Targeting high-cost beneficiaries in the medium-term with predictive modelling

By D Shapiro, B Childs and C Getz

Presented at the Actuarial Society of South Africa's 2013 Convention
31 October–1 November 2013, Sandton Convention Centre

ABSTRACT

Individuals with high healthcare cost make up a small section of the population, but may be responsible for a large portion of all costs. Identifying high-cost individuals for care management can be effective in managing costs and improving health outcomes. However, mean reversion of costs means that current high-cost beneficiaries should not all necessarily be the focus of care management. The purpose of this research is to describe the development and evaluation of a model in predicting high-cost beneficiaries in the South African medical scheme environment. This research extends existing research to consider expenditure over the medium-term. The research revealed that high-cost beneficiaries can be predicted with decreasing accuracy over longer time horizons. Notably, beneficiaries who were previously low-cost and transition to become high-cost can be predicted with some degree of accuracy. The extension of cost measurement from individual years to aggregated costs of multiple years increased the predictive accuracy of models markedly. This suggests a longer-term view to care management may allow for better targeted care management and hence better outcomes.

KEYWORDS

Care management; predictive modelling; medium-term; medical scheme

CONTACT DETAILS

Daniel Shapiro, CareGauge (Pty) Ltd. 400 16th Avenue, Midrand
Tel: +27(0)11 540 0923; E-mail: daniels@caregauge.net

1. INTRODUCTION

1.1 Care management seeks to provide the most appropriate care for patients in order to manage costs and improve health outcomes. Resources and personnel with which to implement care management programmes are limited and there is a need to identify interventions that represent the greatest opportunity to achieve beneficial outcomes. High-cost beneficiaries make up a small section of the population, but are typically responsible for a large proportion of aggregate costs. Treatment of high-cost patients presents the greatest opportunity to efficiently direct resources and maximise financial outcomes (Lynch et al., 2000). Predicting patients' healthcare utilisations can be used to identify candidates for enrolment in care management programmes and in developing and evaluating care management programmes. Predictions can be used together with economic assumptions to analyse the economic value of care management programmes (Cousins & Liu, 2003; Duncan, 2011).

1.2 Efficient management and delivery of healthcare may be needed in the South African private healthcare sector where cost escalation pressures are significant. In the South African private healthcare sector, medical schemes constitute the predominant type of healthcare plan. Medical schemes are governed by Medical Schemes Act,[1] which prescribes that schemes operate on the basis of social solidarity. Membership is open to any individual and contributions are community rated such that premiums may not be differentiated based on demographic or risk rating factors. Legislation thus prohibits the use of prediction for underwriting and premium rating on an individual level. However, given the risks posed to medical schemes by the inability to exclude or underwrite high-risk beneficiaries or charge higher contributions for higher risks, there may a heightened need for care management. Furthermore, the Medical Schemes Act dictates that schemes must provide benefits for members that are underpinned by a set of minimum benefits, which must be paid for at full costs,[2] which entails an open ended liability to medical schemes.

1.3 Care management programmes may also be seen as important for South Africa's public sector. South Africa has a 'quadruple burden of disease' characterised by a combination of prevalent conditions that result in high levels of morbidity and costly care. Furthermore, South Africa's financial and human resource constraints in healthcare delivery require the allocation of scarce healthcare resources. Current efforts to revitalise the South African public healthcare system may consider care management of high-risk beneficiaries as a means to achieve this.

1 Medical Schemes Act, 1998, Act no. 131 of 1998 as amended, Republic of South Africa

2 Full costs refers to the amount billed the provider. See Judgement in the matter between Board of Healthcare Funders and South African Municipal Workers Medical Scheme (versus the Council for Medical Schemes and eleven others, 2011

1.4 Research in identifying and predicting costs of high-cost individuals is typically based on a single year projection. This is in line with schemes' typical budgetary time horizons. The nature of certain diseases is that health deteriorates, and high costs may come at the end of a long and gradual build-up. Thus, care management may need to consider individuals who pose risks of high medium- or long-term costs. Furthermore, care management may be aimed at long-term financial sustainability. In the private sector restricted schemes have an interest in keeping closed membership pools healthy and sustainable in the medium- and long- terms. In the public sector long-term sustainability of financing is a key goal. To the authors' knowledge, current literature does not subsume predicting high-cost individuals over the medium- and long-terms.

1.5 The purpose of this research is to describe the development, validation and evaluation of models that identify future high-cost beneficiaries in the South African medical scheme environment. In particular, this research aims to extend the analysis to the medium- and long-terms.

1.6 Before proceeding to the analysis, Section 2 reviews the literature of predictive modelling in the context of identifying and predicting costs of high-cost individuals, as well as prediction of high-cost individuals over varying time horizons. Section 3 describes the data used for the research and their preparation. The methodology used in this research is described in Section 4, which includes a review of the modelling approach that was used. The results of the analysis are set out in Section 5. Section 6 discusses the findings of the research and their implications.

2. LITERATURE REVIEW

2.1 Predictive Modelling of High-Cost Individuals

2.1.1 Cousins et al. (2002) outline predictive modelling in the context of stratification of individuals in terms of health costs. They describe the process of constructing models for risk stratification in terms of a risk score and introduce sensitivity and specificity as measures of correctly identifying high risk and non-high risk members.

