Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
0(0) 1–20

DOI: 10.1177/0962280214568110
smm.sagepub.com
SAGE

Bayesian regression models for the estimation of net cost of disease using aggregate data

Nicholas Mitsakakis[1,2,3] and George Tomlinson[4,5]

Abstract

Estimation of net costs attributed to a disease or other health condition is very important for health economists and policy makers. Skewness and heteroscedasticity are well-known characteristics for cost data, making linear models generally inappropriate and dictating the use of other types of models, such as gamma regression. Additional hurdles emerge when individual level data are not available. In this paper, we consider the latter case were data are only available at the aggregate level, containing means and standard deviations for different strata defined by a number of demographic and clinical factors. We summarize a number of methods that can be used for this estimation, and we propose a Bayesian approach that utilizes the sample stratum specific standard deviations as stochastic. We investigate the performance of two linear mixed models, comparing them with two proposed gamma regression mixed models, to analyze simulated data generated by gamma and log-normal distributions. Our proposed Bayesian approach seems to have significant advantages for net cost estimation when only aggregate data are available. The implemented gamma models do not seem to offer the expected benefits over the linear models; however, further investigation and refinement is needed.

Keywords

Aggregated data, Bayesian methods, gamma regression, random effects, net cost of disease

1 Introduction

In health technology assessment and health economics, we are often interested in the cost that a specific illness or health condition imposes to the health system, or otherwise the ''net cost'' due to

[1]Toronto Health Economics and Technology Assessment (THETA) Collaborative, University of Toronto, Toronto, ON, Canada
[2]Leslie Dan Faculty of Pharmacy, University of Toronto, ON, Canada
[3]Department of Anesthesia and Pain Management, Toronto General Hospital, Toronto, ON, Canada
[4]Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada
[5]Department of Medicine, University Health Network and Mt Sinai Hospital, ON, Canada

Corresponding author:
Nicholas Mitsakakis, Toronto Health Economics and Technology Assessment (THETA) Collaborative, Leslie L. Dan Pharmacy Building, University of Toronto, 6th Floor, Room 658, 144 College Street, Toronto, Ontario, M5S 3M2, Canada.
Email: n.mitsakakis@theta.utoronto.ca

this condition. Net costs are used in studies of cost of illness or burden of disease that offer important information to health service providers and influence policy decision making.[1–4] This cost is assumed to be accumulated over a specific time horizon (e.g. 1 year after diagnosis) or up to a specific event (e.g. until death), under specific environmental conditions (e.g. in hospital). Often individual patient data are available and this problem can be addressed with the use of a (generalized) linear model, taking into account such characteristics of the cost data as skewness, heteroscedasticity and heterogeneity. However, due to privacy issues, centers developing and maintaining bases of administrative data refrain from releasing individual level data to scientists and researchers. Instead, only aggregate level data, summarizing individual patient level data, but making impossible the identification of any specific patient, can be issued. Traditional modeling approaches have limited applicability to this type of cost data. Here, we examine some alternative methods of estimating the net cost of illness in the situation where individual patient level data are not available, but instead summary data are available for diseased and healthy patients across a range of strata.

Usually, individual patient data have the following format: for each patient i, C_i denotes the cost, $X_i^1, \ldots, X_i^n$ denote covariates that affect cost, which we assume to be categorical with X_i^j taking the values $\{1, \ldots, k_j\}$, and Y_i denotes the level of disease for patient i, with possible values $\{0,1,\ldots,L\}$, with 0 indicating the control condition. We are assuming that the cost is accumulated over time, that the level of disease remains unchanged throughout the cost accumulation, and that the covariates affecting the accumulated cost are measured at baseline (i.e. at the beginning of the cost accumulation time period) and as such as are not time-dependent. Situations where disease conditions progress over time, potentially depending upon time-varying covariates, and also affecting the cost accumulation process, are not uncommon, but they are not going to be considered in this study. We can define as $W_{f_1,\ldots,f_n} = \{i | X_i^1 = f_1, \ldots, X_i^n = f_n\}$ the stratum with $f_1, \ldots, f_n$ covariate values. If we further limit to a specific disease level l, the aggregated data have the format of the triad $(\bar{C}_{f_1,\ldots,f_n,l}, S_{f_1,\ldots,f_n,l}, N_{f_1,\ldots,f_n,l})$, denoting the sample mean cost, sample deviation of the cost and size of the "cell" $\{i \in W_{f_1,\ldots,f_n}, Y_i = l\}$ respectively.

The example that motivated the development presented in this study regards the estimation of the "net" or "attributable" cost associated with pressure ulcers using available aggregate level data. The original data contain strata-specific mean and standard deviations of costs for 1351 patients with pressure ulcers developed in hospital (cases), as well as for 180,092 patients without pressure ulcers but who have been hospitalized (controls). Strata were defined by the combinations of age group, sex, co-morbidity, and most responsible diagnosis. Cases (patients with pressure ulcers) are also characterized by the stage of the ulcer (II, III, IV, or unstageable). Stage-specific net cost estimates are of interest and were the objective of a recent study.[5]

Using the full individual level data, net cost can be estimated using the appropriate generalized linear model for the cost outcome and disease level variable Y_i as predictor, adjusting for $X_i^1, \ldots, X_i^n$. Assuming that the scale of interest and the scale of estimation are additive[6] and that the appropriate reference level is chosen, net cost estimates are given by the coefficient estimates of Y_i. If the only available data have the aggregated form described in the previous paragraph, reasonable approaches for net cost estimation could be those used in meta-analysis or a weighted regression analysis, accounting for heteroscedasticity. In meta-analysis of comparative studies, a treatment effect and its variance (on a logarithmic scale for ratio data) are estimated from each study's data. The most widely used summary estimates in meta-analysis[7] are weighted averages of study-specific effects where the weights are the estimated variances. By analogy, the net cost estimation problem we present here can be seen as a meta-analysis, where each stratum acts as a study and the within-stratum difference in mean cost between the control group $W_{f_1,\ldots,f_n,l=0}$ and groups with higher disease

levels, for example $W_{f_1,..,f_n,l=1}$, acts as a study-specific treatment effect. This suggests the use of closed-form estimators from the meta-analysis literature. Here, we also propose an alternative method inspired by approaches that have been used in Bayesian meta-analysis. We also develop new estimators based on a model for the distribution of aggregate data (mean and standard deviation) from where the raw data arise from a gamma distribution.

The structure of the manuscript is as follows: within ''Materials and Methods'', the first subsection defines formally the problem of estimation of disease specific net cost and discusses estimators using patient level data. Subsequently, the structure of aggregated data we are dealing with is described in detail and existing appropriate applicable methods of analysis are presented. The next subsection describes the proposed linear and gamma mixed effects Bayesian methods. The final subsection describes the design of a simulation study evaluating the performance of the proposed methods. In ''Results'' section the results of the simulation study are presented, while in the following section the analysis of real pressure ulcer cost data using those methods is presented. Finally, ''Discussion and conclusions'' discusses advantages and disadvantages of the proposed method, offering some concluding remarks and suggesting future research steps.

2 Methods

2.1 Models and approaches

In this section, we first define net cost for a specific disease or condition, and then discuss a number of previously used approaches for net cost estimation. We then focus on methods appropriate for data where subjects are stratified on a number of factors.

We formally define the *net cost* (NC) associated with the exposure of a specific level $l = l_1$ compared to the ''control'' level $l = 0$ as the difference of the mean costs given a set of covariates x,i.e.

$$NC(l = l_1|x) = E[C|l = l_1, x] - E[C|l = 0, x]$$

Note here that NC depends on the values of the covariates x. We extend this definition to the *mean net cost*, MNC, which is defined as the expected value of NC over the distributions of the covariates x, in the population of interest, $MNC(l = l_1) = E_x[NC(l = l_1|x)]$.

