International Journal of Technology Assessment in Health Care, 30:3 (2014), 325–332.

doi:10.1017/S026646231400021X

EXPLORING TWO COST-ADJUSTMENT METHODS FOR SELECTION BIAS IN A SMALL SAMPLE: USING A FETAL CARDIOLOGY DATASET

Hema Mistry
Warwick Medical School, University of Warwick

Objectives: In economic evaluations of healthcare technologies, situations arise where data are not randomized and numbers are small. For this reason, obtaining reliable cost estimates of such interventions may be difficult. This study explores two approaches in obtaining cost estimates for pregnant women screened for a fetal cardiac anomaly.
Methods: Two methods to reduce selection bias in health care: regression analyses and propensity scoring methods were applied to the total mean costs of pregnancy for women who received specialist cardiac advice by means of two referral modes: telemedicine and direct referral.
Results: The observed total mean costs of pregnancy were higher for the telemedicine group than the direct referral group (4,918 versus 4,311 GBP). The regression model found that referral mode was not a significant predictor of costs and the cost difference between the two groups was reduced from 607 to 94 GBP. After applying the various propensity score methods, the groups were balanced in terms of sizes and compositions; and again the cost differences between the two groups were smaller ranging from -62 (matching "by hand") to 333 GBP (kernel matching).
Conclusions: Regression analyses and propensity scoring methods applied to the dataset may have increased the homogeneity and reduced the variance in the adjusted costs; that is, these methods have allowed the observed selection bias to be reduced. I believe that propensity scoring methods worked better for this dataset, because after matching the two groups were similar in terms of background characteristics and the adjusted cost differences were smaller.

Keywords: Costs and cost-analysis, Congenital heart disease, Pregnancy, Nonrandomized

In economic evaluations of healthcare technologies, situations commonly arise where there are no randomized data and only a small number of observations exist. Estimates of the likely cost-effectiveness of such interventions may still be desirable to inform resource allocation decisions. This study presents two approaches to deal with situations where reliable evidence on cost-effectiveness of such interventions is difficult to obtain because of the small numbers of available patients and the observational nature of the data. The study focuses on cost estimates obtained from a nonrandomized study, which evaluated two referral methods for obtaining specialist advice in the field of fetal cardiology.

In the United Kingdom, it is estimated that there are approximately 4,600 babies born each year with congenital heart disease (CHD); 1 in every 145 births (1). Of these, the estimated incidence of complex CHD is 1.5 per 1,000 live births; moderate CHD the incidence is 0.9 per 1,000 live births; and for simple CHD the incidence is 4.5 per 1,000 live births (1). Fetal cardiology is concerned with the diagnosis and management of pregnant women with a fetal heart anomaly. In England and Wales, almost all women at approximately 18–20 weeks in their pregnancy are screened by trained sonographers using an ultrasound anomaly scan to detect major congenital heart problems or other structural anomalies (2). Prenatal detection rates for fetal cardiac anomalies vary across the United Kingdom and have stayed around 23 percent (3;4). For the great majority of women, no heart defects are found. For those few women, detection of a heart defect at the anomaly scan allows parents to choose whether to terminate or to continue with the pregnancy.

As part of an economic evaluation of the role of telemedicine in pediatric and fetal cardiology, only Medway hospital in Gillingham used the telemedicine equipment for fetal cardiology (5). Telemedicine offers an alternative referral strategy for fetal cardiology and can lead to a significant decrease in time to diagnosis compared with sending a patient to a specialist hospital. The study and sample of women for this study have been described in detail elsewhere (5;6). In brief, the analysis covered all pregnant women who were referred for detailed fetal heart ultrasound examination with a perinatal cardiologist after

I am grateful to Dr Robin Dowie, Professor Martin Buxton, Professor Stephen Morris and Dr Karla Hemming for all their assistance and in reading earlier drafts of this manuscript. The original dataset and research formed part of the TelePaed project, which was funded by the Department of Health and the Charitable Funds Committee of the Royal Brompton and Harefield NHS Trust. This work formed part of my PhD thesis, which was undertaken at the Health Economics Research Group at Brunel University. The views expressed in the publication are mine and not necessarily those of the Department of Health. Disclosure of interests: None. Contribution to authorship: H.M. conducted all the economic analyses and wrote the study.

a routine anomaly scan. Sonographers and obstetricians decided on how the women would be assessed by a specialist. The referred women formed two groups: a telemedicine group where a prerecorded videoed anomaly scan was relayed to the specialist in the absence of the women and a direct referral group where women were seen face-to-face for a detailed assessment by the specialist. Women were followed up from time of the anomaly scan until they delivered or in a few cases, after termination of pregnancy (5).