2.1.2 Dove et al. (2003) developed and validated a predictive model that identified future high-cost managed care programme members from a population of members who were initially low-cost. They ranked patients according to their probabilities of incurring costs above a predetermined threshold and found that the predictive accuracy of their model was significantly better than randomly assigning patients for care management. They found that few patients were consistently high-cost across the years of the study and that a large number of high-cost members in the prediction year were low-cost in the year prior to the prediction year.

2.1.3 The model of Dove et al. (2003) was based on a statistical methodology. Clinical grouper-based models have also been used to predict healthcare utilisation. Meenan

et al. (2003) examined the ability of a range of grouper models to identify high-cost individuals. They evaluated models on the sensitivity and specificity of identifying high-cost individuals and found certain grouper models to be more successful in identifying high-cost beneficiaries. Zhao et al. (2005) examined the ability of various grouper models to identify individuals with high future costs. The models were validated on data from the year following calibration. High-cost beneficiaries were evaluated as individuals whose actual costs were above a set threshold. Winkelman and Mehmud (2007) analysed a range of grouper models and their abilities to predict costs. They included an analysis of the predictive ability of the models at a range of different risk score levels and found all grouper models to significantly under-predict the actual costs of high-risk individuals. Winkelman and Mehmud (2007) further analysed the predictive performance for members who had costs lower than the median cost of all individuals. They found that costs were over-predicted for these beneficiaries, indicating high variability of costs at high levels.

2.2 Predictive Modelling over Different Time Horizons

Zhao et al. (2005) and Winkelman and Mehmud (2007) performed analyses in which costs were forecast using concurrent models, in which costs were in the same year as the explanatory diagnosis and demographic data, and prospective models, in which costs were in the year following the diagnosis and demographic data. Both found that concurrent models performed significantly better than prospective models. This may indicate that analyses using more immediate costs are likely to be more accurate than models of longer term costs.

3. DATA

3.1 This research was performed using beneficiary data from a South African medical scheme administrator. The dataset consisted of 100,077 beneficiaries who were continuously enrolled from 2007 to 2012. The variables contained in the dataset are elaborated upon in Section 3.2 and are listed in Appendix A.

3.2 The predicted variable was the amount paid by the scheme per life per year (PLPY) in Rands for each beneficiary. In- and out-hospital claims were recorded separately to provide response variables for out-hospital and in-hospital expenditures. Out-hospital visits were defined as the use of the same provider on the same day, and in-hospital visits were recorded as unique hospital admissions. Trauma and neonatal expenditure were excluded from costs, given the highly unpredictable nature of these costs. Data trimming was not undertaken as this would exclude the very individuals that the models attempt to identify. Rosenberg and Farrell (2008) found that for healthcare cost distributions with long tails, high costs form part of the tail of the distribution and should not be interpreted as outliers.

3.3 Independent variables were derived from 2007 data. Age and gender were included as demographic explanatory variables. Province was used as a proxy for access to healthcare services. The choice of plan within the scheme was included, as the choice of plan may exogenously determine healthcare utilisation due to the different levels of benefits offered, and may also directly affect beneficiary health-seeking behaviour. Benefit options of beneficiaries were taken at the beginning of 2008. The benefit option selected by the member, as well as the benefits within each option, may have changed during the course of the five-year prediction horizon. Potential changes in plan were not taken into account in the analysis. However, the effect of incorrectly predicting high-cost beneficiaries based solely on plan design is likely to be minor (Dove et al, 2003).

3.4 Beneficiary conditions were included as binary variables indicating the presence or absence of conditions for each beneficiary in the 2007 year. The CCS Grouper[3] was used to collate clinical information into 64 categories for each beneficiary. The 64 categories are listed in Appendix A. Clinical information was derived from hospital authorisation diagnoses, diagnoses arising with respect to approved chronic medication and ICD-10 diagnoses claimed for by all remaining provider types. The clinical identification algorithm employed required one out-hospital event and one hospital or chronic medicine authorisation in order to indicate the presence of a condition. This algorithm was tested by increasing the number of diagnoses events required. The differences in results compared to the analysis reported on in this paper were negligible.

3.5 The variables used in this research were limited by the extent of information recorded by medical scheme administrators and do not constitute a comprehensive list of variables that could be used in modelling healthcare utilisation. Mehmud (2013) showed that adding non-traditional variables in statistical models of healthcare costs can have significant effects on the amounts transferred under risk equalisation formulae. Notably, income, which is not included in this research, is suggested as a significant variable by economic theory (Grossman, 1972).

4. METHODOLOGY

4.1 Definition of High-Cost

This research required a definition of 'high-costs' with which to evaluate high-cost beneficiaries. In order to simplify the analysis a binary definition of low- and high-cost beneficiaries was adopted. Annual expenditure above $2000 per year can be considered to be high-cost (Dove et al., 2003). This amount, in 2008 Rand terms, approximately corresponds to the 90th percentile of 2008 costs. Therefore the 90th percentile of costs

3 Clinical Classifications Software (CCS) was developed at the Agency for Healthcare Research and Quality (AHRQ). It is open source and is available at www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp

was taken to be the high-cost threshold. This level is tested for sensitivity in section 5.2.4. The high-cost threshold was set to be a percentile of the cost distribution rather than an absolute amount as annual tariff increases mean that costs at different terms will be dissimilar with respect to a set absolute threshold, whereas the 90% threshold can be used consistently across time horizons.

