Journal of Evaluation in Clinical Practice
International Journal of Public Health Policy and Health Services Research

Journal of Evaluation in Clinical Practice ISSN 1365-2753

Identifying spin in health management evaluations

Ariel Linden DrPH MS

President, Linden Consulting Group, LLC, Hillsboro, OR, USA

Keywords
cost savings, disease management, health management, primary outcomes, spin

Correspondence
Ariel Linden
Linden Consulting Group
LLC
Hillsboro
OR 97124
USA
E-mail: alinden@lindenconsulting.org

Accepted for publication: 1 November 2010

doi:10.1111/j.1365-2753.2010.01611.x

Abstract

Decision-makers in payor and provider organizations rely on the peer-reviewed literature as a source of ideas for new interventions to control costs and improve the quality of health care. However, recent evidence has emerged that spin tactics are regularly employed in the description of results from randomized controlled trials (RCTs) that actually do not show statistically significant outcomes. Observational studies, which are more commonly used to evaluate health management interventions than RCTs, offer greater opportunity for spin because evaluators control more aspects of the evaluation approach. This paper provides a detailed description of how to critically review study outcomes from health management interventions and identify spin. Our emphasis on health management is motivated by the tremendous discrepancy between the large financial savings reported by commercial vendors and the savings reported in RCTs and systematic reviews that indicate little-to-no financial benefit. We use unpublished data from a medical home pilot project to demonstrate how easily statistically non-significant findings can be portrayed in a favourable light, either through error, omission or intentional spin. We then describe additional techniques that should be utilized in order to present outcomes in an accurate and comprehensive manner. The step-by-step approach described here will hopefully assist readers in becoming more critical consumers of outcomes reported in scholarly journals or the popular media by identifying when spin tactics are used to camouflage ineffective interventions.

Introduction

Decision-makers in payor and provider organizations rely on the peer-reviewed literature as a source of ideas for new interventions to control costs and improve the quality of health care. The majority of these individuals have clinical or managerial backgrounds and do not have extensive knowledge of research methods and statistics. As a result, they place their trust in the peer-reviewed process to ensure that the findings and associated reporting are unbiased and accurate. Unfortunately, there is evidence to suggest that the peer-reviewed process is failing to fulfill such expectations [1]. In a recent study published in the Journal of the American Medical Association, Boutron *et al*. [2] reviewed 72 journal papers reporting results from randomized controlled trials (RCTs) with non-statistically significant primary results. They sought to determine if 'spin' was employed in the description of results. Spin was defined as:

> . . . use of specific reporting strategies, from whatever motive, to highlight that the experimental treatment is beneficial, despite a statistically nonsignificant difference for the primary outcome, or to distract the reader from statistically nonsignificant results. [2] (p. 2059)

The paper identified various spin tactics, such as claiming efficacy with no consideration of the statistically non-significant primary outcome; acknowledging statistically non-significant results for the primary outcome but emphasizing the beneficial effect of treatment; and recommending the use of the treatment. Generally, there was a higher prevalence of spin in the abstract than in the main body of the manuscript and a subset of papers used spin in the title [2]. This is of particular concern as many readers do not have access to full papers or read them in detail, and thus may only rely on the title and/or abstract to shape decisions.

Perhaps most disconcerting was that such a high degree of spin was found in RCTs for which spin tactics should be most difficult to employ because of the inherent rigour in this evaluation design. Observational studies, which are more commonly used to evaluate health management interventions than RCTs, offer greater opportunity for spin because evaluators control more aspects of the evaluation approach. For example, evaluators have complete control over the method used to form a comparison group in addition to the adjustment techniques used to control for bias and confounding. To help readers identify spin in the disease management (DM) literature, Linden and Roberts [3] provided practical guidance on how to critically assess reported programme

outcomes. The authors' focus on DM was motivated by the tremendous discrepancy between the large financial savings reported by commercial vendors [4,5] and the savings reported in RCTs and systematic reviews that indicate little-to-no benefit [6–13].

This paper builds upon the concepts presented in Linden and Roberts [3] by providing a more detailed description of how to critically review study outcomes from health management interventions and identify spin. We rely on unpublished data from a medical home pilot project, which was chosen for two reasons. First, the intervention employed an observational study design, thereby requiring the reader to assess additional steps in the analytic process than would be necessary in an RCT. Second, there was a statistically non-significant difference in the primary outcome, the change in medical costs, between the treatment and control group (following the inclusion criteria in Boutron *et al*. [2]). In the next section we describe the setting in more detail and then discuss the components of the evaluation. For each component, we explain the current approach taken, highlight how spin could be employed and then describe additional techniques that should be utilized in order to present outcomes in an accurate, comprehensive manner. We conclude with final thoughts to further assist the reader in becoming a more critical consumer of health management outcomes reported in scholarly journals or the popular media.

