

# Going Beyond the Mean in Healthcare Cost Regressions: a Comparison of Methods for Estimating the Full Conditional Distribution

Andrew M. Jones <sup>a</sup>      James Lomas <sup>a,b,\*</sup>      Nigel Rice <sup>a,b</sup>

<sup>a</sup> *Department of Economics and Related Studies, University of York, YO10 5DD, UK*

<sup>b</sup> *Centre for Health Economics, University of York, YO10 5DD, UK*

May 31, 2014

---

## Summary

Understanding the drivers of the data generating process behind healthcare costs remains a key empirical issue. Much research to date has focused on the prediction of the conditional mean cost, although this can potentially miss important features of the distribution for policymakers. We conduct a quasi-Monte Carlo experiment using English NHS inpatient data to compare 14 approaches to modelling the distribution of healthcare costs: nine of which are parametric in nature and have been proposed for fitting healthcare costs and five less parametric methods designed specifically for constructing a counterfactual distribution. Our results indicate that there is no one method is clearly dominant and that there is a trade-off between bias and precision. The method proposed by Chernozhukov et al. (2013), designed specifically for constructing a counterfactual distribution, seems to perform relatively well according to both metrics, ranked best in terms of bias, but not in terms of precision. While this method and its related approaches show considerable promise, in particular with larger sample sizes, they cannot be used to extrapolate outside of the observed range of the sample.

*JEL classification:* C1; C5

*Key words:* Healthcare costs; heavy tails; counterfactual distributions; quasi-Monte Carlo

---

\*Corresponding author: Tel.: +44 1904 32 1411.

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

# 1 Introduction

There is a great deal of interest in modelling healthcare costs within a number of areas of research aimed at informing healthcare policy through understanding the drivers of healthcare expenditures. Econometric models of healthcare costs have diverse and many uses including: to estimate key parameters for populating decision models in cost-effectiveness analyses (Hoch et al., 2002); to adjust for healthcare need in resource allocation formulae in publically funded healthcare systems (Dixon et al., 2011); to undertake risk adjustment in insurance systems (Van de ven and Ellis, 2000) and to assess the effect on resource use of observable lifestyle characteristics such as smoking and obesity (Johnson et al., 2003; Cawley and Meyerhoefer, 2012; Mora et al., 2014).

The distribution of healthcare costs poses substantial challenges for econometric modelling. Healthcare costs are non-negative, highly asymmetric and leptokurtic, and often exhibit a large mass point at zero. The relationships between covariates and costs are likely to be non-linear. Basu and Manning (2009) provides a useful discussion of these issues. The relevance and complexity of modelling healthcare costs has led to the development of a wide range of econometric approaches, and a description of commonly found approaches can be found in Jones (2011).

Much of the focus in comparisons of regression methods for the analysis of healthcare cost data has centered on predictions of the conditional mean of the distribution,  $E(y|X)$  (Deb and Burgess, 2003; Veazie et al., 2003; Basu et al., 2004; Buntin and Zaslavsky, 2004; Gilleskie and Mroz, 2004; Manning et al., 2005; Basu et al., 2006; Hill and Miller, 2010; Jones, 2011; Jones et al., 2013, 2014). Applied researchers commonly model cost data using generalised linear models (GLMs) (Blough et al., 1999). This framework offers a relatively simple way to incorporate non-linearities in the relationship between the conditional mean and observed covariates. Furthermore, GLMs allow for heteroskedasticity through a choice of a ‘distribution’ which specifies the conditional variance as a function of the conditional mean. GLMs use pseudo-maximum likelihood estimation where the researcher is required only to specify the form of the mean and the variance. Unlike maximum likelihood estimation, where consistency requires that the whole likelihood function is correctly specified, pseudo-maximum likelihood is consistent so long as the mean is correctly specified with the choice of distribution affecting the efficiency of estimates. Whilst the GLM framework has attractive properties for researchers concerned only with  $E(y|X)$ , there are important limitations with this method. GLMs have been found to perform badly with heavy-tailed data (Manning and Mullahy, 2001), and they implicitly impose restrictions on the entire distribution. For example, whatever distribution is adopted, the skewness must be directly proportional to the coefficient of variation and the kurtosis is linearly related to the square of the coefficient of

variation (Holly, 2009). Whilst they may be well placed to estimate  $E(y|X)$  (and  $Var(y|X)$ ), they cannot produce estimates of  $F(y|X)$  or  $Pr(y > k|X)$ .

While the mean is an important feature of a distribution, which is essential when the analysis is concerned with the expected total cost, it is generally not the only aspect that is interesting to policymakers (Vanness and Mullahy, 2007). Analysis based solely on the mean misses out potentially important information in other parts of the distribution (Bitler et al., 2006). As a result, a growing literature in econometrics has developed techniques to model the entire distribution,  $F(y|X)$ , thus ‘going beyond the mean’ (Fortin et al., 2011). In health economics there is a particular emphasis on identifying individuals or characteristics of individuals that lead to very large costs and there is a demand for empirical strategies to “target the high-end parameters of particular interest” including tail probabilities,  $Pr(y > k)$  (Mullahy, 2009).

In this paper we conduct a quasi-Monte Carlo experiment to compare fit of the entire conditional distribution of healthcare costs using competing approaches proposed in the economics literature. We therefore consider approaches which offer greater flexibility in terms of their potential applications by estimating  $F(y|X)$ , imposing fewer restrictions on skewness and kurtosis and allowing for a greater range of estimated effects of a covariate.

We first consider developments in the use of flexible parametric distributions for modelling healthcare costs (Manning et al., 2005; Jones et al., 2014), which have been applied to healthcare costs principally in order to overcome the challenge posed by heavy-tailed data. Unlike the GLM framework, these models impose a functional form for the entire distribution with estimation by maximum likelihood. As a result, an estimate of  $f(y|X)$  is produced, which can then be used to calculate  $E(y|X)$ ,  $Var(y|X)$ <sup>1</sup> and  $Pr(y > k|X)$  as required. By using flexible distributions, the restrictions on skewness and kurtosis can be relaxed somewhat (McDonald et al., 2011), which is likely to lead to a better fit of the full distribution (Jones et al., 2014).<sup>2</sup>

A related development is the use of finite mixture models (FMM), which allow the distribution to be estimated as a weighted sum of distribution components (Deb and Trivedi, 1997; Deb and Burgess, 2003). These are also estimated using maximum likelihood, but are often referred to as semi-parametric, since the number of components could, in principal, be increased to approximate any distribution. In this paper we will group FMM with the fully parametric distributions given the similarities to these approaches, especially since we use a fixed number of components.

Other developments regarding the estimation of  $f(y|X)$  for healthcare costs are less parametric, they typically involve dividing the outcome vari-

---

<sup>1</sup>Note that population moments may not be defined for all ranges of parameter estimates (Mullahy, 2009).

<sup>2</sup>It is also possible to specify multiple parameters as functions of regressors to allow for more complex covariate effects.

able into discrete intervals and estimating parameters for each of these intervals. Gilleskie and Mroz (2004) propose using a conditional density approximation estimator for healthcare costs to calculate  $E(y|X)$  and other moments, where the density function is approximated by a set discrete hazard rates. To implement this Jones et al. (2013) use an approach based on Han and Hausman (1990), where  $F(y|X)$  is estimated by creating a categorical variable that describes the cost interval into which each observation falls, and running an ordered logit with this as the dependent variable. This implementation is slightly different from what is proposed by Gilleskie and Mroz (2004), but has the advantage of being conceived in order to fit  $F(y|X)$  and ties into a related literature on semi-parametric estimators for conditional distributions (Han and Hausman, 1990; Foresi and Peracchi, 1995; Chernozhukov et al., 2013). While the ordered logit specification used in the Han and Hausman (1990) method allows for flexible estimation of the thresholds in the latent scale, methods such as Foresi and Peracchi (1995) instead estimate a series of separate logit models. More recently, Chernozhukov et al. (2013) propose that a continuum of logits should be estimated (one for each unique value of the outcome variable) to allow for an even greater range of estimates for the effect of a covariate. In an application to Dutch health expenditures, de Meijer et al. (2013) use the Chernozhukov et al. (2013) method to decompose changes in the distribution of health expenditures between two periods. The authors find that the effect of covariates varies across the distribution of health expenditures, which would have been missed if analysis had focused solely on the mean. They also find that pharmaceutical costs were growing mainly at the top of the distribution due to structural effects, whereas growth in hospital care costs is observed more in the middle of the distribution and can be explained by changes in the observed determinants of expenditure.