4.2 Analysing High-Cost Beneficiaries

Dove et al. (2003) and Duncan (2011) underscore the presence of mean reversion in beneficiary costs, in which beneficiaries who are high-cost in a given year may not remain high-cost beneficiaries in a subsequent year, while beneficiaries who are low-cost in a given year may become the high-cost beneficiaries in the subsequent year. The phenomenon was evident in the dataset used in this research. Of the beneficiaries classified as high-cost in 2007, only 40.1% were classified as high cost in 2008, 34.3% in 2010 and 31.2% in 2012. The proportion of total costs attributable to high-cost beneficiaries from 2007 was 54.5%, compared to just 31.0% for the same beneficiaries in 2008, 27.5% in 2010 and 26.3% in 2012. Consequently, the majority of high-cost beneficiaries and a large proportion of costs in years post-2007 come from beneficiaries who were classified as low-cost in 2007. This research thus aimed to identify the beneficiaries that transition to become high-cost beneficiaries ahead of the time by predicting costs based on identifiable characteristics of future high-cost beneficiaries.

4.3 Modelling Approach

4.3.1 The literature review suggested statistical modelling and grouper-based models as the two main methods for predictive modelling of healthcare costs. Statistical models present the advantages of being easily available, as opposed to groupers which are, in general, proprietary. Grouper models give the advantages of easier implementation, clearer understanding due to the interpretability of risk groups and stability with respect to analysing different populations. However, the models are based on the definitions of the developer and there may be lack of transparency. This research employed statistical modelling, based on its availability for the purpose of the research.

4.3.2 Descriptive statistics of the dataset showed that in- and out-hospital median costs were greater than mean costs. Furthermore, costs at extreme percentiles increased substantially, particularly for in-hospital costs. This indicated right-skewed and heavy-tailed distributions. Approximately 83% of beneficiaries did not incur in-hospital costs in 2008. The proportion for out-hospital costs was considerably less at 7%. Based on the different distributions for in- and out-hospital costs, the two sets of costs were modelled separately and combined to give the total cost.

4.3.3 The choice of statistical model should be able to capture the skewed, long-tailed distribution of costs and the point mass of costs at zero. Basu et al. (2004) describe

alternative statistical modelling approaches for healthcare expenditure data, including Ordinary Least Squares regression (OLS), transformations to OLS and Generalised Linear Model (GLM) variants. Buntin and Zaslavsky (2004) similarly compare the performance of eight alternative models, including OLS and GLMs in predicting Medicare costs. They found that four of the alternative models produced very similar results. The two-part model was proposed for modelling healthcare expenditure data by Duan et al. (1983) and Manning et al. (1987). The model decomposes costs into two parts: one part indicating whether or not a claim has occurred, and one part for severity, indicating the amount of the expenditure given that a claim has occurred. Frees, et al. (2011) extended the two part model by modelling the frequency of visits as the first part of the model and the expenditure given the number of visits as the second. The second part of the model included a variable representing the number of events to predict the expenditure per event, capturing dependencies between the first and second parts. The model reflects the characteristics from each visit rather than rather than only modelling use or non-use during a period. Frees et al. (2011) found that the approach provided better point predictions of expenditure than competing models and provided a plausible distribution of costs where there were many zeroes.

4.3.4 Given the high number of beneficiaries with zero a two-part model was used in which the frequency of utilisation and cost per utilisation are modelled.

4.3.4.1 The frequency of out-hospital visits and hospital admissions required a discrete response distribution that is able to accept observations of zero. A Generalised Linear Model (GLM) with Negative Binomial response distribution was used, similarly to Frees et al. (2011), who used the model for the number of in- and out-hospital utilisations.

4.3.4.2 Ordinary Least Squares was considered for the model of the cost per utilisation. Costs were log-transformed in order to remove skewness and heteroscedasticity in the cost data. Predictions of transformed costs produce results on the transformed scale and require a retransformation to the original cost scale. This may introduce bias and creates the need for a bias adjustment for estimating costs (Manning, 1998; Manning & Mullahy; 2001; Mullahy, 1998). If the distribution of the transformed data is normally distributed, predictions can be derived using a log-normal retransformation. If the error term is not normally distributed, this prediction may be biased. The Duan smearing estimator (Duan, 1983) is a non-parametric adjustment that can provide a better-fitting retransformation adjustment. However, heteroscedasticity in the data may result in the Duan adjustment producing biased results (Manning,1998, Manning & Mullahy, 2001). This research fitted models to log-transformed costs using both the log-normal and Duan adjustments. However, both adjustments produced results that were heavily biased.