Various general guidelines regarding types of data and statistical methods have been previously suggested for the estimation of net cost for a particular disease from individual patient level data. Barlow[1] suggests the use of matched data, where each case is matched to one (or more) controls, i.e. patients without the disease, but with similar characteristics (demographic and clinical covariates). Net cost can then be estimated from the cost differences between cases and matched controls, where statistical methods for matched data need to be employed.[1,8,9] Alternatively, the estimation can be performed after fitting appropriate multivariable linear or generalized linear models with cost as dependent variable, and case-control indicator, along with a number of important or confounding variables, as independent variables. Depending on the model, proper transformation of the data (log, Cox–Box, power) or link function (e.g. log) and distribution (e.g. gamma, Generalized gamma, log-normal, etc.) may be used under a frequentist[10–12] or a Bayesian framework.[13]

As a special case we consider data where each subject (case or control) has been stratified based on a number of variables. We also assume that within strata, the subjects are separated according to their level of ''exposure'' which we denote with l, Here, we assume that level 0 indicates the absence of the condition or disease, hence the control data, and levels >0 indicate the case data. Assuming a number of stratifying factors $f_1, \ldots, f_n$, where f_i has k_i levels, we consider the following three

different approaches of generalized linear models for modeling the effect of the condition, and as a consequence the net cost:

(a) Assuming a fixed effect of the condition on the (possibly transformed) mean cost outcome, the cost increment due to disease does not depend on the factor values, so $C|l,f_1,\ldots,f_n \sim F$, with

$$g(\mu_{l,f_1,\ldots f_n}) = \alpha + \sum_{j=1}^{L} I(j=l)\cdot\beta_j + \sum_{i=1}^{n}\sum_{j=1}^{k_i-1} I(f_i=j)\cdot\beta_{ij}$$

where $\mu_{l,f_1,..f_n} = E[C|l,f_1,\ldots f_n]$. This model needs the estimation of $L-n+1+\sum_{i=1}^{n} k_i$ parameters plus any parameters needed for the estimation of the variance. In this model, in the simplest case of the link function g being the identity (e.g. in linear regression), the coefficient β_l represents the net cost of the disease for the level l in comparison to the control, 0 level. Here, $g(\cdot)$ is the chosen link function, for example the identity function or logarithm function.

(b) Assuming fixed effects as in (a) but this time with the cost increment depending on the factor covariates. The implied model has second degree interaction terms between the exposure and each one of the covariate factors, with

$$g\left(\mu_{l,f_1,\ldots,f_n}\right) = \alpha + \sum_{j=1}^{L} I(j=l)\cdot\beta_j + \sum_{i=1}^{n}\sum_{j=1}^{k_i-1} I(f_i=j)\cdot\beta_{ij} + \sum_{m=1}^{L}\sum_{i=1}^{n}\sum_{j=1}^{k_i-1} I(m=l,\quad f_i=j)\cdot\beta_{mij}$$

In this model, we need to estimate $L=1+(L+1)\cdot\left(-n+\sum_{i=1}^{n} k_i\right)$ parameters, many of which will represent the interactions between factor covariates and the incremental mean cost due to illness.

(c) The impact of the factor covariates on the cost outcome is not considered, except as a method for stratification. Each combination of possible values of the factors defines one stratum, and as a result we end up with a total of $\prod_{i=1}^{n} k_i$, possibly empty, strata. We use random effect model assuming that each stratum has different intercept (cost in the non-diseased group within the stratum) and increment coefficient (cost increment due to disease in the stratum) sampled from a distribution. The model then becomes

$$g\left(\mu_{l,j}\right) = \alpha_j + \sum_{i=1}^{L} I(i=l)\cdot\beta_{ij}$$

where $\alpha_j \sim F_a, \left(\beta_{1j,\ldots,ij}\right) \sim F_\beta$ are the random effects and $\mu_{l,j}$ is the mean cost for exposure level l and stratum j. This model assumes that each stratum has been sampled from a larger population and the data consist of a random sample of strata. This differs from model in (b) where the strata are assumed to constitute our population of interest, and the effect of each particular covariate to cost is of interest.

Each one of these models can be further characterized by the condition of homoscedasticity. Depending on the type of the underlying distribution assumed, the errors are inheritably homoscedastic (e.g. in normal models) or heteroscedastic (e.g. in gamma regression).

We need to notice here that even if we assume the homogeneous model (a), without the presence of interactions, the net cost will depend on the values of $f_1, \ldots, f_n$, if the link function is not the identity. Therefore the calculation of MNC will require us to "integrate out" the covariates $f_1, \ldots, f_n,$.

An exception on the above approach is when model (c) is considered, where the effect of the covariates to the cost is not captured, except through the random effects. In that case, net cost for a specific combination of n covariate values cannot be calculated. Instead, in a Bayesian context we can define the random quantity NC_j that depends on the value of the random effect for stratum j

$$NC_j\,(l = l_1) = E[C|l = l_1, \alpha_j, \beta_{1j}, \ldots, \beta_{Sj}] - E[C|l = 0, \alpha_j, \beta_{1j}, \ldots, \beta_{Sj}]$$

and by taking the expected value over the random effects we obtain the mean net cost

$$MNC(l = l_1) = E_{\alpha_j, \beta_{1j}, \ldots, \beta_{Sj}}[NC_j\,(l = l_1)]$$

It is apparent that the choice of the approach for the net cost estimation will depend on the availability of the data, the assumptions over the homogeneity of the effect of the treatment level over the different values of the covariates, and the choice for adopting a random effects model or not.

2.2 Aggregated data and appropriate model choices

Statistical methods have been developed and can be applied for the models described so far, assuming that individual patient level data are available. Here, we describe some of the existing methods that could be used for net cost estimation, when only aggregate level data are available.

We assume that the available data come in an aggregate form, as means, standard deviations and sizes of the different strata, as described in the Introduction section. In a previous section, we described approaches that can be taken for the estimation of net cost. One important way to classify these approaches has to do with the way the values of the stratifying factors are affecting the result. In some of the approaches, the actual values of the factors are not important, and the method is not estimating the effect of the factor on the cost outcome. Instead, the model is taking into account the heterogeneity among the potential strata, and the combinations of the factor values are collapsed into a one-dimensional index that enumerates the existing strata. Here we adopt the latter approach and intention, assuming the arrangement and notation in modeling approach (c), where we enumerate all non-empty strata resulting from each combination from the values of the stratifying factors $f_1, \ldots, f_m$. Assuming that we have a total of S strata, each number s from 1 to S corresponds to a combination of values for the $f_1, \ldots, f_m$ factors that specifies a non-empty stratum. If l is the disease level and s the stratum, the available data are of the form $C_{l,s}, s_{l,s}, N_{l,s}$, for mean, standard deviations, and sample sizes, respectively. If $N_{l,s}$ is equal to 1, $s_{l,s}$ cannot be defined and it is missing, while if $N_{l,s}$ is 0, both $C_{l,s}$ and $s_{l,s}$ are missing.

We now present a number of approaches that can be followed for the estimation of the net cost, using only aggregate data in this form. These methods are of more general applicability or they have been primarily used for other research application (e.g. meta-analysis). Even though we are not

aware of previous application of any of these methods to aggregate data of the form described here, we present them as plausible choices for analysis:

(1) One approach is borrowed from meta-analysis. For the net cost comparing condition level l with 0 (control), we define the stratum specific estimate of mean net cost to be equal to

$\hat{\Delta}_s = C_{ls} - C_{0s}$, and the pooled stratum-wide estimate of the mean net cost to be given by the weighted mean difference, $\bar{\Delta}_\cdot = \frac{\sum_{s=1}^{S} w_s \hat{\Delta}_s}{\sum_{s=1}^{S} w_s}$.

The variance of this estimator is minimized when $w_s = \frac{1}{Var(\hat{\Delta}_s)}$. $Var(\hat{\Delta}_s)$ can be estimated by $s_{ps}^2 \cdot \left(\frac{1}{N_{ls}} + \frac{1}{N_{0s}}\right)$, where $s_{ps}^2 = \frac{(N_{ls}-1)\cdot s_{ls}^2 + (N_{0s}-1)\cdot s_{0s}^2}{N_{ls}+N_{0s}-2}$ denotes the pooled variance, estimate of the unknown true variance assumed common for cases and controls. When this assumption is not valid, the variance of the stratum specific cost difference can be estimated by $\frac{s_{ls}^2}{N_{ls}} + \frac{s_{0s}^2}{N_{0s}}$. More details can be found in Sutton et al.[14]

