UNIVERSITY OF TECHNOLOGY SYDNEY

ECONOMICS DISCIPLINE GROUP

UTS BUSINESS SCHOOL

WORKING PAPER NO.1 May 2012

Explaining Health Care Expenditure Variation: Large-sample Evidence Using Linked Survey and Health Administrative Data

Randall P. Ellis
Denzil G. Fiebig
Meliyanni Johar
Glenn Jones
Elizabeth Savage

ISSN: 2200-6788
http://www.business.uts.edu.au/economics/

Explaining Health Care Expenditure Variation: Large-sample Evidence Using Linked Survey and Health Administrative Data

Randall P. Ellis[a], Denzil G. Fiebig[b], Meliyanni Johar[c], Glenn Jones[c*], Elizabeth Savage[c]

[a]Department of Economics, Boston University, Boston MA USA
[b]School of Economics, University of New South Wales, Sydney, Australia
[c]Economics Discipline Group, University of Technology Sydney, Sydney, Australia

Abstract

Explaining individual, regional, and provider variation in health care spending is of enormous value to policymakers, but is often hampered by the lack of individual level detail in universal public health systems because budgeted spending is often not attributable to specific individuals. Even rarer is self-reported survey information that helps explain this variation in large samples. In this paper, we exploit the linkage of a survey of over 267,000 Australians age 45 and over to several years of hospital, medical and pharmaceutical records. After calculating total health care cost for each survey respondent, we examine health expenditures due to health shocks and those that are intrinsic to an individual. We find that high fixed-effects are positively associated with age, especially older males, poor health, obesity, smoking, cancer, stroke and heart conditions. Hospital admissions are the largest component of fixed effects. High time-varying expenditures are associated with speaking a foreign (not English) language at home, low income and low education, suggesting greater exposure to adverse health shocks. For these individuals, health expenditure is comprised mainly of out-of-hospital medical services and drugs.

Acknowledgements

This research uses data from the 45 and Up Study which is managed by the Sax Institute in collaboration with major partner Cancer Council New South Wales; and partners the Heart Foundation (NSW Division); NSW Ministry of Health; *beyondblue: the national depression initiative*; Ageing, Disability and Home Care, NSW Family and Community Services; and Australian Red Cross Blood Services. This project was undertaken by the University of Technology Sydney and utilised MBS and PBS data supplied by the Department of Human Services. Data linkage for the project was undertaken by the Centre for Health Record Linkage. The project has ethics approval from the NSW Population and Health Services Research Ethics Committee. The study findings are those of the authors and do not necessarily represent the views of the Commonwealth of Australia, represented in this instance by the Department of Health and Ageing and the Department of Human Services. The project is funded by an ARC Discovery Project grant (DP110100729).

*Correspondence to: E-mail: Glenn.Jones@uts.edu.au
Keywords: Health expenditure, health insurance, panel data
JEL Codes I10, I13, C23

1. Introduction and Background

Understanding patterns of health care spending is of enormous policy and research interest, but is often hampered by the available data. Researchers are often forced to choose between large samples of insurance-related administrative data or smaller samples of survey based information, with the former lacking individual demographic and self-reported information and the latter often lacking detailed spending and diagnostic data. In this paper, we exploit the linkage of a large survey of Australian individuals aged 45 years and over, who consume the bulk of health care, to four years of comprehensive administrative health utilisation data. The administrative records cover hospital admissions, emergency department presentations, claims for out-of-hospital medical services (e.g. general practitioner and specialist visits, diagnostic testing) and claims on prescription subsidised drugs by each survey respondent.[1] We use this data to calculate an estimate of individual total health expenditure and then identify variations in the health expenditures that are due to exogenous changes in health conditions and variations that are specific to the individual. Exploiting the richness of the survey data and health provider information, we examine the influence of background characteristics and socioeconomic factors, lifestyle, existing health conditions and the use of a regular general practitioner on these sources of variations in individual health expenditure. The results not only fill a gap in the understanding of individual health expenditures in publicly funded health systems that often lack individual-level health expenditure data, but also improve our understanding of how linking administrative and survey information enhances predictions over time, a key topic for many health plan and provider payment systems.