4.3.4.3 Generalised Linear Models (GLMs) have frequently been used to predict healthcare expenditure (Blough et al., 1999; Buntin & Zaslavsky, 2004; Manning et al., 2005). Studies using GLMs for healthcare costs have commonly used a gamma distribution as the response distribution (Diehr et al., 1999; Blough et al., 1999; Manning et al., 2005). This specification of the GLM is further able to capture heteroscedasticity in the data. Generalised Linear Models with Gamma response distribution and log-link were used to model the cost per admission and cost per out-hospital visit. The frequency of utilisations is included as a variable in the cost model, which allows predicted cost to be influenced by whether utilisation was concentrated or spread over time.

4.3.5 Models were fitted using the predictor variables from 2007 to predict in-hospital and out-hospital costs. Beneficiaries were randomly assigned into equally-sized training and validation sub-samples in order to avoid over fitting. The dataset was set up as a panel dataset and all variables, including time, were included as fixed effects in the models. Predictions of the in-hospital and out-hospital models were added to arrive at total predicted costs. Risks scores were then assigned to every beneficiary by ranking beneficiaries from high to low based on costs.

5. RESULTS

Costs were measured for three time horizons: one year, three years and five years. These correspond to costs from 2008, 2010 and 2012, and were termed 'short term', 'medium term' and 'long term' respectively. In addition to predicting costs for individual years, predictions were made for costs aggregated over multiple years. For example, 'medium-term aggregated' costs included all costs from 2008, 2009 and 2010, and 'long-term aggregated costs' included costs from 2008, 2009, 2010, 2011 and 2012. Thus, five separate analyses were performed:

— Short-term
— Medium-term
— Long-term
— Medium-term aggregated
— Long-term aggregated

5.1 Model Fit

5.1.1 The R-squared is popularly used as a summary measure of overall model fit for healthcare cost data. However, the R-squared tends to be sensitive to prediction error for beneficiaries with very high costs (Cousins et al., 2002) and should only be calculated after truncating large medical expenses (Cumming et al, 2002). The Mean Absolute Percentage Error (MAPE) is an alternative measure that is less influenced by large costs. The MAPE is defined as $\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\frac{|\hat{y}_i - y_i|}{y_i}$, where $\hat{y}_i$ denotes the

predicted cost for beneficiary i in the validation dataset, y_i the actual cost and n the number of beneficiaries.

5.1.2 Both the R-squared and MAPE were evaluated. A greater R-squared and a lower MAPE indicate a better model fit. The results of the model fits are given in Table 1. The R-squared decreases and the MAPE increases for models with costs from longer time horizons. This indicates less successful prediction of costs for longer time horizons.

5.1.3 The R-squared and MAPE for model of aggregated costs are significantly better than for single years, indicating relatively good model fit for costs over multiple years. The model fit increases most substantially from short- to medium-term.

Table 1. R-squared and MAPE of fitted models

	R-squared	MAPE
Short-term	11.7%	0.850
Medium-term	9.6%	0.869
Long-term	9.1%	0.909
Medium-term aggregated	18.2%	0.701
Long-term aggregated	22.9%	0.653

5.2 Predicting High-Cost Beneficiaries

5.2.1 Accuracy in predicting high-cost beneficiaries was evaluated using the proportion of beneficiaries who were correctly predicted to be high-cost. The accuracies of predictive models were considered for the population of all beneficiaries, as well as for the subsets of beneficiaries who were predicted to be high-cost that arose from beneficiaries who were low- and high-cost in 2007. The purpose of splitting beneficiaries into previously low- and high-cost categories was to split the analysis into subsets of beneficiaries that transition from low-cost to high-cost and beneficiaries who are repeatedly high-cost.

5.2.2 The proportion of all high-cost beneficiaries that were correctly identified was 40.1% for short-term, and decreased for medium- and long-term. The decrease was driven by the decrease in accuracy of predicting repeat high-cost beneficiaries. The proportion of beneficiaries who transition from low- to high-cost that are correctly identified did not decline for longer terms. The proportions of correctly identified previously low-cost beneficiaries are lower than the proportions of repeat high-cost beneficiaries, but are significantly greater than randomly selecting low-cost beneficiaries.

5.2.3 High-cost beneficiaries were predicted more accurately for aggregated costs. In particular, the accuracy pertaining to previously low-cost beneficiaries shows a large

proportionate increase, which is felt most keenly when increasing the time horizon from short- to medium-term.

TABLE 2 Proportion of correctly identified high-cost beneficiaries

	2007 Low-cost beneficiaries	2007 High-cost beneficiaries	All beneficiaries
Short-term	28.1%	53.4%	40.1%
Medium-term	27.2%	48.6%	37.3%
Long-term	28.3%	43.3%	35.1%
Medium-term aggregated	33.1%	60.0%	46.0%
Long-term aggregated	36.3%	60.1%	47.8%

5.2.4 The sensitivity of the choice of 90% high-cost threshold was tested by repeating the analysis at alternative levels. Figure 1 presents results with thresholds defined at the 85th and 95th percentiles of costs. The definition of high-cost has a substantial effect on the proportion of high-cost beneficiaries that are correctly identified. Lower thresholds produce higher accuracy and higher thresholds lower accuracy. However, the patterns of relative accuracy for the models of different time-horizons remain the same as for the 90% threshold and thus may hold in general for all high-cost thresholds.