Background

Setting

We examined the evaluation of a primary care-based medical home pilot programme that invited patients to enroll if they had a chronic illness or were predicted to have high costs in the following year. The goal of the pilot was to lower health care costs for programme participants by providing intensified primary care that was intended to reduce unnecessary emergency department visits and hospitalizations.

Data

A substantial number of data were available for the evaluation: demographic characteristics (age and gender); health services utilization (primary care visits, other outpatient visits, laboratory tests, radiology tests, prescriptions filled, hospitalizations, emergency department visits and home-health visits); and total medical costs (the amount paid for all these health services). These data were retrieved to support an evaluation of the outcome of interest – the change in total medical costs from the 12 months prior to the programme (baseline period) to the 12 months after programme initiation (programme period). There were 374 programme participants for whom data were available for this 24-month period. Based on the analytic approach described in the following section, data were also retrieved for a group of non-participants who had 24 months of continuous insurance eligibility and served as controls.

Analytic approach: selecting a control group

In an RCT, individuals are randomly assigned to receive either treatment or control, thereby giving each person an equal probability to be chosen for the intervention. This process is intended to ensure that individuals assigned to either group are comparable on both known and unknown characteristics. Accordingly, any differences found in outcomes between the study groups can be attributed to the programme intervention and not biased by baseline differences in group characteristics or confounders, such as illness severity, co-morbidities, motivation to change health behaviours, etc.

In contrast, observational studies are characterized by their participants electing to participate in the intervention, usually with the knowledge of what the intervention will entail. Thus, individuals selecting to engage in the intervention are likely to be quite different than otherwise similar individuals who elect not to participate (referred to as *Selection Bias*) [14]. There is considerable evidence supporting the myriad factors, such as belief systems, enabling factors and perceived need, that explain why and how individuals access health care and make health-related decisions, such as participating in a health management programme [15,16]. Therefore, in order to make causal inferences about the effect of the intervention in observational studies, the evaluator should attempt to emulate the randomization process of an RCT as closely as possible by finding (or creating) a control group that is approximately equivalent to the treatment group on known pre-intervention characteristics. Subsequently, they can assume that the remaining unknown characteristics are inconsequential and will not bias the results [17].

For the current analysis, a propensity score matching technique was employed. The propensity score, defined as the probability of assignment to the treatment group given the observed characteristics [18], controls for pre-intervention differences between treated and non-treated groups. Propensity scores are generally derived from a logistic regression equation that reduces each participant's set of covariates to a single score. It has been demonstrated that in large samples, when treatment and control groups have similar distributions of the propensity score, they generally have similar distributions of the underlying covariates used to create the propensity score. This means that observed baseline covariates can be considered independent of treatment assignment (as if they were randomized), and therefore will not bias the treatment effects [18]. A comprehensive discussion on the application of propensity scoring techniques in health management programmes is provided elsewhere [19–21].

In the current study, the propensity score was estimated using logistic regression to predict programme participation status using pre-intervention demographic, utilization and cost covariates described above and presented in Table 1. An optimal matching algorithm [22] was then employed to match pairs (one participant to one non-participating control) on the estimated propensity score resulting in 276 matched pairs. Controls were selected from the population of non-participants ($n = 1628$) that had 24 months of continuous insurance eligibility and were never exposed to the intervention.

Identifying and addressing spin in comparing treatment and control groups

Regardless of the type of study (RCT or observational), a table presenting the pre-intervention characteristics of the treatment and comparison groups should be presented to enable the reader to

Table 1 Comparison of baseline characteristics of programme participants and the untreated population

	Participants ($n = 374$)	Non-participants ($n = 1628$)	Standardized differences	*P*-value*
Demographic characteristics				
Age	54.9 (6.7)	43.4 (12.0)	1.704	<0.001
Female	211 (56.4%)	807 (49.6%)	0.138	0.017
Utilization and cost				
Primary care visits	11.3 (7.3)	4.6 (4.3)	0.914	<0.001
Other outpatient visits	18.0 (16.6)	7.2 (10.6)	0.647	<0.001
Laboratory tests	6.1 (5.3)	2.4 (3.3)	0.705	<0.001
Radiology tests	3.2 (4.5)	1.3 (2.5)	0.424	<0.001
Prescriptions filled	40.6 (30.0)	11.9 (17.1)	0.956	<0.001
Hospitalizations	0.2 (0.5)	0.1 (0.3)	0.326	<0.001
Emergency department visits	0.4 (1.0)	0.2 (0.5)	0.226	<0.001
Home-health visits	0.1 (0.9)	0.0 (0.4)	0.083	0.012
Total costs	8236 (9830)	3047 (5817)	0.528	<0.001