All Australians have access to a universal public health system. The Australian Medicare program fully covers public inpatient care in public hospitals and heavily subsidises prescription drugs and most private medical care provided both in-hospital and out-of-hospital. General practitioners and specialists are paid on a fee for service basis and their fees are unregulated. Over the past decade the share of the public contribution to total health expenditure has been constant at around 70% (Australian Institute of Health and Welfare, 2010). In 2009/2010, the total health expenditure was $124 billion, of which $116 million was recurrent. Of the recurrent expenditure, 31% was spent on public hospitals, 19% on medical services and 14% on pharmaceuticals. Because of the relatively

[1] The major excluded services are dental, optometry and allied health, non-subsidised drugs and outpatient hospital services.

complex health expenditure subsidy structure, an accurate measure of individual total health expenditure can only be obtained through careful cleaning and cost attribution algorithms applied to the integrated administrative health records. We briefly describe how the dependent variable, total annual expenditure, was calculated, with further details in an appendix.

For our statistical analysis, we adopt a forward-looking or prospective model of health expenditure in which the dependent variable is *next year's* health expenditure. The prospective model emphasises systematic variations in health expenditure due to chronic conditions as opposed to acute conditions or expensive one-off events such as surgery which are emphasised in a concurrent model in which the dependent variable is expenditure in the current period. Furthermore, the prospective model is more useful than the concurrent model for payment or budgeting purposes as it can forecast the payers' future financial obligations. In addition, from an econometric perspective, the prospective model reduces the problem of reverse causality bias which arises if health conditions are affected by changes in health expenditure.

The availability of four years of health expenditure data permits a three year prospective panel sample and allows us to estimate the model using a fixed effects approach. An individual's health expenditure is a function of his health conditions, use of pharmaceutical and medical services, background and economic characteristics, location, macroeconomic conditions, lifestyle and doctor's characteristics. Many of these potential predictors are constant over time (or slow changing), so their effects are not identified by the standard fixed effect approach. We therefore conduct a two-step analysis. In the first step, we estimate a model that includes both individual specific fixed effects and all of the time-varying predictors. The time-varying portion of the model defines the variation in predicted "time-varying" health expenditures, while the individual "fixed effects" capture the variation of health expenditure that is intrinsic (time invariant) to a person. Standard fixed effects estimation provides a decomposition of these two components. In step two, we regress each of these expenditure components on time-invariant predictors.

We find that time-varying factors and fixed effects each explain about half of the variation in individual health expenditures. Large fixed effects are associated with age, especially older males, disability, poor health, smoking, and chronic conditions, especially cancer, heart disease, stroke, and obesity. We do not find strong evidence that high fixed effects are systematically related to the

consistent use of a general practitioner nor with the fee setting behaviour of a regular general practitioner, even though in Australia, general practitioners are free to set their own fees. In contrast, large time-varying expenditures are associated with speaking a foreign language (not English) at home, low education and low income. Older individuals, obese individuals, those in poor health and those with diabetes and hypertension have high fixed effects and time-varying expenditures. These individuals may be characterised as experiencing further health deterioration from an already poor health state and they rely heavily on the hospital sector.

2. Model

In our prospective model, the dependent variable is next year's expenditure. Let y_{it} be health expenditure of individual i=1, ... ,n in survey year t=1, ... ,T. Our baseline specification is the following linear model estimated by ordinary least squares (OLS):

(1) $y_{it} = \beta_0 + x_i'\beta_1 + z_{it-1}'\beta_2 + u_{it}$,

where x_i is a vector of time invariant (or slow changing) covariates (e.g. background and socio-economic characteristics, location, general practitioner (GP) characteristics), z_{it} is a vector of time-varying covariates including chronic conditions, various pharmaceuticals and variables capturing medical service use, β_k are conformable parameter vectors to be estimated and u_{it} is the error term.