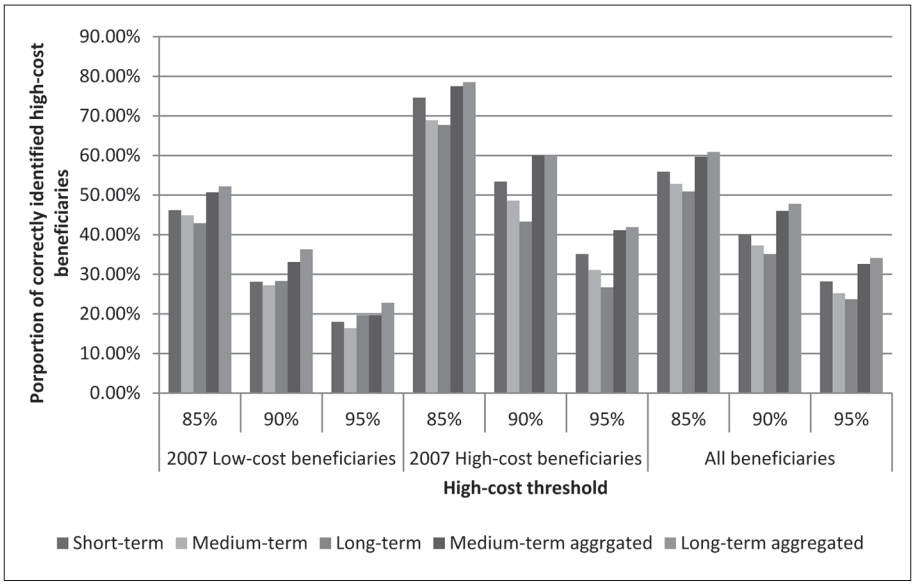

FIGURE 1 Correctly identified high-cost beneficiaries at different high-cost thresholds

5.3 Characteristics of Predicted High-Cost Beneficiaries

This section compares the characteristics of predicted high-cost beneficiaries to the characteristics of predicted low-cost beneficiaries. Results are compared by time-horizon as well as by whether beneficiaries were low- or high- cost in 2007. Analyses of beneficiary characteristics using aggregated costs did not show large differences from analyses using costs from single years and hence are not shown.

5.3.1 Gender

Table 5 shows the prevalence of females amongst subsets of beneficiaries. The proportion of females in the full dataset is 56.6%. Thus, high-cost beneficiaries are over-represented by females. The proportion of beneficiaries who transition from low- to high-cost is predominantly female, which increases further for longer-term time horizons. The proportion of repeat high-cost beneficiaries is predominantly female, but the extent declines for longer-term time horizons.

Table 3 Prevalence of females among beneficiary subsets

	Short term		Medium term		Long term	
	Low-cost	High-cost	Low-cost	High-cost	Low-cost	High-cost
2007 Low cost	56.0%	58.4%	56.1%	57.3%	55.9%	60.9%
2007 High cost	59.6%	62.0%	61.8%	59.3%	63.2%	57.6%

5.3.2 Age

The age distribution of beneficiaries shows a prevalence of high-cost beneficiaries between ages 55 and 75. The age distribution does not vary with time horizon and is not influenced by whether beneficiaries were previously high- or low- cost. Low-cost beneficiaries have a younger predominance, concentrated around ages 0 to 40 for previously low-costs beneficiaries and 30 to 55 for previously high-cost beneficiaries.

5.3.3 Disease Prevalence

5.3.3.1 Table 4 contains the prevalence of the five most predominant acute and chronic diseases for beneficiaries that were low-cost in 2007. The prevalence is shown separately for beneficiaries predicted to be low- and high- cost respectively. All diseases show greater prevalence among beneficiaries predicted to be high-cost, compared to predicted low-cost beneficiaries. Acute bronchitis and disorders of the teeth and decline in prevalence over long time horizons among the subsets of high-risk beneficiaries, while eye disorders, skin disorders and disorders of the central nervous system have no discernible change in prevalence.

5.3.3.2 Chronic conditions show a significantly greater prevalence among beneficiaries predicted to be high-cost compared to low-cost. In particular, the prevalence of hypertension, hyperlipidaemia, heart conditions and diabetes are all in excess of seven