In observational studies, such as this evaluation, the assignment of women to the telemedicine and direct referral groups is not random. As a result, the costs which have been evaluated may be biased and these cost differences may be due to preexisting differences between the groups rather than the intervention itself. A range of methods exist to eliminate or reduce this bias. This study presents two of these analytical methods to obtain more reliable cost estimates for these two groups of women.

METHODS

Selection Bias in Healthcare

Selection bias refers to systematic differences in comparison groups (7). Selection bias arises as a result of the interaction of treatments and omitted or unobserved patient characteristics that may influence treatment choice, but independently affect health outcomes: in other words, the participants in the intervention group have different characteristics from those allocated to the control group (and these differences affect outcomes) (6).

Despite the growing use of nonrandomized studies to evaluate healthcare technologies, there is currently no "gold standard" approach to control for selection bias in nonrandomized studies. This study presents two of these methods for obtaining cost estimates: regression analyses and propensity scoring methods.

Regression Analyses

As cost data were skewed a generalized linear regression model was used to examine the relationship between total costs of pregnancy and referral mode, controlling for other variables which were prespecified as important variables for determining costs of pregnancy: age, parity, gestation in weeks at anomaly scan, pregnancy duration which takes into account the length of pregnancy (8), and cardiac risk factors such as diabetes, Down's syndrome, epilepsy, and family history of CHD (9). The link test was used to check whether the model is well specified.

Propensity Scores

Rosenbaum and Rubin defined a propensity score as, "the conditional probability of assignment to a particular treatment given a vector of observed covariates" (10). In other words, a probability model is fitted to predict the likelihood that women are assigned to the telemedicine group, compared with the direct referral group, based on the values of the observed variables. A propensity score summarizes all the background covariates for each woman into a single-index variable (the propensity score). The generation of a single score for each women in each group allows an assessment of whether the background variables are sufficiently similar (that is they overlap). When such overlap is present, the propensity score approach allows calculation of the estimated treatment versus control effects that reflect adjustment for differences in all observed background characteristics (11). So the propensity score is a balancing score, that is, the telemedicine and direct referral groups have similar distributions on all observed covariates, as in a randomized experiment, and so observed selection bias is removed when comparisons are made between groups with the same propensity scores (6).

(a) Estimating propensity scores. Propensity scores were estimated using the "pscore" program (12), which uses a logit model to generate a propensity score for each women. The dependent variable was the method of referral to specialist (telemedicine = 1 and direct referral = 0) and the independent variables were: age, parity, gestation in weeks at anomaly scan, pregnancy duration, and cardiac risk factors (diabetes, Down's syndrome, epilepsy, and family history of CHD).

(b) Applying propensity scores using the "pscore program." Once calculated, propensity scores were applied in three ways to help reduce bias:

(I) Matching method - orders the telemedicine and direct referral groups by propensity scores and then matches each patient who receives telemedicine to a direct referral patient with a similar propensity score. The estimated propensity scores can be used to obtain estimates of the average treatment effect on the treated (ATT) using various methods: (i) *nearest neighbor matching*: if a match happens to be equally good as determined by the Stata program, then there are two feasible options: the *random draw* program randomly draws either the forward or backward matches; whereas, the *equal weights* program gives equal weight to the groups of forward and backward matches (12); (ii) *radius matching*, each telemedicine patient is matched only with the direct referral patients whose propensity score falls in a predefined radius (12); (iii) *kernel matching*, all telemedicine patients are matched with a weighted average of all direct referral patients with weights that are inversely proportional to the distance between the propensity scores of telemedicine and direct referrals (12); and (iv) *stratification matching*, consists of dividing the range of variation of the propensity score in intervals, so that within each interval telemedicine and direct referral patients have on average the same propensity score (12).

