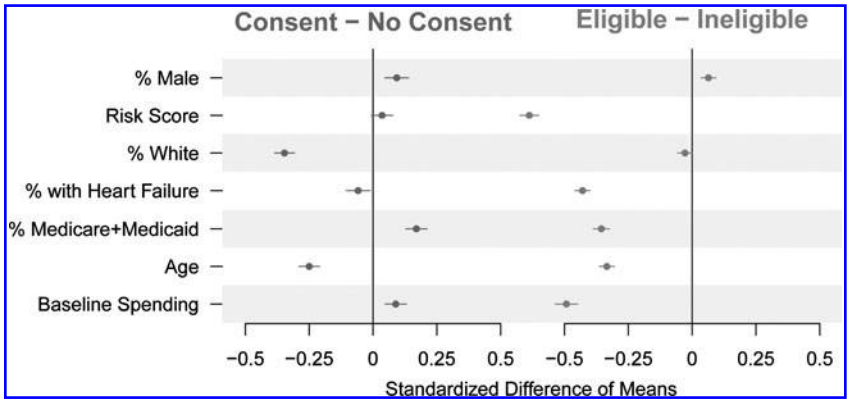

FIG. 4. Bias induced by consenting (*left panel*) and eligibility requirements (*right panel*). Standardized differences in means and 95% confidence intervals.

This procedure thus induces post-treatment bias with high probability. To take one example, suppose a beneficiary moves to a hospice because of information received from the treatment regime offered by the chronic care management company. This and all the other inclusions and exclusions are post-treatment. It may well be that these patients become ineligible for treatment but, to avoid post-treatment bias, the outcome variable needs to include them all or else a different subset of beneficiaries must be randomized in the first place.

Missing data

Information on baseline medical spending is missing from each of the data sets CMS made available to us. For example, in the most recent data set representing 36 months post-treatment, this information is missing for 571 treateds and 261 controls. These beneficiaries were in the experiment, but their medical spending information is missing from the data set. Part of the reason for this issue may be the delay in applying eligibility criteria post randomization, and part appears to be because CMS withheld medical spending information about these individuals.

This delay in implementing the biased procedure of deleting ineligible individuals also produces more potential for bias. If missing data cannot be recovered, then the problem here is deleting the missing observations listwise. Better methods of treating the missing data are available that can avoid some of these issues if missingness is unavoidable,[12] but it is almost always better to avoid such issues via ex ante design than ex post adjustment, when feasible.

Constructing the outcome variable

Adjustments made to the outcome variable in the course of the experiment induced serious post-treatment bias in 2 ways.

First, the basic measure of the outcome variable used in the MHS experiment is the average per beneficiary per month Medicare expenditure. It is calculated by computing total Medicare payments for all days in which a beneficiary meets the inclusion criteria and does not meet the exclusion criteria, divided by 30.42 to convert the days eligible into fractional months in which these conditions are met. Because these decisions are post treatment, the time periods from which the outcome variable is calculated are not even defined until after randomization. As such, the outcome variable constructed in this way will strongly induce post-treatment bias.

This first problem is considerably more serious because the adjustment for noneligibles also leads to dollar figures unrelated to actual health care costs. For example, consider 2 individuals who are eligible on the first day of a month, but died at the end of the day, and so are ineligible for the rest of the month. Suppose also that the first dies at home, incurring no medical costs, whereas the other dies in a hospital after $10,000 in costs. The value of the per beneficiary per month variable is the cost on that first day multiplied by 30, which is $0 for the first beneficiary and $300,000 for the second. This makes no sense, of course, because there is no reality under which the second patient would cost $10,000 per day for 30 days. This is especially the case as the largest costs are often incurred just prior to death, rather than repeatedly over time. And once a person has died, there are no additional costs to Medicare, and so the actual cost of this person, as distinct from the value implied by the way the outcome variable is coded, should simply be $10,000 for the entire month. The immediate biases in calculating this variable are therefore substantial on their own, but the biases then multiply because eligibility is determined post treatment.

