HEDG
HEALTH, ECONOMETRICS AND DATA GROUP

THE UNIVERSITY of York

WP 13/30

A quasi-Monte Carlo comparison of developments in parametric and semi-parametric regression methods for heavy tailed and non-normal data: with an application to healthcare costs

Andrew M. Jones, James Lomas, Peter Moore, Nigel Rice

October 2013

york.ac.uk/res/herc/hedgwp

A quasi-Monte Carlo comparison of developments in parametric and semi-parametric regression methods for heavy tailed and non-normal data: with an application to healthcare costs

Andrew M. Jones [a] James Lomas [a,b,*] Peter Moore [c] Nigel Rice [a,b]

[a] *Department of Economics and Related Studies, University of York, York, YO10 5DD, UK*
[b] *Centre for Health Economics, University of York, York, YO10 5DD, UK*
[c] *Oxford Outcomes, 688 W. Hastings Street, Suite 450, Vancouver, British Columbia, V6B 1P1, Canada*

October 14, 2013

Summary

We conduct a quasi-Monte Carlo comparison of the recent developments in parametric and semi-parametric regression methods for healthcare costs against each other and against standard practice. The population of English NHS hospital inpatient episodes for the financial year 2007-2008 (summed for each patient: 6,164,114 observations in total) is randomly divided into two equally sized sub-populations to form an estimation and a validation set. Evaluating out-of-sample using the validaton set, a conditional density estimator shows considerable promise in forecasting conditional means, performing best for accuracy of forecasting and amongst the best four (of sixteen compared) for bias and goodness-of-fit. The best performing model for bias is linear regression with square root transformed dependent variable, and a generalised linear model with square root link function and Poisson distribution performs best in terms of goodness-of-fit. Commonly used models utilising a log-link are shown to perform badly relative to other models considered in our comparison.

JEL classification: C1; C5

Key words: Health econometrics; healthcare costs; heavy tails; quasi-Monte Carlo

*Corresponding author: *E-mail address:* james.lomas@york.ac.uk

1 Introduction

The distribution of healthcare costs provides many challenges to the applied researcher: they are non-negative (often with many observations with costs of zero), heteroskedastic, positively skewed and leptokurtic. While these, or similar, challenges are found within many areas of empirical economics, the large interest in modelling healthcare costs has driven the development of an expanding array of estimation approaches and provides a natural context to compare methods for handling heavy-tailed and non-normal distributions. Econometric models of healthcare costs include applications to risk adjustment in insurance schemes (Van de ven and Ellis, 2000); in devolving budgets to healthcare providers (e.g. Dixon et al., 2011); in studies calculating attributable healthcare costs to specific health factors or conditions (Johnson et al., 2003; Cawley and Meyerhoefer, 2012) and in identifying treatment costs in health technology assessments (Hoch et al., 2002).

In attempting to capture the complex distribution of healthcare costs two broad modelling approaches have been pursued. The first consists of flexible parametric models - distributions such as the three-parameter generalised gamma and the four-parameter generalised beta of the second kind. This approach is attractive because of the range of distributions that these models encompass, whereas models with fewer parameters are inherently more restrictive, especially in regard to the assumptions they impose upon higher moments of the distribution (e.g. skewness and kurtosis). The second is semi-parametric models including extended estimating equations, finite mixture models and conditional density estimators. The extended estimating equations model (EEE) adopts the generalised linear models framework and allows for the link and distribution functions to be estimated from data, rather than specified a priori. Finite mixture models introduce heterogeneity (both observed and unobserved) through mixtures of distributions. Conditional density estimators are implemented by dividing the emprical distribution into discrete intervals and then decomposing the conditional density function into 'discrete hazard rates'. Despite the burgeoning availability of healthcare costs data via administrative records together with an increased necessity for policymakers to understand the determinants of healthcare costs, it is surprising that no study, as yet, compares comprehensively the models belonging to these two strands of literature. In this paper we compare these approaches both to each other and against standard practice: linear regression on levels, and on square root and log transformations of costs and generalised linear models (GLM).

Traditional Monte Carlo simulation approaches would not be appropriate for such an extensive comparison, as we are interested in a very large number of permutations of assumptions underlying the distribution of the outcome variable. In addition, such studies are prone to affording advantage to certain models arising from the chosen distributional assumptions used for generating data. Instead, using a large administrative database consisting of the population of English NHS hospital inpatient users for the year 2007-2008 (6,164,114 unique patients), we adopt a quasi-Monte Carlo approach where regression models are estimated on observations from one sub-population and evaluated on the remaining sub-population. This enables us to evaluate the regression methods in a rigourous and consistent manner, whilst ensuring results are not driven by overfitting to rare but influential observations, or traditional Monte Carlo distributional assumptions, and are generalisable to hospital inpatient services.

This paper compares and contrasts systematically these recent developments in semi-parametric and fully parametric modelling against each other and against standard practice. There is no comprehensive comparison existing in the literature at the moment, and given the number of choices available for modelling heavy-tailed, non-normal data, this study makes an important contribution towards forming a ranking of possible approaches (for a similar study comparing propensity score methods, see Huber et al. (2013)). The focus of the paper is the performance of these models in terms of predicting the conditional mean, given its importance in informing policy in healthcare and its prominence in comparisons between econometric methods in healthcare cost regressions[1]. Given our focus, we analyse bias, accuracy and goodness of fit of forecasted conditional means. We find that no model performs best across all metrics of evaluation. Commonly used models, such as linear regression on levels of and log-transformed costs, gamma GLM with log-link, and the log-normal distribution, are not among the four best performing models with any of our chosen metrics. Our results indicate that models estimated with a square root link function perform much better than those with log- or linear-link functions. We find that linear regression with a square root-transformed dependent variable is the best performing model in terms of bias, the conditional density estimator (using multinomial logit) for accuracy, and in terms of goodness of fit the best performer is a Poisson GLM with a square root link.

2 Previous comparative studies

A number of studies have compared the performance of subsets of regression based approaches to modelling healthcare cost data where model performance is assessed on either actual costs, that is costs with an unknown true distribution (Deb and Burgess, 2003; Veazie et al., 2003; Buntin and Zaslavsky, 2004; Basu et al., 2006; Hill and Miller, 2010; Jones et al., 2013)

[1] If the policymaker has a sufficiently large budget, Arrow and Lind (1970) argue that the policymaker should focus on mean outcomes. Other features of the distribution may be of interest (Vanness and Mullahy, 2007), especially when the policymaker has a smaller budget to allocate to healthcare.

or simulated costs from an assumed distribution (Basu et al., 2004; Gilleskie and Mroz, 2004; Manning et al., 2005). Using actual costs preserves the true empirical distribution of cost data, and all of its complexities, while simulating costs provides a benchmark using the known parameters of the assumed distribution (classic Monte Carlo) against which models can be compared.

Studies based on the classic Monte Carlo design are ideally suited to assessing whether or not regression methods can fit data when specific assumptions, and permutations thereof, are imposed or relaxed. The complexities in the observed distribution of healthcare costs are such that a comprehensive comparison of modelling approaches would require an infeasibly large number of permutations of distributional assumptions used to generate data to make a classic Monte Carlo simulation worthwhile. Choosing a subset of the possible permutations of assumptions is prone to bias the results in favour of some methods over others. A reliance on actual data instead requires large datasets so that forecasting is evaluated on sufficient observations to credibly reflect all of the idiosyncratic features of cost data. With this approach, however, it is difficult to assess exactly what aspect of the distribution of healthcare costs is problematic for each method under comparison.

2.1 Studies using cross-validation approaches

With improvements in computational capacity, there has recently been a number of papers using large datasets to perform quasi-Monte Carlo comparisons across regression models for healthcare costs. Quasi-Monte Carlo comparisons divide the data into two groups, with samples repeatedly drawn from one group and models estimated, while the other group is used to evaluate out-of-sample performance (using the coefficients from the estimated models).

Deb and Burgess (2003) examined a number of models to predict total healthcare expenditures using a quasi-Monte Carlo approach with data from US Department of Veterans Affairs (VA) comprised of approximately 3 million individual records. From within these observations a sub-group of 1.5 million individual records was used as an 'estimation' group and another sub-group of 1 million records formed a 'prediction' group. Deb and Burgess (2003) examined the predictive performance of models across different sizes of sample drawn from the 'estimation' group. For each sample drawn model predictive performance was assessed on the full set of observations in the 'prediction' group according to mean prediction error (MPE), root mean squared error (RMSE), mean absolute prediction error (MAPE) and absolute deviations in mean prediction error (ADMPE). Using this methodology they were able to show that models based on a gamma density had better performance in forecasting individual costs than standard linear regression, with the most accurate individual forecasts coming from a finite mixture model with two gamma density components. In terms of bias, linear regression on level and square root transformed costs perform the best. Deb and Burgess (2003) also note that the performance of finite mixture models in forecasting individual costs improves with increasing sample size, with MAPE between 10-15% lower than linear regression from sample sizes as large as 20,000 observations. Their results highlight a trade-off between bias and precision, and the need for caution surrounding the use of finite mixture models at smaller sample sizes.

Jones et al. (2013) focus exclusively on parametric models and suggest the use of generalised beta of the second kind as an appropriate distribution for healthcare costs. Their quasi-Monte Carlo design compares this distribution together with its nested and limiting cases, including the generalised gamma. Using data from Hospital Episode Statistics (HES) split into 'estimation' and 'validation' sets, they find little evidence of performance of models varying with sample size, but find variation between models in their ability to forecast mean costs with generalised gamma the most accurate, and beta of the second kind the least biased.

Hill and Miller (2010) and Buntin and Zaslavsky (2004) also use cross-validation techniques so that models are estimated on samples of data and evaluated on the remaining observations. Samples for estimation and the remaining data for evaluation differ across replications such that, unlike a quasi-Monte Carlo design, individuals may fall into either the estimation sample or the validation sample at each replication. This approach is less data intensive and providing sufficient replications should produce sufficient replications should produce sufficient information in the evaluation exercise to judge model performance. The approaches are similar in that they both replicate the sampling process to ensure there is no 'lucky split' and guard against overfitting by evaluating out of sample.

Hill and Miller (2010) use the first eight waves of the Medical Expenditure Panel Survey (MEPS) dataset (from 1996-1997 to 2003-2004) to compare linear regression on untransformed and log transformed dependent variables, Poisson and gamma GLMs with log link, EEE and a generalised gamma model. They examine four outcomes: total and prescription expenditures for privately insured adults (28,579 and 22,011 observations respectively) and elderly adults (12,547 and 11,671 observations respectively). For each outcome, 1,024 half samples were created for estimation and validation. Models with log link were found to perform well in only one of these: total expenditures for privately insured, nonelderly adults, and with this outcome the gamma GLM and generalised gamma model performed well (in terms of MPE and MAPE). They show that the flexible link function of EEE improved goodness of fit, without inducing overfitting, in all four outcomes. In this way, Hill and Miller (2010) represents the first paper to compare common practice with the semi-parametric EEE model and the non-nested, fully parametric, generalised gamma model.

