Journal of Evaluation in Clinical Practice, **12**, 2, 140–147 doi:10.1111/j.1365-2753.2005.00612.x

Strengthening the case for disease management effectiveness: un-hiding the hidden bias

Ariel Linden Dr.PH MS,[1,2] John L. Adams PhD[3] and Nancy Roberts MPH[4]
[1]President, Linden Consulting Group, Portland, OR, USA
[2]Oregon Health and Science University, School of Medicine, Department of Preventive Health/Preventive Medicine, Portland, OR, USA
[3]Senior Statistician, RAND Corporation, Santa Monica, CA, USA
[4]Regional Director of Integrated Performance/Six Sigma Champion, Providence Health System, Portland, OR, USA

Correspondence
Ariel Linden
President, Linden Consulting Group
6208 NE Chestnut Street
Hillsboro, OR 97124
USA
E-mail: alinden@lindenconsulting.org

Keywords: disease management, hidden bias, observational study design, sensitivity analysis, unobserved covariates

Accepted for publication:
10 March 2005

Abstract

As is the case with most health care program evaluations, disease management (DM) programs typically follow an observational study design, indicating that randomization to treatment or control was not performed. The foremost limitation of observational studies, compared to randomized studies, is that the only biases that can be controlled for are those associated with observed variables. Hidden bias refers to all those unobserved covariates that may distort the conclusions of the study. This paper introduces a sensitivity analysis that is used to determine the magnitude of hidden bias necessary to alter the conclusion that a DM program intervention was indeed effective.

Introduction

Randomization is the great equalizer of research study design. The key idea is that by allowing each suitable individual equal probability of receiving treatment or control, all variability is distributed equally between the two groups (e.g. Cochran 1965; Rosenbaum 2002; Wilson & MacDowell 2003). Variability comes in two forms; *Observed covariates* or characteristics are tangible data points available to the analyst via sources such as claims, medical records, member files, or survey reports. *Unobserved covariates* are all other characteristics not captured or recorded.

While observed covariates are used for ensuring that subjects assigned to the two groups are similar on baseline characteristics (i.e. age, sex, disease status, etc.), it is left to the process of randomization to ensure that unobserved characteristics are similar in both groups as well. Why is it so important that treatment cases and controls be comparable on all baseline characteristics? It is simply to ensure that any differences found in outcome measures be attributable to the program intervention and not biased by baseline differences in study group characteristics.

Disease management (DM), as defined by the Disease Management Association of America (2004), is a system of coordinated interventions and communications for populations with conditions in which patient self-care efforts are significant. DM programs were developed under the assumption that by augmenting the traditional episodic medical care system with services and support between doctor visits, the overall cost of health care could be reduced. For many chronic diseases, such as diabetes, asthma, and congestive heart failure, there is much opportunity to improve the quality and consistency of care. DM programs were developed to assist doctors and their patients in identifying and closing those gaps in care.

Although disease management has been in existence for over a decade, there is still much uncertainty as to its effectiveness in improving health status and reducing medical cost. The vast majority of DM program evaluations follow an observational study design, without randomization to treatment or control. Given that DM programs are limited to the use of these designs owing to cost, time and logistical constraints, it is imperative that the inherent threats to validity are reduced, controlled for, or at the very least, the magnitude of their impact estimated (e.g. Linden *et al*. 2003a). Several research-based techniques with applicability to DM have recently been introduced to the DM evaluation discussion (Linden *et al*. 2003b,c, 2004a,b,c,d,e, 2005a,b,c,d). The prevailing theme in these papers is that there are many observational study designs and statistical methods available for reducing or controlling threats to validity that can be applied to the evaluation of DM programs. However, as indicated earlier, the limitation of observational studies compared to randomly assigned studies is that the only biases that can be controlled for are those that can be observed. The more observed covariates that can be found to match cases and controls on, the more confident one can be that the two groups are comparable. However, if important characteristics have not been included in the matching process there may be concern that the groups are not comparable, and thus, outcomes may be distorted because of these unobserved covariates. This threat to validity is referred to as hidden bias. While it is impossible to control for hidden bias, estimation of its magnitude should be considered common practice in DM program evaluations. This type of assessment provides the consumer with more confidence in the study findings.