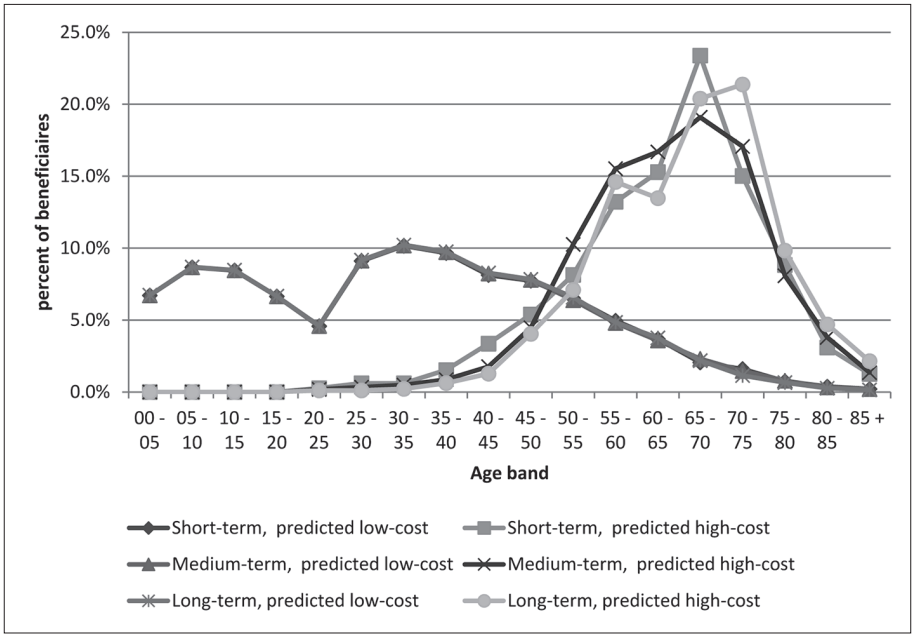

FIGURE 2 Age distributions of 2007 low-cost beneficiaries

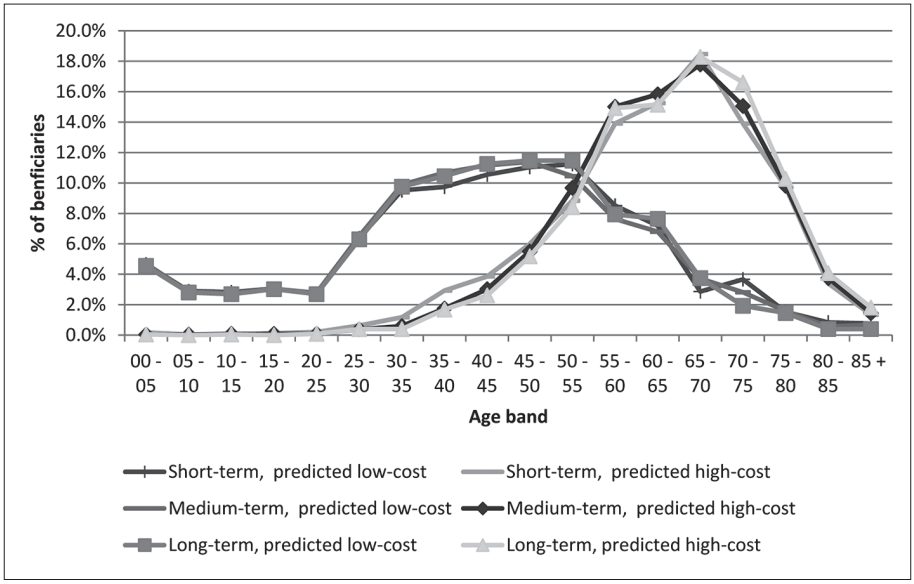

FIGURE 3 Age distributions of beneficiaries of 2007 high-cost beneficiaries

Table 4 Disease prevalence diseases among beneficiaries who were low-cost in 2007

	Short-term		Medium-term		Long-term	
	Low-cost	High-cost	Low-cost	High-cost	Low-cost	High-cost
Acute diseases						
Acute bronchitis and URI	37.0%	48.1%	37.2%	44.8%	37.4%	42.5%
Disorders of teeth and jaws	28.6%	47.0%	28.6%	46.3%	28.7%	44.7%
Other eye disorders	19.5%	39.1%	19.3%	41.5%	19.4%	39.2%
Skin disorders	12.4%	28.8%	12.4%	28.4%	12.4%	28.4%
Other central nervous system disorders	6.7%	21.8%	6.7%	22.6%	16.5%	31.0%
Chronic diseases						
Hypertension	8.4%	60.2%	8.4%	58.3%	8.1%	62.6%
COPD, asthma	15.2%	32.1%	15.4%	27.5%	15.4%	27.3%
Hyperlipidaemia	3.5%	32.6%	3.8%	26.2%	4.0%	23.1%
Heart conditions	3.6%	26.5%	3.6%	26.0%	3.6%	24.4%
Diabetes mellitus	1.7%	21.5%	1.4%	25.1%	1.5%	22.7%

Table 5 Disease prevalence among beneficiaries who were high-cost in 2007

	Short-term		Medium-term		Long-term	
	Low-cost	High-cost	Low-cost	High-cost	Low-cost	High-cost
Acute diseases						
Acute bronchitis and URI	43.3%	48.3%	45.2%	46.3%	45.7%	45.8%
Disorders of teeth and jaws	43.9%	53.0%	43.6%	53.5%	44.5%	52.9%
Other eye disorders	33.4%	39.5%	31.7%	41.5%	33.2%	40.0%
Skin disorders	19.6%	32.5%	20.1%	32.2%	20.4%	32.3%
Other central nervous system disorders	16.5%	31.0%	17.1%	30.5%	16.9%	31.2%
Chronic diseases						
Hypertension	18.7%	58.3%	18.8%	58.9%	18.8%	60.5%
COPD, asthma	23.9%	38.1%	25.6%	36.3%	25.6%	36.8%
Hyperlipidaemia	8.5%	29.6%	10.5%	27.7%	12.2%	26.4%
Heart conditions	12.5%	37.6%	13.1%	37.4%	13.3%	38.0%
Diabetes mellitus	4.3%	20.9%	3.3%	22.4%	4.0%	22.2%

times more prevalent in predicted high-cost beneficiaries, with hypertension present in 60.2% of predicted high-cost beneficiaries. The prevalence of hypertension, heart conditions and diabetes shows no discernible trend for medium- and long-terms. Hyperlipidaemia decreases in prevalence among predicted high-cost beneficiaries for longer time horizons, which may seem as counterintuitive and analyses using even longer time-horizons may be needed to confirm the presence of this trend.