Buntin and Zaslavsky (2004) examined the performance of eight alternative estimators, comparing the performance of models with transformed dependent variables and GLMs (with log link). The authors used data from the 1996 Medicare Current Beneficiary Survey (MCBS), taking 10,134 observations in total. This was split in half to form an estimation group and a validation group and repeated 100 times in total. They found that predictive performance was improved with careful consideration of the nature of heteroskedasticity. A GLM with variance proportional to the mean and using two smearing factors in a transformed dependent variable model were both found to be good choices for their application in terms of lower MAPE and mean squared forecast error (MSFE).

In Veazie et al. (2003) 500 half samples are drawn repeatedly from a dataset consisting of 8,495 observations from MCBS (risk adjusters from 1993 and expenditures from 1994). In addition, they compare models estimated on the years 1992-1993 (7,450 observations), and evaluated out-of-sample on the 1993-1994 observations, with the full samples bootstrapped 500 times to derive results. They found that with linear regression, a square root transformed dependent variable can reduce MAPE, but not necessarily MPE compared to using the level of costs.

Finally, Basu et al. (2006) compared EEE to linear regression with log transformed dependent variable and GLM with log link and gamma variance using data from Medstat's MarketScan database with a final sample of 7,428 observations. Performance was mainly assessed in-sample, where the EEE performed well in terms of MPE across deciles of covariates. Split-sampling was used to perform tests of overfitting (Copas, 1983), where they found little evidence for this in the case of EEE compared to other models.

2.2 Recent developments in semi-parametric and fully parametric modelling

Figure 1 outlines the literature comparing regression models for healthcare costs as described above. As shown, there is no study that comprehensively and systematically evaluates all recent developments in approaches. In addition, any synthesis of the existing literature would be inconclusive in terms of which method is most appropriate for an application. Amongst the semi-parametric methods, EEE and finite mixture models have never been directly compared in a rigourous evaluation. They have both separately been compared against standard practice (transformed dependent variable regression and GLM) in Basu et al. (2006); Hill and Miller (2010) and Deb and Burgess (2003), for EEE and finite mixture models respectively. The conditional density estimator, as yet, has not been compared with other healthcare cost regression models using actual data, although evidence from Monte Carlo studies suggests it to be a versatile approach (compared with standard practice methods) (Gilleskie and Mroz, 2004). Jones et al. (2013)

introduce the use of the generalised beta of the second kind distribution with healthcare cost regressions, the most flexible amongst the fully parametric distributions used to date with healthcare cost regressions, and compares against the generalised gamma which is a limiting case of the generalised beta of the second kind. Given an increasing interest in modelling healthcare costs for resource allocation, risk adjustment and identifying attributable treatment costs, together with the proliferation of data through administrative records, a comprehensive and systematic comparison of available approaches would appear timely. The results of which will have resonance beyond healthcare costs and should be of interest to empirical applications to other right skewed, leptokurtic or heteroskedastic distributions such as income and wages.

	Studies using Monte Carlo			Studies using cross-validation				Studies using quasi-Monte Carlo		
	Basu et al. (2004)	Gilleskie and Mroz (2004)	Manning et al (2005)	Veazie et al (2003)	Buntin and Zaslavsky (2004)	Basu et al (2006)	Hill and Miller (2010)	Deb and Burgess (2003)	Jones et al (2013)	This paper
linear regression				■	■		■	■		■
linear regression (log)			■		■	■	■	■		■
linear regression (square root)				■	■			■		■
log-normal	■				■				■	■
gaussian GLM					■					(a)
Poisson					■		■			■
gamma	■	■	■		■	■	■	■	■	■
extended estimating equations						■	■			■
Weibull	■		■						■	(b)
generalised gamma			■				■		■	■
GB2									■	■
finite mixture of gammas								■		■
conditional density estimator		■								■

Figure 1: Models included in recent published comparitive work

(a) Not commonly used and problematic in estimation for our data in preliminary work.

(b) A special case of generalised gamma and generalised beta of the second kind which are included in our analysis.

3 Specification of models

We compare 16 different models applicable to healthcare cost data. Each makes different assumptions about the distribution of the outcome (cost) variable. Each regression utilises the same vector of covariates X_i, although the precise way in which they enter the distribution varies across models. All models specify at least one linear index of covariates $X_i'\beta$. In addition, linear regression methods with transformed outcome require assumptions surrounding the form of heteroskedasticity (modelled as a function of X_i), in order to retransform predictions onto the natural cost scale (Duan, 1983). Within the GLM family, we explicitly model the mean and variance functions as some transformation of the linear predictor (Blough et al., 1999). Fully parametric distributions, such as the gamma- and beta-family of models, assume the form of the entire distribution, where a scale parameter is a function of the linear index. Finite mixture models allow for multiple densities, each a function of the covariates in linear form. For conditional density estimator models, the empirical distribution of costs is divided into intervals, and functions of the independent variables predict the probability of lying within each interval.

Beginning with linear regression, we estimate three models using ordinary least squares: the first is on the level of costs, the second and third use a log and square root transformed dependent variable respectively (log transformation is more commonly used in the literature Jones (2011)). With these approaches, predictions are generated on a transformed scale, and it is necessary to calculate an adjustment in order to retransform predictions to their natural cost scale. This is done by applying a smearing factor, which varies according to covariates in the presence of heteroskedasticity (Duan, 1983).

Given the complications in retransforming in the presence of heteroskedasticity, researchers more frequently use methods that estimate on the natural cost scale and explicitly model the variance as a function of covariates. The dominant approach that achieves these aims is the use of GLM (Blough et al., 1999). There are two components to GLM, the first is a link function that relates the index of covariates to the conditional mean, and the second is a distribution function that describes the variance as a function of the conditional mean. These are estimated simultaneously, using pseudo- or quasi-maximum likelihood, leading to estimates that are consistent providing the mean function is correctly specified. Typically, the link function in applied work takes the form of a log or square root function. In this paper we consider two types of distribution function, each a power function of the conditional mean. In the Poisson case, the variance is proportional to the conditional mean function of covariates and in the gamma case the variance is proportional to the conditional mean squared. Two of the combinations of link functions and distribution families are associated with commonly used distributions. The GLM with a log link and Poisson variance is associated with the Poisson model (see discussion in Mullahy, 1997), and often a GLM with log link and gamma variance is applied to healthcare costs.

3.1 Flexible parametric models

Within the GLM and OLS approaches, much focus is placed on heteroskedacity and the form that it takes. However, recent developments have been made where the modeling of higher moments, skewness and kurtosis, is tackled explicitly. Our approach focuses on estimating the entire distribution using maximum likelihood, which requires that the distribution is correctly specified for consistent results, but if correctly specified, then estimates are efficient.

3.1.1 Generalised gamma

We estimate two models from within the gamma-family, which have typically been used for durations, but also have precedent in the healthcare costs literature (Manning et al., 2005): log-normal and generalised gamma distribution. Each of these is estimated with a scale parameter specified as an exponential function of covariates and estimated using maximum likelihood. The probability density function and conditional mean for the generalised gamma distribution are given below:

$$f(y_i|X_i) = \frac{\kappa\left(\kappa^{-2}\left(\frac{y_i}{\exp\left(X_i'\beta\right)}\right)^{\kappa/\sigma}\right)^{\kappa^{-2}} exp\left(-\kappa^{-2}\left(\frac{y_i}{\exp\left(X_i'\beta\right)}\right)^{\kappa/\sigma}\right)}{\sigma y_i \Gamma\left(\kappa^{-2}\right)} \tag{1}$$

$$E(y_i|X_i) = \left(\exp\left(X_i'\beta\right)\right)\left(\kappa^{2\sigma/\kappa}\right)\frac{\Gamma\left(\kappa^{-2}+\frac{\sigma}{\kappa}\right)}{\Gamma\left(\kappa^{-2}\right)} \tag{2}$$

where σ is a scale parameter, κ is a shape parameter and $\Gamma(.)$ is the gamma function

When $\kappa \to 0$ the generalised gamma distribution approaches the limiting case of the log-normal distribution, for which the probability density function and conditional mean are:

$$f(y_i|X_i) = \frac{1}{\sigma y_i\sqrt{2\pi}}\exp\left(\frac{-\left(\ln y_i - X_i'\beta\right)^2}{2\sigma^2}\right) \tag{3}$$

$$E(y_i|X_i) = \left(\exp\left(X_i'\beta\right)\right)\exp\left(\frac{\sigma^2}{2}\right) \tag{4}$$

3.1.2 Generalised beta of the second kind

We also include the generalised beta of the second kind, which has yet to be compared with a broad range of regression models.[2] Beta-type models, as gamma-type models, require assumptions about the form of the entire distribution. Until recently, they have been used largely in actuarial applications, as well as modelling incomes (Cummins et al., 1990; Bordley et al., 1997). However, they have been suggested for use with healthcare costs owing to their ability to model heavy tails, for example in Mullahy (2009), and used with healthcare costs in Jones et al. (2013). We include the generalised beta of the second kind, since all beta-type (and gamma-type) distributions are nested or limiting cases of this distribution. It therefore offers the greatest flexibility in terms of modelling healthcare costs amongst the duration models used here; see for example the implied restrictions on skewness and kurtosis (McDonald et al., 2013). The probability density function and conditional mean are:

$$f(y_i|X_i) = \frac{a y_i^{ap-1}}{b(X_i)^{ap} B(p,q)[1 + (\frac{y_i}{b(X_i)})^a]^{(p+q)}} \tag{5}$$

$$E(y_i|X_i) = b(X_i)\left[\frac{\Gamma(p+\frac{1}{a})\Gamma(q-\frac{1}{a})}{\Gamma(p)\Gamma(q)}\right] \tag{6}$$

where a and b are scale parameters, p and q are shape parameters and $B(p,q) = \Gamma(p)\Gamma(q)/\Gamma(p+q)$ is the beta function

We parameterise the generalised beta of the second kind with the scale parameter b as two different functions of covariates: a log-link and a square root link.