This paper introduces methods for assessing the consequences of hidden bias in DM program, evaluations. Examples with discussion will be provided so that these techniques can be easily replicated in DM program evaluations.

Study design and hidden bias

In order to better explain the impact of hidden bias in observational studies, a comparison of three relevant research designs is presented in Table 1 and briefly explained below:

Randomized controlled trial (RCT)

The randomized controlled trial (RCT) is an experimental design in which distribution of observed covariates occurs by screening subjects to meet study eligibility criteria, and then randomly assigning those eligible individuals to either treatment or control before study commencement. Similarly, the random assignment acts to evenly distribute unobserved covariates, and as a result, the outcomes can be considered a relatively unbiased estimate of the true treatment effect (assuming that no confounding effects occurred during the intervention, influencing the results). This type of study design is rarely used to evaluate disease management program effectiveness (Linden & Roberts 2005).

Total population-based approach

This method is the most ubiquitous evaluation design currently used in DM (e.g. Linden *et al.* 2003a). In contrast to the RCT, all individuals suitable for participation in the program are invited to enrol, and

Table 1 Characteristics of the randomized trial and two observational study designs

Characteristic	*Randomized controlled trial*	*Population-based approach*	*Case–control study*
Design type	Experimental	Observational	Observational
Assignment to treatment	Prospective/random	Retro-/non-random	Retro-/non-random
Distribution of observed covariates	Balanced	Unbalanced	Balanced
Control of unobserved covariates?	Yes	No	No
Susceptibility to biased outcomes due to selection?	Unbiased	High	Low

none are assigned to a comparable control group. As a result, participants may differ significantly in some unobserved way from those who were not asked to participate, refused to participate, or not identified as suitable for program inclusion. To further confuse the issue, all members of the population with the given disease are used in the analysis, regardless of whether they were program participants or not. This population's change in the outcome measure is compared to the population's experience in the year prior to program commencement. Therefore, for the purpose of analysis, 'assignment' to the diseased population occurs retrospectively, when all members of the population have been identified (along with their corresponding prior year 'control' population).

Critics may argue that comparing the entire population, pre- and post-population (as compared to just program participants), distributes covariates equally among the groups. However this approach does not control for bias because of the following reasons; (a) if the program is thought to be the driver of improved outcomes for the entire diseased population, then bias in selecting program participants does, in fact, impact outcomes, and (b), population characteristics may change year after year, as a result of aging, health plan turnover, etc., thereby adding another level of bias to the study. At the end of the day, there may be enough bias to raise concerns about the validity of the study findings. (Linden *et al*. 2003a).

Case-control study

The case-control study, in which program participants are matched to suitable non-program members, remains the best practical method for controlling for observed bias (Linden *et al*. 2005b). Using this technique, program participants are matched retrospectively to controls on several observed baseline characteristics. Even though innumerable variables can be used in the matching process, validity of the outcomes may still be questioned because of the concern that hidden bias may be large enough to nullify those results.

Estimating the magnitude of hidden bias

While it is impossible to measure the scope or impact of hidden bias (what is unknown cannot be measured), it is certainly possible, and recommended, to provide an estimate of the magnitude of hidden bias necessary to invalidate the study findings. Cornfield *et al*. (1959) has been credited with being the first study to measure sensitivity to hidden bias (e.g. Rosenbaum 2002). In this 1959 investigation, the authors developed a model estimating the size of hidden bias necessary to invalidate the relationship between smoking and lung cancer. They concluded that hidden bias alone would have to be a near perfect predictor of lung cancer and about 10 times more prevalent in smokers than non-smokers. That a variable such as this exists is highly unlikely.

Many procedures and permutations for estimating hidden bias have been developed since that benchmark study was first introduced, although fundamentally they are all very similar, and only a few are noted here (Bross 1967; Schlesselmann 1978; Rosenbaum 1987; Manski 1990; Gastwirth *et al*. 1998). In the present paper, Rosenbaum's (1987; 2002) technique for estimating sensitivity to hidden bias in matched pairs with continuous outcome variables will be presented because most outcome measures in DM are continuous (e.g. costs, hospital admissions, emergency department visits, etc.). Rosenbaum (1988, 1989, 1991, 2002; Rosenbaum & Krieger 1990) have similarly developed tests for other types of outcome variables such as for unmatched groups with continuous outcomes, censored survival times, paired binary data, matched binary data, and matched binary data with multiple controls.