Because health expenditures are highly skewed and often include zero observations, reliance on a linear model estimated by OLS may seem inappropriate. To confirm its suitability we tested the use of a range of more flexible models reviewed in Buntin and Zaslavsky (2004), including log-transformed models and generalized linear models (GLM) assuming gamma and negative binomial distributions. To account for zero expenditure, we also estimated a Tobit model with left-truncation at $0 and a two-part version of the linear, log and GLM models with a logit model as the first part to estimate the probability of having a positive expenditure. However, perhaps because only 3% of the sample has zero expenditure, the improvement in fit of the two-part models was very marginal. We find that the linear model is far superior to the other models in terms of fit, as measured by prediction errors, both in and out of sample.[2] The log and GLM models substantially reduce the weight on observations with very high expenditure, and when retransformed results in severe overestimation at

[2] Results are available from the authors.

the upper expenditure tail. The mean prediction absolute error in raw levels from the log model is twice that from OLS while the mean errors from the GLM models are about 20% larger. On the same criterion, the Tobit model performs worse than the GLM models. The out-of-sample performances are similar to the in-sample performances, which is expected given the large sample size; hence we focus on within-sample measures in this paper.

While the linear model produces the best fit in a comparison with a range of commonly used alternatives, it does not deal with potential endogeneity due to omitted variables. Specifically, u_{it} may not be truly random. For example, "frail" individuals may always have higher expenditure than others because they are more prone to illness. Another possibility leading to systematically high expenditures could be a doctor's tendency to order many diagnostic tests or set high fees. In equation (1), these systematic variations are captured in the disturbance term and to the extent they are correlated with the existing set of covariates will result in biased and inconsistent estimates of the β's. Most previous studies of modelling health care expenditures such as Buntin and Zaslavsky (2004) and Manning, Basu and Mullahy (2005) rely on large cross sections of individuals and hence are limited in how they can address such threats to inference. In our case, the availability of panel data means we are able to control for some types of omitted variables in estimating β_k through the use of linear fixed effects methods.

The standard assumption used in the fixed effects model is that u_{it} can be decomposed into individual-specific effects α_i and a random component ε_{it} as follows:

(2) $u_{it} = \alpha_i + \varepsilon_{it}$.

Under this specification, we are able to consistently estimate β_2, the coefficients of the time-varying parameters, irrespective of any correlation between the covariates included in equation (1) and time invariant omitted variables. The fixed effects estimator can be thought of as applying OLS to the within transformed model (all variables are expressed in terms of deviations from their sample means calculated over time for each individual) or, equivalently, to a model that includes individual specific constants for all *n* individuals in the data. While providing estimates for β_2, the parameters associated with the time invariant parameters, β_1, will not be identified. However, it is possible to recover estimates for the individual effects. Implementing this procedure, the individual effects are restricted

to sum to zero or equivalently are parameterised as deviations from their overall mean which appears as the estimate of β_0 in the fixed effects estimation. Therefore

(3) $\hat{\alpha}_i = \bar{y}_i - \hat{\beta}_0 - \bar{z}_i'\hat{\beta}_2$

where $\hat{\beta}_0$ and $\hat{\beta}_2$ are fixed effects estimates and $\bar{y}_i$ and $\bar{z}_i$ are sample means averaged over the time series observations for each individual.

Post estimation, predictions of individual expenditures using the fixed effects results will comprise two components; the first we term "time-varying" expenditure because its predictors vary over time and is defined by:

(4) $\tilde{y}_{it} = z_{it-1}'\hat{\beta}_2.$

$\tilde{y}_{it}$ may be interpreted as changes in expenditure arising from health shocks or other information that changes over our three year panel. The second "fixed effects" expenditure component is defined for each individual as $\hat{\alpha}_i + \hat{\beta}_0$ and captures all time invariant effects, including socioeconomic variables, access and provider effects, chronic health problems and long term effects of preventative care. The two predicted components of expenditure relate to the observed data in the following way:

(5) $y_{it} = \tilde{y}_{it} + (\hat{\alpha}_i + \hat{\beta}_0) + \hat{\varepsilon}_{it}.$

The fixed effect results provide estimates and predictions with minimal assumptions regarding the fixed effects and their relationship with included covariates but they come at the cost of not being able to identify the impact of time invariant variables that contain a considerable amount of information on individuals available from the survey and provider information available in the medical services expenditure data. In order to recover this rich information we conduct a second stage analysis using the two predicted components of expenditures as the outcome variables in the following auxiliary regressions:

(6) $\tilde{y}_{it} = \delta_0 + x_i'\delta_1 + \nu_{it},$
(7) $(\hat{\alpha}_i + \hat{\beta}_0) = \theta_0 + x_i'\theta_1 + \varpi_i.$

x_i includes the typical background and socio-economic characteristics of individuals, but also we make use of the information on provider identifier and out-of-pocket costs to construct a measure of the strength of the relationship between patient and GP as well as the GP's fee setting behaviour. This

second set of variables allows us to test the hypothesis that systematically high expenditures could be due to high cost consultations due to high fees or multiple (duplicate) diagnostic tests.

