Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Systematic review of statistical methods for analyzing healthcare cost in administrative data

Mary Dooley (mizzelm@musc.edu)
Medical University of South Carolina https://orcid.org/0000-0003-4454-8732
Kit N. Simpson
Medical University of South Carolina
Heather S. Bonilha
Medical University of South Carolina

Research article

Keywords: Cost data, multivariable regression, generalized linear models, ordinary least squares models, administrative data, healthcare cost, healthcare cost analysis, skewed data

Posted Date: June 25th, 2019

DOI: https://doi.org/10.21203/rs.2.10625/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Abstract

BACKGROUND Healthcare costs are increasing at alarming rates in the United States (US) putting a heavy burden on the healthcare reimbursement system. Cost and cost savings have become an important focus as health policy administrators are tasked with determining the most effective allocation of limited resources. The availability of large databases, such as administrative data, comes with many challenges for analyses, including: skewed data, inflated zero counts, and potential selection bias among comparison groups. Thus, it is imperative that they are evaluated correctly. There are many different methods currently being used to estimate costs including: generalized linear models with a log link, natural logarithm transformed costs, gamma distribution, median regression, two-part models, and Bayesian models. This systematic review will identify which methods are statistically and mathematically appropriate for large claims data. METHODS Scopus and Ovid were searched for potential statistical method papers using multivariable modelling of cost that were published up to the end of February 2018. Inclusion criteria required either a comparison of two or more statistical methods to analyze cost or one statistical method performed on two or more different types of cost data. This systematic review follows the guidelines according to Preferred Reposting Items for Systematic Reviews and Meta-Analysis (PRISMA). RESULTS The review identified 1,048 potential papers, of which, 80 met the inclusion criteria for a full article review. There was a total of 9 papers included in the systematic review; one paper included simulations and eight papers assessed real cost data. There were 28 models assessed across the nine papers with ordinary least squares (OLS) and generalized linear models (GLM) being the most common. CONCLUSIONS GLM using the gamma distribution was included in all but two of the comparisons. Most other models that were compared to the GLM Gamma distribution with log link found it to be the superior model in both simulated data and real administrative data.

Background

Cost and cost savings are becoming increasingly important for the US healthcare system. During the first 5 months of 2019, over one hundred papers with a key word of “cost savings” were published in major US journals; 5 in the New England Journal of Medicine, 4 in JAMA, 3 in PLoS ONE, and the balance in other journals aimed at decision makers. This overwhelming focus on cost savings makes it imperative that we begin to standardize our approach to cost analysis, because different approaches have different underlying statistical assumptions and often result in different outcomes.

Appropriate health care cost estimation is crucial as it is used to guide evidence-based health policy implementation. Health policy makers rely on costs to drive their decisions[1]. The estimation for costs associated with a disease, such as diabetes, can influence the allocation of resources for the prevention and treatment of the disease[2]. As numerous administrative data sources have become available for analysis, it is essential that these data are analyzed properly. These claims data, or billing data, are usually observational data sources, such as Medicare billing data, that are often assessed for health care cost outcomes. Inaccurate cost analysis can lead policy makers to make sub-optimal decisions.

Analysis of large datasets of observational data come with its own challenges. Data are often skewed and group comparisons often biased. Cost data are commonly positively skewed with a few outlying patients, often those with complications or who are very ill, disproportionately consisting of much of the total costs[3,6,9,11]. Administrative data can often include zero costs or zero observations (non-users) that may make up a large proportion of the data[9,11]. Further challenges arise when analysis of cost data also need to consider censored data within the research time frame[3]. Healthcare cost can vary according to region, health care system, population, and payer; thus, adjustment of costs is needed for comparisons across these factors. There are many different methods currently being used to estimate costs including: generalized linear models with a log link, natural logarithm transformed costs, gamma distribution, median regression, two-part models, and Bayesian models. As cost drives many health care policies, inaccurate analyses of cost can have serious consequences.

A review of approaches that compare cost analysis methods, which are conducted statistically and mathematically, is essential to provide needed evidence as to which methods are the most appropriate and valid for the evaluation of claims data. Therefore, we conducted such a systematic review to identify what methods of cost analysis are statistically and mathematically appropriate to use with large claims data and, specifically, determine whether one method could be considered optimal. The knowledge gained from this review can be used to guide evidence-based cost analysis and properly assist policy makers’ decisions.