The above methods seek to estimate the full distribution, by modelling  $F(y|X)$  for different values of  $y$  (interval thresholds) and imposing varying degrees of flexibility on the covariate effects for these. An alternative is to construct  $F(y|X)$  through the inverse of the distribution function, the quantile function  $q_\tau(X)$ .<sup>3</sup> We consider two methods which estimate a range of quantiles separately as functions of the covariates to allow for flexibility as to the estimated effects of each regressor across the full range of the distribution. The first was proposed by Machado and Mata (2005) and Melly (2005) and uses a series of quantile regressions to estimate the full range of quantiles across the distribution. Quantile regressions have been used where the outcome variable was healthcare costs for analysing the varying effects of race at different points of the distribution (Cook and Manning, 2009). However we were unable to find any applications of the Machado and Mata (2005) and Melly (2005) method to construct a complete estimate of  $F(y|X)$

---

<sup>3</sup> $\tau \in (0, 1)$  denotes the quantile being considered.

with healthcare costs as the outcome variable, although the applications in the original papers were to wages, which share similar distributional characteristics. Quantile functions can alternatively be estimated using recentered-influence-function (RIF) regression (Firpo et al., 2009), where the outcome variable is first transformed according to the recentered-influence-function and then regression used to model the effects of covariates.

This paper provides a systematic comparison of parametric and distributional methods<sup>4</sup> for fitting the full distribution of healthcare costs using real data in a quasi-Monte Carlo experiment. As such, it is novel in two ways: firstly, it provides a methodology for comparing the distributional fit of models which are neither nested/limiting cases nor estimated using the same procedure, and secondly it is the first paper to compare competing econometric approaches for modelling the distribution of healthcare costs. We find that distributional methods demonstrate significant potential in modelling tail probabilities, particularly with larger sample sizes where the variability of predictions is reduced. Parametric distributions such as log-normal, generalised gamma and generalised beta of the second kind are found to estimate tail probabilities with high precision, but with varying bias depending upon the cost threshold being considered.

The study design is described in the next section, followed by a detailed description of the methods compared. Then we discuss the results, and place these in the context of related research, some of the limitations of our study and possible extensions for future work.

## 2 Methodology and Data

### 2.1 Overview

Rather than comparing competing approaches for estimating  $E(y|X)$ , which is the focus of most empirical work in this area (Mullahy, 2009), we assess performance in terms of tail probabilities,  $Pr(y > k)$ , for varying levels of  $k$  to assess the fit of the entire distribution,  $F(y|X)$ . We compare a number of different regression methods, each with a different number of estimated parameters. Since more complex methods may capture idiosyncratic characteristics of the data as well as the systematic relationships between the dependent and explanatory variables, there is a concern that better fit will not necessarily be replicated when the model is applied to new data (Bilger and Manning, 2014). To guard against this affecting our results we use a quasi-Monte Carlo design where models are fitted to a sample drawn from an ‘estimation’ set and performance is evaluated on a ‘validation’ set. This means that methods are assessed when being applied to new data.<sup>5</sup>

---

<sup>4</sup>This term was used in Fortin et al. (2011).

<sup>5</sup>There are substantial precedents for using split-sample methods to evaluate different regression methods for healthcare costs, for example Duan et al. (1983); Manning et al.

Each method can be used to produce an estimate of the whole distribution  $F(y|X)$  for each sample. We use the estimated parameters to produce a counterfactual distribution given the covariates in the ‘validation’ set. Our interest is in the survival function,  $Pr(y > k)$ , and use round values of  $k$  to assess the performance of each method, where the values of  $k$  are unknown to the researcher at the point of estimating each method.

## 2.2 Data

Our data comes from the English administrative dataset the Hospital Episode Statistics (HES)<sup>6</sup> for the financial year 2007-2008. We have excluded spells which were primarily mental or maternity healthcare and all spells taking place within private sector hospitals.<sup>7</sup> The remaining spells constitute the population of all inpatient episodes, outpatient visits and A&E attendances that were completed within 2007-2008 for all patients who were admitted to English NHS hospitals (where treatment was not primarily mental or maternity healthcare). Spells are costed using tariffs from 2008-2009<sup>8</sup> by applying the relevant tariff to the most expensive episode within the spell (where a spell can be thought of as a discrete admission).<sup>9</sup> Our analysis is undertaken at the patient level and so we sum costed spells for each patient to create the dependent variable, giving us 6,164,114 observations in total. The empirical density and cumulative distribution of the outcome variable can be seen in Figure 1 and descriptive statistics are found in Table 1.<sup>10</sup>

In order to tie in with existing literature on comparisons of econometric methods for healthcare costs we use a set of morbidity characteristics which is the same for each regression method. In addition, we control for age and sex using an interacted, cubic specification, which leaves us with a set of regressors similar to a simplified resource allocation formula where health expenditures are modelled as a function of need proxied using detailed socio-demographic and morbidity information (Dixon et al., 2011). In total we use 24 morbidity markers, adapted from the ICD10 chapters (WHO, 2007), which are coded as one if one or more spells occur with any diagnosis within

---

(1987).

<sup>6</sup>HES is maintained by the NHS Information Centre.

<sup>7</sup>This dataset was compiled as part of a wider project considering the allocation of NHS resources for secondary care services. 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. In addition, 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 ‘needs’ (morbidity) characteristics, and accordingly for the setting of healthcare budgets are determined using alternative mechanisms.

<sup>8</sup>Reference costs for 2005-2006, which were the basis for the tariffs from 2008-2009, were used when 2008-2009 tariffs were unavailable.

<sup>9</sup>This follows standard practice for costing NHS activity.

<sup>10</sup>Costs above £10,000 are excluded in these plots to make illustration clearer.

<b>N</b>	6,164,114	
<b>Mean</b>	£2,610	
<b>Median</b>	£1,126	
<b>Standard deviation</b>	£5,088	
<b>Skewness</b>	13.03	
<b>Kurtosis</b>	363.18	
<b>Minimum</b>	£217	
<b>Maximum</b>	£604,701	
	% observations	% of total costs
> £500	82.96%	97.20%
> £1,000	55.89%	89.80%
> £2,500	27.02%	72.35%
> £5,000	13.83%	54.65%
> £7,500	6.92%	38.67%
> £10,000	4.09%	29.35%

Table 1: Descriptive statistics for hospital costs

the relevant subset of ICD10 chapters (during the financial year 2007-2008) and zero otherwise.

To give some illustration of the features of the data conditional upon these covariates we construct an index using these regressors and divide the data from the ‘estimation’ set into five quantiles (quintiles) according to the value of the index.<sup>11</sup> For each quintile we display the empirical distribution of log-costs<sup>12</sup> in Figure 2, and in particular pick out those that exceed  $\ln(\pounds 10,000)$ . It’s clear from Figure 2 that the conditional distributions of log-costs (and thus costs) vary dramatically by quintile of covariates in terms of their shape, range and number of high cost patients, with 17% of observations with greater annual costs than  $\pounds 10,000$  in the most morbid patients, compared to a population average of 4.09% (and 0.14% in the least morbid quintile). An analysis looking only at the mean of each quintile would overlook these features of the data.