In equation (6), we seek to explain the time-varying expenditures. Because there is a separate prediction for each time period for each individual, there is a choice as to how this estimation proceeds. In what follows, (6) is estimated for the survey year in order to better match characteristics and time-varying expenditures.

Analogously, equation (7) provides evidence on the factors associated with the fixed effects. More formally, this procedure produces Hausman and Taylor (1981) estimates of β_1, in the original specification. They result from making a stronger assumption about the omitted unobservables, namely that any correlation between omitted variables is confined to the time-varying variables and are uncorrelated with the time invariant variables. Under these conditions the estimates of θ_1 obtained from (7), are consistent estimates of β_1 in equation (1).

Finally, we exploit the combination of predicted time-varying and fixed effect expenditures in one further way. Observations are categorised into 4 mutually exclusive types, namely observations with: (i) large time-varying and large fixed effects (type I), (ii) large time-varying and small fixed effects (type II), (iii) small time-varying and large fixed effects (type III), and (iv) small time-varying and small fixed effects (type IV). For each type, we examine the mix of health services use.

3. Data

Our sample is derived from the 45 and Up Study of over 267,000 New South Wales (NSW) residents aged 45 and over (45 and Up Study collaborators, 2008). NSW is the most populous state of Australia. While the 45 and Up data covers a sub population, the population group over 45 incurs 62% of Australia's total health expenditure (AIHW, 2010)[3]. The survey data is linked at the individual level, with participants' consent, to the following administrative data from 2006 to 2009:

1. NSW Admitted Patient Data Collection (APDC), with one record per separation;
2. NSW Emergency Department Data Collection (ED), with one record per presentation;
3. Medicare Australia Medical Benefits Schedule data (MBS), with one record per claim; and

[3] Excluding expenditure non-admitted patients, high-level residential aged care, over-the-counter pharmaceuticals and other health practitioner services

4. Medicare Australia Pharmaceutical Benefits System data (PBS), with one record per prescription.

The data linkage for the APDC and the EDDC was performed by the Centre for Health Record Linkage (CHeReL) using a probabilistic matching on first name, surname, date of birth and address. The linkage of the MBS and PBS data was performed by the Sax Institute.[4] The linked, de-identified data is released under ethics approval from the NSW Population and Health Services Research Ethics Committee.

The first data set is used to price hospital separations. Every separation, public or private, has an Australian-Refined Diagnosis Related Group (AR-DRG) code. The AR-DRG is a patient classification scheme based on an algorithm of hierarchies of diagnoses and procedures that relates the number and types of patients treated in a hospital to the resources required by the hospital.[5] A cost weight is attached to each AR-DRG measuring its relative cost compared to the average cost of all AR-DRGs; the average cost of all AR-DRGs has a cost weight of 1. Because Australian states manage their own public hospitals, the cost weights vary by state. Here, we apply the cost weights published in the *Costs of Care Standards 2009/10* by the NSW Department of Health. We follow the guidelines in the Standards to adjust the cost weight to the characteristics of each separation. This adjustment depends on hospital type, type of care (overnight, same day, transfer, in mental health unit, non- or sub-acute care units such as rehabilitation), length of stay, ICU hours and the use of ventilation machine. Similarly, for emergency department presentations, the Standards outline variation in emergency department cost by hospital type, triage category (more urgent category is more expensive) and whether the patient is subsequently admitted. Further details of the expenditure imputation are provided in the Appendix.