3.2 Semi-parametric methods

3.2.1 Extended estimating equations

A flexible extension of GLM is proposed by Basu and Rathouz (2005) and Basu et al. (2006), known as the extended estimating equations (EEE). It approximates the most appropriate link using a Box-Cox function, where $\lambda = 0$ implies a log link and $\lambda = 0.5$ implies a square root link:

$$E(y_i|X_i) = (\lambda X_i'\beta + 1)^{\frac{1}{\lambda}} \tag{7}$$

as well as a general power function to define the variance with constant of proportionality θ_1 and power θ_2:

$$var(y_i|X_i) = \theta_1(E(y_i|X_i))^{\theta_2} \tag{8}$$

[2] In Jones et al. (2013), beta-type models are limited to comparison with gamma-type distributions.

Suppose that the distribution of the outcome variable is unknown, but has mean and variance nested within (7) and (8). An incorrectly specified GLM mean function[3] yields biased and inconsistent estimates, while estimates from EEE should be unbiased, providing the specification of regressors is correct. A well-specified mean function combined with an incorrectly specified distribution form will be inefficient compared to EEE. If the distribution is known to be a specific GLM form, the EEE is less efficient than the appropriate GLM, but both are unbiased.

3.2.2 Finite mixture models

Finite mixture models have been employed in health economics in order to allow for heterogeneity both in response to observed covariates and in terms of unobserved latent classes (Deb and Trivedi, 1997). Heterogeneity is modelled through a number of components, C, each of which can take a different specification of covariates (and shape parameters, where specified), written as $f_j(y_i|X_i)$, and where there is a parameter for the probability of belonging to each component, π_j. The general form of the probability density function of finite mixture models is given as:

$$f(y_i|X_i) = \sum_{j}^{C} \pi_j f_j(y_i|X_i) \tag{9}$$

We use two gamma distribution components in our comparison.[4] In one of the models used, we allow for log-links in both components (10), and in the other we allow for a square root link (11). In both, the probability of class membership is treated as constant for all individuals and a shape parameter, α_j, is estimated for each component.

$$f_j(y_i|X_i) = \frac{y_i^{\alpha_j}}{y_i\Gamma\left(\alpha_j\right)\exp\left(X_i'\beta_j\right)^{\alpha_j}} \exp\left(-\left(\frac{y_i}{\exp\left(X_i'\beta_j\right)}\right)\right) \tag{10}$$

$$f_j(y_i|X_i) = \frac{y_i^{\alpha_j}}{y_i\Gamma\left(\alpha_j\right)\left(X_i'\beta_j\right)^{2\alpha_j}} \exp\left(-\left(\frac{y_i}{\left(X_i'\beta_j\right)^2}\right)\right) \tag{11}$$

The conditional mean is given for the log-link specification and for the square root link by (12) and (13) respectively:

$$E(y_i|X_i) = \sum_{j}^{C} \pi_j \alpha_j \exp\left(X_i'\beta_j\right) \tag{12}$$

[3] In common usage GLM mean functions are limited to standard forms such as log and square root link function.

[4] Preliminary work showed that models with a greater number of components lead to problems with convergence in estimation. Empirical studies such as Deb and Trivedi (1997) provide support for the two components specification for healthcare use.

$$E(y_i|X_i) = \sum_{j}^{C} \pi_j \alpha_j (X_i' \beta_j)^2 \tag{13}$$

Unlike the models in the previous section, this approach can allow for a multi-modal distribution of costs. In this way, finite mixture models represent a flexible extension of parametric models (Deb and Burgess, 2003). Using increasing numbers of components, it is theoretically possible to fit any distribution, although in practice researchers tend to use few components (two or three) and achieve good approximation to the distribution of interest (Heckman, 2001).

3.2.3 Conditional density estimators

We use two additional models that are based on the conditional density estimator proposed by Gilleskie and Mroz (2004). Their method is an extension of the two-part model that is frequently used to deal with zero costs, in that the range of the dependent variable is divided into K intervals ($k = 1, ..., K$), where the lower and upper threshold values of an interval k are y_{k-1} and y_k[5]. The probability of an observation falling into interval k is:

$$P(y_{k-1} \leq y_i < y_k|X_i) = \int_{y_{k-1}}^{y_k} f(y_i|X_i) dy_i \tag{14}$$

The conditional density function is approximated using a hazard rate decomposition. The hazard rate is defined as the probability of lying in interval k conditional on not lying in intervals $1, ..., k-1$:

$$\lambda(k, X_i) = P(y_{k-1} \leq y_i < y_k|X_i, y_i \geq y_{k-1}) = \frac{\int_{y_{k-1}}^{y_k} f(y_i|X_i) dy_i}{1 - \int_{y_0}^{y_{k-1}} f(y_i|X_i) dy_i} \tag{15}$$

where:

$$P(y_{k-1} \leq y_i < y_k|X_i) = \lambda(k, X_i) \prod_{j=1}^{k-1} [1 - \lambda(j, X_i)] \tag{16}$$

The conditional mean, $E(y|x) = \int_{-\infty}^{\infty} y f(y|x) dy$, is approximated by:

$$\widehat{E(y_i|X_i)} = \sum_{k=1}^{K} \overline{y}_k \widehat{\lambda(k, X_i)} \prod_{j=1}^{k-1} \left[1 - \widehat{\lambda(j, X_i)}\right] = \sum_{k=1}^{K} \overline{y}_k P_{ik}(X_i) \tag{17}$$

where $\overline{y}_k$ is the mean of y within the interval. The hazard rates for each separate interval could be estimated as separate logit models but Gilleskie

[5] y_0 is the lowest observed cost.

and Mroz (2004) suggest a flexible smooth approximation that involves estimating a single logit model, augmented by a constructed covariate that takes values which vary across the intervals. Here we adopt a convenient alternative, following the approach of Han and Hausman (1990), we use an ordered logit specification to estimate the discrete hazard function and hence $P_{ik}(X_i)$. In addition, we estimate a second alternative that uses a multinomial logit model to estimate the $P_{ik}(X_i)$ terms in (17).

For our application we use 15 equally sized intervals across all samples. Gilleskie and Mroz (2004) find that between 10 and 20 intervals result in a good approximation in their application to healthcare costs. We found 15 intervals to result in good convergence performance in our preliminary work.

4 Data and Choice of Variables

Our study uses individual-level data from the English Hospital Episode Statistics (HES) (for the financial year 2007-2008)[6]. This dataset contains information on all inpatient episodes, outpatient visits and A&E attendances for all patients admitted to English NHS hospitals (Dixon et al., 2011). For our study, we exclude spells which were primarily mental or maternity healthcare, as well as private sector spells.[7] HES is a large administrative dataset collected by the NHS Information Centre, with our dataset comprising 6,164,114 separate observations, representing the population of hospital inpatient healthcare users for the year 2007-2008. Since data is taken from administrative records, we only have information on users of inpatient NHS services, and therefore can only model strictly positive costs[8].

The cost variable used throughout is individual patient annual NHS hospital cost for all spells finishing in the financial year 2007-2008. In order to cost utilisation of inpatient NHS facilities, tariffs from 2008-2009[9] were applied to the most expensive episode within the spell of an inpatient stay (following standard practice for costing NHS activity). Then, for each patient, all spells occuring within the financial year were summed. The data

[6] In our dataset, episodes are grouped into spells, which can be thought of as discrete admissions for a patient.

[7] This dataset was compiled as part of a wider project considering the allocation of NHS resources to primary care providers. Since a lot of mental healthcare is undertaken in the community and with specialist providers, and hence not recorded in HES, the data is incomplete, and also since healthcare budgets for this type of care are constructed using separate formulae. Maternity services are excluded since they are unlikely to be heavily determined by morbidity characteristics, and accordingly for the setting of healthcare budgets are determined using alternative mechanisms.

[8] Zeros are typically handled by a two-part specification and the main challenge is to capture the long and heavy tail of the distribution rather than the zeros.

[9] Reference costs for 2005-2006, which were the basis for the tariffs from 2008-2009, were used when 2008-2009 tariffs were unavailable.

are summarised in Table 1 and Figure 2.

	Level	**Square root**	**Logarithm**
N	6,164,114		
Mean	£2,610	43.18	7.25
Median	£1,126	33.56	7.03
Standard deviation	£5,088	27.30	1.00
Skewness	13.03	2.84	0.74
Kurtosis	363.18	19.62	2.99
Maximum	£604,701	777.63	13.31
99th percentile	£19,015	137.89	13.31
95th percentile	£8,956	94.64	9.10
90th percentile	£6,017	77.57	8.70
75th percentile	£2,722	52.17	7.91
25th percentile	£610	24.70	6.41
10th percentile	£446	21.12	6.10
5th percentile	£407	20.16	6.01
1st percentile	£347	18.63	5.85
Minimum	£217	14.73	5.38

Table 1: Descriptive statistics for hospital costs

The challenges of modelling cost data are clearly observed in Table 1 and in Figure 2[10]: the observed costs are heavily right-hand skewed (even after log transformation), with the mean far in excess of the median, and are highly leptokurtic (although roughly mesokurtic following log transformation). Whilst transforming the data clearly reduces skewness, neither transformation results in a completely symmetric distribution, which implies that a flexible link function could be useful. The distribution may also be multi-modal, or at least noisy with many spikes, which can most clearly be seen in Figure 2 on the histogram of the log transformed costs.

We construct a linear index of covariates and divide the data into quantiles according to this, to analyse conditional (on X) distributions of the outcome variable [11]. First, we plot the variances of each quantile against their means (Figure 3). This gives us a sense of the nature of heteroskedasticity and feasible assumptions relating these aspects of the distribution. From Figure 3, we can see that there is evidence against homoskedasticity (where there would be no visible trend), and evidence for some relationship between the variance and the mean.

We also carry out a similar analysis, where we plot the kurtosis of each quantile against their skewness. Parametric distributions impose restric-

[10] Costs above £30,000 were excluded, for this figure only, to make the graphs clearer.

[11] This is done by regressing the outcome variable on the set of covariates we include in our regression models using OLS.