Conceptually, the basic premise of the sensitivity analysis is that subjects in observational studies differ from those in RCTs in their recruitment to the treatment group. While in a RCT all individuals have a 50/50 chance of being assigned to the treatment group, selection bias is commonplace in observational studies. A perfect example of this bias in DM occurs in the enrolment process. Typically, individuals at highest risk or with highest severity of illness are targeted for program participation, thereby creating a treatment group dissimilar from the population from whence they were drawn. Moreover, those individuals that choose to enrol may be different from their peers with the same level of risk or severity, owing to higher motivation or differing belief systems. In short, it is safe to presume that program participants differ from non-participants in both observed and unobserved ways. The sensitivity analysis provides

estimates for how far these hidden characteristics must diverge from the 50/50 split of an RCT to raise concerns about the validity of the study findings. It also implies that this unknown variable is highly correlated with the outcome measure, so that varying levels of this bias estimated in the treatment group will impact the outcome accordingly.

Rosenbaum (1987, 2002) used the parameter ω to represent the odds of receiving treatment. In an RCT, all subjects have the same odds of receiving treatment, so $\omega = 1$. Conversely, in an observational study, a $\omega = 2$ indicates that one subject is twice as likely as another to receive treatment, and so on. As outside of an RCT ω is an unknown quantity, an array of ω estimates are provided in order to give a sense of the magnitude needed to explain away the relationship between participation in the program and achievement of the desired outcome (i.e. lowered admit rate, ED visit rate, costs, etc.). In the sensitivity analysis, *P*-values are given to indicate the upper and lower bounds for each measure of ω. Presenting a range of significance levels is necessary to estimate the scope of any potential hidden bias. As the value of ω increases, the bounds of *P*-values widen. The ω at which an upper bound *P*-value meets the cut-off level for significance (typically set at $\alpha > 0.05$) is considered the minimum size of hidden bias that would be required to invalidate the study findings.

Applying sensitivity analysis to hidden bias in disease management

In this section, an example will be used to illustrate the methodology for performing a sensitivity analysis to hidden bias. As the outcome variable in this example is continuous (dollars), the Wilcoxon signed-rank statistic is used. This test is a good choice for cost data which may have a skewed distribution. A simulated data set was created to represent a hypothetical evaluation of a DM program effectiveness in reducing costs. Table 2 presents the steps necessary to obtain the *T*-value of the positive ranks, required for calculating both the Wilcoxon signed-rank statistic (Wilcoxon 1945) as well as the bound estimates for values

Table 2 A hypothetical comparison of changes in cost between DM program participants and matched controls using the procedure to obtain the *T*-score for positive ranks

Pair	*DM case*	*Control*	*Difference*	*Absolute difference*	*Absolute rank*	*Relative rank*	*Positive rank*
1	47 367	9 230	38 137	38 137	19	19	19
2	47 831	18 759	29 072	29 072	17	17	17
3	2 403	4 370	–1 967	1 967	3	–3	
4	7 537	5 092	2 445	2 445	5	5	5
5	30 721	4 181	26 540	26 540	16	16	16
6	18 228	17 349	879	879	1	1	1
7	23 785	2 034	21 751	21 751	13	13	13
8	294	2 869	–2 575	2 575	6	–6	
9	45 003	1 507	43 496	43 496	20	20	20
10	29 159	10 419	18 740	18 740	12	12	12
11	25 042	1 388	23 654	23 654	14	14	14
12	40 628	5 721	34 907	34 907	18	18	18
13	8 387	5 538	2 849	2 849	7	7	7
14	28 912	16 394	12 518	12 518	10	10	10
15	3 020	1 208	1 812	1 812	2	2	2
16	18 120	9 113	9 007	9 007	8	8	8
17	17 883	19 938	–2 055	2 055	4	–4	
18	36 559	18 729	17 830	17 830	11	11	11
19	4 737	16 334	–11 597	11 597	9	–9	
20	30 016	3 822	26 194	26 194	15	15	15
Means	23 282	8 700	14 582				
T–positive scores							188

of ω. The data are for 20 matched pairs of program participants and their non-program controls. The number and type of baseline characteristics used in the matching process is not a factor in performing this sensitivity analysis to hidden bias. Linden *et al*. (2005b) provides additional guidance on the development of non-program matched control groups. The values represent the difference in costs between baseline and the end of the first measurement year. To make the example more dramatic, a large disparity was intentionally created in the outcomes to show that the DM program was successful in reducing costs as compared to controls (mean savings was $23 282 in the program cases as opposed to only $8700 in the matched controls).