5.3.3.3 Table 5 presents the prevalence of conditions among beneficiaries who were high-cost in 2007. Beneficiaries who are predicted to be high-cost from this sub-group constitute predicted repeat high-cost beneficiaries. The prevalence of diseases among repeat high-cost beneficiaries is similar to that of beneficiaries who are predicted to transition from low- to high-cost. Chronic conditions show a significantly greater prevalence among beneficiaries predicted to be high-cost compared to beneficiaries predicted to be low-cost.

6. DISCUSSION AND CONCLUSIONS

6.1 Mean reversion in beneficiary costs means that beneficiaries who are currently high-cost should not necessarily be the focus of care management. This research found that there is significant flux in high-cost beneficiaries. Beneficiaries who are currently high-cost may not be high-cost in future and current low-cost beneficiaries may become high-cost A key finding of this research was that low-cost beneficiaries who transition to high-cost can be targeted before high costs arise with a reasonable amount of accuracy. Furthermore, the degree of accuracy did not decrease with longer time horizons. This is in contrast to predictions of repeat high-cost beneficiaries which became less accurate with a longer time horizon.

6.2 The extension of costs from costs of individual years to aggregated costs of multiple years increased the predictive accuracy of models substantially. The authors regarded this to be a significant finding. The increase in accuracy was felt most keenly when increasing the number of years of aggregated costs from 1 to 3. This suggests that a relatively small extension of the time horizon can result in markedly greater accuracy and suggests that a medium-term approach be adopted in predicting high-cost beneficiaries. Medical schemes have budgetary horizons of a single year, which may necessitate one-year predictions. However, where long-term sustainability is the aim of care management, a longer-term view may allow better targeted care management programmes and hence better outcomes.

6.3 Low-cost beneficiaries who were predicted to transition to high-cost presented a distinctively different disease burden from repeat low-cost beneficiaries. Disease prevalence was significantly higher, most especially for chronic conditions. Hypertension presented the greatest prevalence. This suggests that beneficiaries with chronic diseases present the greatest opportunity for care management. Disease

prevalence among predicted high-cost beneficiaries reduced for some acute diseases over longer time horizons. Additionally, beneficiaries predicted to be high-cost were predominantly elderly, and were more likely to be female.

6.4 This research suggested that predictive modelling can be of value in identifying beneficiaries for care management. Given the current South African healthcare environment, predictive modelling may be a tool that can be used in achieving gains in healthcare efficiency. In practice, care management may be directed at prevalent diseases that are amenable for intervention and cost-effective treatment. Detailed planning, economic evaluation and implantation are needed for the development of care management programmes. These aspects are not considered in this research. This research thus represents one step in the wider task of planning patient management and interventions.

REFERENCES

Basu, A, Manning, WG & Mullahy, J (2004). Comparing alternative models: log vs proportional hazard? *Health Economics*, **13**(8):74965

Blough, DK, Madden, CW & Hornbrook, MC (1999) Modelling risk using generalized linear models. *Journal of Health Economics*, 18:153–71

Buntin, MB & Zaslavsky, AM (2004). Too much ado about two-part models and transformation? Comparing methods of modeling Medicare expenditures. *Journal of Health Economics*, 23:525–42

Cousins, MS & Liu, Y (2003). Cost savings for a preferred provider organization population with multi-condition disease management: Evaluating program impact using predictive modeling with a control group. *Disease Management*, **6**(4):207–17

Cousins, MS, Shickle, LM & Bander, JA (2002). An introduction to predictive modeling for disease management risk stratification. *Disease Management*, 2002(5):157–67

Cumming, RB, Knutson, D, Cameron, BA & Derrick, B (2002). A comparative analysis of claims-based methods of health risk assessment for commercial populations. A research study sponsored by the Society of Actuaries

Diehr, P, Yanez, D, Ash, A, Hornbrook, M & Lin, DY (1999) Methods for analyzing health care utilization and costs. *Annual Review of Public Health*, 20:125–44

Dove, HG, Duncan, I & Robb, A (2003). A Prediction Model for targeting low-cost, high-risk members of managed care organizations. *The American Journal of Managed Care*, **9**(5):381–9

Duan, N (1983). Smearing estimate: a nonparametric retransformation method. *Journal of the American Statistical Association*, 78:605–10

Duan, N, Manning, WG, Morris, CN & Newhouse, JP (1983). A comparison of alternative models for demand for medical care. *Journal of Business & Economic Statistics*, **1**(2):115–26

Duncan, I (2011). *Healthcare Risk Adjustment and Predictive Modelling*. Winsted, Connecticut, Actex Publications.