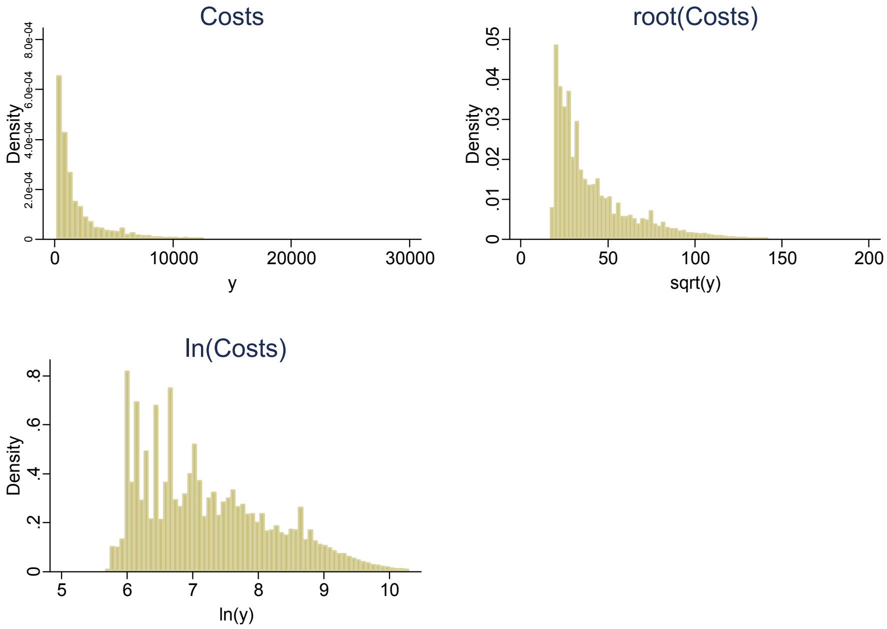

Figure 2: Histogram plots of costs

tions upon possible skewness and kurtosis: one parameter distributions are restricted to a single point (e.g. normal distribution imposes a skewness of 0 and a kurtosis of 3), two parameter distributions allow for a locus of points to be estimated, and distributions with three or more parameters allow for spaces of possible skewness and kurtosis combinations. Figure 4 shows that the data is non-normal and provides motivation for flexible methods since they appear better able to model the higher moments of the conditional distributions of the outcome variables analysed here.

All of the models in the quasi-Monte Carlo comparison use a specified vector of covariates, and have at least one linear index of these. This vector mirrors the practice in the literature regarding comparing econometric methods for healthcare costs, allowing models to control for age (as well as age squared and age cubed), gender (interacted fully with age terms), and morbidity characteristics (from ICD classifications).[12] Each of the 24 morbidity markers indicates the presence or absence, coded 1 and 0 respectively, of one or more spells with any diagnosis within the relevant subset of ICD10 chapters, during the financial year 2007-2008 (see Appendix A). We do not use a fully interacted specification, since morbidity is modelled with a separate intercept for presence of each type of diagnosis (and not interacted with

[12] Morbidity information is available through the HES dataset, adapted from the ICD10 chapters (WHO, 2007) - see Appendix for further details.

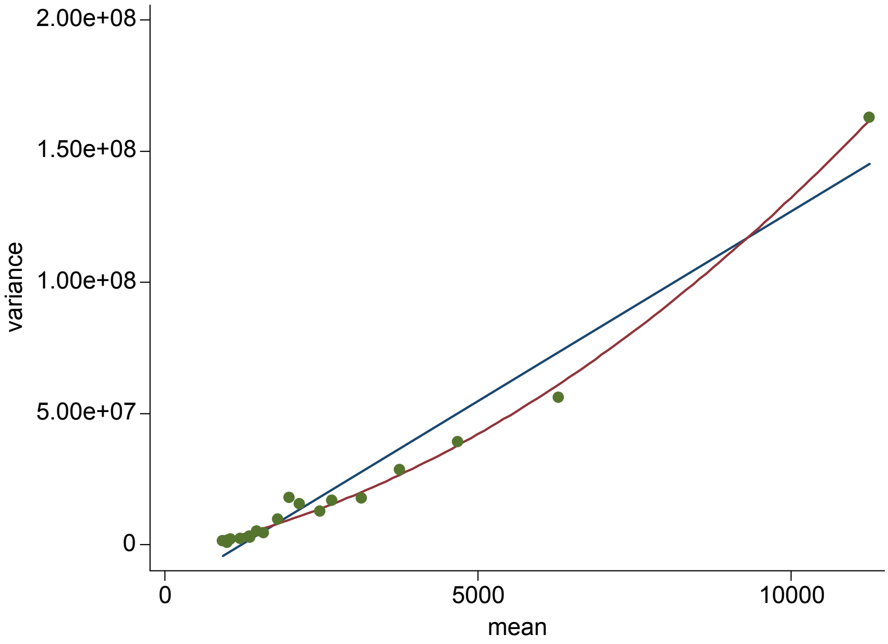

Figure 3: Variance against mean for each of the 20 quantiles of the linear index of covariates

Note:

The data were divided into twenty subsets using the deciles of a simple linear predictor for healthcare costs using the set of regressors introduced later. Figure 3 plots the means and variances of actual healthcare costs for each of these subsets.

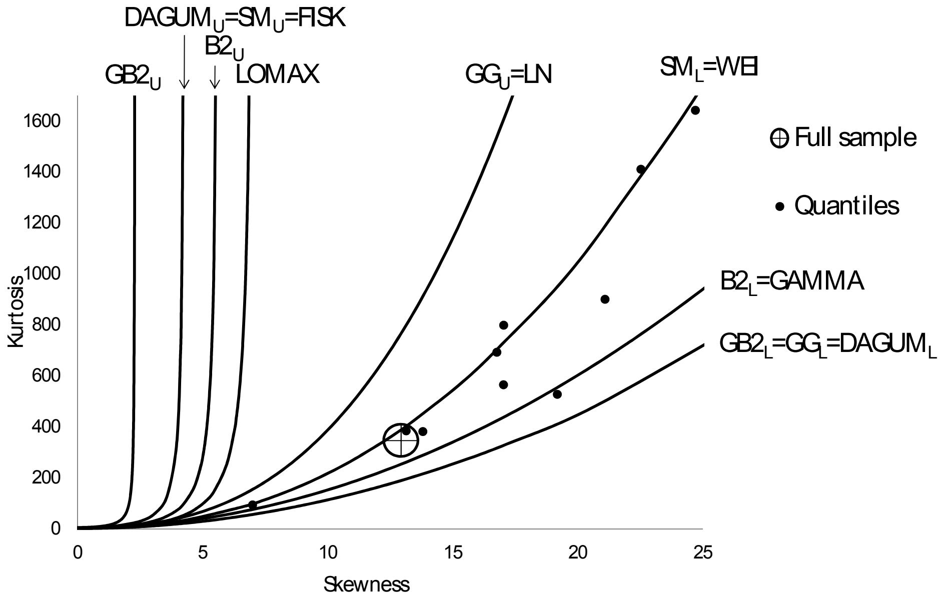

Figure 4: Kurtosis against skewness for each of the 10 quantiles of the linear index of covariates, adapted from McDonald et al. (2013)

Note:

The dots shown on Figure 4 were generated as follows: the data were divided into ten subsets using the deciles of a simple linear predictor for health-care costs using the set of regressors used in this paper. Figure 4 plots the skewness and kurtosis coefficients of actual healthcare costs for each of these subsets, the skewness and kurtosis coefficient of the full estimation sub-population (represented by the larger circle with cross) and theoretically possible skewness-kurtosis spaces and loci for parametric distributions considered in the literature.

age or gender). However, we do allow for interactions between age and its higher orders and gender. This means that we are left with a specification close to those used in the comparative literature as well as a parsimonious version of the set of covariates used to model costs in Person-Based Resource Allocation in England, for example Dixon et al. (2011). In addition, making the specification less complicated aids computation and results in fewer models failing to converge.

5 Methodology

5.1 Quasi-Monte Carlo Design

By using the HES data we have access to a large amount of observations representing the whole population of English NHS inpatient costs. To exploit this, we use a quasi-Monte Carlo design similar to Deb and Burgess (2003).[13] The population of observations (6,164,114) is randomly divided into two equally sized sub-populations: an 'estimation' (3,082,057) and a 'validation' set (3,082,057).[14] From within the 'estimation' set we randomly draw, 100 times with replacement, samples of size N_s ($N_s \in$ 5,000; 10,000; 50,000; 100,000). The models are estimated on the samples and performance then evaluated on both the sample drawn from the 'estimation' set and the full 'validation' set. Figure 5 illustrates our study design in the form of a diagram, note the subscript m denotes the model used, N_s the sample size used, and r the replication number.

In order to execute this quasi-experimental design, we automate the model selection process for each approach: for instance, with the conditional density estimator, we specify a number of bins to be estimated, a priori, rather than undergoing the investigative process outlined in Gilleskie and Mroz (2004). Similarly, all models have been automated to some extent, since we set a priori: the specification of regressors (all models), the parameters that vary with covariates (generalised gamma and generalised beta of the second kind), and the number of mixtures to model (finite mixture models). Our specification of regressors was based on preliminary work, which showed alternative specifications to give similar results, but with worse convergence performance[15].

[13] Using a split-sample to evaluate models has precedent in the comparitive literature on healthcare costs, see Duan et al. (1983); Manning et al. (1987).

[14] Given the size of the dataset, any sub-optimality resulting from the proportions allocated to each set is likely to be minimal. To ensure the results are replicable, we set a fixed seed for splitting the dataset and for randomly drawing samples.

[15] For example, one alternative specification featured a count of the number of morbidities instead a vector of morbidity markers.

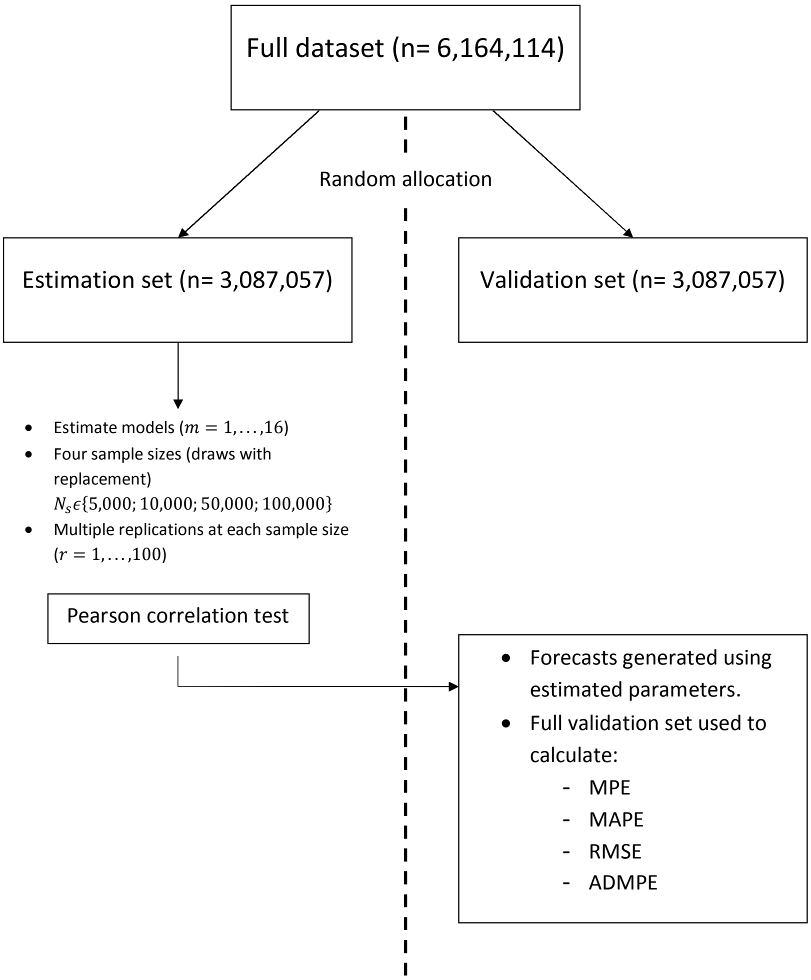
Full dataset (n= 6,164,114)
Random allocation
Estimation set (n= 3,087,057)
Validation set (n= 3,087,057)
Estimate models ($m = 1,\ldots,16$)
Four sample sizes (draws with replacement) $N_s \epsilon \{5{,}000; 10{,}000; 50{,}000; 100{,}000\}$
Multiple replications at each sample size ($r = 1,\ldots,100$)
Pearson correlation test
Forecasts generated using estimated parameters.
Full validation set used to calculate:
MPE
MAPE
RMSE
ADMPE