Obtaining the *T*-value of the positive ranks requires the following steps: (1) calculate the difference in values between each pair, (2) convert the differences into absolute values, (3) determine the rank of each absolute value in the data set, (4) restore the original positive or negative sign to the ranked value, and (5) calculate the sum of the positive ranks (which is the *T*-value).

Upon determination of the *T*-value, minimum and maximum significance levels for the signed-rank test at each values of ω can be calculated using the following formulae (Rosenbaum 1987, 2002):

$$p_{lower} = \frac{1}{1+\omega},\ p_{upper} = \frac{\omega}{1+\omega} \tag{1}$$

$$E = p\frac{n(n+1)}{2} \tag{2}$$

$$V = p(1-p)\frac{n(n+1)(2n+1)}{6} \tag{3}$$

$$Z = \frac{T-E}{\sqrt{V}} \tag{4}$$

p_{lower} and p_{upper} represent the probability range of being assigned to the treatment group. In an RCT $\omega = 1$, therefore Eq. 1 results in both lower and upper probabilities of 0.50. In other words, any given individual has a 50/50 chance of being assigned to the treatment group. At higher values of ω, the range of probabilities expands accordingly. *E* is the expected *T*-value assuming the null hypothesis of no difference in the matched pairs, V is the variance of the *T*-value, and *Z* is the statistic used to test the null hypothesis that there is no difference in matched pairs.

Table 3 illustrates the sensitivity analysis derived from the above formulae for the hypothetical 20 matched pairs. The upper and lower *P*-value when $\omega = 1$ is 0.001 (as the probability of receiving treatment is 50/50, both upper and lower values are the same). This *P*-value is identical to that of an RCT, as tests of statistical significance presume that samples are randomly drawn from a population and thereby assume a normal distribution.

Following the significance levels for each value of ω, we see that the upper-bound *P*-value surpasses the conventional 0.05 level somewhere between $\omega = 2$ and $\omega = 3$ (upper-bounds are 0.029 and 0.094, respectively). This analysis suggests that our study findings becomes sensitive to hidden bias somewhere in the range of $\omega = 2$ and 3. Stated another way, these results suggest that DM program participants would need to be 2 to 3 times more likely to possess hidden traits or factors than their matched controls in order to change our conclusion that the program intervention lead to significant cost savings. This indicates

Table 3 Illustration of a sensitivity analysis performed on a sample of 20 matched pairs with *T* of positive signed-ranks = 188

	$\omega = 1$		$\omega = 2$		$\omega = 3$		$\omega = 4$		$\omega = 5$	
Range	l	U	l	U	l	u	l	u	l	u
p	0.50	0.50	0.33	0.67	0.25	0.75	0.20	0.80	0.17	0.83
E	105	105	70	140	52.5	157.5	42	168	35	175
V	717.5	717.5	637.8	637.8	538.1	538.1	496.0	496.0	398.6	398.6
Z	3.10	3.10	4.67	1.90	5.84	1.31	6.56	0.90	7.66	0.65
P (1 tailed)	0.001	0.001	0.000	0.029	0.000	0.094	0.000	0.185	0.000	0.257

p, probability of assignment to treatment; *E*, expected *T*-value under the null hypothesis; *V*, variance of the *T*-value; *Z*, *z* statistic; and *P*, the significance level. The range of values is given for lower (*l*) and upper (*u*) values of each ω.

that these findings are insensitive to small amounts of bias and require moderately high levels of bias to alter our conclusions.