We also carry out a similar analysis, this time using untransformed costs and dividing the ‘estimation’ set into 10 quantiles (deciles) of the linear index of covariates, where we plot the kurtosis of each decile against their skewness. Parametric distributions impose restrictions 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 possi-

<sup>11</sup>This is constructed by regressing cost against the regressors using OLS and taking the predicted cost.

<sup>12</sup>A log-transformation is used to make the whole distribution easier to illustrate and  $Pr(y > k) = Pr(\ln(y) > \ln(k))$  since it is a monotonic transformation.

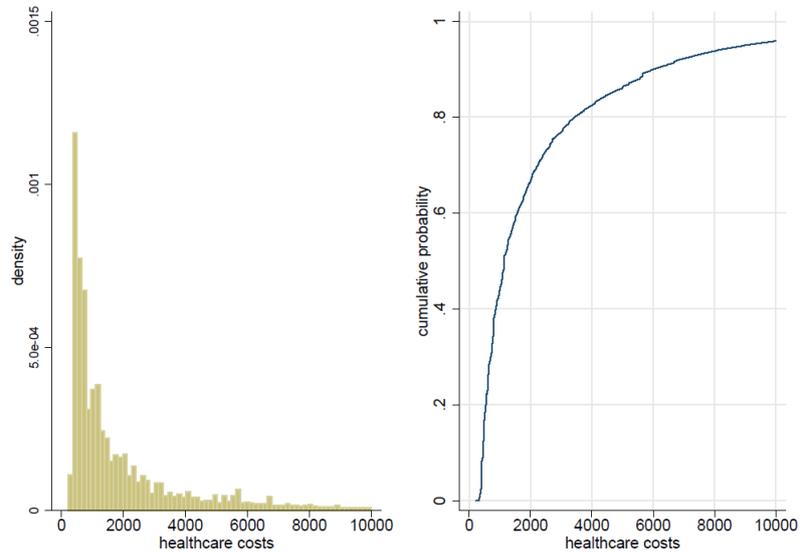


Figure 1: Empirical density and cumulative distribution of healthcare costs

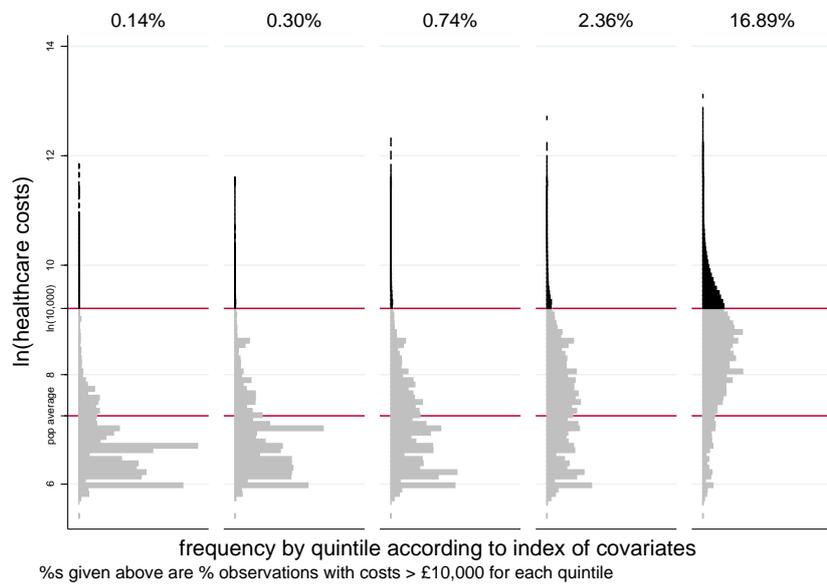


Figure 2: Empirical distribution of log-costs for each of the 5 quantiles of the linear index of covariates

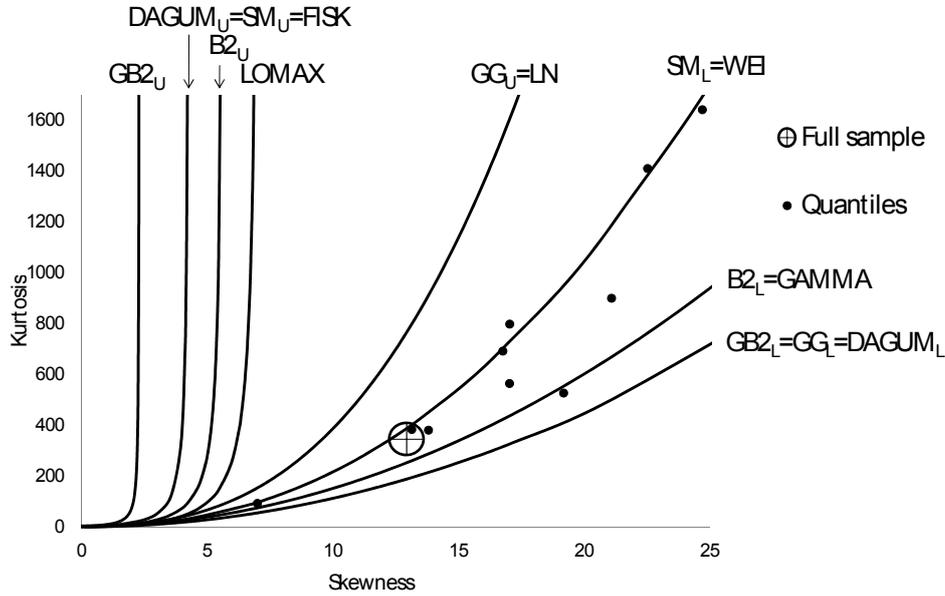


Figure 3: Kurtosis against skewness for each of the 10 quantiles of the linear index of covariates

*Note: Taken from Jones et al. (2014) and adapted from McDonald et al. (2011). The dots shown on Figure 3 were generated as follows: the data were divided into ten subsets using the deciles of a simple linear predictor for healthcare costs using the set of regressors used in this paper. Figure 3 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.*

ble skewness and kurtosis combinations. Figure 3<sup>13</sup> 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 variable analysed here. We haven't represented the less parametric approaches used in this paper in this Figure, since they discretise the distribution and/or estimate several models, and the effects on implied skewness and kurtosis is not clear.

### 2.3 Quasi-Monte Carlo design

In order to fully exploit the large dataset at our disposal, before we undertake analysis we randomly divide the 6,164,114 observations into two equally sized groups: an 'estimation' set and a 'validation' set (each with

<sup>13</sup>Key for abbreviations: GB2 - generalised beta of the second kind, SM - Singh-Maddala, B2 - beta of the second kind, GG - generalised gamma, LN - log-normal, WEI - Weibull.

3,082,057 observations). Because researchers using observational data from social surveys typically have fewer observations in their datasets than are present in our ‘estimation’ set, we draw samples from within the ‘estimation’ set. On these samples we estimate the regressions that will later be evaluated using the ‘validation’ set data. In total we randomly draw 300 samples with replacement: 100 samples of each size  $N_s$  ( $N_s \in 5,000; 10,000; 50,000$ ), where samples with  $N_s = 5,000$  or  $10,000$  may be thought of as having a similar number of observations as small to moderately sized datasets (Basu and Manning, 2009). We estimate 14 methods using the outcome and regressor data from each sample, where each method can be used to construct a counterfactual distribution of costs  $F(y|X)$  (more details on each method are found in the Empirical Models section). Then using all 3,082,057 observations in the ‘validation’ set, we use the covariates from the data (but not the outcome variable) to construct  $F(y|X)$  for each method. Depending upon which method is being considered we can either directly obtain  $Pr(y > k|X)$  which we then integrate out over values of  $X$  to produce an estimate of  $Pr(y > k)$ , or we can use  $F(y|X)$  which we integrate out over values of  $X$ , to give  $F(y)$ , to then estimate  $Pr(y > k)$ . Once the estimate of  $Pr(y > k)$  is produced for the ‘validation’ set, using either method, it can be compared to the observed empirical proportion of costs in the data that exceeds the threshold  $k$ .<sup>14</sup> In this paper we choose round values for  $k$  throughout the distribution of the outcome variable (numbers in brackets correspond to % of population mean):  $k \in \text{£}500$  (19%);  $\text{£}1,000$  (38%);  $\text{£}2,500$  (96%);  $\text{£}5,000$  (192%);  $\text{£}7,500$  (287%);  $\text{£}10,000$  (383%).<sup>15</sup> Results displayed look at performance across each replication for given method with a given sample size. We construct a ratio of predicted to observed  $Pr(y > k)$  and look at the average of this across all replications, as well as the variability of the ratio from replication to replication using the average absolute deviation from the average computed ratio for that method, their standard deviation and their range.