Second, for some calculations, CMS allowed the medical cost variable to be capped so that no beneficiary's costs were higher than the top 1% of all beneficiaries. The reasoning here is presumably that sometimes reinsurance for catastrophic expenditures protects the government and the MHS Organizations. Although reinsurance for catastrophic expenditures may be a viable option in the private health care space, there is no reinsuring entity for the government. But whatever the justification, the threshold for when the cap was implemented was defined after the treatment and as such leaves the potential for additional post-treatment bias. Fixing the cap, if there is one, to an exogenous amount, chosen prior to randomization and applied to both the treated and control groups, could have been used to avoid this particular source of bias, but eliminating it entirely probably would have been preferable.

Ex Post Adjustment?

Ex post adjustment in experimental data involves any statistical modeling or other adjustment for confounders that occurs during data analysis. This is necessary in some cases and wasteful or inefficient in others.

Useful adjustments: Reducing bias or inefficiency

A carefully designed intent-to-treat experiment can be analyzed via a simple difference in means between the treated and control groups, and will not benefit from ex post statistical adjustment. For example, a unit-randomized matched pair design will not benefit from adjusting on variables that defined the pairs because the groups are already well balanced in the sample.

When a proper experimental design is not used or is used but not implemented correctly, a good scientific strategy is to repeat the experiment; after all, correcting statistical analyses via design decisions that enable one to avoid assumptions is usually preferable to ex post statistical corrections that make assumptions. Of course in the real world, and especially with large-scale evaluations, this strategy is typically infeasible or undesirable. In that case the randomization cannot then be relied upon to avoid assumptions entirely, and so the data from the broken experiment must be regarded as effectively observational. In these situations, 2 types of adjustments may be relevant, both of which apply to the MHS experiment.

First, and most obviously, statistical adjustments should be applied if the analysis without them will lead to bias. In this situation, the advantages of randomization are already lost and the researcher might as well try to recover some information from the data in whatever condition they are in. In modern data analyses, matching methods are typically preferable in these situations.[5,13]

The second condition involves choosing to adjust even if there is no bias, such as for the purpose of reducing the variance. This is a more difficult decision because we would be intentionally giving up the assumption-free benefits of randomization, and thus risking bias, in return for the

possibility of a larger variance reduction. For example, a researcher may wish to control for variables they neglected to block on ex ante, such as via matching or regression. This decision is usually only justified if the confidence intervals are not narrow enough for substantive purposes, as may be the case with the MHS experiment.

Unnecessary or harmful adjustments

Our analysis in Figure 2 shows that the MHS experiment appears to have been randomized as described, at least with respect to the variables we had available to evaluate. It was far from the most efficient method of randomizing, but it was nevertheless valid. As such, there was no immediate need for adjustment after randomization. However, the analysis methods chosen were flawed: As described above, the method of computing the outcome variable, the procedure for consenting subjects, the method of dealing with missing data, and the protocol for including and excluding beneficiaries after the experiment began all likely induced post-treatment or other biases. The point of the ex post adjustments, then, was only to reduce the bias unnecessarily introduced by these flawed analysis procedures. Drop these procedures, and one may be able to drop the adjustments, run an appropriate statistical analysis, and recover some of the model-free benefits of randomization.

We describe here 3 of the ex post adjustments used in analyzing the MHS experiment.

First, the analysts used a difference-in-difference procedure to estimate the causal effect.[10] This strategy involves subtracting from the usual estimate (based on the difference in means between the treated and control groups after an intervention period) the difference in means at baseline. This is a reasonable strategy in many observational contexts, or in experimental designs in which certain types of measurement error may be present,[14] but it is highly inefficient and thus ill advised when randomization succeeds and the variables are measured well.

To see the problem with the difference-in-difference analysis in a randomized experiment like MHS, let $\overline{Y}_T$ and $\overline{Y}_C$ denote the average values of the outcome variable for the treated and control groups, respectively, at follow-up. Then the simple estimate of the causal effect is the difference in means: $d = \overline{Y}_T - \overline{Y}_C$. Let $\overline{Y}_T^\circ$ and $\overline{Y}_C^\circ$ denote the average values of the treated and control groups at baseline. The difference-in-difference estimator is based on the fact that the true causal effect at baseline (before the program is implemented) is known to be zero if no problems occurred. To check for problems, we merely estimate the bias directly by computing the difference in means at baseline: $B = \overline{Y}_T^\circ - \overline{Y}_C^\circ$. If B is systematically different from zero, and we assume that d also has this same degree of bias, then we can simply subtract out the estimated bias at baseline $D = d - B$, which gives us the difference-in-difference estimator.