Figure 5: Diagram setting out study design

5.2 Evaluation of Model Performance

5.2.1 Estimation Sample

Researchers modelling healthcare costs will typically carry out multiple tests to establish the reliability of their model specification. These tests are carried out in sample, and help to inform the selection of models that will then be used for predictive purposes. They are commonly used to build the specification of the 'right hand side' of the regression: the covariates used and interactions between them. In addition, researchers working with healthcare costs use these tests to establish the appropriate link function between covariates and expected conditional mean and other assumptions about functional form. We include results from the Pearson correlation coefficient test, which is simple to carry out and has intuitive appeal.[16] In order to carry out the Pearson correlation coefficient test, residuals (computed on the raw cost scale) are regressed against predicted values of cost. If the slope coefficient on the predicted costs is significant, then this implies a detectable linear relationship between the residuals and the covariates, and so evidence of model misspecification.

5.2.2 Validation Set

We use our models to estimate forecasted mean healthcare costs over the year for individuals ($\hat{y}_i^v = \widehat{E(y_i^v|X_i^v)}$[17], v denotes that the observation is from the 'validation' set) and evaluate performance on metrics designed to reflect the bias (mean prediction error, MPE), accuracy (mean absolute prediction error, MAPE) and goodness of fit (root mean square error, RMSE) of these forecasts. MPE can be thought of as measuring the bias of predictions at an aggregate level, where positive and negative errors can cancel each other out, while MAPE is a measure of the accuracy of individual predictions. RMSE is similar to MAPE in that positive and negative errors do not cancel out, however larger errors count for disproportionately more, since they are squared. In addition, we evaluate the variability of bias across replications (absolute deviations of mean prediction error, ADMPE). These are all evaluated on the full 'validation' set. Formulae for calculating these metrics are provided below, where m denotes the model used, for ease of exposition we denote the size of the subsample of estimation data used for each metrics as s, and r is the replication number.

[16] We also carried out Pregibon link, Ramsey RESET and modified Hosmer-Lemeshow tests in preliminary work although only include results from the Pearson correlation coefficient tests, since they were found to display the same pattern more clearly (with the other tests there was smaller variation in rejection rates across the different models).

[17] This is computed using coefficients from models estimated on the 'estimation' set, e.g. for linear regression $\widehat{E(y_i^v|X_i^v)} = \hat{\alpha}^E + \hat{\beta}^E X_i^v$

$$MPE_{msr} = \frac{\sum \left(y_i^v - \hat{y}_i^v\right)}{N_s} \tag{18}$$

$$MAPE_{msr} = \frac{\sum \left|y_i^v - \hat{y}_i^v\right|}{N_s} \tag{19}$$

$$RMSE_{msr} = \sqrt{\frac{\sum \left(y_i^v - \hat{y}_i^v\right)^2}{N_s}} \tag{20}$$

$$ADMPE_{msr} = \left| MPE_{msr} - \frac{\sum_{r=1}^{R} MPE_{msr}}{R} \right| \tag{21}$$

Only replications where all 16 models are successfully estimated on the sample are included for evaluation, and model performance according to each criterion is calculated as an average over all included replications, e.g. $MPE_{ms} = \frac{\sum_{r=1}^{R} MPE_{msr}}{R}$.[18]

In order to get a greater insight into the performance of different distributions, we evaluate forecasted conditional means at different values of the covariates. In practice this is done by patitioning the fitted values of costs into deciles. We assess MPE and MAPE for deciles of predicted costs, since there is concern that models perform with varying success at different points in the distribution, e.g. models designed for heavy-tails might be expected to perform better in predicting the biggest costs. This also represents a desire to fit the distribution of costs for different groups of observations according to their observed covariates.

We combine the results that we obtain from different sample sizes (N_s), and attempt to find a pattern in the way in which models perform as sample size varies. To do this we construct response surfaces (e.g. Deb and Burgess (2003)). These are polynomial approximations to the relationship between the statistics of interest and the sample size of the experiment, N_s. For our purposes, we estimate the following regression for each model and for each metric of performance (illustrated below for the mean prediction error).

$$MPE_{msr} = \alpha_m^{MPE} + \beta_m^{MPE} \frac{1}{N_s} + u_{msr}^{MPE} \tag{22}$$

We specify the relationship between MPE and the inverse of the sample size, reflecting that we expect reduced bias as the number of observations increases. In particular, the value of α_m^{MPE} represents the value of MPE to which the model approaches asymptotically with increasing sample size, testing whether or not this is statistically significant from zero gives an

[18] All models estimated successfully every time, except for CDEM and EEE. CDEM could not be estimated on two of the 100 replicates with samples of 5,000 observations. EEE could not be estimated on four, four, six and four of the 100 replicates with sample sizes of 5,000, 10,000, 50,000 and 100,000 observations, respectively.

indication of whether the estimator is consistent. Here, u_{msr}^{MPE} represents the error term from the regression. For the metrics that cannot be negative, we use the log function of the value as the dependent variable, for example in the case of mean absolute prediction error we estimate:

$$\ln\left(MAPE_{msr}\right) = \alpha_m^{LMAPE} + \beta_m^{LMAPE}\frac{1}{N_s} + u_{msr}^{LMAPE} \qquad (23)$$

With the log specification, differences in estimates are to be interpreted as percentage differences, as opposed to absolute differences.

6 Results and Discussion

To begin with, we consider the results from the smallest samples that we draw from the 'estimation' set (5,000 observations). Results from larger samples are analysed by way of the response surfaces which we present later. Table 2 is a key for the labels we use for each model in discussion of the results.

OLS	linear regression
LOGOLSHET	transformed linear regression (log), heteroskedastic smearing factor
SQRTOLSHET	transformed linear regression ($\sqrt{.}$), heteroskedastic smearing factor
GLMLOGP	generalised linear model, log-link, Poisson-type family
GLMLOGG	generalised linear model, log-link, gamma-type family
GLMSQRTP	generalised linear model, $\sqrt{}$-link, Poisson-type family
GLMSQRTG	generalised linear model, $\sqrt{}$-link, gamma-type family
LOGNORM	log-normal
GG	generalised gamma
GB2LOG	generalised beta of the second kind, log-link
GB2SQRT	generalised beta of the second kind, $\sqrt{}$-link
FMMLOGG	two-component finite mixture of gamma densities, log-link
FMMSQRTG	two-component finite mixture of gamma densities, $\sqrt{}$-link
EEE	extended estimating equations
CDEM	conditional density estimator (multinomial logit)
CDEO	conditional density estimator (ordered logit)

Table 2: Key for model labels

6.1 Estimation Sample Results

We first conduct tests of misspecification across the models used. Researchers use these tests to inform the specification of regressors, and the appropriateness of distributional assumptions, in particular the link function. Since we use the same regressors in all models, our tests are used to inform choices of distributional assumptions. The Pearson correlation coefficient test is able to detect if there is a a linear association between the estimated residuals and estimated conditional means, where the null hypothesis is no association. A lack of this kind of association suggests evidence against misspecification. However it is also possible that the relationship between the error and covariates is non-linear which this test cannot detect. Linear regression estimated using OLS, by construction, generates residuals orthogonal to predicted costs, and so the Pearson test cannot be applied to this model.

Model	**Pearson**
OLS	N/A
LOGOLSHET	99%
SQRTOLSHET	0%
GLMLOGP	11%
GLMLOGG	99%
GLMSQRTP	0%
GLMSQRTG	13%
LOGNORM	95%
GG	89%
GB2LOG	96%
GB2SQRT	85%
FMMLOGG	85%
FMMSQRTG	82%
EEE	48%
CDEM	7%
CDEO	1%

Table 3: % of tests rejected at 5% significance level, when all converged, 94 converged replications, sample size 5,000

Table 3 shows that according to this test, there is less evidence of misspecification when the model is estimated using a square root link function compared to other possible link functions, when all other distributional assumptions are the same. This is also the case in the GLM family of models, where the link and distribution functions can be flexibly estimated using EEE, with results indicating that there is less evidence of misspecification with GLMSQRTP and GLMSQRTG than the flexible case (on average across replications with sample size 5,000, the estimated λ coefficient in EEE was 0.28 with standard deviation of 0.07, indicating a link function between logarithmic and square root). Whilst EEE should be better specified on the scale of estimation (following effective transformation of dependent variable), the re-transformation may lead to increased evidence of misspecification on the scale of interest which is the subject of this comparison (levels of costs). Introducing more flexibility in terms of the whole distribution, generally, appears to have mixed effects upon results from this test. In the case of LOGNORM and GLMLOGG which are special cases of GG, there is the least evidence of misspecification from the most complicated distribution amongst the three. There is also evidence of less misspecification with

FMMLOGG compared to GLMLOGG, which it nests. Conversely, GG and LOGNORM are special cases of GB2LOG, for which there is the most evidence of misspecification from these three models. Looking at the rejection rates above for FMMSQRTG and GLMSQRTG, there is more evidence of misspecification in the more flexible case. Finally, the results from CDEM and CDEO are promising, with little evidence of misspecification compared to other models tested. This may be because there is no retransformation process onto the scale of interest for these models.

6.2 Validation Set Results

All tests in the above section are carried out on the estimation sample. Given the practical implementation of the models considered here, a researcher may be more interested in how models perform in forecasting costs out-of-sample. Results based on the estimation sample may arise from overfitting the data. Therefore, our main focus is the forecasting performance out-of-sample, that is evaluation on the 'validation' set.

We look first at performance of model predictions on the whole 'validation' set. Then we consider how well the models forecast for different levels of covariates throughout the distribution, by analysing performance by decile of predicted costs. Finally, we analyse the out-of-sample performance with increasing sample size by constructing response surfaces.