(2) Similar to the approach above, a weighted least squares model[15] can be fitted to the aggregate data. Each mean cost per stratum can be treated as an observation, where the observations have unequal variance error based on which weights are assigned. Depending on the assumption of heteroscedasticity or not, the weights can be defined as mathematical functions of the sizes of the strata and the aggregate standard deviations. If we assume that individual cost data follow a distribution with mean value μ_s that depends on the covariates (i.e. the stratum they belong to) and some variance σ^2, then the sample mean has asymptotically normal distribution with mean μ and variance σ^2/N, where N is the sample size. If we assume that σ^2 is the same for all strata, then the heteroscedasticity for the sample mean data is introduced only because of the different sample sizes of the strata. In that case, if weighted least squares is used, sample sizes can be used as weights, i.e. the weight for disease level-stratum (l, s) is equal to $N_{l,s}$. In this way, larger strata are given larger weights, as the estimates in those strata are more precise. On the other hand, if we assume that there is heteroscedasticity in the individual person data, then this is transferred in the sample mean data, and it needs to be accounted for. In that case, the weights are set equal to ratios of sample size over sample variance, with the weight for the stratum (l, s) being equal to $\frac{N_{l,s}}{s_{l,s}^2}$. Using matrix notations, with $\mathbf{Y}$ indicating the mean cost data, $\mathbf{X}$ being the matrix of indicator variables indicating disease level, and $\mathbf{V}$ a diagonal matrix with inverse of weights in the diagonal, the weighted least squares estimate is given by $\mathbf{b} = \left(\mathbf{X'V^{-1}X}\right)^{-1}\mathbf{X'V^{-1}Y}$.[15] For the case where the analysis considers only one level of the disease, the net cost estimate is given by

$$NC = \frac{\sum_{i\in I_d} w_i C_i - \left(\sum_{i\in I_d} w_i\right)\left(\sum_{i\in I_d\cup I_c} w_i C_i\right) \Big/ \left(\sum_{i\in I_d\cup I_c} w_i\right)}{\sum_{i\in I_d} w_i - \left(\sum_{i\in I_d} w_i\right)^2 \Big/ \left(\sum_{i\in I_d\cup I_c} w_i\right)}$$

where I_d, I_c indicate indices for the observations for the cases (diseased) and controls respectively.[16]

(3) Following a Bayesian framework one could propose the model

$$\begin{aligned}
\bar{x}_{sl} &\sim N\left(\mu_{sl}, \frac{s_{sl}^2}{n_{sl}}\right) \\
\mu_{sl} &= \alpha_s + \sum_{k=1}^{L} \beta_{ks} \cdot I(k = l) \\
(\alpha_s, \beta_{1s}, \ldots, \beta_{Ls})^t &\sim MVN((\alpha_0, \beta_{10}, \ldots, \beta_{L0})^t, \Sigma)
\end{aligned}$$

using uninformative priors for $\alpha_0, \beta_{10}, \ldots, \beta_{L0}, \Sigma,$ where α_s is the random "intercept" for stratum s, and therefore indicating the stratum s specific control cost, while similarly β_{ks} indicates the stratum specific *net cost* for disease level k. The *mean net cost* for disease level k is then given by β_{k0}.

In this normal-based model, which we call model N1, we assume that the sample variances are good estimates of the true variances for the particular stratum, as in weighted mean difference and Weighted Least Squares methods. These conditions typically hold in situations where the strata sample sizes are large, such as meta-analysis, where strata are replaced by studies and where this type of model is traditionally used.[17] However, this may not be accurate if the strata are of small sizes. To tackle this shortcoming, we propose an alternative method that does not make this assumption. According to this method, if for stratum s and disease level l we observe the sample mean, sample standard deviation and sample size, $(\bar{x}_{sl}, s_{sl}^2, n_{sl})$ we can use the random effects model

$$\begin{aligned}
\bar{x}_{sl} &\sim N\left(\mu_{sl}, \frac{\sigma^2}{n_{sl}}\right) \\
\mu_{sl} &= \alpha_s + \sum_{k=1}^{L} \beta_{ks} \cdot I(k = l) \\
s_{sl}^2 &\sim \Gamma\left(\frac{n_{sl} - 1}{2}, \frac{n_{sl} - 1}{2\sigma^2}\right) \\
(\alpha_s, \beta_{1s}, \ldots, \beta_{Ls})^t &\sim MVN((\alpha_0, \beta_{10}, \ldots, \beta_{L0})^t, \Sigma)
\end{aligned}$$

while using as in model N1 uninformative priors for the parameters (including σ^2). We call this second normal-based model N2. This method was also previously used in meta-analysis applications;[18] however, we are not aware of its application to aggregate data with the format described here.

It is important to note here that in model N2 above, the sample variance and sample mean are independent. This is a unique property of the normal model, which does not apply to other distributions,[19] and it has important implications to the implementation of the model, which will be illustrated in the following sections.

These models are valid under the assumption that individual patient-level costs follow a normal distribution. In that case the sample variance follows a gamma distribution. When the normality assumption of the cost outcome does not hold (which is reasonable to expect), the appropriateness of this model is doubtful. Under a large sample size, we could rely on asymptotic normality for the sample mean, but it is not clear whether the gamma distribution could be used for the sample

variance. Additionally, the above model does not account for heteroscedasticity in the cost data, the variance of which is assumed to be equal to σ^2 (conditional on the random effects) regardless of the stratum the patient belongs to.

To better understand some of these issues, we ran a preliminary simulation study to examine the performance of the simple weighted mean estimator using stratum-specific sample mean costs, standard deviations and sample sizes under various assumptions about the underlying data generation process. Full details and results from this investigation are in Supplementary Appendix SA1 (available at http://smm.sagepub.com). The main findings were that (1) when the underlying distribution is a gamma with substantial skewness (small shape) and when the sample sizes of the individual strata are small, the weighted mean estimator is negatively biased and the usual 95% confidence interval has poor coverage; (2) increasing shape and sample size reduce bias to acceptable levels but do not result in 95% coverage; and (3) when the underlying distribution is normal, estimates are always unbiased but the smallest stratum sizes lead to slightly reduced coverage. Given the skewness of individual cost data and the small sizes of many of our costing groups, these results suggested that the models above might result in biased estimates of mean costs and spuriously high estimates of precision.

2.3 Proposed gamma models

In the previous section, we described existing methods that could potentially be used for net cost estimation from aggregate data (although we are not aware of their application to this problem yet). Here, we introduce a novel modeling approach that does not rely on normality assumptions for individual costs or approximate normality of the sample mean cost. The method is based on the assumption that the underlying distribution of the cost outcome is gamma, and the proposed model is based on gamma regression. This is a valid option that has been previously adopted[12] for the analysis of individual level data as it accommodates issues such as right skewness and heteroscedasticity. It is motivated by the fact that the distribution of the sample mean of gamma-distributed random variables is also gamma (with shape and rate parameters equal to the shape and rate of the original gamma multiplied by the sample size). This property is not shared by other skewed distributions such as log-normal, where sample means do not have recognized distributions.

2.3.1 Basic gamma model

A proposed model assuming an underlying gamma distribution with constant shape a across strata could be

$$
\begin{aligned}
\bar{x}_{sl} &\sim Gamma(a_{sl}, a_{sl}/\mu_{sl}) \\
a_{sl} &= a \cdot n_{sl} \\
\mu_{sl} &= \exp\left(\alpha_s + \sum_{k=1}^{L} \beta_{ks} \cdot I(k = l)\right) \\
s^2_{sl} &\sim F(a, \mu_{sl}, n_{sl}) \\
(\alpha_s, \beta_{1s}, \ldots, \beta_{Ls})^t &\sim MVN((\alpha_0, \beta_{10}, \ldots, \beta_{L0})^t, \Sigma)
\end{aligned}
$$

and uninformative priors for $\alpha_0, \beta_{10}, \ldots, \beta_{L0}, \Sigma, a$. Note that for the gamma distribution we use the parameterization and notation using shape and rate parameters.

In the model above we are faced with an inherent difficulty as the distribution of the sample variance is not known. There have been some efforts in determining this distribution in its exact[20] or approximate form,[21,22] but each of these approaches poses significant difficulties to the implementation and computation of the required posterior distributions.

Additionally, the proposed model uses both $\bar{x}_{ij}$ and s^2_{ij} as stochastic nodes in order to estimate the parameters of the underlying gamma model. The BUGS-type samplers require conditional independence between the stochastic nodes in the model.[23] It is known that sample means and sample variances are independent only if the underlying distribution is normal,[24] therefore stochastic nodes $\bar{x}_{ij}$ and s^2_{ij} in the aforementioned model are not conditionally independent and as a result, the implementation of such a model using BUGS-type software is problematic.