Methods

Potential papers were identified by searching two databases, Scopus and Ovid, using the following search terms: [“methodology" or "simulations" or "bootstrap" or "bootstrapped" or "model comparison" or "compare models" or "markov" or “statistical model”] and [“cost data" or "claims data" or "billing data" or "health insurance data" or "health insurance claims" or "billing claims" or "administrative data”] and [“cost analysis" or "health care costs" or "hospital costs”] in the title, abstract, or keywords. Searches were conducted for all articles published up to the end of February 2018.

Only statistical method papers using multivariable modelling of cost, with or without methods controlling for selection bias, were considered eligible given they met one of the following inclusion criteria: 1) a comparison of two or more statistical methods to analyze cost or 2) one statistical method performed on two or more different types of cost data. Since administrative cost data can be very different from cost data collected in a randomized controlled trial, papers that used only data from randomized controlled trials and not administrative data were excluded.

Additional exclusions included not being published in English and review papers. Bayesian methods and joint modelling methods were considered beyond the scope of this review. Data elements extracted from eligible papers include: paper's first author, year of publication, statistical methods being evaluated, data types used in analysis, year(s) the data was collected, simulation approach used, sample size of data, distribution of the data, the results of the method performance and the author's recommendations based on their results.

The titles and abstracts of all papers found utilizing the search terms in both databases were reviewed for potential eligibility. Among those titles and abstracts meeting potential eligibility, the full text article was reviewed for eligibility to be included in the systematic review. The reference lists of the included papers were also assessed to identify additional eligible papers. A second reviewer verified the initial potential eligibility of papers by reviewing 20% of the full list of the titles and abstracts originally obtained from the search keywords in the Scopus database. This systematic review follows the guidelines according to Preferred Reposting Items for Systematic Reviews and Meta-Analysis (PRISMA) [12]. A meta-analysis was not assessed due to the nature of the results of this review.

Results

Potential papers were identified by searching two databases, Scopus and Ovid, using the following search terms: ["methodology" or "simulations" or "bootstrap" or "bootstrapped" or "model comparison" or "compare models" or "markov" or "statistical model"] and ["cost data" or "claims data" or "billing data" or "health insurance data" or "health insurance claims" or "billing claims" or "administrative data"] and ["cost analysis" or "health care costs" or "hospital costs"] in the title, abstract, or keywords. Searches were conducted for all articles published up to the end of February 2018.

Only statistical method papers using multivariable modelling of cost, with or without methods controlling for selection bias, were considered eligible given they met one of the following inclusion criteria: 1) a comparison of two or more statistical methods to analyze cost or 2) one statistical method performed on two or more different types of cost data. Since administrative cost data can be very different from cost data collected in a randomized controlled trial, papers that used only data from randomized controlled trials and not administrative data were excluded. Additional exclusions included not being published in English and review papers. Bayesian methods and joint modelling methods were considered beyond the scope of this review. Data elements extracted from eligible papers include: paper's first author, year of publication, statistical methods being evaluated, data types used in analysis, year(s) the data was collected, simulation approach used, sample size of data, distribution of the data, the results of the method performance and the author's recommendations based on their results.

The titles and abstracts of all papers found utilizing the search terms in both databases were reviewed for potential eligibility. Among those titles and abstracts meeting potential eligibility, the full text article was reviewed for eligibility to be included in the systematic review. The reference lists of the included papers were also assessed to identify additional eligible papers. A second reviewer verified the initial potential eligibility of papers by reviewing 20% of the full list of the titles and abstracts originally obtained from the search keywords in the Scopus database. This systematic review follows the guidelines according to Preferred Reposting Items for Systematic Reviews and Meta-Analysis (PRISMA) [12]. A meta-analysis was not assessed due to the nature of the results of this review.

Discussion

The systematic review aimed to identify what methods of cost analysis are statistically and mathematically appropriate to use with large claims data. While most papers used established methods, there were three papers that introduced new methods. Kurtz et al. found their suggested Tweedie method could be an alternative approach to analyze cost with a small number of zero costs, though, existing methods still performed superior. Garrido et al. compared only novel models to each other, further research will be needed to compare these to existing accepted methods.