	Bias	Accuracy	Goodness of fit
	MPE (£)	**MAPE (£)**	**RMSE**
OLS	-1.56	1833.49	4475.49
LOGOLSHET	-140.53	1816.63	4960.08
SQRTOLSHET	**0.11**	1725.95	**4432.94**
GLMLOGP	**-1.44**	1748.43	4557.19
GLMLOGG	-147.33	1818.06	4984.86
GLMSQRTP	**0.26**	1704.77	**4426.24**
GLMSQRTG	46.71	**1689.28**	**4454.25**
LOGNORM	64.25	1734.10	4825.51
GG	44.60	1750.79	4754.22
GB2LOG	-63.96	1796.91	4873.13
GB2SQRT	134.84	**1686.48**	4483.35
FMMLOGG	-3.19	1758.06	4782.69
FMMSQRTG	121.80	**1690.28**	4477.10
EEE	-42.31	1727.28	4508.03
CDEM	**0.89**	**1683.40**	**4444.85**
CDEO	-10.13	1725.53	4474.84

Table 4: Results of model performance, when all converged, sample size 5,000; averaged across 94 replications

Looking at the results in Table 4, where the four best performing models in each category (MPE, MAPE and RMSE) are emboldened, it is clear that some of the most commonly used models: OLS, LOGOLSHET, GLMLOGG, and LOGNORM, do not perform well on any metric. CDEM is among the models with top four performance in every category illustrating the potential advantages of this approach for analysts concerned with any of bias, accuracy or goodness of fit. Generally speaking, the results also indicate that a square root link function is the most appropriate of those featured.

In terms of bias, models which are mean preserving in sample also perform well out-of-sample in these results. This is evidenced by the strong performance of OLS, GLMLOGP and GLMSQRTP, with absolute levels of mean prediction error of £1.56, £1.44 and £0.26 respectively. All models with a square root link function underpredict costs on average, whereas some log link function models underpredict (LOGNORM and GG) and others overpredict on average (LOGOLSHET, GLMLOGP, GB2LOG and FMMLOGG). SQRTOLSHET and CDEM perform best and third best respectively, and worst performing is GLMLOGG, which overpredicts by £147.33 on average (5.64% of the population mean).

With respect to accuracy and goodness of fit, there is a clear message from the results that the best performing link function is the square root. The ordering of the other link functions varies. For accuracy the flexible link function of EEE is next best, followed by log link function and then OLS. For goodness of fit OLS is second best, followed by EEE and the log link is the worst. There is variation in performance amongst different models with the same link function, which we discuss next when considering the gains to increased flexibility. In addition, CDEM performs very well according to these criteria.

First we consider the gains to using a mixture of gamma distributions, over the nested single gamma distribution models. Looking at the results for the GLMLOGG and FMMLOGG, the mixture improves forecasting performance in terms of bias, accuracy and goodness of fit. This is also observed in results from other sample sizes (see Online Appendix B). As discussed earlier, the gains to this increased flexibility are insufficient for results from FMMLOGG to perform better than relatively simple models using a square root link function (e.g. GLMSQRTP). Comparing results from GLMSQRTG with FMMSQRTG is more complicated, at sample size 5,000, as seen in Table 4, we observe GLMSQRTG to perform better than FMMSQRTG in all metrics. When looking at results using larger samples, FMMSQRTG performs better than GLMSQRTG in terms of accuracy, where FMMSQRTG is actually the best performing model of all 16 compared, but the nested single distribution case still performs better in terms of bias and goodness of fit (see Online Appendix B).

Greater flexibility amongst the fully parametric models has an ambiguous effect on performance of forecasting means. GG is a limiting case of

GB2 and its performance is better across all metrics. On the other hand, LOGNORM, a special case of GG and GB2, performs best of the three in terms of accuracy, and between GG (best) and GB2LOG (worst) in terms of goodness of fit (measured by RMSE), and worst in terms of bias. Using GG or GB2LOG improves performance over special case GLMLOGG based on MPE, MAPE and RMSE. Once again, the best of these four models performs worse than certain models with a square root link function. Comparing GLMSQRTG and GB2SQRT, we can see that there is not a great deal gained from introducing more parameters, since performance is worse for GB2SQRT than GLMSQRTG except in the cases of accuracy at sample sizes 5,000 and 10,000 (the difference is small at all sample sizes analysed).

Crucially, these results only consider performance based on the mean, while some of these models are capable of providing information on higher moments and on other features of the conditional distribution such as tail probabilities[19]. We construct graphs of bias and accuracy by decile of predicted costs. This can be thought of as analysing the fit of models for the mean of distributions of costs conditional on observed variables, since each decile of predicted costs represents a group of observations with certain values of covariates. In previous analysis, we have considered all observations as equal, but it is possible that a policymaker prioritises the prediction error of certain observations over others. There is considerable interest in modelling the outcomes for high-cost patients, since these can be responsible for large proportions of overall costs. The highest costs are likely to be found in the highest decile of predicted costs.

[19] This is a significant qualitative advantage of parametric models over models such as linear regression, where the models have been used to predict probabilities of lying beyond a threshold value, e.g. tail probabilities, see Jones et al. (2013) who find that the GG and LOGNORM distribution perform best for the threshold values they choose.

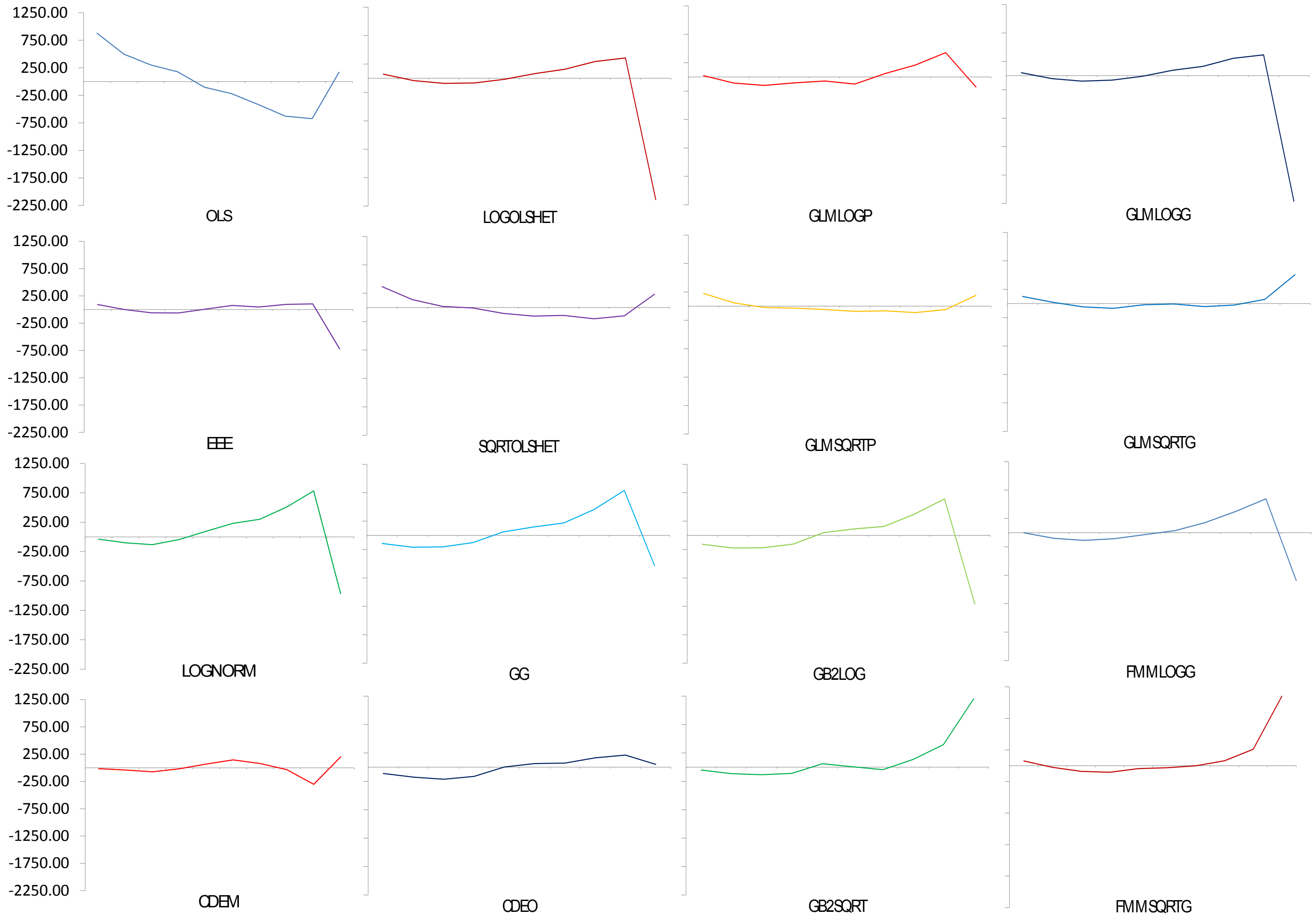

Figure 6: MPE by decile of fitted costs

Models with the same link function follow a largely similar pattern, for example those with square root link functions underpredict in the decile of highest predicted costs, whereas log link models overpredict in the last decile. Results with other link functions: OLS, EEE, CDEM and CDEO all have different patterns. Generally speaking, the first decile of predicted costs from square root models are on average underpredictions (only GB2SQRT overpredicts in the smallest decile), which combined with underpredicted last decile gives them a 'u-shaped' line. The performance of each model varies across the deciles. SQRTOLSHET has a 'u-shaped' line, and while it performs best in predicting costs on average across all deciles, the performance for certain groups may be worse than other models. For example, CDEM performs slightly worse across all ten deciles, but has a smaller range of over/underpredictions. In terms of the highest decile of predicted costs, the model with the lowest MPE is CDEO underestimating on average £48.96, generally this decile tends to be the largest absolute MPE for models with values as large as an average overprediction of £2211.47 in the case of GLMLOGG.

