Parametric modelling of cost data: some simulation evidence

Andrew Briggs[a,*], Richard Nixon[b], Simon Dixon[c] and Simon Thompson[b]
[a] *University of Oxford, UK*
[b] *MRC Biostatistics Unit, Cambridge, UK*
[c] *University of Sheffield, UK*

Summary

Recently, commentators have suggested that the distributional form of cost data should be explicitly modelled to gain efficiency in estimating the population mean. We perform a series of simulation experiments to evaluate the usual sample mean and the mean estimator of a lognormal distribution, in the context of both theoretical distributions and three large empirical datasets. The sample mean is always unbiased, but is somewhat less efficient when the population distribution is truly lognormal. However the lognormal estimator can perform appallingly when the true distribution is not lognormal. In practical situations, where the true distribution is unknown, the sample mean generally remains the estimator of choice, especially when limited sample size prohibits detailed modelling of the cost data distribution. Copyright

Introduction

The appropriate analysis of cost data generated by clinical trials is problematic. While the usual outcome of interest is the population mean cost for a particular treatment, the distribution of cost data is generally highly skew because a few patients incur very large costs.

A number of commentators have made recommendations for the analysis of cost data. Briggs and Gray [1] have argued against the use of standard non-parametric methods for analysing cost data, precisely because of the necessary focus on mean cost. While they considered that there may be some merit in transforming cost data, they emphasised the importance of presenting cost estimates on the untransformed scale.

Thompson and Barber [2] however argued strongly that transformation of cost data is not appropriate. They recommended the use of the sample mean as an estimator, with confidence limits derived either from standard asymptotic theory, or from non-parametric bootstrapping of the sample mean.

In a recent contribution, O'Hagan and Stevens have criticised the unequivocal nature of Thompson and Barber's recommendations [3]. They argue that while the appropriate focus of cost analysis is clearly the population mean cost, where cost data are not normally distributed the sample mean is not necessarily the most efficient estimator. They argue that, if cost data truly follow a lognormal distribution, there will be efficiency gains from using $\exp(\mathrm{lm} + \mathrm{lv}/2)$ as an estimator of the mean cost in the population, where lm and lv are the log

*Correspondence to: Department of Public Health, Health Economics Research Centre, University of Oxford, Old Road Campus, Headington, Oxford OX3 7LF, UK. E-mail: andrew.briggs@dphpc.ox.ac.uk

Received 21 July 2003
Accepted 28 April 2004

scale sample mean and variance, respectively. While O'Hagan and Stevens used an example of apparently lognormally distributed costs, they are careful not to argue for making a lognormal assumption in general. Rather they argue that cost data should be appropriately modelled, leaving something of a question as to what that appropriate distribution might be.

Other commentators have been less circumspect. In particular, Zhou, in a series of articles [4–7], has focused attention on the use of estimators based on assuming that costs follow a lognormal distribution. From the reported applied examples, this assumption is linked to the failure to reject a null hypothesis of normality on the log scale [8], with no consideration of whether there are more appropriate distributions for cost data.

Although the issues related to estimating mean costs for health care interventions have recently come to the fore for health economists conducting economic evaluations alongside clinical trials, these issues are by no means new. Similar issues in the risk-adjustment literature were raised in relation to estimating costs in the RAND insurance experiment in the early 1980s [9,10], where log transformation was found to yield better compliance with modelling assumptions, but where prediction on the untransformed scale was the ultimate goal. This requirement led Duan to propose the 'smearing estimator' that could be used to correct for the bias that occurs if expectations on the transformed scale are back transformed [11]. In developing this initial framework, commentators such as Manning and Mullahy have also focused much attention on the use of transformation as a method for improving the estimation of mean cost relating to health care costs and expenditures [12–15].

The purpose of this paper is to evaluate two alternative estimators of population mean cost. First, we look at their relative performance under a number of assumed parametric distributions for cost. Second, we repeat the comparison in an empirical context using three datasets where large numbers of individual level costs are available. The aim is to highlight not only the potential efficiency gains to be obtained from choosing the appropriate estimator, but also any potential problems of choosing the incorrect estimator. We return to the link between the estimation of costs for cost-effectiveness analysis and the estimation of costs in the risk-adjustment literature in the discussion.