2.3.2 *A first alternative model, based on approximate distribution of the coefficient of variation*

In order to comply with the basic requirements of BUGS-type software, of having conditional independent stochastic nodes,[23] we refrain from using stochastic nodes for the sample variances, but instead we use the sample coefficient of variation, equal to the standard deviation over the mean. It is known that for a gamma distribution, the mean and coefficient of variation are independent.[24] Using an approach similar to Frost et al.,[25] and utilizing facts presented in Hwang and Huang,[26] we approximate the distribution of the squared sample coefficient of variation with a gamma distribution with parameters shape a' and rate r', such that

$$a' = \frac{(n-1)(a+2/n)(a+3/n)}{2a(a+1)}, r' = a'(a+1/n)$$

where a is the shape parameter of the parent gamma distribution and n is the sample size. Details are given in Supplementary Appendix SA2 (available at http://smm.sagepub.com). If we denote with c_{sl}, the sample coefficient of variation for stratum s and treatment l, equal to $s_{sl}/\bar{x}_{sl}$, the model becomes

$$\begin{aligned}
\bar{x}_{sl} &\sim Gamma(a_{sl}, a_{sl}/\mu_{sl}) \\
a_{sl} &= a \cdot n_{sl} \\
\mu_{sl} &= \exp\left(\alpha_s + \sum_{k=1}^{L} \beta_{ks} \cdot I(k = l)\right) \\
c^2_{sl} &\sim Gamma(a'_{sl}, r'_{sl}) \\
(\alpha_s, \beta_{1s}, \ldots, \beta_{Ls})^t &\sim MVN((\alpha_0, \beta_{10}, \ldots, \beta_{L0})^t, \Sigma)
\end{aligned} \qquad \text{(Model G1)}$$

and uninformative priors for $\alpha_0, \beta_{10}, \ldots, \beta_{L0}, \Sigma, a$.

2.3.3 *A second alternative model, based on the asymptotic distribution of the inverse coefficient of variation*

In an alternative approach, we can use the inverse coefficient of variation, equal to mean over standard deviation. The asymptotic distribution of the inverse coefficient of variation is normal with mean equal to the true inverse coefficient of variation and variance equal to 1 over the sample size.[27]

If for stratum s and treatment l we denote with ic_{sl} the sample inverse coefficient of variation, equal to $\bar{x}_{sl}/s_{sl}$, the model becomes

$$\begin{aligned}
\bar{x}_{sl} &\sim Gamma(a_{sl}, a_{sl}/\mu_{sl}) \\
a_{sl} &= a \cdot n_{sl} \\
\mu_{sl} &= \exp\left(\alpha_s + \sum_{k=1}^{L} \beta_{ks} \cdot I(k = l)\right) \\
ic_{sl} &\sim N(\sqrt{a}, 1/n_{sl}) \\
(\alpha_s, \beta_{1s}, \ldots, \beta_{Ls})^t &\sim MVN((\alpha_0, \beta_{10}, \ldots, \beta_{L0})^t, \Sigma)
\end{aligned} \qquad \text{(Model G2)}$$

and uninformative priors for $\alpha_0, \beta_{10}, \ldots, \beta_{L0}, \Sigma, a$.

2.4 Simulation experiments

The main objective of this study is to evaluate the usefulness of the gamma models G1 and G2, and compare them with simpler normal-based linear models N1 and N2. This was done with a series of simulation experiments. This section describes the design of the experiments and the methods of evaluation of the models.

2.4.1 Design of simulation experiments

The proposed methods for the estimation of net cost were evaluated through a number of simulation experiments. In an iterative procedure, a data set of individual level costs was generated assuming an underlying skewed distribution (such as gamma or lognormal) and then aggregated to stratum specific sample means and variances, as described in the previous sections. Subsequently, the gamma random effects models were fitted, along with models N1 and N2 described in previous section. After a number of iterations, the models were evaluated based on (a) how well they estimated the true net cost based on the data, and (b) for models G1 and G2 only, how close the coefficient estimates were to the true coefficients of the underlying data generation model.

The performance of the models was evaluated under different values for the total sample size and the shape of the underlying gamma distribution. As the total sample size increases, the normal linear models are expected to perform better, due to the central limit theorem. It is of interest to see whether the proposed gamma models outperform the linear models for smaller sample sizes. Additionally, it is of interest to evaluate how robust the model is as the shape of the underlying gamma or log-normal distribution varies, and more specifically for models G1 and G2 whether a larger shape parameter (which corresponds to a larger proximity to normality) results in better estimation.

The simulation was performed in three different steps:

Step 1: Determining the size of the strata

The sample size N_{sl} for each disease level l and stratum s was determined stochastically, assuming a fixed total sample size N (using three different values: 5000, 15,000, and 50,000), three disease levels (including control) and a number of $K = 200$ strata per disease level. The procedure accounted for

different probabilities of occurrence for each disease level and for heterogeneity of stratum sizes across disease levels. The stratum and disease distributions were considered independent. More details can be found in Supplementary Appendix SA3 (available at http://smm.sagepub.com).

Step 2: Sampling from a gamma or lognormal distribution for each stratum assuming random effects

The main characteristic of the proposed gamma models is that they assume a constant shape parameter and therefore a constant coefficient of variation across all strata (for gamma distribution shape is equal to the inverse of the square of the coefficient of variation). This assumption, although strong, is not rejected by the previously published data on pressure ulcer costs we use as motivating example. We make the same assumption for the data generating procedures. For the gamma models, we assume a constant shape parameter across all strata. We follow a full factorial design using for the shape parameter the values 0.5, 1 and 2 and for the total sample size the values $N = 5000$, 15,000, 50,000.

Shape values of 0.5, 1 and 2 correspond to coefficient of variation values of $1/\sqrt{0.5}, 1, 1/\sqrt{2}$, respectively.

For the log-normal model, the coefficient of variation is a function of the parameter σ. We select values of σ that give similar coefficient of variation *cv* as with the gamma generating process. Following simple calculations (see Supplementary Appendix SA3, available at http://smm.sagepub.com) we chose the values $\sqrt{\log(3)}$, $\sqrt{\log(2)}$, and $\sqrt{\log(1.5)}$ or 1.0481471, 0.8325546, 0.6367614, respectively.

For the gamma based simulated data the procedure is

$$
\begin{aligned}
&(\alpha_s, \beta_{1s}, \beta_{2s})' \sim N\big((\alpha_0, \beta_{10}, \beta_{20})', \Sigma\big), \quad s = 1, \ldots, S \\
&\mu_{sl} = \exp(\alpha_s + \beta_{1s} \cdot I(j = 1) + \beta_{2s} \cdot I(j = 2)) \\
&X_{sli} \sim Gamma(shape, shape/\mu_{sl}), \quad s = 1, \ldots, 200, l = 0, 1, 2, i = 1, \ldots, N_{sl}
\end{aligned}
$$

where $I(\cdot)$ denotes the indicator function, *shape* is the shape parameter for the gamma distribution, and the parameterization used above involves the rate parameter, equal to the reciprocal of the scale. The parameter values used for the simulations are $\alpha_0 = 2, \beta_{10} = 1, \beta_{20} = 1.3$, with matrix $\Sigma = \begin{bmatrix} 0.25 & 0 & 0 \\ 0 & 0.25 & 0.125 \\ 0 & 0.125 & 0.25 \end{bmatrix}$, i.e. the standard deviation of each random effect is 0.5, random effects between intercept and "increment coefficient" are uncorrelated (higher control costs does not assume higher proportional case vs. control costs), and random effects for each case level have correlation 0.5.

Similarly, the log-normal-based sampling procedure is

$$
\begin{aligned}
&(\alpha_s, \beta_{1s}, \beta_{2s})' \sim N\big((\alpha_0, \beta_{10}, \beta_{20})', \Sigma\big), \quad s = 1, \ldots, S \\
&\mu_{sl} = \exp(\alpha_s + \beta_{1s} \cdot I(j = 1) + \beta_{2s} \cdot I(j = 2)) \\
&X_{sli} \sim Lognormal\,(\log(\mu_{sl}) - \sigma^2/2, \sigma), \quad s = 1, \ldots, 200, \quad l = 0, 1, 2, i = 1, \ldots, N_{sl},
\end{aligned}
$$

Step 3: Summarizing the data

Sample means can be calculated when $N_{sl} > 0$, while sample variances when $N_{sl} > 1$, and their values are given by

$$\bar{X}_{sl} = \frac{1}{N_{sl}} \sum_{i=1}^{N_{sl}} X_{sli}, \ \textit{if } N_{sl} > 0$$

$$S^2_{sl} = \frac{1}{N_{sl} - 1} \sum_{i=1}^{N_{sl}} (X_{sli} - \bar{X}_{sl})^2, \ \textit{if } N_{sl} > 1$$

The triplet $\left(\left\{ \bar{X}_{sl} \right\}_{sl}, \left\{ S^2_{sl} \right\}_{sl}, \{N_{sl}\}_{sl} \right)$ constitutes the observed data to be modeled.