(II) Stratification method - this consists of grouping subjects into strata, so each stratum contains patients from both groups, determined by observed background covariates. Once the strata are defined (based on the propensity scores), telemedicine and direct referral patients who are in the same group or stratum are compared directly (13); and

https://doi.org/10.1017/S026646231400021X Published online by Cambridge University Press

Table 1. Results from the Regression Analyses

	Coefficient	Standard error	z statistic	p-value
Generalized linear model: Total costs of pregnancy				
Referral mode	94.16	420.50	0.22	0.823
Mother's age	17.98	29.65	0.61	0.544
Gestation	− 37.44	101.29	− 0.37	0.712
Pregnancy duration				
1. Termination (set as base case)	–	–	–	–
2. Pre-term (babies born before 37 weeks)	4179.46	796.19	5.25	< 0.001
3. Full-term (babies born between 37 and 42 weeks)	3750.74	739.94	5.07	< 0.001
Parity	− 411.09	403.10	− 1.02	0.308
Diabetes	1375.66	577.60	2.38	0.017
Downs	− 545.45	687.03	− 0.79	0.427
Epilepsy	− 430.69	603.23	− 0.71	0.475
Family history	− 639.18	497.29	− 1.29	0.199
Constant	1758.14	2485.85	0.71	0.479

(III) Regression method, the propensity scores are then added as an independent variable into the regression model, along with the other independent variables. Including propensity scores, as a covariate takes into account the likelihood for treatments and the component of correlation which is due to the assignment process can be eliminated (14).

(c) Using propensity scores to match "by hand." Using the "pscore program" not all patients *we*re matched. So the task of matching telemedicine cases to direct referral cases was undertaken (by hand) using the nearest neighbor approach without replacement (5, 15) for the estimated propensity scores. For this approach, all patients are first sorted by their estimated propensity score, and then matching for telemedicine patients is done by hand by searching forward and backward for the direct referral patient(s) with the closest score. Rather than matching each patient on a one-to-one basis, which would have resulted in losing information from at least 28 patients from the telemedicine cohort, a one-to-many, or many-to-one, matching approach was used (16). This approach meant that a telemedicine patient could be matched to more than one direct referral patient with a similar propensity score or a direct referral patient could be matched to more than one telemedicine patient. If no exact match was found for a patient (i.e., a propensity score to 5 decimal places), then matching was based within 4 decimal places, then within 3 decimal places, then within 2 decimal places and finally within 1 decimal place, similar to an approach used by Gum and colleagues (17).

Statistical Analysis

Detailed information on resource use, cost data collection and methods have been presented in detail elsewhere (5;6). Cost data have been adjusted to 2009/2010 prices using UK Hospital and Community Health Service indices (18). As the cost data were skewed, nonparametric bootstrapping was used whereby the distribution of costs are generated by repeated sampling of the data (to stabilize the mean and to generate confidence intervals around the mean estimates), with replacement and, in the absence of any other data from the population, gives a guide to its distribution (19). Bootstrapping was performed by taking 5,000 iterations of the data. All statistical analyses were conducted in Stata version 10 (20) and a p-value $\leq .05$ was considered to be statistically significant for the comparative analyses.

RESULTS

During the period 1st May 2001 to 31st July 2002, a total of 76 pregnant women were referred for specialist opinion following a routine anomaly scan: 52 (68.4 percent) were assessed by means of the telemedicine link, and 24 women saw a specialist in London. The overall total mean costs of pregnancy (which consist of the costs of antenatal care from the time of the anomaly scan and also maternal delivery costs [except for the few women who had a termination of pregnancy]) for the telemedicine group were higher than direct referral group (4,918 versus 4,311 GBP), a difference of 607 GBP which was not significant ($p = .202$).

Regression Analysis

Table 1 shows the generalized linear regression model results to examine the relationship between costs and referral mode controlling for all other variables. Pregnancy duration and diabetes were significant predictors of the total costs of pregnancy. Women assessed by telemedicine had higher costs than direct

https://doi.org/10.1017/S026646231400021X Published online by Cambridge University Press

Table 2. Propensity Score Logistic Regression Model

	Odds ratio	Standard error	z statistic	*p*-value
Pseudo $R^2 = 0.223$, Likelihood ratio $x_9^2 = 21.13$, p = 0.012				
Mother's age	0.9505	0.0484	− 1.00	.318
Gestation	0.5899	0.1555	− 2.00	.045
Pre-term birth	4.2297	5.4104	1.13	.260
Full-term birth	11.5210	14.6854	1.92	.055
Parity	2.2990	1.5599	1.23	.220
Diabetes	1.0032	1.0331	0.00	.998
Downs	0.1317	0.1601	− 1.67	.095
Epilepsy	3.6454	4.8615	0.97	.332
Family history	0.5296	0.4803	− 0.70	.483

referral women (an extra 94 GBP); this is in the same direction as the observed cost results, but of a much smaller magnitude and is not significant. The *p*-value from the link test was not significant ($p = .743$), indicating that the regression model was well specified.