### 3 Empirical models

#### 3.1 Overview

In total we compare the performance of 14 different estimators, which we will describe in terms of two groups: parametric methods and distributional methods. First we describe each of the parametric distributions and

---

<sup>14</sup>It is worth noting that the practice of comparing observed versus empirical probabilities forms the basis of the Andrews (1988) chi-square test, although this is designed for use with parametric methods only, and as such is not implemented in this paper, where we are interested in the performance of both parametric and semi-parametric approaches.

<sup>15</sup>Table 1 gives the proportion of observations in the population that exceed these thresholds.

provide its conditional probability density function ( $f(y|X)$ ), the equation to calculate  $Pr(y > k|X)$ , as well as the procedure for integrating over  $X$  in order to produce an estimate of  $Pr(y > k)$ . For the remaining 5 methods, the procedure is more varied and complex, so we provide a detailed account of the steps required to produce estimates of  $Pr(y > k)$  for all of these distributions. Table 2 provides a key for the abbreviations used for each method throughout the remainder of the paper.

GB2_LOG	generalised beta of the second kind (log-link)
GB2_SQRT	generalised beta of the second kind ( $\sqrt{\cdot}$ -link)
GG	generalised gamma (log-link)
GAMMA	two-parameter gamma (log-link)
LOGNORM	log-normal (log-link)
WEIB	Weibull (log-link)
EXP	exponential (log-link)
FMM_LOG	two-component finite mixture of gamma densities (log-link)
FMM_SQRT	two-component finite mixture of gamma densities ( $\sqrt{\cdot}$ -link)
HH	Han and Hausman
FP	Foresi and Peracchi
CH	Chernozhukov, FernandezVal and Melly (linear probability model)
MM	Machado and Mata - Melly (log-transformed outcome)
RIF	recentered-influence-function regression (linear probability model)

Table 2: Key for method labels

### 3.2 Parametric methods

All nine of the parametric approaches that we consider, including two variants of finite mixture models<sup>16</sup>, are estimated by specifying the full conditional distribution of healthcare costs using between one and five parameters. While it is possible in principle to allow shape parameters to vary with covariates, preliminary work showed that this produced unreliable and uninterpretable results, so in all cases we only specify location parameters as functions of covariates. This means that all models have only one parameter depending upon covariates, except FMM\_LOG and FMM\_SQRT which have scale parameters in each component that are allowed to vary with covariates. All other parameters are estimated as scalars. In Table 3 we give the conditional probability density function and the conditional survival function for each model we compare.

---

<sup>16</sup>These are elsewhere considered to be semi-parametric, since the number of components can vary, but we fix the number of components as two, meaning that they are essentially parametric.

Model	$f(\mathbf{y} \mathbf{X}) =$	$\Pr(\mathbf{y} > \mathbf{k} \mathbf{X}) =$
GB2.LOG	$\frac{ay^{ap-1}}{\exp(X\beta)^{ap}B(p,q)[1+(\frac{y}{\exp(X\beta)})^a]^{p+q}}$	$1 - I_Z(p, q)^*$ where $z = \left(\frac{k}{\exp(X\beta)}\right)^a$
GB2.SQRT	$\frac{ay^{ap-1}}{(X\beta)^{2ap}B(p,q)[1+(\frac{y}{(X\beta)^2})^a]^{p+q}}$	$1 - I_Z(p, q)^*$ where $z = \left(\frac{k}{(X\beta)^2}\right)^a$
GG	$\frac{\kappa}{\sigma y \Gamma(\kappa-2)} \left(\kappa^{-2} \left(\frac{y}{\exp(X\beta)}\right)^{\kappa/\sigma}\right)^{\kappa-2} \exp\left(-\kappa^{-2} \left(\frac{y}{\exp(X\beta)}\right)^{\kappa/\sigma}\right)$	if $\kappa > 0$ : $1 - \Gamma(z; \kappa^{-2})^{**}$ if $\kappa < 0$ : $\Gamma(z; \kappa^{-2})^{**}$ where $z = \kappa^{-2} \left(\frac{k}{\exp(X\beta)}\right)^{\kappa/\sigma}$
GAMMA	$\frac{1}{y \Gamma(\kappa-2)} \left(\kappa^{-2} \left(\frac{y}{\exp(X\beta)}\right)\right)^{\kappa-2} \exp\left(-\kappa^{-2} \left(\frac{y}{\exp(X\beta)}\right)\right)$	$\kappa > 0$ : $1 - \Gamma(z; \kappa^{-2})^{**}$ if $\kappa < 0$ : $\Gamma(z; \kappa^{-2})^{**}$ where $z = \kappa^{-2} \left(\frac{k}{\exp(X\beta)}\right)$
LOGNORM	$\frac{1}{\sigma y \sqrt{2\pi}} \exp\left(\frac{-(\ln y - X\beta)^2}{2\sigma^2}\right)$	$1 - \Phi\left(\frac{\ln k - X\beta}{\sigma}\right)$
WEIB	$\frac{1}{\sigma y} \left(\frac{y}{\exp(X\beta)}\right)^{\frac{1}{\sigma}} \exp\left(-\left(\frac{y}{\exp(X\beta)}\right)^{\frac{1}{\sigma}}\right)$	$\exp\left(-\left(\frac{k}{\exp(X\beta)}\right)^{\frac{1}{\sigma}}\right)$
EXP	$\frac{1}{\exp(X\beta)} \left(\frac{-y}{\exp(X\beta)}\right)$	$\exp\left(-\frac{k}{\exp(X\beta)}\right)$
FMM.LOG	$\sum_j^2 \pi_j \frac{y^{\alpha_j}}{y \Gamma(\alpha_j) \exp(X\beta_j)^{\alpha_j}} \exp\left(-\left(\frac{y}{\exp(X\beta_j)}\right)\right)$	$\sum_j^2 \pi_j (1 - \Gamma(z; \alpha_j))^{***}$ where $z = \frac{k}{\exp(X\beta_j)}$
FMM.SQRT	$\sum_j^2 \pi_j \frac{y^{\alpha_j}}{y \Gamma(\alpha_j) (X\beta_j)^{2\alpha_j}} \exp\left(-\left(\frac{y}{(X\beta_j)^2}\right)\right)$	$\sum_j^2 \pi_j (1 - \Gamma(z; \alpha_j))^{***}$ where $z = \frac{k}{(X\beta_j)^2}$

\*where  $I_Z(p, q) = \frac{1}{B(p, q)} \int_0^z \frac{t^{p-1}}{(1+t)^{p+q}} dt$  is the incomplete beta function ratio. \*\*where  $\Gamma(z; \kappa^{-2}) = \frac{1}{\Gamma(\kappa^{-2})} \int_0^z t^{(\kappa^{-2}-1)} \exp(-t) dt$ .

\*\*\*where  $\Gamma(z; \alpha_j) = \frac{1}{\Gamma(\alpha_j)} \int_0^z t^{(\alpha_j-1)} \exp(-t) dt$ .