Step 4: Fitting the Bayesian models

Given the simulated data expressed by the triplet $\left(\left\{ \bar{X}_{sl} \right\}_{sl}, \left\{ S^2_{sl} \right\}_{sl}, \{N_{sl}\}_{sl} \right)$, each one of the three Bayesian models described in previous sections was fitted. To facilitate rerunning models within a simulation, the JAGS sampler[28] and the rjags R package[29] were used for the implementation of the models. Each model used three chains, with the first 10,000 iterations discarded as burn-in. A subsequent 10,000 samples were used for estimates of the posterior distribution of the parameters of interest. In a series of initial trial runs, it was determined that this number of iterations was sufficient for the Markov chain to converge. The parameters of interest estimated for each one of the four models are shown below

$$\begin{aligned} \text{N1, N2} &: C_0, NC_1, NC_2 \\ \text{G}_1, \ \text{G}_2 &: C_0, NC_1, NC_2, \alpha_0, \beta_{10}, \beta_{20}, s \end{aligned}$$

where C_0, NC_1, NC_2 denote the "control cost", and the net cost for levels 1 and 2 respectively. For models N1 and N2 these are given by

$$\begin{aligned} C_0 &= \alpha_0 \\ NC_1 &= \beta_{10} \\ NC_2 &= \beta_{20} \end{aligned}$$

while for models G1 and G2 are given by

$$\begin{aligned} C_0 &= \exp(\alpha_0 + \sigma^2_\alpha/2) \\ NC_1 &= \exp(\alpha_0 + \beta_{10} + (\sigma^2_\alpha + \sigma^2_{\beta 1})/2) - \exp(\alpha_0 + \sigma^2_\alpha/2) \\ NC_2 &= \exp(\alpha_0 + \beta_{20} + (\sigma^2_\alpha + \sigma^2_{\beta 2})/2) - \exp(\alpha_0 + \sigma^2_\alpha/2) \end{aligned}$$

(See Appendix SA4 for details).

The above procedure was repeated 120 times for different values for each of the nine combinations of values for parameters a and N, $\{0.5, 1, 2\} \times \{5000, 15{,}000, 50{,}000\}$.

2.4.2 Assessment of the models

Each one of the three models was evaluated by the proximity of the parameters estimates to the "true" values, using the measures of bias, variance, root mean squared error (RMSE), and 95%

confidence interval coverage. Those measures of accuracy were calculated using the posterior means from the 120 simulation runs. For example, if CC_{true} denotes the true control cost and CC_i denotes the posterior mean for the control cost for the i-th simulation run, the bias is calculated as $bias(CC) = \frac{1}{120}\sum_{i=1}^{120}(CC_i - CC_{true})$. Additionally, the estimates of all the parameters that are only part of the gamma model were compared with the values used for the generation of the data.

2.4.3 Results

Estimation of the bias leads to the following observations: N1 model clearly underestimates the costs, especially the net costs 1 and 2. Model N2 shows very small bias under all conditions. Gamma models overestimate control and net costs. As expected, results improving as total sample size increases and as underlying shape parameter is increasing (i.e. distribution becomes more symmetric). G2 seems to have lower bias than G1. Lognormal data generating distribution results in smaller bias (compared to gamma data). Figure 1 presents the values of the models.

When assessing the variance, we see that N1 model results in smaller values compared to other three models, while N2 has the larger variance in most cases. While very similar, G1 has smaller variance than G2 in most cases. In general variance is decreasing when total sample size increases and underlying distribution becomes more symmetric (shape parameter increases). In most cases estimation of lognormal data shows smaller variances when compared to gamma data. Results are showing in Figure 2.

Regarding RMSE, for control costs model N1 gave in most cases the smaller values. For both net costs model N2 gives always the smaller RMSE, while N1 gives the larger RMSE, with the exception of $s = 1$, $N = 50$ K, when the gamma models show larger RMSE. The two gamma models give similar RMSE. Overall, estimation of lognormal data results in smaller RMSE when compared to gamma data. Results are shown in Supplementary Figure SF1 (available at http://smm.sagepub.com).

Finally, estimates of 95% confidence intervals coverage showed that both normal models had 100% coverage for control cost. N2 model has the highest coverage over all four models. N1 has also 100% coverage for control cost but low coverage for net costs (especially for stage 2). G1 and G2 models have less than nominal coverage (but better that N1 for net costs). The way coverage is affected by the total sample size or the shape of the underlying distribution does not seem to follow a consistent pattern across all four models. Supplementary Figure SF2 (available at http://smm.sagepub.com) shows the results.

We also recorded the elapsed time used by each model for each simulation run, i.e. the time to fit the model to a single dataset. Time was measures in seconds. For all four models we noticed an increasing elapsed time as the total sample size increases and as the shape parameter of the underlying data generating procedure increases. We noticed no difference between the elapsed times for gamma and log-normal simulated data. Median times and inter-quartile ranges were estimated over all combinations of parameters used for the simulation. The estimates for models N1, N2, G1, and G2 were 353.9 (334.3, 377.9), 387.7 (365.6, 409.4), 973.7 (884.5, 1084.2), and 763.1 (724.3, 792.9), respectively. It is clear that the gamma models incur a substantial computational expense in comparison to the linear normal models.

We also assess the estimation of various parameters of the gamma model, such as coefficients, variances and correlations of the random effects and the shape parameter. Tables 1 and 2 present the results for estimates of key parameters of the gamma models, such as coefficients, variances, and correlations of the random effects and the shape parameter. Supplementary Tables ST1 and ST2 (available at http://smm.sagepub.com) contain estimates for data generated based on log-normal distribution.

From those results, we observe that the estimates of the coefficients are quite accurate and their accuracy increases as the sample size increases and as the shape parameter increases. We also