Propensity Score Analysis

Propensity scores were estimated using a logit model (see Table 2). The results showed that only gestation in weeks at time of anomaly scan was a significant predictor in the calculation of propensity scores ($p \leq .05$). The Hosmer-Lemeshow test was not significant ($x_8^2 = 12.92, p = .115$), indicating a good fit for the logistic model. Figure 1 shows that there is some overlap in the estimated propensity scores. The two lines represent the distribution for each group and the histogram shows the combined distribution for the two groups in terms of estimated propensity scores.

(i) *Propensity score matching using "pscore" program.* Supplementary Table 1, which can be viewed online at www.journals.cambridge.org/thc2014xxx, shows using either nearest neighbor matching method, fifty-two telemedicine patients have been matched to fourteen direct referral patients. Both nearest neighbor methods gave the same ATT results, but the 95 percent confidence intervals are different. For the other three methods, fifty-two telemedicine patients have been matched to nineteen direct referral patients. The ATT suggests that by using either of the nearest neighbor matching methods, those patients in the telemedicine group had higher costs (176 GBP higher) compared with patients in the direct referral group; however, the 95 percent confidence interval for the nearest neighbor random draw highlights that the costs can be anywhere between -1,395 and 1,236 GBP. With the other three methods the costs were again higher for the telemedicine group, ranging from 206 (stratification matching) to 333 GBP (kernel matching). Again, there was inherent uncertainty in the different matching methods, as all the 95 percent confidence intervals were wide (although these differences were nonsignificant).

(ii) *Propensity score stratification.* After propensity score stratification (five blocks were used for stratification), the telemedicine group had higher mean costs than the direct referral group (4,298 versus 4,166 GBP) and the difference in mean costs for the two groups after propensity score stratification was reduced from 607 to 132 GBP, and this cost difference was not statistically significant ($p = .855$). For the telemedicine group, the mean cost was much lower than observed mean cost (a difference of 620 GBP: 4,918 versus 4,298 GBP).

(iii) *Propensity score regression method.* With the addition of the propensity scores as an independent variable in the regression model only preterm birth and diabetes were significant predictors of the total costs of pregnancy (see Supplementary Table 2, which can be viewed online at www.journals.cambridge.org/thc2014xxx). Women assessed by telemedicine had higher costs than direct referral women (an extra 64 GBP). This is similar to the results from the generalized linear model (see Table 1). The *p*-value from the link test was again not significant ($p = .771$), indicating that the regression model was well specified.

(iv) *Propensity score matching "by hand".* All fifty-two telemedicine patients were matched to twenty-four patients who were seen by direct referral. After weightings were applied, the number of cases in each group was thirteen. To make a direct comparison between the telemedicine case(s) and the direct case(s), for example, for the total costs of pregnancy for the patients forming the cases, they were adjusted in accordance with their "weights". So if two telemedicine patients formed a case, then the cost for each patient was multiplied by the weight (0.5) and summed together to obtain a final cost per case. After propensity score matching, the telemedicine group had lower mean costs than the direct referral group (4,527 versus 4,589 GBP) and the difference in mean costs for the two groups was reduced to -62 GBP (observed incremental cost was 607 GBP), and again this cost difference was not statistically significant ($p = .908$).