The MBS and PBS data reports the expenditure on subsidised medical services and pharmaceutical items. We aggregate these expenditures on an annual basis and adjust them to constant $2009 (In December 2009, A$1=US$0.90). Individual annual total health expenditure is calculated as the sum of three components: (1) hospital costs (admission and emergency presentation); (2) charges

[4] For details on the APDC and EDDC linkage procedure and quality, see http://www.cherel.org.au/ and for details of the 45 and Up study see http://www.45andup.org.au/ .

[5] For details on the development of AR-DRG classification system, see http://www.health.gov.au/internet/main/publishing.nsf/Content/Casemix-1/$File/Final_Report_November_2009.pdf

for out-of-hospital MBS items; and (3) prices paid to suppliers of out-of-hospital PBS drugs (where the price is greater than the co-payment) in any given year. We these imputed expenditures have a 0.94 correlation with the official (AIHW) statistics.

The hospital and PBS data contains diagnoses and drug codes which predict an individual's health expenditure. Each hospital separation record has one primary diagnosis and up to 55 secondary diagnoses. These diagnoses are coded according to the over 25,000 ICD10-AM codes. To help organise this diagnostic information into a more manageable number of diagnoses groups, without loss of information on co-morbidities, we use a US-based risk adjustment software called DxCG Risk Solution developed by Verisk Health, which has been extensively applied on US data (e.g., Einav et al., 2011; Zhao et al., 2005; Ash et al., 2001).[6] The software groups diagnoses into 394 'hierarchical condition categories' (HCCs), which group together clinically related diagnoses according to their current and future costs so that cost is primarily driven by the most severe manifestation of a given diagnosis; a new diagnosis that adds a related, but less serious medical problem, does not increase cost, while unrelated diagnosis contributes cumulatively to cost. From the HCCs, the software offers a more aggregated grouping into 117 'related condition categories' (RCCs). For our purpose we find that RCCs are sufficient.[7] The RCCs are not mutually exclusive, e.g., a person with two serious conditions like heart and liver failure will have two RCCs. Similarly, for pharmaceutical information, the software provides a mapping from the most detailed level (7-digit) Anatomical Therapeutic Chemical (ATC) drug codes into 164 non-over-the-counter Rx Groups.

Of the original sample, 262,293 respondents (98.2%) are included in our analysis. We exclude respondents who, were surveyed in 2010, have an invalid age, volunteered rather than were randomly selected to participate in the survey and those who died during the study period. Since the health expenditure data extends for four years, we have a balanced panel sample of 262,293 respondents and 1,049,172 (262,293 x 4) person-year observations. Given our prospective approach, the first

[6] The software, which extends the classification system used by the US Medicare program for paying competing health plans, organises ICD-10 diagnoses information into a large number of non-mutually exclusive categories and imposes hierarchies on diseases so that more serious or expensive conditions take precedence over less serious or expensive conditions. Similarly it groups ATC codes into 165 drug groups based on therapeutic class, active ingredients and doses and strength. The software also performs a number of data-cleaning steps to identify illegal (e.g. coding errors) or invalid (e.g. male pregnancies) diagnoses.

[7] The predictive power gain from using HCCs is small relative to the hundreds of extra parameters to be estimated.

observation year of health expenditure, 2006, is not used in estimation, and the estimation sample size is 786,879 person-years.

Table 1 reports the summary statistics of total health expenditure across years. Less than 3% of the sample has zero expenditure, and spending increases over time. As commonly found in most health expenditure series, spending is skewed to the right with the median much lower than the mean, which is closer to the 75th percentile. The bottom 5% of the sample has less than $200 in annual total health cost whilst 5% at the top have more than $15,000 in health expenditure. The coefficient of variation (the standard deviation/mean) ranges from 2.05 to 2.35, notably lower than in samples from the US (Ash et al, 2001) where the CV is often 3 or more. The skewness measure is also moderate, averaging around ten. These features help explain the higher measures of R-square found in our sample compared with US samples, and may help explain the time pattern of predictiveness we observe below.