However, consider what happens in a randomized experiment (with no confounding factors or other biases). Begin by denoting the true causal effect of the program as θ. In this situation, the simple difference in means estimator is unbiased, $E(d) = \theta$ (meaning that on average, across repeated randomizations, the expected value of the estimator d gives us θ), and the baseline correction is zero on average across experiments, $E(B) = 0$. This means that the correction RTI applied will do nothing on average across repeated runs of the same experiment: $E(d) = E(D) = \theta$. However, the variance of a difference-in-difference estimator across experiments is larger, $V(D) = V(d) + V(B) > V(d)$, usually about twice as large. A larger variance with no bias means that in the one run of the experiment actually conducted, d is likely to be closer to the true θ than B. This can be seen if standard errors or confidence intervals are computed correctly because both will be a good deal larger than if the simple difference in means had been used. Put differently, RTI's choice of adjusting via the difference-in-difference estimator was equivalent to discarding about half of the observations collected and research funds used.

However, the problem is worse because RTI did not perform a classic difference-in-difference analysis. Instead, they chose to adjust further by weighting the difference in spending between baseline and follow-up by the beneficiary's fraction of eligible days during the evaluation period.[10] Because eligibility is a post-treatment decision, this adjustment would likely induce further post-treatment bias, which in turn would be further magnified if the rate and timing of ineligibility are not equivalent between the control and treated groups.

The second CMS ex post analytic change to the simple difference in means involved an "actuarial adjustment" that attempted to counterbalance drift in the treatment and control groups after randomization but prior to the start date. CMS identified the cause of the drift as "due to the analytic approach used by RTI in which length of eligibility is factored into" the per beneficiary per month calculation of the outcome variable (CMS, unpublished data, January 9, 2008). Again, dropping this analytic approach would obviate the need for adjustment. Making the adjustment, then, will have an indeterminate effect on bias and efficiency.

There also is no reason to think that an adjustment, even if it were necessary, should be constant over the entire sample. In all likelihood, it would add other biases. For example, it is highly likely that beneficiaries with the highest medical spending have the largest causal effects and so would require the largest adjustments. Ignoring this basic feature of the data and shoehorning all effects into a constant adjustment strategy is not advisable and is unlikely to solve the problem.

A final ex post adjustment method was a multivariate statistical adjustment used by RTI in the Second Report to Congress.[10] A multivariate linear regression was used to statistically adjust for baseline covariates. (The specific regression specification was not given and so is not replicable. This adjustment also apparently included some type of "regression-to-the-mean" adjustment, but exactly what that is also was not specified.) This adjustment had large effects results, with the program administered by Healthways costing the government $26 per beneficiary per month to saving it $18. The fact that this adjustment made such a large difference is a direct indication of a serious problem with the experiment. Assignment rules generated randomly are, by design, unrelated to potential pretreatment confounders and all other variables. Controlling for variables like these, unrelated to the treatment variable, should have no effect on the outcome. What likely happened is that the inappropriate post-treatment adjustments introduced bias and unnecessarily induced a relationship between the treatment and these control variables.

Recommendations for MHS Data Analysis

So far as it is possible to tell without access to the complete data and better reporting on what was done, the MHS experiment was designed highly inefficiently with several serious biases. For the same expenditure and the same number of beneficiaries randomized, simple changes to the design could have produced considerably narrower confidence intervals and far more informative and less biased conclusions about MHS services. These design inefficiencies were then compounded by bias resulting from the changing consent process and post-treatment biases induced by a series of analytical mistakes. The biases were then confounded further by flawed adjustment procedures that were, in effect, designed to correct for the biases that need not have been induced in the first place.

A new experiment designed from scratch is attractive from a scientific perspective, but with the ever-growing challenge in the quality and affordability of health care, we probably do not have the luxury of either the cost or time of repeating the experiment. Fortunately, a great deal of information does exist in the data from this experiment. Thus, even if rerunning this large-scale, long-term, and expensive experiment were feasible, learning as much as possible from the data already produced would be prudent. Thus, we now offer 5 recommendations designed to facilitate this process.