Comparison of Results

Table 3 shows a comparison of the difference in mean results (95 percent confidence intervals) obtained from the various methods. The observed results showed that telemedicine had higher costs than the direct referral group (607 GBP higher). After applying all the methods (except propensity score matching by hand), the costs for the telemedicine group were still higher than the direct referral group, but of a much smaller magnitude. The results from the generalized linear model showed that the cost difference between the two groups was reduced to 94 GBP. Using the various propensity score matching methods to estimate the ATT, the cost differences between the two groups ranged from 176 to 333 GBP. These differences may be due to the way the matching method works and the number of cases they select for matching. For example, using the "pscore" program with the nearest neighbor matching all telemedicine cases are found a match, so in essence some of these matches may be poor; whereas, with the stratification method a telemedicine case may not be included if there are no direct referral cases to match with (12). Both the propensity score stratification and regression adjustment results showed that the cost differences between the two groups were much smaller than the difference between the observed costs. Finally, using the estimated propensity scores and matching "by hand" to use "all patients," again showed that the mean difference in costs between the two groups was reduced and telemedicine group costs were lower than the direct referral groups' costs. All methods applied to this dataset reduced the cost differences between the two groups. This

https://doi.org/10.1017/S026646231400021X Published online by Cambridge University Press

Table 3. Comparison of results from the different methods

Results	Difference in total costs of pregnancy	95% Confidence intervals
Observed	607 GBP	−202 to 1,520 GBP
Regression	94 GBP	−730 to 918 GBP
Propensity score matching ('pscore') – ATT (difference)		
· Nearest neighbour – random draw	176 GBP	−1,395 to 1,236 GBP
· Nearest neighbour – equal weights	176 GBP	−793 to 1,505 GBP
· Kernel	333 GBP	−1,052 to 1,535 GBP
· Stratification	206 GBP	−725 to 1,271 GBP
· Radius	299 GBP	−502 to 1,306 GBP
Propensity score stratification	132 GBP	−1,365 to 1,286 GBP
Propensity score regression	64 GBP	−767 to 895 GBP
PS matching 'by hand'	−62 GBP	−1,150 to 941 GBP

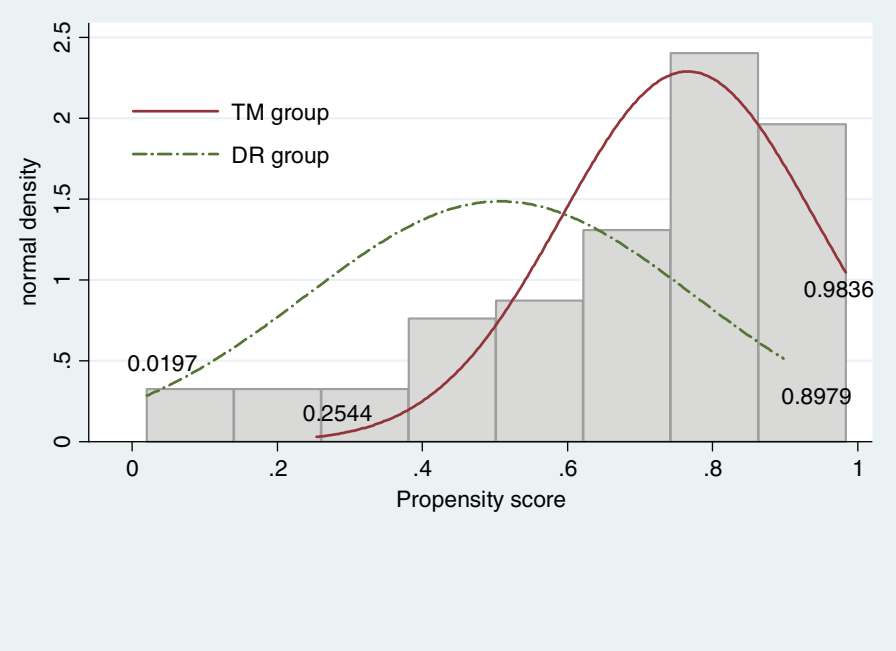

Figure 1. Estimated propensity scores.

indicated these methods may have increased the homogeneity and reduced the variance in the adjusted costs; that is, these methods may have accounted for the observed selection bias between the two groups for this dataset.

DISCUSSION

To obtain unbiased estimates of cost differences, large, adequately powered randomized controlled trials are needed. However, this is not always possible. This study has contrasted two methods that might be used to obtain reliable cost estimates within a nonrandomized study with a small sample size. First, a generalized linear model was used to see whether referral mode is a significant predictor of costs. Second, the various propensity score methods were used to balance the sizes and compositions of the two groups to reduce the element of bias in the estimation of costs.