The most commonly assessed models were OLS and Gamma distribution models. As seen in Table 2, almost all the models that made comparison with the GLM Gamma distribution with log link found this model to be the superior model. Kuwornu et al., who compared eight different models including OLS and Gamma distributions, stated that the OLS Log model with normal retransformation had superior performance based on the highest average R^2 over 10-fold cross-validation; however, while the average R^2 was highest at 18.77, the values varied from 6.73 to 24.76 and were not consistent. Both the GLM Gamma with identity and log links had consistent R^2 values over the 10 replications with average values of 17.04 and 14.52, respectively. The same inconsistency in values can be seen with the recommendations based on lowest RMSE values for the OLS model of untransformed costs. Again, the Gamma with identity link performed just as well as the OLS model. OLS model precision was found to approach that of the GLM Gamma models as sample size increased in simulations conducted by Malehi et al.

SIMULATED VS. REAL DATA

All but one paper used real data to demonstrate the performance of their models. While simulations are a useful tool to assess the accuracy of model performance, the use of only simulated data can lead to biased results as simulations can be structured to fit a particular statistical model for superior performance. Assessment of models in real data is necessary to evaluate performance in uncontrolled environments. Some papers provided incomplete rationales for decisions regarding data, such as excluding zero costs, which may have biased the model assessment.

LIMITATIONS

To focus the scope of this review on the most common use of administrative cost data, methods assessed in randomized controlled trial data only, those that evaluated the effect of missing data, and those that assessed joint modeling methods were excluded. This focus limited the evaluation of numerous methods in various data distributions and, thus, limited the number of model comparisons.

Conclusions

The GLM Gamma distribution with Log link performed most consistently as the superior model in comparisons using both simulated data and real administrative data. The literature review suggests that this is the most appropriate model to use with administrative data. Some caution is suggested when dealing with heteroscedastic data or data with high proportion of zero costs (or non-users). The Tweedie distribution is an emerging new method that may be useful in future research.

Abbreviations

PRISMA: Preferred Reposting Items for Systematic Reviews and Meta-Analysis; OLS: ordinary least squares; GLM: generalized linear model; NLR: normal linear regression; RMSE: root mean square error; MAE: mean absolute error; EEE: extended estimating equations; 2SLS: 2-stage least squares, FIMSL: full information maximum simulated likelihood, LATE: local average treatment effect, ATET: average treatment effect on the treated, GAMLSS: generalized additive models for location, scale, and shape, MAPE: mean absolute predication error, MSE: mean square error, FE: fixed effects.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and material

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Funding

Not applicable.

Authors' contributions

MD conducted and drafted the systematic review under guidance of HB and KS. HB advised the design of the systematic review and evaluated the adherence to PRISMA guidelines. KS advised intellectual content, design of the systematic review, and acted as second reviewer to verify the initial potential eligibility of papers of 20% of the abstracts. HB and KS provided critical review and comments of the manuscript. All authors read and approved the final manuscript.

Acknowledgements

Not applicable.

Authors' information

MD is a Research Associate in the College of Nursing and current doctoral student in the College of Health Professions, under the supervision of KS, at the Medical University of South Carolina. KS is a Professor and Department of Healthcare Leadership & Management Director in the College of Health Professions at the Medical University of South Carolina. HB is an Associate Professor and Co-Director of Health and Rehabilitation Sciences Doctoral Program in the College of Health Professions and Medical Director of Speech-Language Pathology in the Evelyn Trammell Institute for Voice and Swallowing at the Medical University of South Carolina.