*A two-tailed *t*-test for independent samples was used for continuous variables, and a chi-squared test was used for dichotomous variables. Continuous variables are reported as mean (standard deviation) and dichotomous variables are reported as *n* (%).

assess the comparability of the groups. In an RCT with adequate sample size, we would expect most, if not all, of the baseline covariates to be balanced between treatment and control groups. If cohorts are imbalanced on important observed baseline features, they will likely differ on unobserved characteristics as well, and causal inferences about the programme impact will be limited. For observational studies, it is helpful for the reader to see both the original (unadjusted) baseline data and the adjusted data after matching (or other strategy employed to construct the comparison group) in order to assess how well the strategy worked to reduce imbalances on observed covariates.

While there are several methods available to assess covariate balance, the standardized difference is perhaps the most widely used measure of balance and is simple for readers to compute themselves based on data presented in a table of baseline characteristics (modified from the study by Flury and Reidwyl [23]):

$$d = \frac{\left(\bar{X}\ \text{treatment} - \bar{X}\ \text{control}\right)}{(\text{SD treatment} + \text{SD control})/2}$$

where the numerator is the difference in means between the groups, and the denominator is the pooled standard deviation. The appeal of this method is that it is indifferent to the unit of measurement and insensitive to sample size. Normand *et al*. [24] suggest that a standardized difference of less than 0.10 is indicative of good balance. However, there is currently no universally recognized cut-off point.

Table 1 displays the baseline characteristics of programme participants and the non-participant population from which controls were drawn. As evident in the table, the programme group is significantly older, has a higher proportion of females and has higher utilization and costs than the population of non-participants. Except for home-health visits, all covariates have absolute standardized differences over 0.10, which suggests that there are large imbalances between participants and the pool of non-participants.

Spin tactics can be employed when imbalances are observed. Blatant spin would involve showing imbalances without discussing the potential implications. More subtle spin would involve arguing that all remaining bias will be adjusted for later on in the analytic process (by including covariates in the outcomes model for measures on which intervention and control groups differed at baseline). This is problematic for two primary reasons. First, in an RCT, multiple imbalances – especially on covariates that are likely to influence the outcome – are indicative of a failure in the randomization process. If sophisticated analytic methods are required to adjust for major imbalances, the robustness of the RCT design has been lost. In such cases, the reader should be cautious in accepting the reported outcomes. Second, there is no existing statistical method that can ensure all sources of bias or confounding have been completely addressed. Thus, the reader must take it on faith that all sources of bias have been accounted for in the outcomes. Given these concerns, it is essential that an evaluation offers evidence that the adjustment method employed has at least achieved balance on observed baseline covariates and, if not, clearly stated that these differences could account for any observed difference in the outcome (in lieu of an intervention effect).

Table 2 displays the baseline characteristics of the propensity score matched sample of participants and controls. It is evident from reviewing the absolute standardized differences that the matching procedure was successful in reducing imbalances of all observed baseline covariates to under 0.10. The two groups are therefore comparable on all baseline characteristics used in the analysis.

While standardized differences compare the difference in means between treated and untreated subjects, graphic displays of the data offer an alternative way for the reader to assess if balance was achieved, and has the advantage of providing more information about the success of matching across *the entire distribution* of values. Because the standardized difference considers only the means and standard deviations, residual differences at varying points along the continuum will not be captured. It is theoretically possible to have a large spike in one direction and an offsetting spike in the other direction, and yet the standardized difference will be within an acceptable range. Only when reviewing the distributions will these (perhaps meaningful) spikes become evident. Austin [25] suggests using side-by-side boxplots, empirical cumulative distribution functions, quantile-quantile plots, and non-parametric estimates of density functions to compare the

Table 2 Comparison of baseline characteristics of programme participants and their propensity score matched controls

	Participants ($n = 276$)	Matched controls ($n = 276$)	Standardized differences
Demographic characteristics			
Age	54.6 (6.5)	54.0 (6.9)	0.082
Female	152 (55.1%)	150 (54.3%)	0.015
Utilization and cost			
Primary care visits	9.5 (6.5)	9.7 (6.2)	0.022
Other outpatient visits	15.2 (16.2)	15.6 (14.1)	0.029
Laboratory tests	4.8 (5.8)	5.2 (4.5)	0.086
Radiology tests	2.8 (4.4)	2.8 (4.1)	0.009
Prescriptions filled	32.6 (27.8)	34.1 (25.3)	0.058
Hospitalizations	0.2 (0.4)	0.2 (0.4)	0.026
Emergency department visits	0.3 (0.8)	0.3 (0.9)	0.027
Home-health visits	0.1 (0.9)	0.1 (1.0)	0.011
Total costs	6318 (7827)	6748 (7648)	0.056

Continuous variables are reported as mean (standard deviation) and dichotomous variables are reported as *n* (%).

distribution of continuous baseline covariates between treated and untreated subjects in the unmatched and matched samples.