[Table 1 here]

Given that we have hundreds of RCC and Rx dummy variables, for conciseness, we do not present the full summary statistics for time-varying predictors. Among the most common chronic conditions associated with hospital admission are benign neoplasm, hypertension and gastrointestinal conditions. The most common pharmaceuticals are lipid lowering agents, anti-infectives and ulcer/GERD medications. To measure medical services utilisation, we use the total number of GP and specialist consultations in a year. Specifically we use a specific MBS item 23, which is for consultation (not for immunisation) less than 20 minutes by a GP, and item 104 for initial consultation by a specialist excluding an ophthalmologist. These two items make up 44–57% of all out-of-hospital professional attendances in a given year and 37–45% of all out-of-hospital MBS services excluding diagnostic, imaging and pathology. On average, individuals have 7 GP consultations and 1 specialist consultation in a year.

Table 2 provides the summary statistics of survey variables included in equations (6) and (7). About 40% of respondents are over 65, with about 11% over 80. There are slightly more female than male respondents, and the majority of respondents are married with tertiary education. Country of birth, speaking a foreign language (not English) at home and skin colour are included to capture individual background conditions that may affect health care utilisation. Most respondents are either

still working full-time or fully retired. Private health insurance coverage is 65%, higher than the population average of 55%. In addition to demographics and economic variables, the 45 and Up survey also asked its respondents to self-report lifestyle, ever diagnosed chronic conditions and to provide a self-assessment of their health in general. The self-assessed general health suggests that most respondents are in good health or better. However, the incidence of chronic conditions, which may be regarded as a more objective measure of health, signals health care need. Survey responses indicate over 30% of the sample has high blood pressure, 28% have skin cancer, 11% have gender-specific or other cancer, 12% have heart diseases and 9% have diabetes.

[Table 2 here]

To capture the patient-doctor relationship, we use an indicator variable which takes a value of one if the individual has a regular GP who he consulted most of the time. We define a family GP as a GP who provided more than 85% of the total standard consultations (item 23) to a given individual in a given year; this variable indicates consultation to only one GP for those with total number of standard consultations fewer than 7 (the sample mean). The presence of a family GP may indicate a strong patient-GP relationship. Nearly 30% of individuals have a family GP. To measure GP fee-setting behaviour, we use the average revenue from each patient for item 23 which is the average fee above the schedule fee (i.e., out-of-pocket for patients).[8] We make a distinction between family GP's and other GP's fee. The former is switched on when an individual has a family GP, uniquely capturing the fee setting behaviour of this family GP. For those without a family GP, the latter is switched on. However, there can be multiple other GPs in which case the fees are taken to be the average across a number of GPs.

4. Results

In Table 3 we present the fixed effects regression results. The null hypothesis that all $\alpha_i = 0$ is strongly rejected (p-value<0.0005) and the standard Hausman test for comparing fixed and random effects, where the later assumes (2) holds but with zero correlation between α_i and covariates, also favours the use of the fixed effects model ($\chi^2_{232} = 164{,}897$, p-value<0.0005). In comparing

[8] We do not construct these measures for specialists because patients may go to different specialists for different specialty; furthermore, 88% of patients only have 2 specialist initial consultations.

coefficient estimates from fixed and random effects, we find that the latter tend to be overestimates which are consistent with expectations since innate poor health is positively related to hospital admission and use of pharmaceuticals and medical services. From (2), $var(u_{it}) = var(\alpha_i) + var(\varepsilon_{it}) = \sigma_\alpha^2 + \sigma_\varepsilon^2$. Here $\hat{\sigma}_\alpha^2 = 52.49$ and $\hat{\sigma}_\varepsilon^2 = 59.9$ indicating that 47% of the variance in total health expenditure can be attributed to fixed-effects (i.e., variations from the between estimator across individuals); another interpretation of $\hat{\sigma}_\alpha^2/(\hat{\sigma}_\alpha^2 + \hat{\sigma}_\varepsilon^2) = 0.47$ is that it represents the intra-class correlation coefficient measuring the correlation between two observations for the same individual.