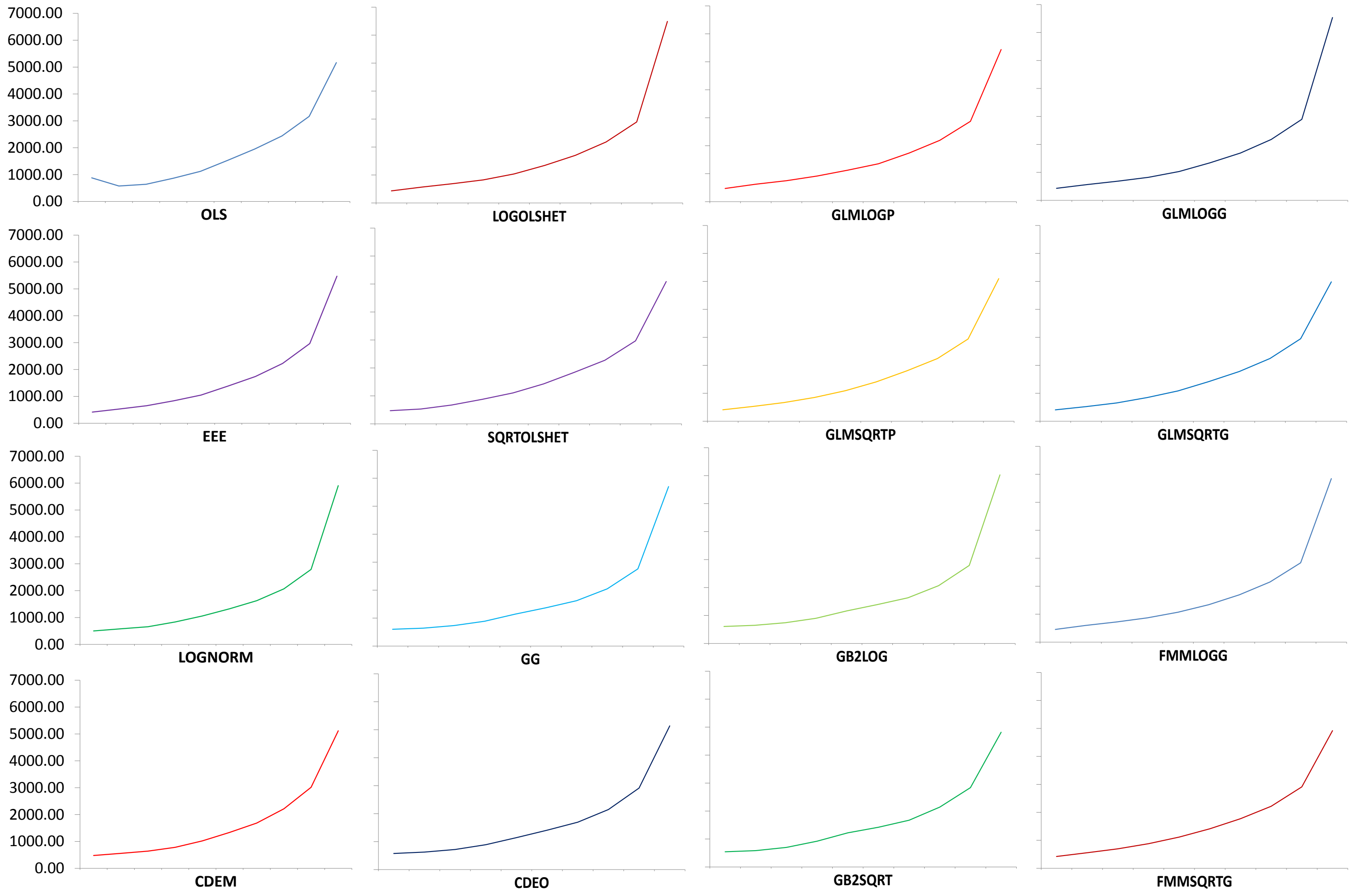

Figure 7: MAPE by decile of fitted costs

When looking at MAPE by decile of predicted cost, it is striking that the pattern across models is very similar. In all models, except OLS, the MAPE is higher in deciles with larger predicted costs. The most inaccurate models in the highest decile are those with a log link function, followed by EEE, then OLS, the conditional density estimators and finally the most accurate are models with a square root link function. Generally, it appears that models that predict larger costs overall are the least accurate in the highest decile, implying that models which estimate the largest range of predicted conditional means will not necessarily perform best in forecasting mean costs for patients most likely to be high cost patients (with lots of observed morbidity). GLMLOGG overpredicts over the whole validation subset by £147.33, has the largest overprediction in the highest decile, and is also the least accurate in this decile with MAPE of £6536.24, over twice the population mean cost.

Figure 8 displays the response surfaces constructed to analyse how each model's performance varied with increasing sample size for the subset of best performing models (those emboldened in Table 4). We have already touched upon this earlier when looking at results regarding accuracy between related distributions. The performance of most estimated models varies little as sample sizes increase above 5,000. There is some evidence of the variability of MPE (measured using ADMPE) reducing as sample size increases, although this happens at a similar rate across all models. Largely, though, the response surfaces for each model are parallel indicating that relative performance of models changes little, and are also flat - evidence of performance not changing for each model with increasing sample size. The exception to this is that the performance of FMMSQRTG varies with increasing sample size: its accuracy improves, and its bias worsens. This suggests that this model behaves differently with samples as small as 5,000 observations, possibly because of the number of parameters that are required. On the whole, though, from samples of 5,000 observations or more, there is little evidence that more flexible models require more observations than less flexible ones.

7 Conclusion

We have systematically evaluated the state of the art in regression models for healthcare costs, using administrative English hospital inpatient data, employing a quasi-Monte Carlo design to ensure rigour and based on out-of-sample forecasting. We have compared recently adopted semi- and fully parametric regression methods that have never before been evaluated against one another, as well as comparing with regression methods that are now considered standard practice in modelling healthcare cost data.

Our results echo other studies, in that there is no single model that

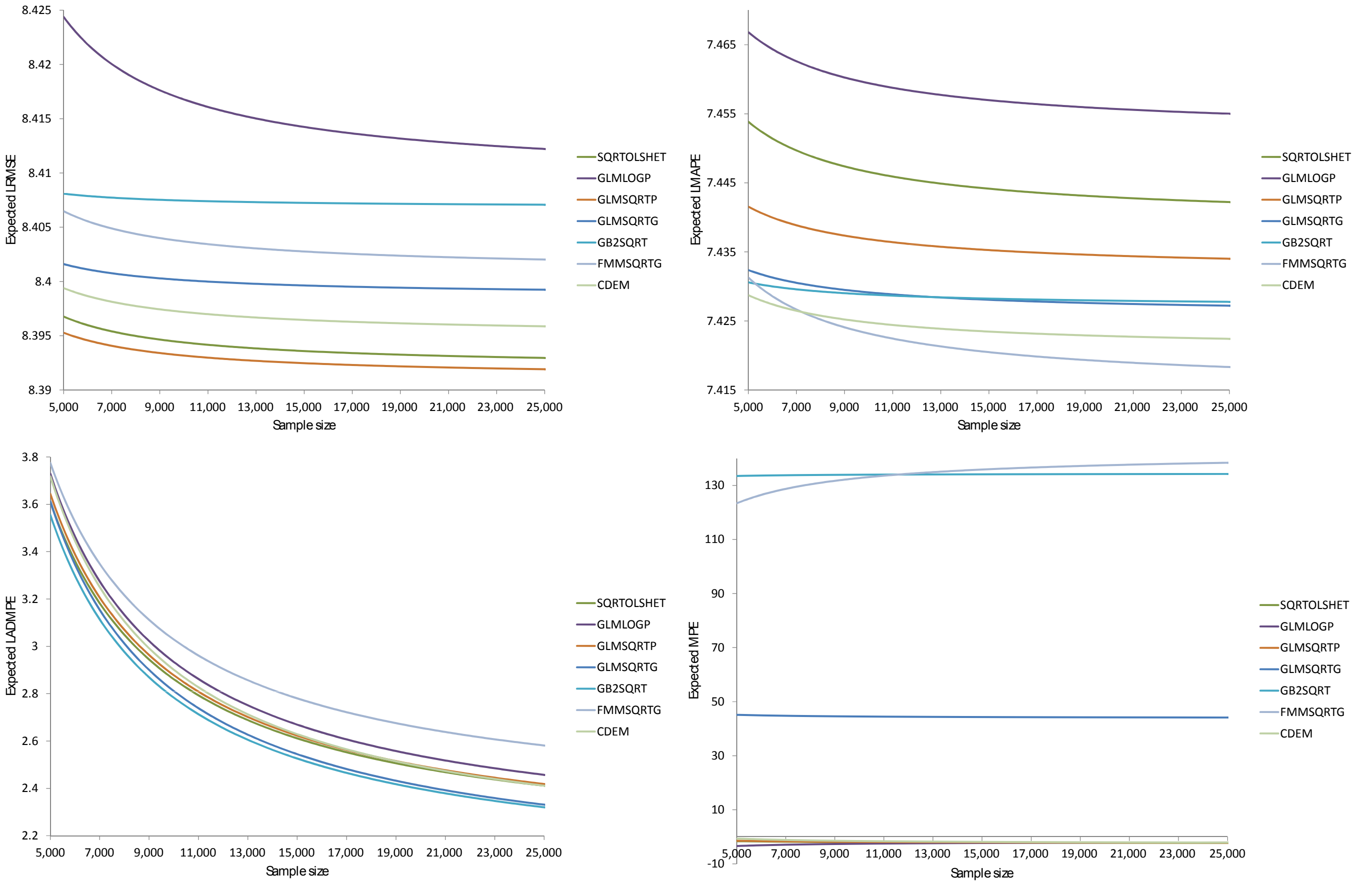

Figure 8: Reponse surfaces for log(RMSE), log(MAPE), MPE, log(ADMPE) (clockwise from top left) against sample size, constructed evaluating performance on 'validation' set

dominates in all respects: SQRTOLSHET is the best performing model in terms of bias, CDEM for accuracy, and in terms of goodness of fit the best performer is GLMSQRTP. Therefore the policymaker has to weigh up these factors in arriving at their preferred model, based upon their loss function over prediction errors. It is worth noting, however, that CDEM performs amongst the best four models for all three metrics. It is also worth noting that four models, commonly employed in regression methods for healthcare costs, do not perform amongst the best four of any of the three metrics (OLS, LOGOLSHET, GLMLOGG and LOGNORM). Our analysis by decile shows the way in which models are sensitive to the choice of link function, with square root link functions underpredicting in the decile of highest predicted costs, and log link models overpredicting in the last decile. Finally, the response surfaces indicate that, on the whole, the more recent developments do not suffer because of using smaller sample sizes (from 5,000 observations).

Acknowledgements

The authors gratefully acknowledge funding from the Economic and Social Research Council (ESRC) under grant reference RES-060-25-0045. James Lomas is thankful for funding from the Royal Economics Society Junior Fellowship Scheme. We are also grateful to members of the Health, Econometrics and Data Group (HEDG) at the University of York for useful discussions, and to participants of the Annual Health Econometrics Symposium, University of Leeds 2013, in particular Sandy Tubeuf, for their insightful comments.