Figure 1 displays side-by-side boxplots and kernel density function estimates for baseline costs in the unmatched and matched samples. The two leftmost graphs present the values of baseline costs for the programme participants compared with population of non-participants. As both graphs illustrate, there is a higher density of values closer to zero in the population of non-participants compared with the participant group. Both groups, however, have several outliers with values over $50 000. The two right-most graphs present the values of baseline costs after matching. Both the boxplot and kernel density function confirm that the matching procedure was successful in reducing the differences in the distribution of baseline costs between the two groups (as evidenced by a near complete overlap of the two distributions) and support the result obtained using the standardized differences. The use of these methods will provide the reader with some reassurance that the appropriate rigour was applied to achieve comparability between the groups, not only in the means, but across the entire distribution of values.

In summary, the reader should expect to see verification that treatment and control groups are balanced on important baseline covariates via empirical or graphical methods, especially for variables that are likely to predict the outcome and therefore have a greater chance of confounding the results. Imbalances on important observed (and very likely on unobserved) baseline covariates indicate that the groups are not comparable, which raises concerns that the outcomes may be biased or reported with spin.

Reporting results

Once the reader is assured that the analytic approach has been successfully executed, they must assess whether outcomes are faithfully reported. According to the 2010 Consolidated Standards of Reporting Trials (CONSORT) guidelines [26], an evaluation should report summary statistics of the outcome (e.g. mean and standard deviation) for each group together with the difference between the groups (effect size). Additionally, confidence intervals (typically calculated at the 95% level) should be provided for all outcomes to indicate the precision (uncertainty) of the estimate, rather than presenting *P*-values alone [26]. Confidence intervals are preferred over *P*-values because they convey information about the magnitude of the estimated effect; they provide an estimated range of values within which we believe that the population mean falls in 95 of 100 hypothetical repetitions of the intervention [3]. Conversely, *P*-values only '. . . define two alternative outcomes – significant or not significant – which is not helpful and encourages lazy thinking' [27] (p. 746).

Following the CONSORT guidelines [26], Table 3 presents the results of the statistical analysis for the outcome variable – the change in total costs. The first line of results was derived using ordinary least squares (OLS) regression with robust standard errors, and the second line of results was derived using quantile regression (for the median value) [28]. As the outcome is costs in the 12-month programme period minus costs in the 12-month baseline, a negative value indicates a decrease in costs while a positive value indicates an increase in costs. The difference score, typically referred to as a difference-in-differences, represents the difference of the treatment group minus the difference of the control group [29]. As evident in Table 3, the participant group shows a mean decrease in costs of –$1039 while the control group shows an increase in costs of $892. Thus, the difference-in-differences mean estimate is –$1932, suggesting that the participant group had a net average decrease in costs of nearly $2000 from the baseline to programme period. The associated *P*-value of 0.11 suggests that this effect does not achieve conventional levels of statistical significance.

Identifying and addressing spin in reporting and interpreting outcomes

The most blatant form of spin would be to claim programme savings by only reporting the negative mean point estimate and de-emphasizing or distracting the reader from the non-significant *P*-value. In our example, this would be particularly misleading given that the confidence interval ranges from –$4321 to $457. In other words, the true mean difference-in-difference value lies between –$4321 and $457 if a series of identical studies were carried out repeatedly on different samples from the same