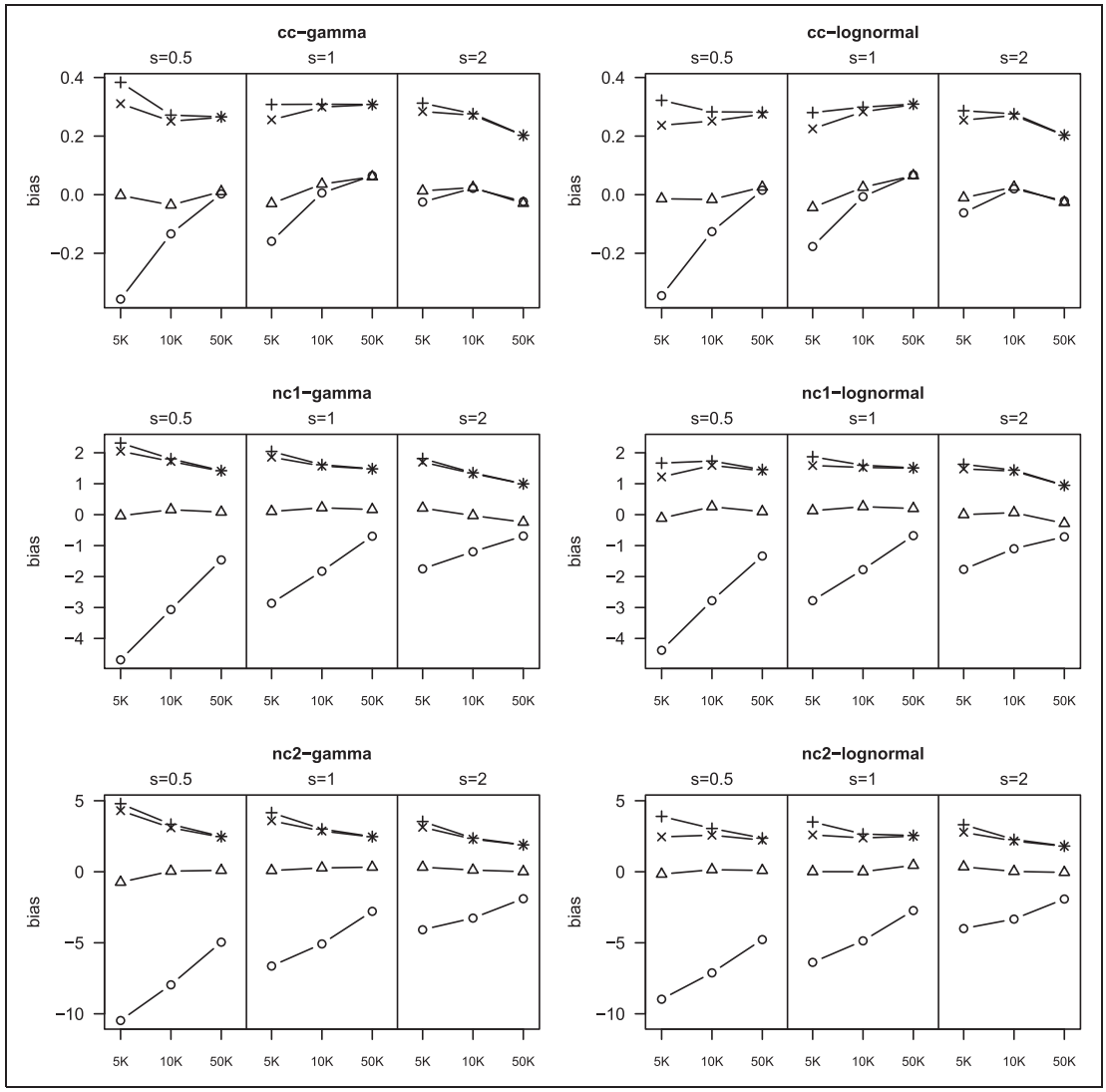

Figure 1. Plots showing the bias for the different models. Symbols ○, Δ, +, and x denote values for models N1, N2, G1, and G2, respectively.

observe that G2 model (which utilizes the asymptotic distribution of the inverse coefficient of variation) heavily overestimates the shape parameter (bias is reduced as sample size increases). This bias results in very poor coverage (0%) in those cases. We looked more carefully into the behavior of model G2, where shape is highly overestimated and also has a very high standard deviation. We found that for few of the 120 generated datasets and estimations, the shape is highly overestimated. This is because the particular simulated datasets contains cells with values of very low standard deviation (i.e. very high inverse coefficient of variation). We hypothesize that these particular extreme data points drive the estimation of the shape parameter to a very high value, due to the fact that the Bayesian model utilizes the asymptotic distribution.

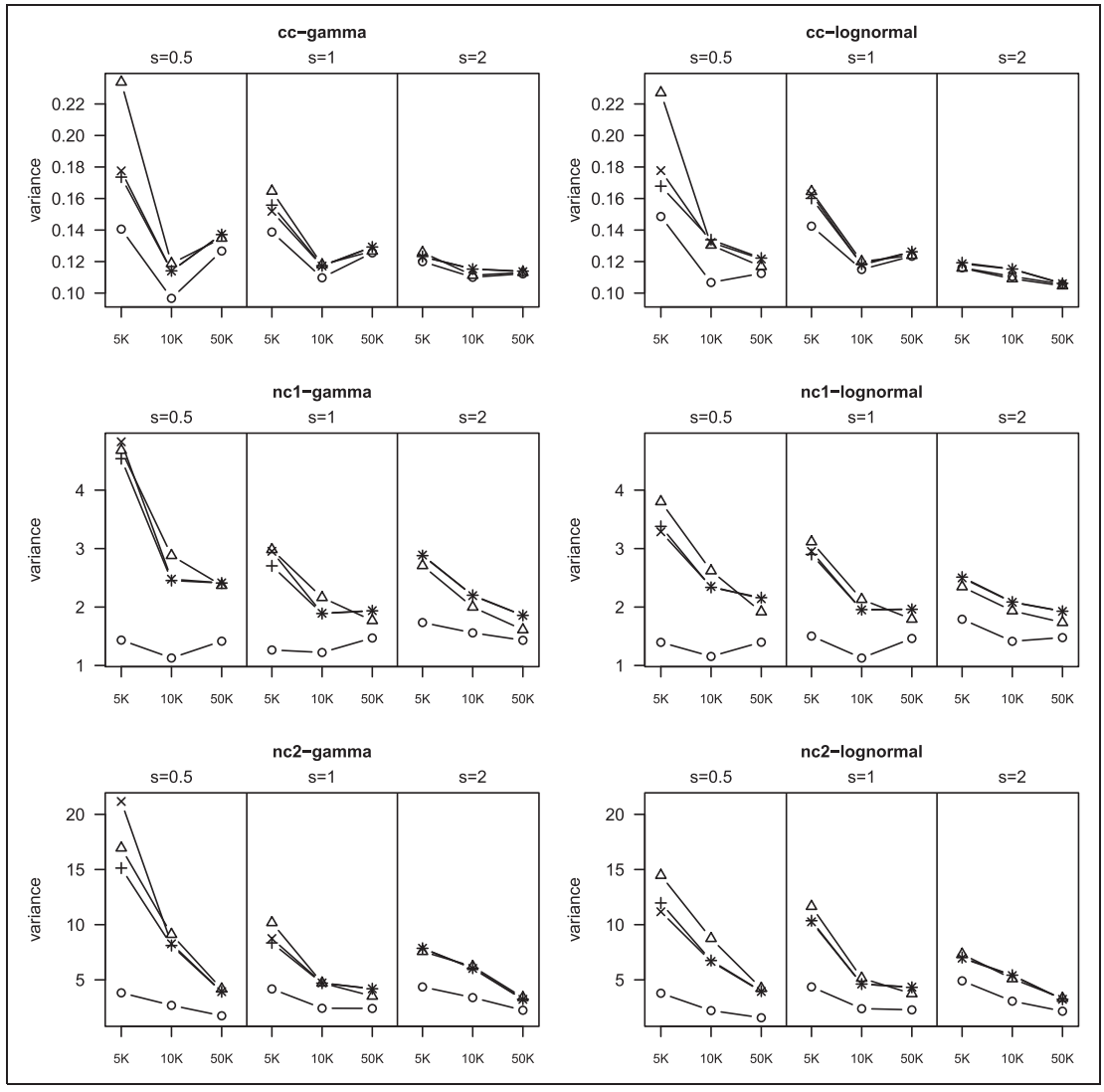

Figure 2. Plots showing the variance for the four different models. Similar notation with Figure 1 is used.

2.5 Analysis of real pressure ulcer cost data

In addition to the assessment based on experiments using simulated data, we consider the cost estimation problem of pressure ulcers described in the Introduction and applied all four Bayesian models. We preprocessed the original data in order to make them more suitable for analysis from all four models. After preprocessing, the data contain controls and two levels of pressure ulcers (stage II pressure ulcers grouped with unstageable and stage III and IV pressure ulcers grouped together, as they have similar costs), resulting in 108 strata based on combinations of age, co-morbidities, and most responsible diagnosis. Most responsible diagnoses with less than 20 cases in total were excluded from the analysis.

Table 1. Parameter estimates for gamma generated data from G1 model.

		$N = 5$ K				$N = 15$ K				$N = 50$ K			
Shape	Parameter	Mean	Median	sd	cov95	Mean	Median	sd	cov95	Mean	Median	sd	cov95
0.5	alpha0	1.986	1.988	0.045	0.967	1.994	1.996	0.037	0.983	2.002	2.001	0.038	0.967
	beta10	0.989	0.992	0.064	0.992	1.002	1.001	0.049	0.975	0.999	0.995	0.037	0.983
	beta20	1.243	1.247	0.093	0.933	1.286	1.287	0.062	0.967	1.301	1.303	0.040	0.983
	Shape	0.527	0.527	0.015	0.517	0.508	0.508	0.010	0.833	0.501	0.501	0.006	0.917
1	alpha0	1.992	1.993	0.042	0.975	2.004	2.000	0.036	0.975	2.006	2.006	0.039	0.950
	beta10	1.004	1.002	0.049	0.983	1.000	1.003	0.043	0.992	0.999	0.996	0.039	0.942
	beta20	1.289	1.295	0.068	0.975	1.296	1.295	0.050	0.967	1.302	1.305	0.038	0.975
	shape	1.029	1.028	0.029	0.742	1.008	1.007	0.015	0.942	1.003	1.003	0.009	0.933
2	alpha0	2.002	2.005	0.038	0.958	2.002	2.001	0.039	0.967	1.998	2.000	0.035	0.967
	beta10	1.003	1.001	0.048	0.958	0.998	0.997	0.042	0.950	0.997	1.002	0.037	0.967
	beta20	1.295	1.294	0.062	0.958	1.294	1.295	0.050	0.967	1.304	1.303	0.038	0.967
	shape	2.029	2.027	0.050	0.933	2.007	2.007	0.025	0.950	2.000	2.000	0.016	0.950

Table 2. Parameter estimates for gamma generated data from G2 model.