Simulating from parametric distributions

Two parametric distributions were employed to generate cost data. These were the lognormal and Gamma distributions, both of which are used in practice to model positively skewed cost data. The appropriate (maximum likelihood) estimator of the population mean of a Gamma distribution is the sample mean, whereas the appropriate estimator of the population mean of a lognormal distribution is $\exp(\mathrm{lm}+\mathrm{lv}/2)$, as described above. Estimation of a confidence interval based on the sample mean estimator is straightforward using standard asymptotic assumptions. Obtaining the confidence interval for the lognormal estimator is a non-trivial problem since it is a function of two transformed sample estimates. The method used in this paper is that outlined in a review paper by Zhou [8].

For each distribution, the population mean was set to be 1000 and five choices of coefficient of variation (CoV, the ratio of the standard deviation to the mean = 0.25, 0.5, 1.0, 1.5, 2.0) were used to define the parameters of the distribution. The resulting distributions (Figure 1) are plausible representations of cost data. These are the population distributions from which samples are drawn in the simulation exercise.

Samples of five different sizes ($n=20$, 50, 200, 500, 2000) were drawn from each distribution and for each coefficient of variation. This represented a total of 50 different simulation experiments. For each experiment, with 10 000 replications, the sample mean and the lognormal estimator were used to estimate the population mean. For each simulation, the bias (average estimate minus 1000) and coverage probability (proportion of 95% confidence intervals containing 1000) were calculated. In order to summarise both bias and precision, the root mean square error (RMSE) was calculated, which equals the square root of the mean $[\text{estimate} - 1000]^2$ in each simulation.

As expected, the RMSE reduces with decreasing coefficient of variation and increasing sample size, no matter which combination of underlying distribution or estimator is chosen (Table 1). When the data are truly from a lognormal distribution, both the lognormal estimator and the sample mean are unbiased, but the lognormal estimator is more precise – all of the cells of the lower-right matrix in Table 1 show a smaller