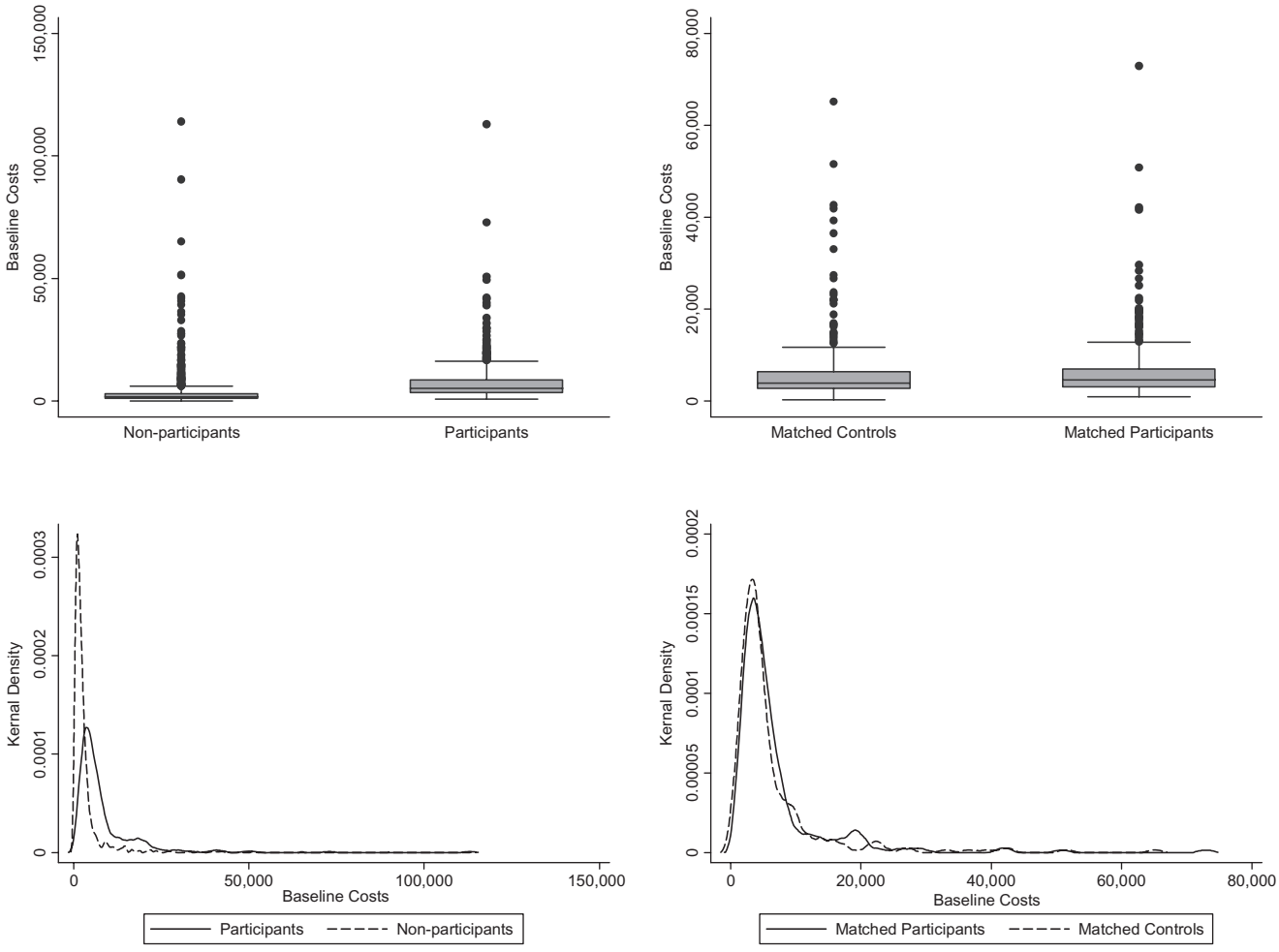

Figure 1 Side-by-side boxplots and kernel density distribution functions for baseline costs. Left panels compare participants to population of non-participants. Right panels compare propensity score matched participants and controls.

Table 3 The change in costs (programme period – baseline) for programme participants and their propensity score matched controls

	Participants ($n = 276$)	Matched controls ($n = 276$)	Difference	*P*-value	95% CI
Change in total costs (mean)	–1039 (9473)	892 (17 848)	–1932 (1216)	0.113	–4321, 457
Change in total costs (median)	–201	–212	11	0.949	–325, 347

Values are reported as mean (standard deviation) for ordinary least squares model and median values for quantile regression. A negative value represents a decrease in costs, and a positive value represents an increase in costs.

population. Because this highlights the potential that costs may have *risen* for participants relative to controls, the reader can immediately see the value of reporting and interpreting confidence intervals relative to only the *P*-value ($P = 0.11$). If the authors favoured a positive outcome, it would be tempting to highlight the average decrease in costs of –$1932, and then generalize that supposed savings to the larger population, or to other populations, settings and outcomes [30]. The 95% confidence interval clearly refutes the validity of such an approach. Spin tactics would therefore be evident in either reporting the confidence intervals and then disregarding them, or not reporting them at all.

A more subtle form of spin would be to suggest that the very large standard deviations in both groups explain why the mean values are not statistically different and stop there. However, there is more that can be done to assess the extent to which this is true. A visual display of the groups' respective distributions clarifies where the variability arises. Figure 2 presents side-by-side boxplots of the change in cost outcome for the propensity score matched participants and controls. It is evident that the control group has three outliers with increases in cost of greater than $100 000, and the participant group has one low outlier with a decrease in costs of –$59 000.