Frees, E, Gao, J & Rosenberg, M (2011). The frequency and amount of inpatient and outpatient healthcare expenditures. *North American Actuarial Journal*, 15:377–92

Grossman, M (1972). On the concept of health capital and the demand for health. *Journal of Political Economy*, 80:223–55

Lynch, JP, Forman, SA, Graff, S & Gunby, FC (2000). High-risk population health management-achieving improved patient outcomes and near-term financial results. *American Journal of Managed Care*, 6:781–9

Manning, WG, Newhouse, JP, Duan, N, Keeler, B, Leibowitz, A & Marquis, S (1987). Health insurance and the demand for medical care: evidence from a randomized experiment. *American Economic Review*, **77**(3):251–77.

Manning, WG (1998). The logged dependent variable, heteroscedasticity, and the retransformation problem. *Journal of Health Economics*, **17**(3):283–95

Manning, WG & Mullahy, J (2001). Estimating log models: to transform or not to transform? *Journal of Health Economics*, 20:461–94

Manning, WG, Basu, A & Mullahy, J (2005). Generalized modeling approaches to risk adjustment of skewed outcomes data. *Journal of Health Economics*, 24:465–88

Meenan, RT, O'Keeffe-Rosetti, C, Hornbrook, MC, Bachman, D & Fishman, P (2003). The sensitivity and specificity of forecasting high-cost users of medical care. *Medical Care*, 37:815–23

Mehmud, SM (2013). Nontraditional variables in healthcare risk adjustment. A research project sponsored by Society of Actuaries

Mullahy, J (1998). Much ado about two: reconsidering retransformation and the two-part model in health econometrics. *Journal of Health Economics*, 17:247–81

Rosenberg, MA & Farrell, PM (2008). Predictive modeling of costs for a chronic disease with acute "High-Cost" episodes. *North American Actuarial Journal*, **12**(1):1–19

Winkelman, R & Mehmud, S (2007). A comparative analysis of claims-based methods of health risk assessment for commercial populations. A research study sponsored by the Society of Actuaries

Zhao, Y, Ash, AS, Ellis, RP, Ayanian, JZ, Pope, GC, Bowen, B & Weyuker, L (2005). Predicting pharmacy costs and other medical costs using diagnoses and drug claims. *Med Care*, **3**(1):34–43

APPENDIX A

Table A1 Variables

Variable name	Description
PLPY	The cost per beneficiary per year in Rands
Age	The age of the beneficiary, categorised in five-year age bands
Sex	Sex of the beneficiary
Plan	The plan to which the beneficiary belongs at the beginning of the first year of prediction
Province	Province in which the beneficiary lives
Chronic Status	Indicator of the presence of one or more chronic diseases
CCS1-64	Indicator of the presence of CCS grouper aggregated conditions

Table A2 CCS grouper aggregated categories

CCS category	CCS category description
CCS1	Acute bronchitis and URI
CCS2	Allergic reactions
CCS3	Anaemia and other deficiencies
CCS4	Appendicitis
CCS5	Back problems
CCS6	Cancer
CCS7	Cataract
CCS8	Cerebrovascular disease
CCS9	CNS infection
CCS10	Coma, brain damage
CCS11	Complications of pregnancy and birth
CCS12	Complications of surgery or device
CCS13	Congenital anomalies
CCS14	Chronic Obstructive Pulmonary Disease (COPD), asthma
CCS15	Diabetes mellitus
CCS16	Disorders of mouth and oesophagus
CCS17	Disorders of teeth and jaws
CCS18	Disorders of the upper GI
CCS19	E Codes: All (external causes of injury and poisoning)
CCS20	Epilepsy and convulsions
CCS21	Female genital disorders, and contraception
CCS22	Gallbladder, pancreatic and liver disease
CCS23	Glaucoma
CCS24	Headache
CCS25	Heart conditions

CCS26	Haemorrhagic, coagulation and disorders of white blood wells
CCS27	Hereditary, degenerative and other nervous system disorders
CCS28	Hernias
CCS29	HIV
CCS30	Hyperlipidaemia
CCS31	Hypertension
CCS32	Infectious diseases
CCS33	Influenza
CCS34	Intestinal infection
CCS35	Male genital disorders
CCS36	Mental disorders
CCS37	Non-malignant breast disease
CCS38	Non-malignant neoplasm
CCS39	Normal birth/live born
CCS40	Osteoarthritis and other non-traumatic joint disorders
CCS41	Other bone and musculoskeletal disease
CCS42	Other care and screening
CCS43	Other circulatory conditions arteries, veins and lymphatics
CCS44	Other CNS disorders
CCS45	Other endocrine, nutritional and immune disorder
CCS46	Other eye disorders
CCS47	Other GI
CCS48	Other kidney disease
CCS49	Other stomach and intestinal disorders
CCS50	Other urinary
CCS51	Otitis media
CCS52	Paralysis
CCS53	Perinatal conditions
CCS54	Pneumonia
CCS55	Poisoning by medical and non-medical substances
CCS56	Renal failure
CCS57	Residual codes
CCS58	Skin disorders
CCS59	Symptoms
CCS60	Systemic lupus and connective tissues disorders
CCS61	Thyroid disease
CCS62	Tonsillitis
CCS63	Trauma-related disorders
CCS64	Urinary tract infections