References

1. Power, E; Eisenberg, J. Are we ready to use cost-effectiveness analysis in health care decision-making? A health services research challenge for clinicians, patients, health care systems, and public policy. *Medical Care*. 1998, 36(5): MS10-MS17.
2. Fukuda, H; Ikeda, S; Shiroiwa, T; Fukuda, T. The effects of diagnostic definitions in claims data on healthcare cost estimates: evidence from a large-scale panel data analysis of diabetes care in Japan. *PharmacoEconomics*. 2016, 34: 1005-1014. DOI 10-1007/is40273-016-0402-3.
3. Dodd, S; Bassi, A; Bodger, K; Williamson, P. A comparison of multivariable regression models to analyse cost data. *Journal of Evaluation in Clinical Practice*. 2006, 12(1): 76-86.
4. Basu, A; Arondekar, B; Rathouz, P. Scale of interest versus scale of estimation: Comparing alternative estimators for the incremental costs of a comorbidity. *Health Economics*. 2006, 15: 1091-1107. DOI 10-1002/hec.1099
5. Garrido, M; Deb, P; Burgess, J, Penrod, J. Choosing models for health care costs analyses: Issues of nonlinearity and endogeneity. *Health Services Research*. 2012, 47(6): 2377-2397. DOI 10-1111/j.1475-6773.2012.01414.x
6. Bohl, A; Blough, D; Fishman, P; Harris, J; Phelan, E. Are generalized additive models for location, scale, and shape an improvement on existing models for estimating skewed and heteroskedastic cost data? *Health Services Outcomes Res Method*. 2013, 13: 18-38. DOI 10-1007/s10742-012-0086-x
7. Akbarzadeh, A; Pourhoseingholi, A; Zayeri, F; Ashtari, S. Zero inflated statistical count models for analyzing the costs imposed by GERD and dyspepsia. *Arab Journal of Gastroenterology*. 2013, 14: 165-168.
8. Kuwornu,J; Lix, L; Quail, J; Wang, E; Osman, M; Teare, G. A comparison of statistical model for analyzing episodes-of-care costs for chronic obstructive pulmonary disease exacerbations. *Health Services Outcomes Res Method*. 2013, 13: 203-218. DOI 10-1007/s10742-013-0112-7
9. Malehi, A; Pourmotahari, F; Angali, K. Statistical models for the analysis of skewed healthcare cost data: a simulation study. *Health Economics Review*. 2015, 5(11): 1-16. DOI 10-1186/s13561-015-0045-7
10. Fiebig, D; Johar, M. Forecasting with micro panels: the case of health care costs. *Journal of Forecasting*. 2017, 36: 1-15. DOI 10-1002/for.2409
11. Kurz, C. Tweedie distributions for fitting semicontinuous health care utilization cost data. *Medical Research Methodology*. 2017, 17(171): 1-8. DOI 10-1186/s12874-017-0445-y
12. Liberati, A; Altman, D; Tetzlaff, J; Mulrow, C; Gotzsche, P; Ioannidis, J; Clarke, M; Devereaux, P; Kleijnen, J; Moher, D. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. PLoS Med. 2009, 6(7): e10000100. DOI 10.1371/journal.pmed.1000100

Tables

Table 1. Summary characteristics of methods, data, and results of reviewed papers.

First Author	Year of Publication	Methods being evaluated	Data Types	Year(s) data collected	Sample size of Data	Distribution of Data	Simulation approach used	Results of simulation	Author's recommendation
Dodd	2006	NLR NLR Log transformed Bootstrapped NLR Median regression GLM Gamma with Log link	Hospital Administrative Data	2000	N=426	Positive skewed	None	None	GLM Gamma with Log Link
Basu	2006	OLS OLS Log Transformation (D retransformation) GLM Gamma with Log Link EEE	Medstat's MarketScan	1997-2001	N=7428	Positive skewed, no zero cost	None	None	EEE
Garrido	2012	2-Stage Least Squares 2-Stage Least Squares Log Transformed GLM Gamma with Log Link Full Information Maximum Simulated Likelihood (FIMSL)	Veteran's Health Administration	2005-2006	N=3389	Positive skewed	None	None	GLM Gamma with Log Link Approach or FIMSL
Bohl	2013	OLS Log transformed (Normal retransformation) GLM Gamma with Log link GAMLSS	Medicare Advantage	Not Mentioned	N=2000	Positive skewed, no zero cost	Gamma and Generalized inverse gaussian	GLM Gamma with Log Link	GLM Gamma with Log Link
Akbarzadeh	2013	Poisson Regression Negative Binomial Zero-inflated Poisson Zero-inflated Negative Binomial	Survey	2006-2007	N=1929	Not mentioned	None	None	Zero-inflated Negative Binomial
Kuwornu	2013	OLS OLS Log transformed (Normal retransformation) OLS Log transformed (heteroscedastic retransformation) OLS Log transformed (Duan's retransformation) Robust regression	Administrative Data of Saskatchewan, Canada	1996-2010	N=17,480	Positive skewed	None	None	OLS OLS Log transformed (Normal retransformation) Robust regression