First, the outcome variable should be constructed to measure total costs, without bias-inducing post-treatment adjustments. Dollar figures should not be weighted up to monthly values that do not and cannot exist. When someone dies or does not receive services, the costs are zero; adjustments, and especially adjustments that are calculated with information available post treatment, should be avoided. Post-treatment bias can be extremely difficult to correct once induced,[11] and so it is best to avoid the situation in the first place.

Second, data sets should be constructed with all variables measured at the time of randomization and then again at the "go live" point when treatment began to be implemented. These data sets should then be subject to randomization tests like the ones we implemented in Figure 2, but for all available variables (and their interactions). Depending on the results of this step, different steps would be taken next. For example, if the randomization tests suggest that the experiment was implemented as designed, it would be best to analyze the data without the eligibility and other features that induced post-treatment bias and without the 3 ex post adjustments designed to correct them.

Third, even if no bias is found in the many other detailed procedures of design, implementation, administration, and analysis, some assumptions need to be made in order to correct for biases related to the consent process differing from how the companies do this in their normal business practices.

Fourth, an important decision needs to be made regarding the inefficiency of the experiment because of a failure to block on the key background covariates. A principal advantage of random treatment assignment is that, without modeling assumptions, one can assure unbiased causal effect estimates, but one can control statistically ex post for some of the variables that should have been blocked on ex ante. Doing this gives up the model-free aspect of the experiment in return for considerably more efficiency. Therefore, if the confidence intervals are still too wide, we also recommend a modern matching and, if necessary, modeling procedure be applied.

Finally, real-world experiments involve a large number of detailed components, many of which can individually cause randomization to fail or can induce bias or inefficiency in casual effect estimation. Such failures are especially common in large-scale evaluations, which tend to be more complicated to design, implement, and analyze. In this light, the difficulties we illustrate here with the MHS evaluation are not exceptional, even though they appear to be serious. To ensure that research conclusions are valid, however, the details of evaluations and the resulting data must be made public. Science does not merely involve acting scientifically and following methodological advice, such as that offered here; it also involves a community of scholars and researchers competing and cooperating in the pursuit of common goals. Only with access to the same information can that community form and check each other's work, and only then can we all benefit from building on each other's research.

In this light, the MHS experiment and other evaluations ought to follow the emerging replication movement and the replication standard now spreading across many fields of science: "The replication standard holds that sufficient information exists with which to understand, evaluate, and build upon a prior work if a third party can replicate the results without any additional information from the author."[15] Replication in this context means beginning with the data and being able to reproduce the numerical results in the tables and figures that support the conclusions of the study. (Of course, one may also wish to replicate the entire experiment from scratch, but that is a separate matter that still would benefit from meeting the replication standard discussed here.) To meet this standard, the data and all the information necessary to follow the whole chain of evidence from the population of research subjects to the specific numerical conclusions must be made available to the research community. Like much social science data, privacy laws prevent all of this from merely being posted on the Web, but there now exist well-developed and legally vetted procedures that make it possible to properly share all information among researchers. Some information about the MHS experiment has been shared, but the rest also needs to be made available before researchers and the American public can begin to benefit from this large-scale, extraordinary experimental evaluation.[16]

A key methodological conclusion of this review is that the small details in large-scale experiments can matter enormously. As such, other corrections may be necessary in the MHS experiment, aside from those we have identified here, once the research community learns about the remaining features of this experiment and has access to the data.

Concluding Remarks

An attempt to randomly assign treatment in a formal experimental design does not imply that reasonable inferences will be drawn from the data that result. To ensure that possibility, an experimental design must provide efficient use of available resources in a manner robust to problems that may arise. The implementation and administration must go as planned, and the analysis of the data that results must not induce biases and inefficiencies afterward. Only by ensuring each of these will we be able to reap the benefits of random treatment assignment.

Proper design, implementation, administration, and analysis are much more difficult to ensure in large-scale evaluation experiments because so many more possibilities exist for problems to arise. However, although the risks are higher, the potential rewards are significantly greater, too. For centuries, large governmental and other programs have been implemented without the benefits of modern experimental design. The increasing prevalence of large-scale randomized evaluations promises to bring science to bear on improving numerous types of large-scale programs.