The observed cost differences between the telemedicine group and the direct referral group in the costs may have been partly due to the additional cost of a teleconsultation (the cost difference between a specialist scan in London and a teleconsultation was approximately 90 GBP); also, some of the women in the telemedicine group were scanned earlier in the second trimester of the pregnancy, and their antenatal care over the remaining months would include one or two extra antenatal visits; and finally, four women in the direct referral group had

https://doi.org/10.1017/S026646231400021X Published online by Cambridge University Press

terminations compared with only two women in the telemedicine group, for these women the cost of their antenatal care would have been lower.

The analysis cannot prove which of these methods is more accurate for this dataset, but I believe that propensity score matching may be a more reliable way of obtaining cost estimates, because after matching the groups were similar in terms of background characteristics (i.e., "balanced"). I cannot be sure with the regression method, whether the covariates were balanced among the groups; hence the two groups may not be similar. Regression models can indicate differences in costs between a dependent variable (e.g., referral method) and other covariates, whereas the propensity score technique cannot indicate differences between the dependent variable and individual covariates, because all covariates are collapsed into a single index variable, possibly obscuring important interactions. An advantage of using propensity score matching is that matching does not have to assume linearity (i.e., assume a constant relationship between an outcome and the covariate within each treatment group), whereas regression analyses do (21). However, both regression analyses and propensity score methods only controlled for observed variables and not for unobserved variables.

Instrumental variables analysis and sample selection models are two other techniques that can reduce selection bias in both observed and unobserved differences. First, the instrumental variable technique aims to find a variable or variables that have two essential properties: (a) the instrumental variable should be statistically correlated with the treatment variable; the higher the correlation, the better the instrument; and (b) the instrument should be uncorrelated with the outcome (or error term). The instrumental variable is a device that aims to achieve pseudo randomization, that is, the instrument assigns subjects to either to the intervention or control using an assignment mechanism that is independent of outcome. For example, McClellan and colleagues applied this technique to assess whether more aggressive use of invasive cardiac procedures improved outcomes for elderly patients with acute myocardial infarction (22). The instrumental variable in this case was the differential distance. They concluded that there was a lower mortality rate among elderly patients who received catheterization than among those treated more conservatively.

Second, sample selection models are conducted in two stages (23). In the first stage, a probit model of treatment selection is estimated. The estimated probabilities from this model, are used to calculate an "adjustment factor" for each patient, which is the probability of not receiving the treatment given that the individual was "at risk" of receiving the treatment. In the second stage, the outcome of interest is predicted and the adjustment factor is included as one of the independent variables in the outcome model. The adjustment factor permits a direct test of whether selection bias is present and if so, what the direction of its impact is (24). For example, Crown et al. applied this method to estimate the effects of alternative antidepressant therapies on a variety of cost measures (25). They concluded if selection bias was not controlled for, the cost estimates in the expenditure equation would have been biased.

Neither method was appropriate for this dataset, as the dataset did not have any variables that satisfied the two main conditions required for a valid instrumental variable estimation and also in relation to the sample selection method, there were no good "identifier" variables; that is, variables that could be used as proxy variables for unobserved variables which may have caused selection bias in this dataset.

Even though the dataset maybe classed as "old," the referral numbers for the two groups over the last 10 years have been very similar (6) and the dataset has purely been used for illustration purposes to demonstrate the application of these methods to a small, nonrandomized dataset. Furthermore, the analysis was confined to patient-related observed variables that were recorded routinely in hospital records. The variables included in this dataset were the most relevant, in terms of identifying women with elevated risk factors for fetal CHD and were critical in helping the clinicians to decide whether a patient was assessed by means of telemedicine or by direct referral. There may be other characteristics which were not recorded in this particular dataset such as body mass index, social class, education, income, ethnicity, or smoking which may affect the cost results. Other unobserved variables which may have a possible impact on costs, but were not included in the dataset include: a patient's and/or a clinician's preference for referral method and the quality of care which the patient receives at the specialist or local hospital (6).