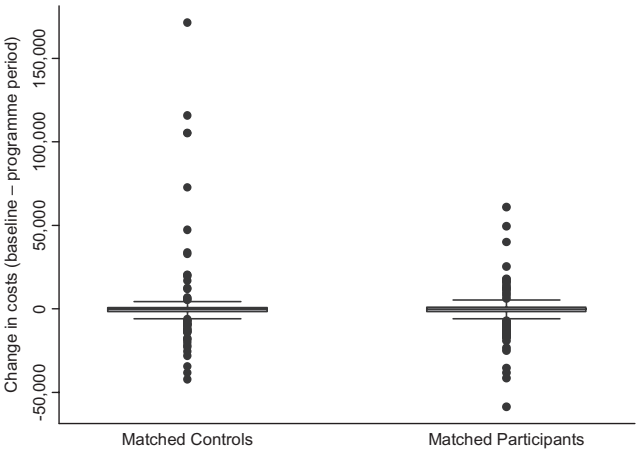

Figure 2 Side-by-side boxplots of the outcome variable – change in costs (baseline – programme period) for the propensity score matched participants and controls. A negative value indicates a reduction in costs and a positive value indicates an increase in costs.

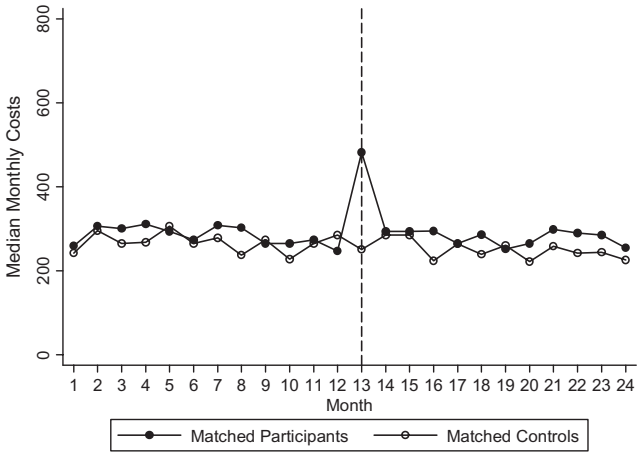

Figure 3 Median monthly costs for propensity score matched participants and controls. Months 1–12 represent the baseline period, and months 13–24 represent the programme period.

There are different ways that evaluators can handle outliers, including removing them or 'winsoring' them (replacing the outlier values with the next closest value counting inward from the extreme) [31]. It is quite common to see health intervention studies report that the outliers have been removed, but not describe the method used to establish how outliers were identified, or how their removal altered the results. Obvious care must be taken when dealing with outliers as this represents an additional opportunity to manipulate the data to portray the results in the most favourable light. Therefore, in studies where outliers have been removed or manipulated, the reader should expect to see outcomes reported both with and without the outlier treatment along with a discussion of the author's rationale for considering such adjustments appropriate within the particular context.

To avoid the outlier issue altogether, alternative analytic approaches such as quantile regression, rank-sum statistics or other non-parametric models should be considered. These approaches generally estimate either differences in medians or median differences and are therefore not influenced by outliers. In Table 3, we present the results of the quantile regression (for the median) on the current data. Quantile regression works like OLS regression, but estimates medians (or other centile values) instead of means [28]. Contrary to the results of the OLS model, the median difference-in-difference indicates an $11 *increase* in costs for the participant group over the matched controls, with confidence intervals ranging from –$325 to $347.

Another common spin tactic used in health management evaluations that do not achieve statistical significance is claiming that it is a function of limited programme duration and/or an insufficient sample size without assessing the plausibility of these assertions. One simple way of testing the extent to which sample size or programme duration is at issue is to visually inspect monthly values of the outcome variable to see if there appears to be a divergence between groups indicating the beginning of a treatment effect and, if so, perform sample size calculations to determine the sample sizes needed to achieve statistical significance.

Figure 3 depicts the monthly median costs for the propensity score matched participants and controls over the course of the 24-month observation period (12 months of baseline and the 12 programme months). As illustrated, it is impossible to detect a meaningful difference in the monthly median cost trend between treatment and control groups starting at any point along the continuum. These data refute the 'insufficient programme duration' argument. It is also impossible to see how an increased sample size would play a factor in demonstrating an intervention effect because there is simply no evidence of an effect and therefore no amount of additional sample size would change the trend.

Another way of reviewing the outcome data is to estimate the proportion of individuals within each matched group who achieved cost savings (indicated by a difference score of less than zero). In total, 153 of the 276 controls (55.43%) had lower costs in the programme year than in the baseline year, while 152 of the 276 programme participants (55.07%) had lower costs in the programme year than in the baseline year. By presenting the results in this manner, we see that the matched control group had one additional individual with lower costs than the treatment group (supporting the results of the quantile regression). Given this finding, a sample size calculation is unwarranted as the treatment group would first need to be 'directionally correct' (higher proportion of individuals with cost savings compared with the control group).