Table 3: Forms of density functions and survival functions for parametric distributions

The generalised beta of the second kind<sup>17</sup> is a four-parameter distribution that was applied to modelling healthcare costs by Jones (2011) specifying the location parameter as a linear function of covariates using software developed by Jenkins (2009). Jones et al. (2014) estimated the distribution with a log-link (GB2\_LOG) making it more comparable with commonly used approaches. For example with this specification GG as proposed by Manning et al. (2005) becomes a limiting case of GB2\_LOG. Jones et al. (2013) also compared GB2\_SQRT as well as GB2\_LOG against a broad range of models, finding that the GB2\_SQRT performed particularly well in terms of accurately predicting mean individual healthcare costs. GG has been compared more extensively in terms of predicting mean healthcare costs, having been found to out-perform a GLM log-link with gamma-distribution in the presence of heavy tails using simulated data (Manning et al., 2005), and a number of models within the GLM framework when a log-link is appropriate using American survey data; the Medical Expenditures Panel Survey (Hill and Miller, 2010). GB2\_LOG, GG and LOGNORM were compared in Jones et al. (2014), with some indication that GB2\_LOG better fits the entire distribution with lower AIC and BIC, although LOGNORM better predicted tail probabilities associated with the majority of high costs considered. We also consider further special cases of GG (and GB2\_LOG) with two parameters: GAMMA and WEIB, and with one parameter: EXP.

Finite mixture models have been used 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  $f_j(y|X)$ , and where there is a parameter for the probability of belonging to each component,  $\pi_j$ . The general form of the probability density function of finite mixture models is given as:

$$f(y|X) = \sum_j^C \pi_j f_j(y|X) \quad (1)$$

We use two gamma distribution components in our comparison.<sup>18</sup> In one of the models used, we allow for log-links in both components (FMM\_LOG), and in the other we allow for a square root link (FMM\_SQRT). In both, the probability of class membership is treated as constant for all individuals. Unlike the other parametric methods, this approach can allow for a multimodal distribution of costs. In this way, finite mixture models represent a flexible extension of parametric models (Deb and Burgess, 2003). Using

<sup>17</sup>Also known as generalised-F, see Cox (2008).

<sup>18</sup>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.

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).

Once we have obtained estimates of location parameters (all  $\beta$ s for each regressor) and shape parameters for each distribution, these are stored in memory and then used to generate estimates of  $Pr(y > k|X)$ , where values for  $X$  are the observed covariates in the ‘validation’ set. These estimated conditional tail probabilities will vary across each possible combination of  $X$ , and hence individual  $i$ , and so we take the average in order to ‘integrate out’ these to provide us with a single estimate of  $Pr(y > k)$  for each method and replication, which can be compared to the proportion of costs empirically observed to exceed  $k$ . We then take the average across all replications of  $Pr(y > k)$  for each method in order to assess bias and analyse the variability across replications as an indicator of precision.

### 3.3 Distributional methods

### 3.4 Methods using the cumulative distribution function

Of the remaining five methods that we compare, three involve estimation of the conditional distribution function and two operate through the quantile function. First we consider the methods which estimate the conditional distribution function  $F(y|X)$ . Han and Hausman (1990) adopts a proportional hazards specification, where the baseline hazard is allowed to vary non-parametrically across  $D_{HH}$  intervals of a discretised continuous outcome variable. The logarithm of the integrated baseline hazard for each of the  $D_{HH} - 1$  intervals (one is arbitrarily omitted for estimation) is estimated as a constant  $\delta_{D_{HH}}$ . The effects of covariates are estimated using a particular functional form, which is typically linear. This approach is similar to the semi-parametric Cox proportional hazard model (Cox, 1972), but differs in that the baseline hazard is not regarded as a nuisance parameter and is better suited to data with many ties of the outcome variable (or in the case of a discrete outcome). In order to implement this method, we construct a categorical variable for each observation, indicating the interval into which the value of the outcome variable falls. This is then used as the dependent variable in an ordered logit regression against the covariates. The cut-points are estimates of the baseline hazard within each interval  $\delta_{D_{HH}}$ . The authors argue that given a large sample size, finer intervals should improve the efficiency of the estimator, without providing guidance on a specific number of intervals to be used. As a result we carried out preliminary work to establish the largest number of intervals that could be used for each sample size whilst maintaining good convergence performance,<sup>19</sup> which resulted in a

---

<sup>19</sup>This was taken to mean that the model converges at least 95 times out of the 100

maximum of 33 intervals for sample sizes 5,000 and 10,000, and 36 intervals for a sample size of 50,000.

Foresi and Peracchi’s (1995) method is similar to Han and Hausman’s (1990) in that it divides the data into a set of discrete intervals. Rather than using an ordered logit specification, Foresi and Peracchi (1995) estimate a series of logit regressions. For each upper boundary of the  $D_{FP} - 1$  intervals (the highest value interval is excluded), an indicator variable is created which is equal to one if the observation’s observed cost is less than or equal to the upper boundary, and zero otherwise. These are then used as dependent variables in  $D_{FP} - 1$  logit regressions each using the full set of regressors. In their application to excess returns in their paper they use zero, as well as the 10th, 15th, 20th, ... , 80th, 85th and 90th percentiles as boundaries. While we do not have information on patients with zero costs in our dataset, we base our intervals on their specification of the dependent variables by using the 5th, 10th, 15th, ... , 85th, 90th and 95th percentiles (vigiciles).

The third approach that we compare is an extension of Foresi and Peracchi (1995) and is described in Chernozhukov et al. (2013). The crucial difference between the methods is that Chernozhukov et al. (2013) argue that a logit regression should be used for each unique value of the outcome variable. A continuum of indicator variables needs to be generated and then regression models are used to construct the conditional distribution functions for each value. Given the computational demand of this approach, and lack of variation in the indicator variables at low and high costs, de Meijer et al. (2013) use linear probability models in place of logit regressions. We also adopt this approach in our comparison, since preliminary work showed that, where it was possible to estimate both logit and linear probability models, there was little difference between the methods.

All of these methods are similar in that they can produce estimates of  $Pr(y > k * |X)$ , where  $k*$  represents one of the boundaries of the intervals generated using either Han and Hausman (1990) or Foresi and Peracchi (1995), or any cost value observed in the sample when implementing Chernozhukov et al. (2013). Since models are estimated without knowing what thresholds ( $k$ ) the policymaker might be interested in, it is not always the case that  $k* = k$ . Therefore, for all three methods described above, we use a weighted average of  $Pr(y > k * |X)$  for the nearest two values of  $k*$  to  $k$  when  $k* \neq k$ . Our weight is based on a simple linear interpolation:

$$Pr(y > k|X) = Pr(y > k * _a |X) + \left( \frac{k - k * _a}{k * _b - k * _a} \right) (Pr(y > k * _b |X) - Pr(y > k * _a |X)) \quad (2)$$

where  $k * _a$  and  $k * _b$  represent the thresholds analysed in estimation closest below and closest above  $k$ , respectively.<sup>20</sup>

---

samples.

<sup>20</sup>This should work well when there are a large number of  $k*$  spaced throughout the

Since we end up with an estimate for each observation of  $Pr(y > k|X)$ , we carry out the same procedure as with the parametric distributions. This means that we take the average of  $Pr(y > k|X)$ , thus ‘integrating out’ over all possible combinations of  $X$  and giving us an estimate of  $Pr(y > k)$  to be compared against the empirical proportion.