While there is no standard 'cut-off' point for ω in which one can declare unequivocally that the concern of hidden bias is large enough to invalidate the study findings, Rosenbaum (1987) provides the following guidance:

> Informally, in testing in an observational study against a one-sided alternative that treatment A is superior, an unobserved covariate U would need to increase the odds of assignment to treatment A by more than 50%, that is $\omega = 1.5$, before altering the qualitative impression that treatment A is superior; however, that impression would be open to question if it were plausible that an unobserved U exists which doubled ($\omega = 2$), or tripled ($\omega = 3$) the odds of assignment to treatment A.

In relative terms, the present findings represent a departure from a RCT on the magnitude of 2 to 3 times (with $\omega = 1$ equalling an RCT). Intuitively, this is far enough away from 1 to lessen our concerns about the influence of hidden bias on the study results. Conversely, if the invalidating hidden bias was found to be $\omega < 1.5$, we would be more inclined to believe that hidden bias poses a serious threat to the validity of the study interpretation. While not providing a definitive answer, performing this analysis quantifies the risk that the results achieved were caused not by the intervention, but by some unaccounted for differences between program participants and non-participants.

Developing a table for estimating the impact of hidden bias is good for illustration purposes, but a more precise method would be to perform an iterative process to identify the exact ω at which the upper-bound *P*-value approaches or equals 0.05. A sample algorithm written for visual basic is provided in Appendix I. Using this method, the exact value of ω in our example was determined to be $\omega = 2.23$ (<0.0001 and 0.05, for lower and upper *P*-values respectively).

Conclusions

The impact of hidden bias may be substantial in DM where suitable individuals are not randomly assigned to participation or non-participation in the program. This paper introduced a sensitivity analysis that is used to determine the magnitude of hidden bias necessary to alter the conclusion that a DM program's intervention was indeed effective. In this analysis it is determined what the odds are that a participant will have a particular unobserved characteristic that their matched control does not have. It is also assumed that this unobserved characteristic is highly predictive of the outcome, so that the more program enrolees having this characteristic the higher the likelihood of achieving the desired outcome.

While it is impossible to control for hidden bias, estimation of its magnitude required to invalidate the conclusions should be considered common practice in DM program evaluations. This type of assessment provides more confidence in the study findings.

References

Addinsoft (2004) XLSTAT, Data analysis solution for Microsoft Excel, Paris, France.

Bross I.D.J. (1967) Pertinency of an extraneous variable. *Journal of Chronic Disease* **20**, 487–495.

Cochran W.G. (1965) The planning of observational studies of human populations (with discussion). *Journal of the Royal Statistics Society, Series A* **128**, 134–155.

Cornfield J., Haenszel W., Hammond E., Lilianfeld A., Shimken M. & Wynder E. (1959) Smoking and lung cancer: recent evidence and a discussion of some questions. *Journal of the National Cancer Institute* **22**, 173–203.

Disease Management Association of America (DMAA) (2004) Definition of disease management. Available at: http://www.dmaa.org/definition.html (retrieved 23 June).

Gastwirth J.L., Krieger A.M. & Rosenbaum P.R. (1998) Dual and simultaneous sensitivity analysis for matched pairs. *Biometrika* **85**, 907–920.

Linden A., Adams J. & Roberts N. (2003a) An assessment of the total population approach for evaluating disease management program effectiveness. *Disease Management* **6** (2), 93–102.

Linden A., Adams J. & Roberts N. (2003b) Evaluating disease management program effectiveness: an introduction to time series analysis. *Disease Management* **6** (4), 243–255.

Linden A., Adams J. & Roberts N. (2003c) Evaluation methods in disease management: determining program effectiveness. Position Paper Commissioned by the Dis-

ease Management Association of America (DMAA). Available at: http://www.dmaa.org/PDFs/Evaluation_Methods_in_DM.pdf

Linden A., Adams J. & Roberts N. (2004a) Using an empirical method for establishing clinical outcome targets in disease management programs. *Disease Management* **7** (2), 93–101.

Linden A., Adams J. & Roberts N. (2004b) The generalizability of disease management program results: getting from here to there. *Managed Care Interface* **17** (7), 38–45.

Linden A., Adams J. & Roberts N. (2004c) Evaluating disease management program effectiveness: an introduction to survival analysis. *Disease Management* **7** (3), 180–190.

Linden A., Adams J. & Roberts N. (2004d) Evaluating disease management program effectiveness adjusting for enrollment (tenure) and seasonality. *Research in Healthcare Financial Management* **9** (1), 57–68.