In this section we have highlighted specific areas of outcomes reporting where spin is often used. This includes: emphasizing the average cost savings estimate without reporting or considering the confidence intervals; arguing that the statistically non-significant result is a function of small sample size and/or insufficient programme duration but not providing supporting evidence; or manipulation of outliers without a detailed explanation of methods and results. Additionally, several methods (both empirical and graphical) have been proposed to assist the reader in challenging these spin tactics. None of these methods require sophisticated statistical knowledge to implement or explain, and they should

therefore be included in any evaluation. When absent, readers should be very cautious in basing their decisions on reported results.

Discussion and conclusion

The medical home example used in this paper illustrates how easily statistically non-significant findings can be portrayed in a favourable light, either through error, omission or intentional spin. Focusing on the point estimate alone (the mean change in costs) would lead the reader to believe that the programme reduced medical cost by $1932 per person. However, this is easily refuted when reviewing the confidence interval (which crosses zero) and the *P*-value (>0.05). Reviewing graphic displays of the distribution of the outcome variable sheds light on the variability around that point estimate and the impact of outliers. After switching to analytic approaches more suitable to this data structure, the point estimate actually changes in favour of the control group. Finally, examining monthly data allows the reader to determine if there is any indication that the intervention group is diverging from the control group in the intended direction. Without this, a claim that more time or more sample size is required suggests the employment of spin.

There is considerable evidence that a large number of research studies use spin tactics to portray statistically non-significant outcomes in a favourable light. Studies conducted on the effectiveness of DM programmes have typically fallen into two categories: those self-reported by vendors which tend to show large positive cost savings, and studies conducted independently (such as those large demonstration projects sponsored by Medicare) which have not consistently demonstrated cost savings. Senior leaders in payor and provider organizations are likely to continue to rely on peer-reviewed literature as a source of ideas for new interventions and programmes to help control the cost and improve the quality of care. The step-by-step approach described here will hopefully assist readers in becoming more critical consumers of outcomes reported in scholarly journals or the popular media by identifying when spin tactics are used to camouflage ineffective interventions.

Concerns about misrepresentation of medical and health care research findings have grown over time, recently highlighted in the lay press in an Atlantic monthly article entitled 'Lies, Damned Lies, and Medical Science' [32]. Business leaders and policymakers alike should be wary of the positive financial and health outcomes reported by studies in the peer-reviewed literature, lay press and websites. With limited resources available in our health care system to provide actual patient care, now more than ever it is critical that rigorous approaches be employed to assess whether these interventions truly improve the quality and reduce the cost of care, or whether they are adding to the growing financial burden of the system.