### 3.5 Methods using the quantile function

Machado and Mata (2005) propose a method for constructing a counterfactual distribution based on a series of quantile regressions using the logged outcome variable. They suggest that a quantile ( $\tau$ ) is chosen at random by drawing from a uniform probability distribution between zero and one. After running the quantile regression for the drawn value, the set of estimated coefficients is used to predict the quantile given the covariate values observed for a randomly selected observation. The authors repeat this process is repeated 4500 times with replacement, generating a full counterfactual distribution. The theoretical motivation for this procedure is that each predicted quantile based on  $q_\tau(X)$  represents a draw from the conditional distribution of healthcare costs ( $f(y|X)$ ). Therefore drawing a random observation and forecasting  $q_\tau$  enough times with random  $\tau$  effectively integrates out  $X$ . Running such a large number of quantile regressions is computationally expensive, and so Melly (2005) suggest running a regression for a fixed number of quantiles spread over the full range of the distribution, e.g. for each percentile, rather than drawing a quantile at random. We use the Melly (2005) approach for the MM method, running quantile regressions for each percentile on the ‘estimation’ set, after log-transforming the outcome variable, and randomly choose one of these quantiles to forecast for each observation in the ‘validation’ set.<sup>21</sup> Once this has been done, the forecasted values represent the counterfactual distribution of healthcare costs belonging to the ‘validation’ set. Therefore to produce an estimate of  $Pr(y > k)$  we observe the proportion of the observations in the counterfactual distribution that exceed  $k$ .

Another method which estimates quantiles of the distribution is developed by Firpo et al. (2009), which employs recentered-influence-function regressions. For a given observed quantile ( $q_\tau$ ), a recentered-influence-function (RIF) is generated, which can take one of two values depending upon whether or not the observation’s value of the outcome variable is less than or equal to the observed quantile:

---

distribution. When interested in high values of  $k$  this linear interpolation may be inappropriate if there are few high values of  $k^*$ , given the often large distances between a high cost and the next highest observed cost.

<sup>21</sup>The prediction is exponentiated to achieve the quantile of the distribution of the levels of healthcare costs.

$$RIF(y; q_\tau) = q_\tau + \frac{\tau - 1 [y \leq q_\tau]}{f_y(q_\tau)} \quad (3)$$

Here,  $q_\tau$  is the observed sample ( $\tau$ ) quantile,  $1[y \leq q_\tau]$  is an indicator variable which takes the value one if the observation's value of the outcome variable is less than or equal to the observed quantile and zero otherwise, and  $f_y(q_\tau)$  is the estimated kernel density of the distribution of the outcome variable at the value of the observed quantile. The recentered-influence-function is then used as the dependent variable in an OLS regression on the chosen covariates, which effectively constitutes a rescaled linear probability model. These estimated coefficients can then be used to predict the quantile being analysed for a given observation's covariates. Following the same thought process as Machado and Mata (2005) and Melly (2005), predictions based on  $q_\tau(X)$  represent a draw from  $f(y|X)$ . This means that we can use the estimated quantile functions to predict a counterfactual distribution in the same way for the RIF method as we do for the MM method.<sup>22</sup>

## 4 Results

When analysing the performance of the methods, we calculate a ratio of the estimated  $Pr(y > k)$  to the actual proportion of costs in the 'validation' set observed to exceed the threshold value  $k$  (see Table 4). Using a ratio allows for greater comparability when looking at performance at different thresholds. We will look at the average ratio across replications (with methods estimated on different samples drawn from the 'estimation' set<sup>23</sup>) as well as the variability of the ratios. The former indicates the bias associated with each method at a given  $k$ , while the latter indicates precision of the method. First we will look at results across methods for a given sample size and threshold cost value:  $N_s = 5,000$  and  $k = \pounds 10,000$ .<sup>24</sup> Second we consider performance for a given sample size, with a range of values for the threshold cost value, since different methods may be better at fitting different parts of the distribution of healthcare costs:  $N_s = 5,000$  and ( $k \in \pounds 500; \pounds 1,000; \pounds 2,500; \pounds 5,000; \pounds 7,500; \pounds 10,000$ ). Lastly performance at different sample sizes is evaluated at a given threshold cost value: ( $N_s \in 5,000; 10,000; 50,000$ ) and  $k = \pounds 10,000$ .

In Figure 4 we present the performance of the 14 methods in predicting the probability of a cost exceeding  $\pounds 10,000$  in the validation set, when sam-

<sup>22</sup>We calculate the recentered-influence-function using the level of costs and so no re-transformation is required unlike when using MM.

<sup>23</sup>Three samples were discarded when  $N_s = 5,000$ , due to being unable to form the categorical variable for HH. Only one sample was discarded when  $N_s = 10,000$  and  $N_s = 50,000$ .

<sup>24</sup>We choose these values of  $N_s$  and  $k$  since they are the smallest and most challenging sample size and the largest and most economically interesting threshold value, respectively.

$k$	% observations in ‘validation’ set $k$
£500	82.93%
£1,000	55.89%
£2,500	27.04%
£5,000	13.84%
£7,500	6.94%
£10,000	4.10%

Table 4: Actual empirical proportion of observations greater than  $k$  in the ‘validation’ set

ples with  $N_s = 5,000$  observations are used. The bars indicate the ratio of estimated to actual probability, and the capped spikes indicate the range of ratios across all of the replications. A ratio of one represents a perfect fit, i.e. the method correctly predicted that 4.10% of observations would exceed £10,000.

From Figure 4, it is clear that performance of the methods is varied both in terms of bias (the point - the average ratio) and precision (the variability of ratios as depicted by the capped spikes showing the range). There is not a clear pattern in terms of parametric versus distributional methods, since in both groups there are methods where the average ratio is seen to be near the desired value of one, as well as methods in both groups where the range of computed ratios does not contain one. In terms of bias, the best method is CH with an average ratio of almost exactly one. It appears that this is not the most precise method for  $k = £10,000$ , however, with a range of ratios:  $0.82 - 1.14$ , that is the fifth largest of all methods compared (the largest belongs to FMM.SQRT). To more clearly represent the tradeoff between bias and precision, see Table 5, which gives the rankings of each method in terms of bias (how far the average ratio is away from one), the range of ratios and also the standard deviation of ratios.

From Table 5 it can be seen that three of the parametric distributions: GB2.SQRT, GG and LOGNORM, demonstrate significant potential in terms of the variability of their predictions as the three methods with the lowest standard deviations of ratios. MM performs consistently well across all three measures of performance, especially when variability is measured by the range of ratios, although the standard deviation is still among the five lowest of methods compared. From these results it’s unclear as to which method is the best for forecasting costs greater than £10,000, since there is no outright winner over the three metrics.

Whilst the results outlined previously give some indication of the methods’ abilities to forecast high costs, we are interested in the performance of the regression methods at all points in the distribution. For this reason we carry out a similar analysis across a range of cost threshold values. To present these results, once again we plot the average ratio and the range of

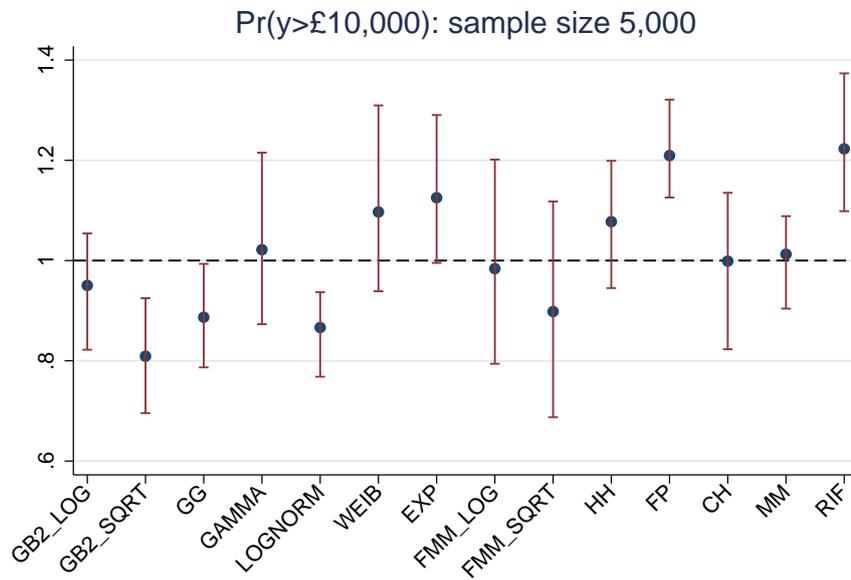


Figure 4: Performance of methods predicting the probability of a cost exceeding £10,000 at sample size 5,000

ratios across the replications. The results presented in Figure 5 are undertaken using samples with 5,000 observations.