References

Arrow KJ, Lind RC. 1970. Uncertainty and the evaluation of public investment decisions. *The American Economic Review* **60**: 364–378.

Basu A, Arondekar BV, Rathouz PJ. 2006. Scale of interest versus scale of estimation: comparing alternative estimators for the incremental costs of a comorbidity. *Health Economics* **15**: 1091–1107.

Basu A, Manning WG, Mullahy J. 2004. Comparing alternative models: log vs cox proportional hazard? *Health Economics* **13**: 749–765.

Basu A, Rathouz PJ. 2005. Estimating marginal and incremental effects on health outcomes using flexible link and variance function models. *Biostatistics* **6**: 93–109.

Blough DK, Madden CW, Hornbrook MC. 1999. Modeling risk using generalized linear models. *Journal of Health Economics* **18**: 153–171.

Bordley R, McDonald J, Mantrala A. 1997. Something new, something old: Parametric models for the size of distribution of income. *Journal of Income Distribution* **6**: 91–103.

Buntin MB, Zaslavsky AM. 2004. Too much ado about two-part models and transformation?: Comparing methods of modeling medicare expenditures. *Journal of Health Economics* **23**: 525–542.

Cawley J, Meyerhoefer C. 2012. The medical care costs of obesity: An instrumental variables approach. *Journal of Health Economics* **31**: 219 – 230.

Copas JB. 1983. Regression, prediction and shrinkage. *Journal of the Royal Statistical Society. Series B* **45**: pp. 311–354.

Cummins JD, Dionne G, McDonald JB, Pritchett BM. 1990. Applications of the GB2 family of distributions in modeling insurance loss processes. *Insurance: Mathematics and Economics* **9**: 257–272.

Deb P, Burgess JF. 2003. A quasi-experimental comparison of econometric models for health care expenditures. *Hunter College Department of Economics Working Papers* **212**.

Deb P, Trivedi PK. 1997. Demand for medical care by the elderly: A finite mixture approach. *Journal of Applied Econometrics* **12**: 313–336.

Dixon J, Smith P, Gravelle H, Martin S, Bardsley M, Rice N, Georghiou T, Dusheiko M, Billings J, Lorenzo MD, Sanderson C. 2011. A person based formula for allocating commissioning funds to general practices in england: development of a statistical model. *British Medical Journal* **343**: d:6608.

Duan N. 1983. Smearing estimate: A nonparametric retransformation method. *Journal of the American Statistical Association* **78**: 605–610.

Duan N, Manning WG, Morris CN, Newhouse JP. 1983. A comparison of alternative models for the demand for medical care. *Journal of Business & Economic Statistics* **1**: 115–126.

Gilleskie DB, Mroz TA. 2004. A flexible approach for estimating the effects of covariates on health expenditures. *Journal of Health Economics* **23**: 391–418.

Han A, Hausman JA. 1990. Flexible parametric estimation of duration and competing risk models. *Journal of Applied Econometrics* **5**: 1–28.

Heckman JJ. 2001. Micro data, heterogeneity, and the evaluation of public policy: Nobel lecture. *Journal of Political Economy* **109**: 673–748.

Hill SC, Miller GE. 2010. Health expenditure estimation and functional form: applications of the generalized gamma and extended estimating equations models. *Health Economics* **19**: 608–627.

Hoch JS, Briggs AH, Willan AR. 2002. Something old, something new, something borrowed, something blue: a framework for the marriage of health econometrics and cost-effectiveness analysis. *Health Economics* **11**: 415–430.

Huber M, Lechner M, Wunsch C. 2013. The performance of estimators based on the propensity score. *Journal of Econometrics* **175**: 1–21.

Johnson E, Dominici F, Griswold M, L Zeger S. 2003. Disease cases and their medical costs attributable to smoking: an analysis of the national medical expenditure survey. *Journal of Econometrics* **112**: 135–151.

Jones AM. 2011. Models for health care. In Clements MP, Hendry DF (eds.) *Oxford Handbook of Economic Forecasting.* Oxford University Press.

Jones AM, Lomas J, Rice N. 2013. Applying beta-type size distributions to healthcare cost regressions. *Journal of Applied Econometrics* In press.

Manning W, Duan N, Rogers W. 1987. Monte carlo evidence on the choice between sample selection and two-part models. *Journal of Econometrics* **35**: 59–82.

Manning WG, Basu A, Mullahy J. 2005. Generalized modeling approaches to risk adjustment of skewed outcomes data. *Journal of Health Economics* **24**: 465–488.

McDonald JB, Sorensen J, Turley PA. 2013. Skewness and kurtosis properties of income distribution models. *Review of Income and Wealth* **59**: 360–374.

Mullahy J. 1997. Heterogeneity, excess zeros, and the structure of count data models. *Journal of Applied Econometrics* **12**: 337–350.

Mullahy J. 2009. Econometric modeling of health care costs and expenditures: a survey of analytical issues and related policy considerations. *Medical Care* **47**: S104–108.

Van de ven WP, Ellis RP. 2000. Risk adjustment in competitive health plan markets. In Culyer AJ, Newhouse JP (eds.) *Handbook of Health Economics*, volume 1. Elsevier, 755–845.

Vanness DJ, Mullahy J. 2007. Perspectives on mean-based evaluation of health care. In Jones AM (ed.) *The Elgar Companion to Health Economics*. Edward Elgar.

Veazie PJ, Manning WG, Kane RL. 2003. Improving risk adjustment for medicare capitated reimbursement using nonlinear models. *Medical Care* **41**: 741–752.

WHO. 2007. International statistical classification of diseases and related health problems 10th Revision.

Appendix

We use the variables shown in Table A1 to construct our regression models. They are based on the ICD10 chapters, which are given in Table A2.

Variable name	Variable description
epiA	Intestinal infectious diseases, Tuberculosis, Certain zoonotic bacterial diseases, Other bacterial diseases, Infections with a predominantly sexual mode of transmission, Other spirochaetal diseases, Other diseases caused by chlamydiae, Rickettsioses, Viral infections of the central nervous system, Arthropod-borne viral fevers and viral haemorrhagic fevers
epiB	Viral infections characterized by skin and mucous membrane lesions, Viral hepatitis, HIV disease, Other viral diseases, Mycoses, Protozoal diseases, Helminthiases, Pediculosis, acaiasis and other infestations, Sequelae of infectious and parasitic diseases, Bacterial, viral and other infectious agents, Other infectious diseases
epiC	Malignant neoplasms
epiD	In situ neoplasms, Benign neoplasms, Neoplasms of uncertain or unknown behaviour and III
epiE	IV
epiF	V
epiG	VI
epiH	VII and VIII
epiI	IX
epiJ	X
epiK	XI
epiL	XII
epiM	XIII
epiN	XIV
epiOP	XV and XVI
epiQ	XVII
epiR	XVIII
epiS	Injuries to the head, Injuries to the neck, Injuries to the thorax, Injuries to the abdomen, lower back, lumbar spine and pelvis, Injuries to the shoulder and upper arm, Injuries to the elbow and forearm, Injuries to the wrist and hand, Injuries to the hip and thigh, Injuries to the knee and lower leg, Injuries to the ankle and foot
epiT	Injuries involving multiple body regions, Injuries to unspecified part of trunk, limb or body region, Effects of foreign body entering through natural orifice, Burns and Corrosions, Frostbite, Poisoning by drugs, medicaments and biological substances, Toxic effects of substances chiefly nonmedicinal as to source, Other and unspecified effects of external causes, Certain early complications of trauma, Comlications of surgical and medical care, not elsewhere classified, Sequelae of injuries, of poisoning and of other consequences of external causes
epiU	XXII
epiV	Transport accidents
epiW	Falls, Exposure to inanimate mechanical forces, Exposure to animate mechanical forces, Accidental drowning and submersion, Other accidental threats to breathing, Exposure to electric current, radiation and extreme ambient air temperature and pressure
epiX	Exposure to smoke, fire and flames, Contact with heat and hot substances, Contact with venomous animals and plants, Exposure to forces of nature, Accidental poisoning by and exposure to noxious substances, Overexertion, travel and privation, Accidental exposure to other and unspecified factors, Intentional self-harm, Assault by drugs, medicaments and biological substances, Assault by corrosive substance, Assault by pesticides, Assault by gases and vapours, Assault by other specified chemicals and noxious substances, Assault by unspecified chemical or noxious substance, Assault by hanging, strangulation and suffocation, Assault by drowning and submersion, Assault by handgun discharge, Assault by rifle, shotgun and larger firearm discharge, Assault by other and unspecified firearm discharge, Assault by explosive material, Assault by smoke, fire and flames, Assault by steam, hot vapours and hot objects, Assault by sharp object
epiY	Assault by blunt object, Assault by pushing from high place, Assault by pushing or placing victim before moving object, Assault by crashing of motor vehicle, Assault by bodily force, Sexual assault by bodily force, Neglect and abandonment, Other maltreatment syndromes, Assault by other specified means, Assault by unspecified means, Event of undetermined intent, Legal intervention and operations of war, Complications of medical and surgical care, Sequelae of external causes of morbidity and mortality, Supplementary factors related to causes of morbidity and mortality classified else
epiZ	XXI

Table A1: Classification of morbidity characteristics

ICD10 codes beginning with U were dropped because there were no observations in the 6,164,114 used. Only a small number (3,170) were found of those beginning with P and so these were combined with those beginning with O - owing to the clinical similarities.

Chapter	**Blocks**	**Title**
I	A00-B99	Certain infectious and parasitic diseases
II	C00-D48	Neoplasms
III	D50-D89	Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism
IV	E00-E90	Endocrine, nutritional and metabolic diseases
V	F00-F99	Mental and behavioural disorders
VI	G00-G99	Diseases of the nervous system
VII	H00-H59	Diseases of the eye and adnexa
VIII	H60-H95	Diseases of the ear and mastoid process
IX	I00-I99	Diseases of the circulatory system
X	J00-J99	Diseases of the respiratory system
XI	K00-K93	Diseases of the digestive system
XII	L00-L99	Diseases of the skin and subcutaneous tissue
XIII	M00-M99	Diseases of the musculoskeletal system and connective tissue
XIV	N00-N99	Diseases of the genitourinary system
XV	O00-O99	Pregnancy, childbirth and the puerperium
XVI	P00-P96	Certain conditions originating in the perinatal period
XVII	Q00-Q99	Congenital malformations, deformations and chromosomal abnormalities
XVIII	R00-R99	Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified
XIX	S00-T98	Injury, poisoning and certain other consequences of external causes
XX	V01-Y98	External causes of morbidity and mortality
XXI	Z00-Z99	Factors influencing health status and contact with health services
XXII	U00-U99	Codes for special purposes

Table A2: ICD10 chapter codes