Shah and colleagues conducted a systematic review to determine whether propensity scores gave different results from regression modeling when adjusting for bias in observational studies (26). They found that both methods produced similar results, although propensity scores gave slightly weaker associations. However, many of the reviewed studies did not implement propensity scores well. Deeks et al. considered different methods (regression, stratification and propensity scoring) for evaluating selection bias in nonrandomized studies and concluded that none of the methods which were applied successfully removed bias in cohort studies (27). They found that most methods applied to reduce selection bias were not standardized and also some covariates are sometimes missing (not at random) which in itself can also lead to bias. Cepeda et al. compared propensity scores with regression models, and found that propensity scores performed better in situations with less than eight cases per covariate (28). Peduzzi et al. stated that usually 10 cases per covariate are required as a minimum for stable estimates in regression models (29). Apart from this specific condition there is little, if any practical guidance for researchers regarding when the use of propensity scores will produce different, and in particular, better estimates compared with regression models (6).

https://doi.org/10.1017/S026646231400021X Published online by Cambridge University Press

One of the main limitations of this analysis is that all the models were conducted on a small sample size; however, in practice these methods are usually applied on bigger sample sizes.

For example, propensity scores work better in larger sample sizes, because achieving an overlap between the two groups in terms of observed characteristics increases with the sample size (11;30). However, it is worth noticing that using matching "by hand" for larger datasets can be difficult and a time-consuming method, and the difficulty grows with the increased number of variables in a dataset. Matching by hand is also subjective and error-prone. Therefore, to address these issues propensity score matching by means of a computer program may be more efficient and effective.

The small sample size may have also created an additional problem for the propensity score matching. Small sample sizes can increase the variance of estimated effects (as seen by the large confidence intervals), and are considered to have a low statistical power, making the identification of significant differences in health outcomes between two groups more difficult; due to these reasons, the interpretation of the tests statistics should also be treated with caution. Also, fewer matches may be available, therefore, by picking distant matches increases the variance.

The small sample size of the dataset and the exclusion of variables such as health status, ethnicity and clinician's choice of referral mode which were not included in the original dataset may have affected the precision of the cost estimates. Nevertheless, there is some confidence in the adjusted cost estimates, as they reduced the difference (incremental) in costs between the two groups (6).

CONCLUSIONS

After adjusting for selection bias, the adjusted cost differences were smaller than the observed differences between the two groups. This means that reviewing the literature from nonrandomized studies and also from small studies, where both types of studies have not been adjusted for selection bias, the results from these studies should be interpreted for caution. This is because the studies may not tell us whether selection bias was present in the dataset and what the direction of the bias is. I believe that the propensity scoring methods worked better for this dataset, because after propensity score matching, the two groups were similar in terms of background characteristics and the adjusted cost differences were smaller. After all, there is no method to check after adjustment for the regression method whether the groups were balanced. Even though the dataset for this patient cohort comes from the United Kingdom, the methods used to account for selection bias are still generalizable to other settings where studies are not randomized and are also based on a small number of observations.

SUPPLEMENTARY MATERIAL

Supplementary Table 1:
http://dx.doi.org/10.1017/S026646231400021X
Supplementary Table 2:
http://dx.doi.org/10.1017/S026646231400021X

CONTACT INFORMATION

Hema Mistry, BA, MSc, PhD (Hema.Mistry@warwick.ac.uk), Assistant Professor in Health Economics, Warwick Evidence, Warwick Medical School, University of Warwick, Gibbet Hill Road, Coventry, UK, CV4 7AL

CONFLICTS OF INTEREST

No conflicts of interest.