Method	Bias rank	Range rank	Standard deviation rank
GB2_LOG	5th	6th	6th
GB2_SQRT	12th	5th	3rd
GG	9th	4th	2nd
GAMMA	4th	11th	11th
LOGNORM	11th	1st	1st
WEIB	7th	12th	12th
EXP	10th	9th	8th
FMM_LOG	3rd	13th	14th
FMM_SQRT	8th	14th	13th
HH	6th	7th	9th
FP	13th	3rd	4th
CH	1st	10th	10th
MM	2nd	2nd	5th
RIF	14th	8th	7th

Table 5: Rankings of methods based on threshold of £10,000 at sample size 5,000

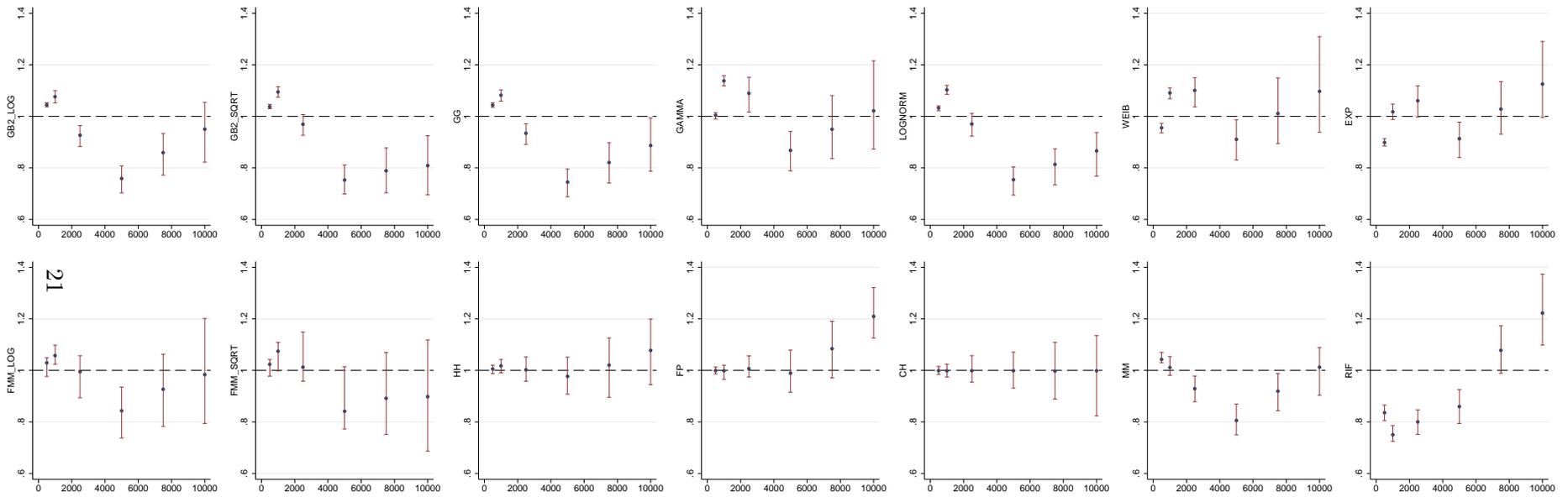


Figure 5: Performance of methods predicting the probability of costs exceeding various thresholds at sample size 5,000

There is a clear pattern in Figure 5 showing that the higher the cost threshold being considered, the greater the variability in ratio of estimated to actual probability. Besides this, the way in which performance varies across different thresholds, including by how much variability increases with higher thresholds, is different for all methods.

Beginning with the parametric distributions, with log-links, there seems to be little difference in the performance of GB2\_LOG and GG, except for that GB2\_LOG performs slightly better at the higher costs considered in terms of bias. Looking at the gamma-type models, LOGNORM demonstrates potential in terms of producing precise estimates of tail probabilities if not in terms of bias. Since FMM\_LOG represents a two-component version of GAMMA, comparing the performance of these methods provides some insight into the returns from using more complex mixture specifications. The pattern of performance at different thresholds is quite similar for these, and the main difference seems to be that FMM\_LOG produces more variable estimates, especially at low cost thresholds. WEIB and EXP seem to perform similarly, with high variability forecasts. It is interesting to note that the  $\sqrt{\cdot}$ -link methods differ from their log-link counterparts, i.e. comparing FMM\_SQRT with FMM\_LOG and GB2\_SQRT with GB2\_LOG, mainly through worse forecasts at the higher costs.

There is considerable variation in performance between the distributional methods. The methods that use the cumulative distribution function seem to vary predominantly according to the number of intervals that are used as opposed to the specification for predicting interval membership. CH is practically unbiased for all cost thresholds, which illustrates the strength of this method, in forecasting  $Pr(y > k)$  for a range of values of  $k$ . As pointed out earlier, though, the variability of the forecasts across replications is larger than the majority of other methods considered in this paper. It seems therefore that much of the bias in HH and FP stems from when  $k*_a$  and  $k*_b$  are not close to the value of  $k$  being investigated, which is more likely to be the case with FP than with HH, since FP has fewer intervals (and is highly unlikely using CH - in our application). This is particularly clear with  $k = \pounds 10,000$ , since with HH and FP in this case  $k*_b$  will often be the highest observed cost in the sample. When this occurs, the linear interpolation that we employ is likely to lead to an overestimation of the forecasted probability (see equation 2 for details). For these three methods the variability of ratios is roughly similar, but when looking also at the methods using the quantile function, it is clear that MM offers an improvement upon the variability, although its performance in terms of bias varies across values of  $k$ . RIF seems to perform badly both in terms of bias and precision.

Finally, we look into how our analysis is affected by the number of observations that are present in the drawn samples. To do this, we return to the style of graph that was produced for Figure 4, but illustrate performances for the three sample sizes analysed ( $N_s \in 5,000; 10,000; 50,000$ ). The results

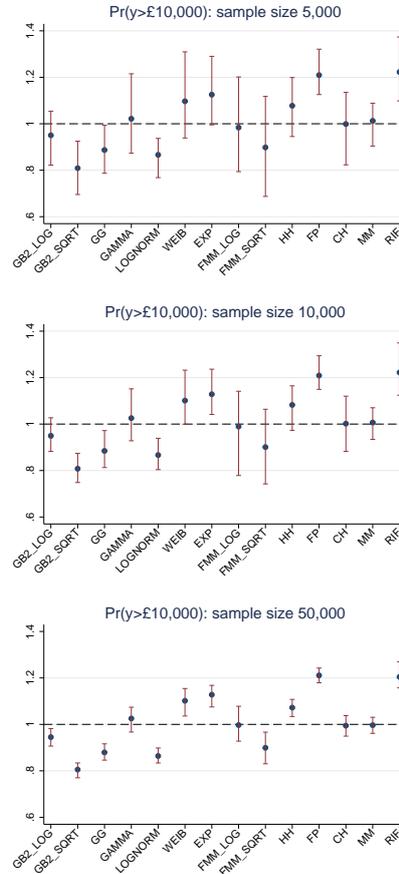


Figure 6: Performance of methods predicting the probability of a cost exceeding £10,000 at all sample sizes

are therefore only for one value of  $k$ , but results at other values followed a similar pattern.

From Figure 6 we can see that there is a clear effect of sample size on the performance of the regression methods fitting the whole distribution. Having more observations does not particularly affect bias of each method, but, as expected, it reduces the variability of the estimates. This therefore means that methods such as CH perform relatively better at bigger sample sizes since it remains unbiased, but forecasts costs with increased precision.