Linden A. & Roberts N. (2004e) Disease management interventions: what's in the black box? *Disease Management* **7** (4), 275–291.

Linden A. & Roberts N. (2005) A users guide to the disease management literature: recommendations for reporting and assessing program outcomes. *American Journal of Managed Care* **11** (2), 81–90.

Linden A., Adams J. & Roberts N. (2005a) Evaluating disease management program effectiveness: an introduction to the bootstrap technique. *Disease Management and Health Outcomes* (in press).

Linden A., Adams J. & Roberts N. (2005b) Using propensity scores to construct comparable control groups for disease management program evaluation. *Disease Management and Health Outcomes* **13** (2), 107–127.

Linden A., Adams J. & Roberts N. (2005c) Evaluating disease management program effectiveness: an introduction to the regression-discontinuity design. *Journal of Evaluation in Clinical Practice* doi:10.1111/j.1365-2753.2005.00573.x.

Linden A. (2005d) Measuring diagnostic and predictive accuracy in disease management: an introduction to receiver operating characteristic (ROC) analysis. *Journal of Evaluation in Clinical Practice* doi:10.1111/j.1365-2753.2005.00598.x.

Manski C. (1990) Nonparametric bounds on treatment effects. *American Economic Review* **80** (2), 319–323.

Rosenbaum P. (1987) Sensitivity analysis for certain permutation tests in matched observational studies. *Biometrika* **74**, 13–26.

Rosenbaum P. (1988) Sensitivity analysis for matching with multiple controls. *Biometrika* **75**, 577–581.

Rosenbaum P. (1989) Sensitivity analysis for matched observational studies with many ordered treatments. *Scandinavian Journal of Statistics* **16**, 227–236.

Rosenbaum P. (1991) Sensitivity analysis for matched case-control studies. *Biometrics* **47**, 87–100.

Rosenbaum P. (2002) *Observational Studies.* 2nd Edn. Springer, New York, NY.

Rosenbaum P. & Krieger A. (1990) Sensitivity analysis for two-sample permutation inferences in observational studies. *Journal of the American Statistical Association* **85**, 493–498.

Schlesselmann J.J. (1978) Assessing the effects of confounding variables. *American Journal of Epidemiology* **108**, 3–8.

Wilcoxon F. (1945) Individual comparisons by ranking methods. *Biometrics* **1**, 80–83.

Wilson T. & MacDowell M. (2003) Framework for assessing causality in disease management programs: principles. *Disease Management* **6**, 143–158.

Appendix I

Sub-routine to derive exact ω-value level where upper bound *P*-value = 0.05

SampleSize = Number of matched pairs
TScore = T of positive sign-ranked values

NewOdds = 1 (starts the loop with ω = 1)

Do Until Phigh ≥0.049

LowP = 1/(1 + NewOdds)
HighP = NewOdds/(1 + NewOdds)
ELow = LowP × ((SampleSize × (SampleSize + 1))/(2)
EHigh = HighP × ((SampleSize × (SampleSize + 1))/(2)
VarLow = (LowP × (1 – LowP)) × (SampleSize × ((SampleSize + 1) × (2 × SampleSize + 1))/6)
VarHigh = (HighP × (1 – HighP)) × (SampleSize × ((SampleSize + 1) × (2 × SampleSize + 1))/6)
DevLow = (Tscore – ELow)/Sqr(VarLow)
DevHigh = (Tscore – EHigh)/Sqr(VarHigh)
PLow = 1 – NormSDist(DevLow)
PHigh = 1 – NormSDist(DevHigh)
NewOdds = NewOdds + 0.001

Loop

Appendix II

Most of the non-parametric statistical tests necessary to conduct sensitivity analyses for hidden bias are readily found in most statistical software packages. The analysis reported in this paper using Wilcoxon signed-rank test statistic was initially generated using XLStat (Addinsoft 2004) for Excel. This is an Excel add-in, similar to the data analysis package that comes built-in to the program. Therefore, users familiar with Excel will find this program easy to use without much instruction. However, there are also several stand-alone programs to choose from, and depending on how many additional functions the analyst requires, the costs vary tremendously.