References

1. Ioannidis, J. P. A. (2005) Why most published research findings are false. *PLoS Medicine*, 2 (8), e124.
2. Boutron, I., Dutton, S., Ravaud, P. & Altman, D. G. (2010) Reporting and interpretation of randomized controlled trials with statistically nonsignificant results for primary outcomes. *Journal American Medical Association*, 303 (20), 2058–2064.
3. Linden, A. & Roberts, N. (2005) A user's guide to the disease management literature: recommendations for reporting and assessing program outcomes. *American Journal of Managed Care*, 11 (2), 81–90.
4. Johnson, A. Disease management: the programs and the promise. *Milliman USA Research Report*. May 2003 Available at: http://publications.milliman.com/research/health-rr/archive/pdfs/Disease-Mangement-Programs-Promise-RR05-01-03.pdf (last accessed 6 October 2010).
5. Shutan, B. (2004) The DM Rx: disease management programs producing fast and meaningful outcomes, impressive ROI. *Employee Benefit News*, 18 (13). Available at: http://www.corsolutions.com/resources/articles/dm/EBNPetit.pdf (last accessed 6 October 2010).
6. Congressional Budget Office (2004) An Analysis of the Literature on Disease Management Programs. Washington, DC: Congressional Budget Office. Available at: http://www.cbo.gov/showdoc.cfm?index=5909&sequence=0 (last accessed 6 October 2010).
7. Ofman, J. J., Badamgarav, E., Henning, J. M., Knight, K., Gano, A. D. Jr, Levan, R. K., Gur-Arie, S., Richards, M. S., Hasselblad, V. & Weingarten, S. R. (2004) Does disease management improve clinical and economic outcomes in patients with chronic diseases? A systematic review. *American Journal of Medicine*, 117 (3), 182–192.
8. Goetzel, R. Z., Ozminkowski, R. J., Villagra, V. G. & Duffy, J. (2005) Return on investment on disease management: a review. *Health Care Financing Review*, 26, 1–19.
9. Mattke, S., Seid, M. & Ma, S. (2007) Evidence for the effect of disease management: is $1 billion a year a good investment? *American Journal of Managed Care*, 13, 670–676.
10. McCall, N., Cromwell, J. & Bernard, S. *Evaluation of phase i of medicare health support (formerly voluntary chronic care improvement) pilot program under traditional fee-for-service medicare*. RTI International, June 2007. Available at: http://www.cms.hhs.gov/Reports/Downloads/McCall.pdf (last accessed 6 October 2010).
11. Brown, R., Peikes, D., Chen, A. & Schore, J. (2008) 15-site randomized trial of coordinated care in medicare FFS. *Health Care Financing Review*, 30 (1), 5–25.
12. Esposito, D., Brown, R., Chen, A., Schore, J. & Shapiro, R. (2008) Impacts of a disease management program for dually eligible beneficiaries. *Health Care Financing Review*, 30 (1), 27–45.
13. Disease Management Purchasing Consortium, Inc (2010) Studies show cost savings in disease management/wellness . . . or do they? DMPC Presents the Intelligent Design Awards. Available at: http://www.dismgmt.com/ida (last accessed 29 October 2010).
14. Linden, A., Adams, J. & Roberts, N. (2003) An assessment of the total population approach for evaluating disease management program effectiveness. *Disease Management*, 6 (2), 93–102.
15. Andersen, R. M. (1968) Behavioral Model of Families: Use of Health Services. Research Series No. 25. Chicago, IL: Center for Health Administration Studies, University of Chicago.
16. Aday, L. & Andersen, R. M. (1981) Equity in access to medical care: realized and potential. *Medical Care*, 19 (12 Suppl), 4–27.
17. Rubin, D. (2007) The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Statistics in Medicine*, 26, 20–30.
18. Rosenbaum, P. R. & Rubin, D. B. (1983) The central role of propensity score in observational studies for causal effects. *Biometrika*, 70, 41–55.
19. Linden, A., Adams, J. & Roberts, N. (2005) Using propensity scores to construct comparable control groups for disease management program evaluation. *Disease Management and Health Outcomes*, 13, 107–127.
20. Linden, A. & Adams, J. L. (2008) Improving participant selection in disease management programmes: insights gained from propensity score stratification. *Journal of Evaluation in Clinical Practice*, 14 (5), 914–918.

21. Linden, A. & Adams, J. L. (2010) Using propensity score-based weighting in the evaluation of health management programme effectiveness. *Journal of Evaluation in Clinical Practice*, 16, 175–179.
22. Rosenbaum, P. R. (1989) Optimal matching for observational studies. *Journal of the American Statistical Association*, 84 (408), 1024–1302.
23. Flury, B. K. & Reidwyl, H. (1986) Standard distance in univariate and multivariate analysis. *The American Statistician*, 40, 249–251.
24. Normand, S. L. T., Landrum, M. B., Guadagnoli, E., Ayanian, J. Z., Ryan, T. J., Cleary, P. D. & McNeil, B. J. (2001) Validating recommendations for coronary angiography following an acute myocardial infarction in the elderly: a matched analysis using propensity scores. *Journal of Clinical Epidemiology*, 54, 387–398.
25. Austin, P. C. (2009) Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in Medicine*, 28, 3083–3107.
26. Moher, D., Hopewell, S., Schulz, K. F., Montori, V., Gøtzsche, P. C., Devereaux, P. J., Elbourne, D., Egger, M. & Altman, D. G. (2010) CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. *British Medical Journal*, 340, c869.
27. Gardner, M. J. & Altman, D. G. (1986) Confidence intervals rather than P values: estimation rather than hypothesis testing. *British Medical Journal*, 292, 746–750.
28. Gould, W. & Rogers, W. H. Quantile regression as an alternative to robust regression. Proceedings of the Statistical Computing Section. Alexandria, VA: American Statistical Association, 1994.
29. Buckley, J. & Shang, Y. (2003) Estimating policy and program effects with observational data: the 'differences-in-differences' estimator. *Practical Assessment, Research & Evaluation*, 8 (24). Available at: http://PAREonline.net/getvn.asp?v=8&n=24 (last accessed 16 October 2010).
30. Linden, A., Adams, J. & Roberts, N. (2004) The generalizability of disease management program results: getting from here to there. *Managed Care Interface*, 17 (7), 38–45.
31. Barnett, V. & Lewis, T. (1994) Outliers in Statistical Data. Chichester: John Wiley.
32. Freedman, D. H. (2010) Lies, damned lies, and medical science. *The Atlantic*. Available at: http://www.theatlantic.com/magazine/archive/2010/11/lies-damned-lies-and-medical-science/8269/2/ (last accessed 17 October 2010).