The dynamic change and growth in programs that deliver and manage health care in the United States comprise a crucial area for improvement. US health care spending, even as a percent of gross domestic product, has soared over the last 4 decades, with Medicare comprising a large proportion of the increases. The vast majority of the increases in Medicare spending are attributable to chronic diseases and their management and treatment.[17] In this light, it is no surprise that the MHS project ranks as one of the largest and most expensive randomized program evaluations to date. Ensuring that appropriate conclusions are drawn, and that they are drawn as soon as possible, from the extensive data generated by the experiment is essential.

Acknowledgment

Thanks to Healthways and the Institute for Quantitative Social Science at Harvard University for research support, and the National Science Foundation for a Graduate Research Fellowship to Mr Nielsen.

Author Disclosure Statement

The authors received funding from Healthways, Inc., one of the subjects of the MHS experiment. Drs. Coberley, Pope, and Wells are current employees of and shareholders in Healthways, Inc.

References

1. King G, Gakidou E, Ravishankar N, et al. A "politically robust" experimental design for public policy evaluation, with application to the Mexican universal health insurance program. *J Policy Anal Manag*. 2007;26:479–506.
2. Cromwell J, McCall N, Burton J. Evaluation of Medicare health support chronic disease pilot program. *Health Care Financ Rev*. 2008;30:47–60.
3. Iacus SM, King G, Porro G. CEM: Coarsened exact matching software. Available at: http://gking.harvard.edu/cem. Accessed December 17, 2010.
4. Box GEP, Hunter WG, Hunter JS. *Statistics for Experimenters*. New York: Wiley-Interscience; 1978.
5. Ho D, Imai K, King G, Stuart E. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Polit Anal*. 2007;15:199–236.
6. Imai K, King G, Nall C. The essential role of pair matching in cluster-randomized experiments, with application to the Mexican universal health insurance evaluation. *Statist Sci*. 2009;24:29–53.
7. Rubin DB. Bias reduction using mahalanobis-metric matching. *Biometrics*. 1980;36:293–298.
8. Cornfield J. Randomization by group: A formal analysis. *Am J Epidemiol*. 1978;108:100–102.
9. Gelman A. Treatment effects in before-after data. In: Gelman A, Meng X-L, eds. *Applied Bayesian Modeling and Causal Inference from an Incomplete Data Perspective*. London: Wiley; 2004:195–202.
10. McCall N, Cromwell J, Urato C, Rabiner D. Evaluation of phase I of the Medicare health support pilot program under traditional fee-for-service Medicare: 18-month interim analysis. Available at: http://www.cms.gov/reports/downloads/MHS_Second_Report_to_Congress_October_2008.pdf. Accessed December 19, 2010.
11. King G, Zeng L. The dangers of extreme counterfactuals. *Polit Anal*. 2006;14:131–159.
12. King G, Honaker J, Joseph A, Scheve K. Analyzing incomplete political science data: An alternative algorithm for multiple imputation. *Am Polit Sci Rev*. 2001;95:49–69.
13. Iacus SM, King G, Porro G. Causal inference without balance checking: Coarsened exact matching. Available at: http://gking.harvard.edu/files/abs/cem-plus-abs.shtml.
14. King G, Gakidou E, Imai K, et al. Public policy for the poor? A randomised assessment of the Mexican universal health insurance programme. *Lancet*. 2009;373:1447–1454.
15. King G. Replication, replication. *Polit Sci Pol*. 1995;28:443–499.
16. Foote SM. Next steps: How can Medicare accelerate the pace of improving chronic care? *Health Affairs*. 2009;28:99–102.
17. Thorpe KE, Ogden LL, Galactionova K. Chronic conditions account for rise in Medicare spending from 1987 to 2006. *Health Affairs*. 2010;29:718–724.

Address correspondence to:
Gary King, Ph.D.
Albert J. Weatherhead III University Professor
Institute for Quantitative Social Science
Harvard University
1737 Cambridge Street
Cambridge MA 02138

E-mail: king@harvard.edu