		GLM Poisson with Log link GLM Gamma with Identity link GLM Gamma with Log link							
Malehi	2015	OLS Log transformed (retransformation not mentioned) GLM Weibull with Log link GLM Gamma with Log link Cox Proportional Hazard	None	None	None	None	Monte Carlo simulated with varying levels of skewness using log-normal, Gamma, and Weibull distributions (no zeros)	GLM models performed better. Weibull performed better than Gamma with higher skewness. As sample size increased, OLS model precision approached GLM precision.	GLM Gamma was superior overall Cox was superior if Weibull distribution
Fiebig	2017	OLS Fixed Effects	Administrative Data of New South Wales	2006-09	N=264,024	Right skewed < 3% zero cost	Monte Carlo simulated with varying correlation between explanatory variable and unobserved time-invariant effects	Uncorrelated: FEP is superior except where there are no unobserved time-invariant effects Correlated: FEP is superior in Post-sample forecast OLSP is superior in out-of-sample forecast	OLSP was superior
Kurtz	2017	Tweedie Tobit Poisson 2-part model: Gamma with Log link Generalized Gamma with Log link	RAND HIE	Not provided	N=3301	Right skewed 18.1% zero cost	Monte Carlo Simulated with varying proportions of non-users/zero cost	<20% zero observations: Tweedie >20% zero observations: 2-part models Gamma was better than Gen Gamma	Tweedie is an alternative to the superior 2-part models

Table 2. Summary Models evaluated and preferred recommendations by author of reviewed paper (denoted as **X***).

MODEL	Dodd 2006	Basu 2006	Garrido 2012	Bohl 2013	Akbarzadeh 2013	Kuwornu 2013	Malehi 2015	Fiebig 2017	Kurtz 2017
NLR	X								
Log NLR	X								
Bootstrap NLR	X								
OLS		X				**X***		**X***	
OLS Log (Heteroscedastic)		X				X			
OLS Log (Normal)				X		**X***			
OLS Log (Duan)						X			
GLM Gamma Log link	**X***	X	**X***	**X***		X	**X***		**X***
GLM Gamma Identity link						X			
GLM Generalize Gamma									**X***
EEE		**X***							
Median Regression	X								
Robust Regression						**X***			
2-Stage Least Squares			X						
2-Stage Least Squares Log			X						
FIMSL			**X***						
GAMLSS				X					
Poisson					X				X
Poisson Log Link						X			
Zero-inflated Poisson					X				
Negative Binomial					X				
Zero-inflated Negative Binomial					X				
Weibull Log							X		
Cox Proportional Hazard							X		
Tweedie									X
Tobit									X
Fixed Effects								X	
Data Type	*Real*	*Real*	*Real*	*Real/ Simulated*	*Real*	*Real*	*Simulated*	*Real/ Simulated*	*Real/ Simulated*

Figures

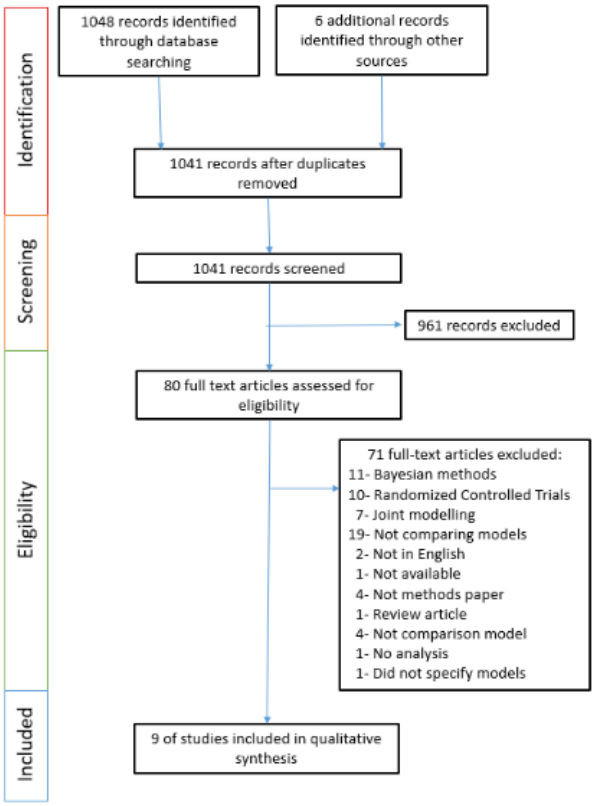

Figure 1

PRISMA flow chart.