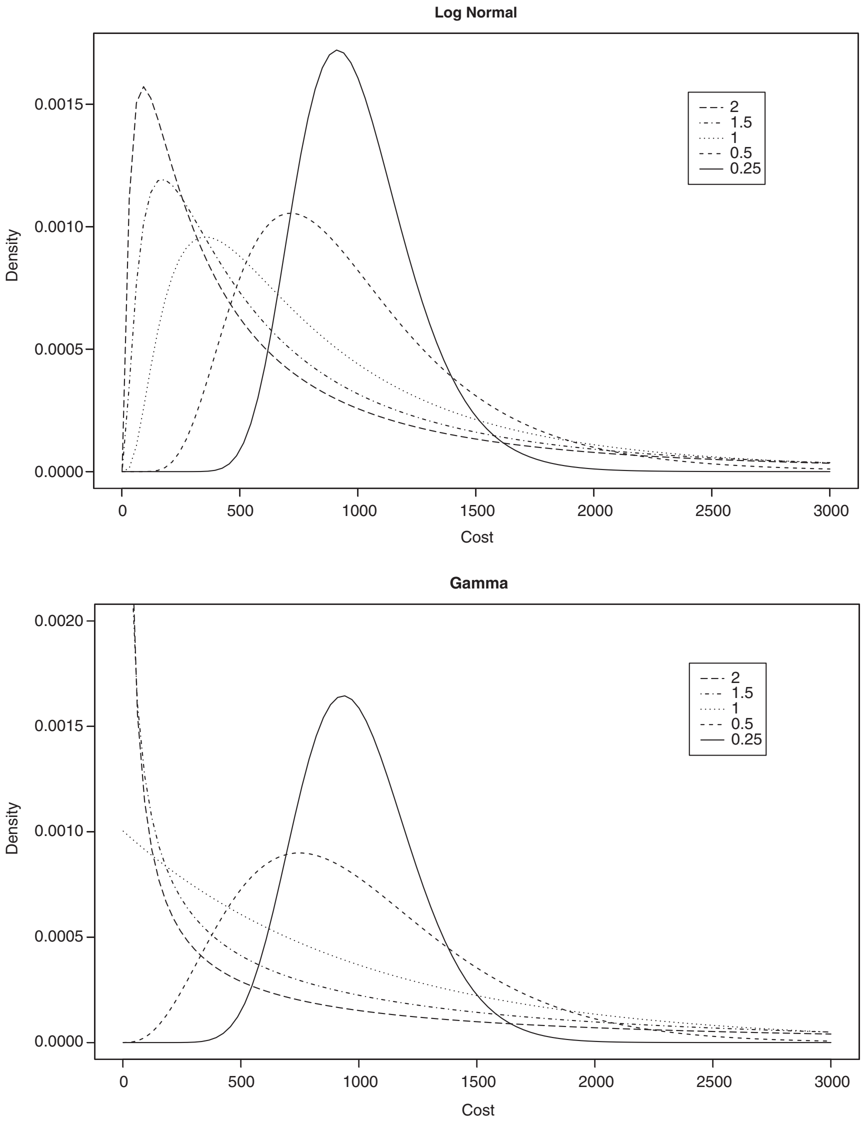

Figure 1. Parametric distributions used to represent population cost distributions: upper panel shows five lognormal distributions and lower panel shows five gamma distributions. Parameters of the distributions are set so that the mean cost is 1000 with coefficients of variation 0.25, 0.5, 1.0, 1.5 and 2.0

RMSE than those in the lower-left matrix. However, the gain in precision is only moderate; the ratio of RMSEs always remains greater than 84% in Table 1. By contrast, when a lognormal estimator is applied to data that follow a Gamma distribution, the results are disastrous. Although, for low coefficients of variation there is little to choose between the estimators, where the coefficient of variation is large, the lognormal estimator performs very poorly. This is because the lognormal estimator is severely biased in these cases, while the sample mean remains unbiased.

These results are echoed by the coverage probabilities of the confidence intervals for the different estimators (Table 2). Increasing coefficients of variation and decreasing sample size generally lead to poorer coverage (less than 95%).

Table 1. Estimated root mean squared error by underlying distribution and estimator for different sample sizes and coefficients of variation

		RMSE for sample mean estimator Simulation sample sizes					RMSE for exp(lm + lv/2) estimator Simulation sample sizes				
Distribution	CoV	20	50	200	500	2000	20	50	200	500	2000
Gamma	0.25	56	35	18	11	6	56	35	18	11	6
	0.50	112	71	35	22	11	114	73	38	25	16
	1.00	221	141	70	44	22	400	304	241	226	218
	1.50	333	214	105	67	34	1388	1097	925	896	878
	2.00	440	284	141	89	45	2663	1914	1510	1420	1378
Lognormal	0.25	56	36	18	11	6	56	36	18	11	6
	0.50	112	71	35	22	11	112	71	35	22	11
	1.00	224	141	72	45	23	221	137	69	43	22
	1.50	336	214	109	67	34	328	197	99	61	31
	2.00	450	288	143	63	45	419	250	122	54	38

RMSE—root mean squared error; lm—log mean; lv—log variance; CoV coefficient of variation.

Again the lognormal estimator exhibits improved performance over the sample mean if the data are truly lognormal. However, the misapplication of the lognormal estimator to Gamma distributed data leads to some very poor coverage results, again due to bias.

Simulations from observed cost datasets

Since it is unlikely that population cost distributions really follow a well-behaved functional form we repeat the comparison of estimators using three large datasets of patient-level cost data. These are treated as population distributions from which samples are drawn in the simulation. The three datasets used for empirical simulations are described below.