## 5 Discussion

The results of this paper are the first to provide a comparative assessment of parametric and distributional methods designed to estimate a coun-

terfactual distribution, and different to most studies concerning econometric modelling of healthcare costs where performance has largely been judged on the basis of the ability to predict conditional means. Limited existing research compares parametric distributions (but not distributional methods) against one another for predicting tail probabilities as well as in-sample fit of the whole distribution based on log-likelihood statistics (Jones et al., 2014). The analysis presented here builds on this work with a range of thresholds for tail probabilities as well as a broader range of parametric distributions including mixture distributions and models with a  $\sqrt{\cdot}$ -link as well as those with a log-link. There is considerable variation in the best performing parametric distributions according to the specific tail probability being considered. When considering costs that exceed £10,000, FMM\_LOG is the least biased parametric method, but is the most imprecise of all methods considered. At other thresholds, the distribution with the best fit on average varies, for example it is WEIB among parametric distributions for costs that exceed £7,500. This means that the preferred parametric distribution would depend upon the decision-maker’s loss function. Some distributions are particularly imprecise at all tails investigated, notably the mixture models: FMM\_LOG and FMM\_SQRT as well as some of the more restrictive distributions: GAMMA, WEIB and EXP. LOGNORM is the most precise and thus demonstrates its potential for modelling the whole distribution of costs. Whilst other papers have focused on the importance of the link function, which seems to have a large impact on performance when it comes to predicting mean healthcare costs (see for example Basu et al., 2006), this paper finds that when we are concerned with predicting tail probabilities the link function is less of an issue than are the distributional assumptions more generally.

The distributional methods show promise for modelling the full distribution of healthcare costs. In particular, CH is practically unbiased in terms of all forecasted tail probabilities considered. The related methods of FP and HH also perform well in terms of bias, but not when considering costs that exceed £10,000, because this is likely to fall in the highest quantile of costs in either method. CH is better placed to model this tail probability, since each unique value of costs that is encountered in the sample is used as the basis for an indicator variable for a separate regression, and using a linear probability model does not require variation across all covariates for each value of the dependent variable. At the smallest sample size of 5,000 observations, these three methods exhibit highly imprecise forecasted probabilities, but this becomes less of an issue at larger sample sizes where the variability was lower for all 14 methods. MM delivers better precision, but its performance on average varies across the different tail probabilities. RIF appears to be the worst among the distributional methods for this data and specification.

For our application, CH demonstrates potential even for forecasting

probabilities of high costs - such as costs that exceed £10,000. A function of the methodology is that CH (as well as HH and FP) is unable to extrapolate beyond the observed sample, and so in applications where sample size is small, or if the decision-maker is interested in the probability of extremely high costs beyond the largest observed, this method would be unable to provide any information on this parameter. This represents a fundamental flaw for this type of method for fitting the distribution of healthcare costs, where the underlying data generating process is heavy-tailed, and any observed sample is unlikely to contain some of the extreme outcomes which are possible.

As mentioned in the methodology section of the paper, some of these methods have been automated in order to make the quasi-Monte Carlo study design feasible. For instance, we only allow location parameters to vary with covariates and we restrict the number of mixtures used in FMM\_LOG and FMM\_SQRT. In practice, analysts are likely to train their model for a given sample - testing the appropriateness of covariates in the specification as well as the number of mixtures that are required etc. Since all methods have been restricted to some degree, e.g. the regressors are the same for all methods, the results of this paper give some indication of the relative performance of these methods and illustrate their pitfalls and strengths.

This paper has analysed the ability of methods to forecast probabilities associated with ranges of costs across the distribution, but there are numerous further challenges that need to be addressed in this area. One aspect of performance might also touch upon the ability to forecast tail probabilities for individual costs. Since it is impossible to observe the empirical probability for an individual cost, it is not clear how this could be done. In addition, constructing an experimental design that allows for this kind of evaluation might be problematic given the estimation procedure for the distributional methods using the quantile function (i.e. MM and RIF). Nevertheless, policymakers could potentially be interested in the performance of each method in this regard and so further research would be useful. Another area for research is the performance of specification tests in terms of discriminating between methods which fit the whole distribution well and those that do not. For researchers fitting the conditional mean function, there is a battery of tests that can be employed, however it's not clear as to whether these would be useful for analysing tail probabilities.

## References

- Andrews DW. 1988. Chi-square diagnostic tests for econometric models: Introduction and applications. *Journal of Econometrics* **37**: 135 – 156.
- 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. 2009. Issues for the next generation of health care cost analyses. *Medical Care* **47**: S109–S114.
- Basu A, Manning WG, Mullahy J. 2004. Comparing alternative models: log vs cox proportional hazard? *Health Economics* **13**: 749–765.
- Bilger M, Manning WG. 2014. Measuring overfitting in nonlinear models: A new method and an application to health expenditures. *Health Economics* In Press.
- Bitler MP, Gelbach JB, Hoynes HW. 2006. What mean impacts miss: Distributional effects of welfare reform experiments. *The American Economic Review* **96**: pp. 988–1012.
- Blough DK, Madden CW, Hornbrook MC. 1999. Modeling risk using generalized linear models. *Journal of Health Economics* **18**: 153–171.
- 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.
- Chernozhukov V, Fernandez-Val I, Melly B. 2013. Inference on counterfactual distributions. *Econometrica* **81**: 2205–2268.
- Cook BL, Manning WG. 2009. Measuring racial/ethnic disparities across the distribution of health care expenditures. *Health Services Research* **44**: 1603–1621.
- Cox C. 2008. The generalized F distribution: An umbrella for parametric survival analysis. *Statistics in Medicine* **27**: 4301–4312.
- Cox DR. 1972. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)* **34**: 187–220.
- de Meijer C, ODonnell O, Koopmanschap M, van Doorslaer E. 2013. Health expenditure growth: Looking beyond the average through decomposition of the full distribution. *Journal of Health Economics* **32**: 88 – 105.
- 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. *BMJ* **343**: d6608.
- 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.
- Firpo S, Fortin NM, Lemieux T. 2009. Unconditional quantile regressions. *Econometrica* **77**: 953–973.
- Foresi S, Peracchi F. 1995. The conditional distribution of excess returns: An empirical analysis. *Journal of the American Statistical Association* **90**: 451–466.
- Fortin N, Lemieux T, Firpo S. 2011. Decomposition methods in economics. volume 4, Part A of *Handbook of Labor Economics*. Elsevier, 1 – 102.
- 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.
- Holly A. 2009. Modeling risk using fourth order pseudo maximum likelihood methods Institute of Health Economics and Management (IEMS), University of Lausanne, Switzerland.
- Jenkins S. 2009. GB2FIT: stata module to fit generalized beta of the second kind distribution by maximum likelihood. *Statistical software components* **S456823**. Boston College Department of Economics.

- 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, Moore P, Rice N. 2013. 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. *Health Econometrics and Data Group Working Paper* .
- Jones AM, Lomas J, Rice N. 2014. Applying beta-type size distributions to healthcare cost regressions. *Journal of Applied Econometrics* **29**: 649–670.
- Machado JAF, Mata J. 2005. Counterfactual decomposition of changes in wage distributions using quantile regression. *Journal of Applied Econometrics* **20**: 445–465.
- 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.
- Manning WG, Mullahy J. 2001. Estimating log models: to transform or not to transform? *Journal of Health Economics* **20**: 461 – 494.
- McDonald JB, Sorensen J, Turley PA. 2011. Skewness and kurtosis properties of income distribution models. *Review of Income and Wealth* .
- Melly B. 2005. Decomposition of differences in distribution using quantile regression. *Labour Economics* **12**: 577 – 590.
- Mora T, Gil J, Sicras-Mainar A. 2014. The influence of obesity and overweight on medical costs: a panel data perspective. *European Journal of Health Economics* In press.
- 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 of *Handbook of Health Economics*, chapter 14. 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*. Elgar Original Reference.
- 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, version for 2007 .