		$s = 0.5$, $N = 5$ K				$s = 0.5$, $N = 15$ K				$s = 0.5$, $N = 50$ K			
Shape	True	Mean	Median	sd	cov95	Mean	Median	sd	cov95	Mean	Median	sd	cov95
0.5	alpha0	1.955	1.961	0.051	0.858	1.988	1.992	0.038	0.975	2.001	2.001	0.038	0.967
	beta10	0.963	0.965	0.065	0.942	0.997	0.996	0.049	0.967	0.998	0.994	0.037	0.983
	beta20	1.162	1.161	0.107	0.675	1.270	1.273	0.065	0.933	1.299	1.301	0.040	0.983
	Shape	1.100	0.773	2.025	0.000	0.663	0.602	0.609	0.000	0.532	0.532	0.005	0.000
1	alpha0	1.974	1.977	0.042	0.950	2.001	1.997	0.036	0.975	2.005	2.005	0.039	0.950
	beta10	0.990	0.991	0.050	0.967	0.997	1.000	0.043	0.975	0.999	0.996	0.039	0.950
	beta20	1.246	1.251	0.075	0.875	1.289	1.290	0.050	0.975	1.301	1.304	0.038	0.975
	shape	1.746	1.485	1.393	0.000	1.165	1.145	0.068	0.000	1.047	1.046	0.010	0.000
2	alpha0	1.993	1.994	0.038	0.967	2.001	2.000	0.039	0.967	1.998	2.000	0.035	0.967
	beta10	0.996	0.993	0.048	0.950	0.996	0.996	0.043	0.958	0.997	1.003	0.037	0.967
	beta20	1.273	1.274	0.063	0.942	1.290	1.291	0.051	0.942	1.303	1.302	0.038	0.967
	shape	3.011	2.883	0.451	0.000	2.323	2.231	0.465	0.000	2.073	2.065	0.069	0.000

The models were evaluated using root mean squared error, mean absolute error, and Spearman rho between the observed and model-predicted values for the mean costs. The predicted values of the sample mean costs per stratum were generated in two ways, following an approach similar to Green et al.[30] The first approach (we call it "full" to follow Green et al.'s[30] notation) uses the predictive distribution conditional on the random effects estimated from the data from the final model, while the second approach ("mixed") relies on the predictive distribution conditional on random effects that are simulated from their underlying distribution at each iteration of the MCMC sampler. The latter predictive distribution method was introduced in Marshall and Spiegelhalter.[31]

Based on this distinction we generate four different estimates for each of the evaluation measures, depending on the type of prediction ("full" or "mixed") and on the strata used (total number or only

Table 3. Cost estimates and measures of predictive accuracy of all four models applied to pressure ulcer cost data. Accuracy measures were calculated using "full" and "mixed" predictive distributions, while using all data and only those with stratum size larger than 1, denoted as "ltd".

Variable	N1 mean (se)	N2 mean (se)	G1 mean (se)	G2 mean (se)
Cost				
Control	17,242 (14)	17,569 (15)	17,273 (7)	17181 (7)
Net 1	39,777 (44)	51,069 (39)	60,457 (150)	63,248 (178)
Net 2	58,061 (155)	72,914 (78)	104,937(655)	113,123 (734)
RMSE				
Mixed (ltd)	1,266,359(576)	1,253,819 (398)	1,603,515 (5353)	2,038,335 (9395)
Mixed	NA	1,515,134 (444)	2,254,761 (9069)	2,606,860 (12,668)
Full (ltd)	724,948 (619)	163,134 (50)	1,177,299 (5569)	1,793,349 (10,260)
Full	NA	251,383 (74)	1,723,210 (8629)	2,168,862 (12,348)
MAE				
Mixed (ltd)	56,403 (20)	59,057 (18)	50,030 (75)	70,708 (175)
Mixed	NA	66,092 (19)	61,304 (115)	65,808 (152)
Full (ltd)	18,862 (8)	6634 (2)	28,936 (58)	57,701 (151)
Full	NA	9338 (2)	39,161 (87)	46,624 (115)
Correlation				
Mixed (ltd)	0.2505 (0.0004)	0.302 (0.0003)	0.4382 (0.0005)	−0.0876 (0.0002)
Mixed	NA	0.2862 (0.0003)	0.376 (0.0005)	0.3302 (0.0006)
Full (ltd)	0.8166 (0.0002)	0.9493 (0)	0.8472 (0.0002)	0.171 (0.0006)
Full	NA	0.9139 (0.0001)	0.7324 (0.0003)	0.6556 (0.0003)

those with size larger than 1). The second distinction was used in order to make model N1 comparable to the other three models. Table 3 describes the posterior means and time-series standard errors of these measures. Posterior median values and interquartile ranges are also given in the Supplementary Table ST3 (available at http://smm.sagepub.com).

We see that although the estimates of the control costs are very similar across the four models, the net cost estimates differ significantly and they follow an increasing order of N1, N2, G1, G2, while model N2 seems to have the smaller standard error and IQR. As expected, the model accuracy measure estimates based on "marginal predictions" are significantly larger than those based on "conditional predictions". For most of those measures, model N2 performed the best, following by G1. G1 performed best on "marginal predictions" based estimates. Measures seem to improve when we exclude strata with size 1. Model G2 seems to collapse when strata from single observations are included. Overall, we can conclude that normal linear model N2 gives the most reliable performance in this data.

3 Discussion and conclusions

In this paper, we examine the problem of estimation of net cost associated with a disease or condition, when only aggregate data are available. Employing a Bayesian framework, we compare the performance of a simple linear mixed model, with a more complex linear model and two proposed gamma mixed models. The three proposed models (the "complex" linear model and the two gamma models) have various advantages over alternative approaches. First, among all methods that could be used for the specific analysis objective using data of the specified aggregate structure, our proposed modeling approach is the only one that we are aware of that uses the sample

variances as data and does not treat them as true variances. We believe that this is important as, especially in smaller strata, sample variances can be far from the true values. The proposed approach provides a conceptual framework that deals with this naturally.

Second, the proposed modeling approach deals with the missing values in a subtle and implicit way. In the proposed models N2, G1, and G2, strata with size 1, for which sample variance is not available, do participate in the part of the model involving the sample means, but not in the part involving the sample variance. This is an advantage over frequentist-based approaches where strata with missing sample variances need to be excluded completely from the model.

Additionally, one would think that normal models would not be able to estimate parameters of the gamma or log-normal generated data. It is somewhat surprising that its performance is quite satisfactory, compared to the more complex and computationally expensive gamma models. One plausible explanation for this behavior relates to the flexibility of the random effects. It appears that the random effects for the intercept (which concerns the ''control'' cost in the normal models) and the ''increment coefficient'' (concerning the net costs) are able to ''mimic'' successfully the ''proportional'' nature of the net costs inherited by the generating gamma model. This was also observed by the high correlation between the random effects posterior means of $\alpha_s, \beta_{1s}, \beta_{2s}$, with values ranging between 0.31 and 0.52, corresponding to different settings in the simulation. Supplementary Figure SF3 (available at http://smm.sagepub.com) presents the relationship of the posterior means for one of these simulation settings.

We also note here that we can also extend to the definitions above to the mean net cost for cases only. This estimate has been previously suggested[6,32] as more appropriate for characterizing the amount of expenditure that could have been potentially saved on average, should the cases have been keeping their demographic and clinical profile, but just switching to being controls. The definitions then become $CMNC(l = l_1) = E_{\alpha_j, \beta_{1j}, \ldots, \beta_{Sj} | l = l_1}[NC_j(l = l_1)]$. It is as yet unclear how to extend our Bayesian models to estimate such a type of mean net cost.

There are a few limitations to this study. One of them regards the choice of measures of goodness of fit for the evaluated models. Deviance information criterion (DIC)[33] has been frequently used for the assessment of goodness of fit for Bayesian models, as well as for model comparison. Although we consider its use for our study we decided against it because it is not clear how it could be used in the specific data and modeling setting. The most obvious choice would be to calculate DIC only based on the sample means data, ignoring the sample variance data. This is imposed by the fact that the four models do not share the same form of the sample variance data (e.g. the gamma model uses the sample coefficient of variation or its inverse instead of the sample variance) and therefore any calculations involving sample variance data would not be consistent. On the other hand, it is unclear if any model assessment and comparison using this version of DIC would reflect the true goodness of the fit of the model. For example, a hypothetical poor fit of the model on the sample variance data may have a serious effect on the estimation of net cost but not be reflected in DIC calculated in the described way.