The CPOU data: These data were constructed for an economic evaluation of a Chest Pain Observation Unit operating within the A and E department of a single large teaching hospital [16]. Patients with acute chest pain, undiagnosed by clinical assessment, electrocardiogram and chest radiograph, were recruited and followed-up to 6 months. Nine hundred and seventy-two patients were recruited to the study. Total cost (2001/2002 prices) was constructed from the initial 6 hours of care, length of stay, parenteral drug therapy, diagnostic tests, reattendances and readmissions, outpatient attendances and cardiology procedures.

The IV fluids data: These cost data were constructed for a randomised controlled trial of two prehospital intravenous fluids protocols for paramedics in patients with serious trauma [17]. 1309 patients were entered into the study from two different Ambulance Trusts and costs were constructed for 1191 patients. Total cost (1997/1998 prices) was calculated for individual patients up to 6 months post-incident, and included ambulance costs, fluid costs, A and E costs, inpatient costs, and ambulatory care.

The Paramedics data: These cost data were constructed from a controlled study comparing ambulance technicians and paramedics in patients with serious trauma [18]. 1852 patients were entered into the study from three different Ambulance Trusts. Total cost (1996/1997 prices) was calculated for individual patients up to 6 months post-incident, and included ambulance costs, ambulance treatment costs, hospital costs, and ambulatory care.

Summary statistics of the per-patient total cost are presented in Table 3 for each example. All three datasets exhibit extreme skewness and kurtosis (compared to the values of 0 and 3 respectively for a normal distribution). Of interest is the coefficient of variation which shows that the standard deviation in the data is roughly twice the mean for all three datasets – this is the upper range of the coefficient of variation examined in the

Table 2. Estimated 95% confidence interval coverage probabilities by underlying distribution and estimator for different sample sizes and coefficients of variation

		Coverage for sample mean estimator Simulation sample sizes					Coverage for exp(lm + lv/2) estimator Simulation sample sizes				
Distribution	CoV	20	50	200	500	2000	20	50	200	500	2000
Gamma	0.25	0.93	0.94	0.95	0.95	0.95	0.93	0.94	0.96	0.96	0.95
	0.50	0.92	0.93	0.95	0.95	0.95	0.94	0.96	0.97	0.95	0.89
	1.00	0.90	0.93	0.95	0.95	0.95	0.97	0.97	0.69	0.18	0
	1.50	0.87	0.91	0.94	0.95	0.94	0.99	0.92	0.14	0	0
	2.00	0.83	0.89	0.93	0.94	0.95	0.99	0.93	0.20	0	0
Lognormal	0.25	0.92	0.93	0.95	0.95	0.95	0.91	0.93	0.95	0.95	0.95
	0.50	0.91	0.93	0.94	0.95	0.95	0.91	0.93	0.94	0.95	0.95
	1.00	0.87	0.91	0.93	0.94	0.95	0.90	0.93	0.95	0.95	0.95
	1.50	0.83	0.88	0.92	0.94	0.94	0.89	0.93	0.94	0.95	0.94
	2.00	0.80	0.86	0.91	0.92	0.94	0.88	0.92	0.94	0.95	0.95

lm—log mean; lv—log variance; CoV coefficient of variation.

previous simulation experiments based on parametric distributions. Taking logs of the cost in each of the datasets appears to make the distributions more normal (Table 4, Figure 2).

The simulation experiment (again with 10 000 replications) involved drawing values at random without replacement from the cost datasets in order to form samples of varying sizes ($n = 20$, 50, 200, 500). On the basis of these samples, the estimators of the population mean (sample mean, and exp(lm + lv/2)) and their corresponding confidence limits were calculated.

The RMSE and coverage probabilities (compared to the sample mean in all the data) are presented in Table 5. As expected, the RMSE decreases with increasing sample size. The RMSEs are similar for the two estimators at the lower sample sizes, but as the *Central Limit Theorem* takes effect, the sample mean becomes the more accurate. This is echoed in the coverage results where the rapid deterioration in the coverage probabilities for the lognormal estimator as sample size increases is striking. This serves to emphasise that while the *Central Limit Theorem* plays an important role in the validity of the sample mean as an estimator, increasing sample size is no guarantee of performance for other estimators based on parametric assumptions if those assumptions turn out not to hold.