REFERENCES

1. Petersen S, Peto V, Rayner M. *Congenital heart disease statistics 2003*. British Heart Foundation Health Promotion Research Group. Oxford: University of Oxford, 2003. http://www.bhf.org.uk/heart-health/statistics.aspx (accessed May 5, 2014).
2. NHS FASP. *NHS Fetal Anomaly Screening Programme 18_{+0} to 20_{+6} Weeks Fetal Anomaly Scan National Standards and Guidance for England*. Exeter: Royal College of Obstetricians and Gynaecologists; 2010.
3. Bull C. Current and potential impact of fetal diagnosis on prevalence and spectrum of serious congenital heart disease at term in the UK. *Lancet.* 1999;354:1242-1247.
4. Central Cardiac Audit Database. *Antenatal diagnosis*. Leeds, 2011. http://www.ccad.org.uk/002/congenital.nsf/vwContent/Antenatal%20Diagnosis?Opendocument (accessed May 5, 2014).
5. Dowie R, Mistry H, Young TA, Franklin R, Gardiner HM. Cost implications of introducing a telecardiology service to support fetal ultrasound screening. *J Telemed Telecare.* 2008;14:421-426.
6. Mistry H. *Economic issues associated with the operation and evaluation of telemedicine*. PhD Thesis, Brunel University, UK: 2011. http://bura.brunel.ac.uk/handle/2438/5830 (accessed May 5, 2014).
7. Cochrane Collaboration. Chapter 6: Assessment of study quality. In: Alderson P, Green S, Higgins JPT, eds. *Cochrane handbook for systematic reviews of interventions*. The Cochrane Library, Issue 1. Chichester: Wiley; 2004.
8. Steer P. The epidemiology of preterm labour. *BJOG.* 2005;112:1-3.
9. Mistry H, Dowie R, Young T, Gardiner H. The costs of maternity care for women with multiple pregnancy compared with high-risk and low-risk singleton pregnancy. *BJOG.* 2007;114:1104-1112.
10. Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. *J Am Stat Assoc.* 1984;79:516-524.
11. Rubin DB. Estimating causal effects from large data sets using propensity scores. *Ann Intern Med.* 1997;127:757-763.
12. Becker SO, Ichino A. Estimation of average treatment effects based on propensity scores. *Stata J.* 2002;2:358-377.
13. D'Agostino RB. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomised control group. *Stat Med.* 1998;17:2265-2281.
14. Rubin DB. Using multivariate matched sampling and regression adjustment to control for bias in observational studies. *J Am Stat Assoc.* 1979;74:318-328.
15. Dehejia RH, Wahba S. Propensity score-matching methods for non-experimental causal studies. *Rev Econ Stat.* 2002;84:151-161.

16. Dowie R, Young T, Mistry H, Weatherburn G. *Economic evaluation of the role of telemedicine in paediatric cardiology*. First report: Paediatric cardiology outpatient services. Final Report to the Department of Health. Uxbridge: Brunel University; 2003.
17. Gum PA, Thamilarasan M, Watanabe J, Blackstone EH, Lauer MS. Aspirin use and all-cause mortality among patients being evaluated for known or suspected coronary artery disease: A propensity analysis. *JAMA;* 2001;286:1187-1194.
18. Curtis L. Pay and prices index. *Unit Costs of Health and Social Care 2010*. Canterbury, UK: Personal Social Services Research Unit, University of Kent at Canterbury; 2010.
19. Manly BFJ. *Randomisation, Bootstrap and Monte Carlo methods in biology* (Texts in Statistical Science), 2nd ed. London: Chapman and Hall; 1997.
20. StataCorp. *Stata Statistical Software: Release 10.0*. College Station, TX: Stata Corporation; 2007.
21. Foster EM. Propensity score matching - An illustrative analysis of dose response. *Med Care.* 2003;41:1183-1192.
22. McClellan M, McNeil BJ, Newhouse JP. Does more intensive treatment of acute myocardial infarction in the elderly reduce mortality? Analysis using instrumental variables. *JAMA.* 1994;272:859-866.
23. Heckman J. Sample selection bias as a specification error. *Econometrica.* 1979;47:153-161.
24. Crown WH. Antidepressant selection and economic outcome: A review of methods and studies from clinical practice. *Br J Psychiatry.* 2001;179:S18-S22.
25. Crown WH, Obenchain RL, Englehart L, et al. The application of sample selection models to outcomes research: The case of evaluating the effects of antidepressant therapy on resource utilization. *Stat Med.* 1998;17:1943-1958.
26. Shah BR, Laupacis A, Hux JE, Austin PC. Propensity score methods gave similar results to traditional regression modelling in observational studies: A systematic review. *J Clin Epidemiol.* 2005;58:550-559.
27. Deeks JJ, Dinnes J, D'Amico, Sowden AJ, Sakarovitch C, Song F, et al. Evaluating non-randomised intervention studies. *Health Technol Assess.* 2003;7:iii-x, 1-173.
28. Cepeda MS, Boston R, Farrar JT, Strom BL. Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. *Am J Epidemiol.* 2003;158:280-287.
29. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. *J Clin Epidemiol.* 1996;49:1373-1379.
30. Pearl J. Section 11.3.5 Understanding propensity scores. *Causality: Models, reasoning and inference.* 2nd ed. New York: Cambridge University Press; 2009.

https://doi.org/10.1017/S026646231400021X Published online by Cambridge University Press