Another limitation is the number of simulation runs (120) used for each combination of model parameters. This was due to the high computational expense of the Bayesian models used (especially the gamma models), which puts constraints on the number of simulation runs. A larger number of runs would have allowed a higher accuracy in the assessment of models' performance.

Despite those limitations, we believe that our study contributes significantly in the investigation of appropriate statistical methods that can be used for the estimation of health care disease specific net cost, when only aggregate level cost data are available. We explored the use of a Bayesian modeling approach that has the main advantage of utilizing the stratum specific sample cost variances as stochastic and we demonstrated its advantage over a simpler modeling approach

where the sample variances are used as the true underlying variances. We found that utilizing more complex models relying on gamma distribution does not improve the results over a simpler and less computationally expensive normal linear model. We hypothesize that the main reason for that are the limitations associated with the use of approximate or asymptotic distributions of the squared or inverse coefficient of variation. A suggested direction of future research could explore alternative approaches of modeling the coefficient of variation (or functions of it) for gamma distributed data. Another possible reason for the suboptimal performance of the gamma models is their complexity and the relation between mean and variance. This could have potential repercussions in the convergence of the MCMC. Although some preliminary examinations we performed did not find any significant issues, this cannot be ensured for the whole series of simulation runs. Finally, we need to also point out a potential sensitivity of the models to the choice of the prior for the covariance (precision) matrix. There has recently been some discussion in the literature suggesting that the choice of the inverse Wishart prior can be problematic.[34] These potential issues may have a larger effect on the gamma model because of the dependence between variance and mean. A thorough investigation of alternative priors for the covariance matrix (or appropriate re-parameterizations) is one of our suggested next research steps.

We conclude that normal-based Bayesian linear mixed models that model sample variances as stochastic (in a way similar to N2 model described in this study) provide a reliable framework for the analysis of aggregate cost data for the estimation of health care net cost, combining desirable properties (minimal bias, efficiency, robustness) at a relatively low computational expense. We therefore recommend its use among all other currently available alternative approaches.

We believe that the findings of our study can contribute to the increase of quality of the estimation of health care net costs overcoming limitations posed by data availability. As a consequence, we believe that our study can have important impact to health care science and services and policy decision making.

Acknowledgments

The authors would like to thank Professors Michael Escobar and Murray Krahn, as well as the anonymous reviewers for their constructive comments on the paper.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Conflict of interest

None declared.

References

1. Barlow WE. Overview of methods to estimate the medical costs of cancer. *Med Care* 2009; **47.7**(Suppl 1): S33.
2. Baker MS, Kessler LG, Urban N, et al. Estimating the treatment costs of breast and lung cancer. *Med Care* 1991; **29**: 40–49.
3. Brown ML, Riley GF, Schussler N, et al. Estimating health care costs related to cancer treatment from SEER-Medicare data. *Med Care* 2002; **40**(8 Suppl): 104–117.
4. Taplin SH, Barlow W, Urban N, et al. Stage, age, comorbidity, and direct costs of colon, prostate, and breast cancer care. *J Natl Cancer Inst* 1995; **87**: 417–426.
5. Chan BC, Ieraci L, Mitsakakis N, et al. Net costs of hospital-acquired and pre-admission pressure ulcers among older people hospitalized in Ontario. *J Wound Care* 2013; **22**(7): 341–342, 344–346.

6. Basu A, Arondekar BV and Rathouz PJ. Scale of interest versus scale of estimation: Comparing alternative estimators for the incremental costs of a comorbidity. *Health Econ* 2006; **15**(10): 1091–1107.
7. DerSimonian R and Laird N. Meta-analysis in clinical trials. *Control Clin Trials* 1986; **7**(3): 177–188.
8. Martin BC, Ricci JF, Kotzan JA, et al. The net cost of Alzheimer disease and related dementia: A population-based study of Georgia medicaid recipients. *Alzheimer Dis Assoc Disord* 2000; **14**(3): 151–159.
9. Zhan C and Miller MR. Excess length of stay, charges, and mortality attributable to medical injuries during hospitalization. *JAMA* 2003; **290**(14): 1868–1874.
10. Mihaylova B, Briggs A, O'Hagan A, et al. Review of statistical methods for analysing healthcare resources and costs. *Health Econ* 2011; **20**(8): 897–916.
11. Manning WG, Basu A and Mullahy J. Generalized modeling approaches to risk adjustment of skewed outcomes data. *J Health Econ* 2005; **24**(3): 465–488.
12. Manning WG and Mullahy J. Estimating log models: To transform or not to transform? *J Health Econ* 2001; **20**(4): 461–494.
13. Cooper NJ, Sutton AJ, Mugford M, et al. Use of Bayesian Markov Chain Monte Carlo methods to model cost-of-illness data. *Med Decis Making* 2003; **23**(1): 38–53.
14. Sutton AJ, Abrams KR, Jones DR, et al. *Methods for meta-analysis in medical research*. New York: John Wiley, 2000.
15. Draper N and Smith H. *Applied regression analysis*, 2nd ed. New York: John Wiley, 1981.
16. Kleinbaum DG and Kupper LL. *Applied regression analysis and other multivariable methods*. Boston: Doxbury Press, 1978.
17. Spiegelhalter DJ, Abrams KR and Myles JP. *Bayesian approaches to clinical trials and health-care evaluation*. Chichester: Wiley, 2004.
18. Stevens JW. A note on dealing with missing standard errors in meta-analyses of continuous outcome measures in WinBUGS. *Pharm Stat* 2011; **10**: 374–378.
19. Geary RC. The distribution of the "Student's" ratio for the non-normal samples. *Suppl J R Stat Soc* 1936; **3**(2): 178–184.
20. Royen T. Exact distribution of the sample variance from a gamma parent distribution. arXiv:0704.1415v1[math.ST], 2007.
21. Mudholkar GS and Trivedi MC. A Gaussian approximation to the distribution of the sample variance for nonnormal populations. *JASA* 1981; **76**(374): 479–485.
22. Solomon H and Stephens MA. An approximation to the distribution of the sample variance. *Can J Stat* 1983; **11**(2): 149–154.
23. Spiegelhalter DJ, Thomas A and Best NG. WinBUGS Version 1.4 user manual. MRC Biostatistics Unit, 2003.
24. Hwang TY and Hu CY. On a characterization of the Gamma distribution: The independence of the sample mean and the sample coefficient of variation. *Ann I Stat Math* 1999; **51**(4): 749–753.
25. Frost J, Keller K, Lowe J, et al. A note on interval estimation of the standard deviation of a Gamma population with applications to statistical quality control. *Appl Math Model* 2013; **37**(4): 2580–2587.
26. Hwang TY and Huang PH. On new moment estimation of parameters of the gamma distribution using its characterization. *Ann Inst Stat Math* 2002; **54**: 840–847.
27. Sharma KK and Krishna H. Asymptotic sampling distribution of inverse coefficient-of-variation and its applications. *IEEE T Reliab* 1994; **43**(4): 630–633.
28. Plummer M. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In: *Proceedings of the 3rd international workshop on distributed statistical computing*, Vienna, Austria, 20–22 March 2003.
29. Plummer M. rjags: Bayesian graphical models using MCMC. 2011. R package version 3-3. http://CRAN.R-project.org/package=rjags.
30. Green MJ, Medley GF and Browne WJ. Use of posterior predictive assessments to evaluate model fit in multilevel logistic regression. *Vet Res* 2009; **40**(4): 30.
31. Marshall EC and Spiegelhalter DJ. Approximate cross-validatory predictive checks in disease mapping. *Stat Med* 2003; **22**: 1649–1660.
32. Imbens GW. Nonparametric estimation of average treatment effects under exogeneity: A review. *Rev Econ Stat* 2004; **86**(1): 4–29.
33. Spiegelhalter DJ, Best NG, Carlin BP, et al. Bayesian measures of model complexity and fit (with discussion). *J Roy Stat Soc B Met* 2002; **64**(4): 583–639.
34. Barnard J, McCulloch R and Meng XL. Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statist Sin* 2000; **10**: 1281–1311.