Table 3. Summary statistics for per patient total cost in the three example datasets

	CPOU	IV Fluids	Paramedics
n	972	1191	1852
Mean	518	2693	4233
Sd	1145	7083	7961
Skewness	5.3	4.8	7.5
Kurtosis	37	32	88
CoV	2.2	2.6	1.9

CoV coefficient of variation.

Table 4. Summary statistics for the natural log of cost for the three example datasets

	CPOU	IV Fluids	Paramedics
n	972	1191	1852
Mean	5.37	6.51	7.70
Sd	1.19	1.32	1.09
Skewness	0.59	1.69	−0.05
Kurtosis	3.73	4.72	4.76
CoV	0.22	0.20	0.14

CoV coefficient of variation.

It is commonly considered that the *Central Limit Theorem* applies for samples with greater than 30 observations, whatever the distribution in the underlying population. However, this rule of thumb does not apply to non-symmetric distributions typical of cost data. Cochran's alternative guideline of $n > 25\eta^2$ for situations where the 'principal deviation from normality consists of marked positive skewness' [19] is based on η, the skewness coefficient in the sample. The guideline was devised such that a 95% confidence interval

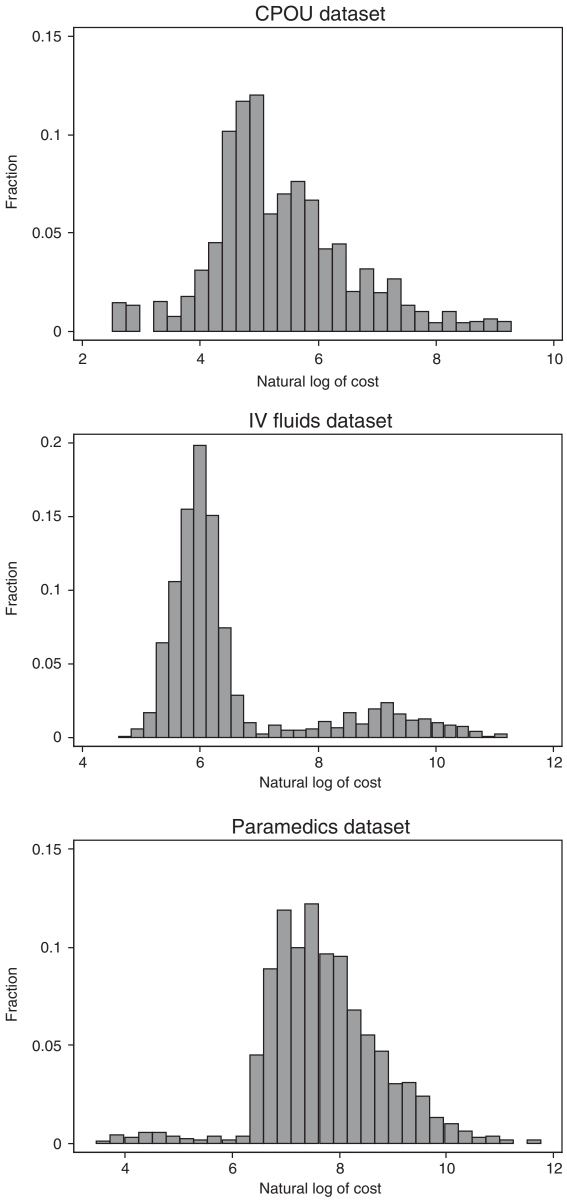

Figure 2. Histograms showing the distribution of log cost in each of the three example cost datasets

will have an error probability no greater than 6%. For the three example datasets this suggests that a minimum sample size of approximately 700 is required for the CPOU data, 560 for the IV fluids data and 1400 for the paramedics data. The results presented in Table 5 for the coverage probabilities in fact indicate the guideline to be rather conservative, since the sample mean has coverage close to 95% at sample sizes of 200, 200 and 500, respectively. Nevertheless, what is clear is that the $n > 30$ rule of thumb is totally inappropriate for cost data.

Comments and conclusions

It was George Box that famously quoted 'All models are wrong' [20] and Nester has added, in his *Applied Statistician's Creed* that 'No data are normally distributed' [21]. For cost data it is highly unlikely that parametric distributions are anything other than a simple approximation to the true distribution. The very construction of cost data – as a weighted sum of different resource counts – emphasises that total cost distributions are really mixtures of many other types of distribution. Even with moderately sized samples drawn from known distributions the form of that distribution can often not be reliably ascertained from the data alone.

The simulation experiments performed here confirm that when the appropriate distributional form of cost data is known, a degree of efficiency can be gained from using the estimator appropriate for that distribution. However, application of estimators based on incorrect parametric assumptions can lead to totally misleading conclusions.

The focus of the simulation experiment was the estimation of mean costs as might be generated alongside a clinical trial. However, in the risk-adjustment literature, similar issues have arisen and other commentators have also presented large scale simulation experiments to address the issue of what form of model/estimator is appropriate for the data. For example, Manning and Mullahy [12] presented a comprehensive simulation experiment to explore the use of log-transformation versus generalised linear modelling (GLM) of cost data. Their general conclusion was that the choice of modelling technique was an empirical issue and they provided an algorithm to help analysts choose the appropriate approach to their data. More recently Deb and Burgess have conducted similar experiments but using real-life health expenditure data and they conclude much more strongly that GLMs based on Gamma densities are more appropriate for risk-adjustment models of cost

Table 5. Simulation results based on drawing samples without replacement from three large datasets

		Sample mean estimator Simulation sample sizes				exp(lm + lv/2) estimator Simulation sample sizes			
Statistic	Dataset	20	50	200	500	20	50	200	500
RMSE	CPOU	253	160	70	37	234	141	95	87
	IV fluids	1583	1015	480	249	1721	1233	1090	1083
	Paramedics	1915	1131	585	321	1863	990	501	334
Coverage	CPOU	0.76	0.83	0.95	0.98	0.80	0.79	0.62	0.28
	IV fluids	0.77	0.86	0.94	0.98	0.60	0.47	0.12	0.00
	Paramedics	0.78	0.84	0.89	0.95	0.86	0.87	0.84	0.84

RMSE—root mean squared error; lm—log mean; lv—log variance; CoV coefficient of variation.

data than transformation of the dependent variable [22].

The link between what might be described as the risk-adjustment literature, where modelling of cost data is widespread, and the health economic evaluation literature is instructive. In general, the risk-adjustment literature has found that careful parametric modelling of cost data leads to more efficient estimators than simply using OLS (equivalent to the sample mean). However, in the risk-adjustment literature, there are typically many more data points (Deb and Burgess [22] estimated models based on 10,000–500,000 observations). In the health economic evaluation literature, where cost observations are generally collected from patients recruited to randomised controlled trials, the sample sizes are typically much less. It is within the context of such studies, where formal modelling of cost data is much more challenging, that we have tried to focus our simulation study. The real danger, as our simulations show, is the use of transformation based on parametric assumptions that cannot be sufficiently tested. For example, Zhou, in a review of analysing cost data generated alongside clinical trials presents a series of examples where he emphasises the use of the log transformation [8]. The only justification provided is failure to reject a hypothesis of normality on the log scale.

In empirical cost analysis for health economic evaluation, the true form of the cost distribution remains unknown. Overall, the sample mean performs well and is unlikely to lead to inappropriate inferences. Only when there are sufficient data to permit detailed modelling and choice of a suitable parametric distributions, is the use of other estimators of the population mean warranted.

Acknowledgements

A previous version of this paper was presented at the third Developing Economic Evaluation Methodology (DEEM) workshop on '*The analysis of cost data*' held at Aberdeen University in April 2003. We are grateful to comments from participants, in particular, Willard Manning and Paul McNamee. The paper was conceived at a '*Focus fortnight on cost data*' organised by the Centre for Bayesian Statistics in Health Economics at Sheffield University in February 2003 and we are indebted to Anthony O'Hagan for providing critical comment on early ideas. Finally, we are grateful to Hia Lee Lee for her assistance with the parametric simulations of cost data and to Steven Goodacre for making available one of the cost data sets for the simulations. All errors remain our own responsibility. Andrew Briggs holds a Public Health Career Scientist award from the UK Department of Health.

References

1. Briggs A, Gray A. The distribution of health care costs and their statistical analysis for economic evaluation. *J Health Services Res Pol* 1998; **3**(4): 233–245.
2. Thompson SG, Barber JA. How should cost data in pragmatic randomised trials be analysed? *Br Med J* 2000; **320**(7243): 1197–1200.
3. O'Hagan A, Stevens JW. Assessing and comparing costs: how robust are the bootstrap and methods based on asymptotic normality? *Health Econ* 2003; **12**: 33–49.

4. Zhou XH, Gao S, Hui SL. Methods for comparing the means of two independent log-normal samples. *Biometrics* 1997; **53**(3): 1129–1135.
5. Zhou XH, Melfi CA, Hui SL. Methods for comparison of cost data. *Ann Intern Med* 1997; **127**(8 Part 2): 752–756.
6. Zhou XH, Tu W. Interval estimation for the ratio in means of log-normally distributed medical costs with zero values. *Comput Statist Data Anal* 2000; **35**: 201–210.
7. Zhou XH, Melfi CA, Hui SL. Methods for comparison of cost data. *Ann Intern Med* 1997; **127**(8 Part 2): 752–756.
8. Zhou XH. Inferences about population means of health care costs. *Statist Methods Med Res* 2002; **11**: 327–339.
9. Manning WG. Health insurance and the demand for medical care: evidence from a randomized experiment. *Am Econ Rev* 1987; **77**: 251–277.
10. Duan N. A Comparison of alternative models for the demand for medical care. *J Business Econ Statist* 1983; **1**: 115–126.
11. Duan N. Smearing estimate: a nonparametric retransformation method. *J Am Statist Assoc* 1983; **78**: 605–610.
12. Manning WG, Mullahy J. Estimating log models: to transform or not to transform? *J Health Econ* 2001; **20**(4): 461–494.
13. Manning WG. The logged dependent variable, heteroscedasticity, and the retransformation problem. *J Health Econ* 1998; **17**(3): 283–295.
14. Mullahy J. Much ado about two: reconsidering retransformation and the two-part model in health econometrics. *J Health Econ* 1998; **17**(3): 247–281.
15. Mullahy J. What you don't know can't hurt you? Statistical issues, and standards for medical technology evaluation. *Med Care* 1996; **34**(12 Suppl): DS124–DS135.
16. Goodacre S, Nicholl J, Dixon S, Cross E, Angelini K, Arnold J *et al*. Cluster randomised controlled trial and economic evaluation of a chest pain observation unit versus routine care. *BMJ* 2004; **328**: 254–260.
17. Turner J, Nicholl J, Webber L, Cox H, Dixon S, Yates D. A randomised controlled trial of prehospital intravenous fluid replacement therapy in serious trauma. *Health Technol Assess* 2000; **4**(31): 1–57.
18. Nicholl J, Hughes S, Dixon S, Turner J, Yates D. The costs and benefits of paramedic skills in pre-hospital trauma care. *Health Technol Assess* 1998; **2**(17).
19. Cochran WG. *Sampling Techniques*. Wiley: New York, 1977.
20. Box GEP. Science and statistics. *J Am Statist Assoc* 1976; **71**: 791–799.
21. Nester MR. An Applied Statistician's Creed. *Appl Statist* 1996; **45**(4): 401–410.
22. Deb P, Burgess JF. *A Quasi-Experimental Comparison of Econometric Models for Health Care Expenditures*. Hunter College Department of Economics Working Papers, 212: 2003.