Turning to the fixed effects coefficient estimates in Table 3 we find most chronic conditions have negative coefficients, which may be explained by the prospective nature of the model. For example, one-off conditions like bone marrow transplants and complications have a large negative effect (-$21,070) because next year this patient is unlikely to have another bone marrow transplant. Other examples are cardiac arrest, respiratory arrest, artificial openings, implant and device complications and amputations. In contrast, conditions like cancer, tumours, congenital heart conditions, kidney inflammation have positive coefficients which may suggest that these patients will require ongoing treatment and subsequently higher expenditures. Pharmaceuticals mostly have positive coefficients because drug use tends to be persistent. Some drugs have large contributions to health expenditure such as octreotides (injection associated with cancer-related side effects), pulmonary hypertension drugs, otic and nasal agents, medication for chronic kidney disease and treatment for endometriosis. The effects of the common pharmaceuticals, lipid lowering agents and anti-infectives medications are imprecisely estimated, probably because a lack of variation over time in the consumption of these drugs. Finally, each additional GP consultation adds about $131, on average to total expenditure next year and each additional specialist consultation adds $972.

[Table 3 here]

Figure 1 shows the distribution of (in a clockwise direction): observed expenditure (y_{it}), predicted expenditure ($\tilde{y}_{it} + \hat{\alpha}_i + \hat{\beta}_0$), predicted time-varying expenditure ($\tilde{y}_{it}$) and predicted fixed effects ($\hat{\alpha}_i + \hat{\beta}_0$). Nothing in the nature of the linear model restricts predictions to be positive and a few negative predictions (4.9%) appear. Overall predicted expenditure exhibits wider variation than the time-varying expenditure component with the difference driven by considerable individual heterogeneity. The mean and standard deviation of predicted expenditure are around $4,423 and

$7,500, respectively, while the mean of the time-varying expenditures is approximately $1,376 and the standard deviation is $2,357. Half of the sample is estimated to have time-varying expenditure between $261 ($25^{th}$ percentile) and $2,274 ($75^{th}$ percentile), and in the upper half of the distribution, 5% of the sample has time-varying expenditure above $5,013 ($95^{th}$ percentile).

[Figure 1 here]

A quarter of individuals have negative fixed effects. The mean is $3,047 but the distribution is skewed to the right indicating the presence of individuals with very large positive fixed expenditures. It can also be seen that the shape of the fixed effects distribution drives the shape of the overall predicted expenditures.

To classify individuals into distinct types, we examine interaction between time-varying expenditures (TV) and fixed effects (FE) in the sample. Each individual has three TV values corresponding to each of the three years of data, but to best match the survey information the survey year TV values are used in what follows. Figure 2 panel (A) shows the sample density over the TV-FE plane. This graph supports the presence of four types of individuals by high or low TV and FE, given by four high-density regions. We find that these four regions are well-captured by defining a high expenditure as expenditure in the top 20%. We classify individuals with large $\hat{\alpha}_i + \hat{\beta}_0$ and $\tilde{y}_{it}$ (i.e., high FE and high TV types) as type I, and label them as representing 'further health deterioration'. Type II (high FE, low TV) represents individuals needing 'ongoing treatment'. Type III (low FE, high TV) represents those with 'adverse health shocks', and finally type IV (low FE, low TV) are those with 'low use of health services'. Type I has the largest average expenditure ($18,350), and they represent 6.6% of the sample. Type II has the next largest average expenditure of $11,111 for 13.6% of the sample. Type III makes up 13.6% of the sample with $4,635 in average expenditure and type IV makes up the remaining 66.2% with $1,546.

In Figure 2 panel (B), we examine the mix of health services used by each type. Types I and II's expenditures are high, driven by hospital admission costs. Since these two types have large fixed effects, this result may reflect selection in the hospitalised population. Type II spends the largest share of their expenditure on inpatient costs. For these two types, we can expect heavy reliance on the hospital sector. In contrast, type III and IV's expenditure comprise mainly of out-of-hospital drugs

(PBS) and medical services (MBS). Type III has the highest share of medical services suggesting increased use of doctors, diagnostics, therapists etc. associated with new health conditions.

[Figure 2 here]

Table 4 presents the estimation results for equations (6) and (7). We first discuss the results for (7) given in the first column. About 15% of variation in $(\hat{\alpha}_i + \hat{\beta}_0)$ can be explained by time invariant survey measures.[9] Those with large fixed effects tend to be older especially older males, never been married, widowed or divorced, not in full time employment, disabled, have fair or poor self-assessed health, suffer from chronic conditions, especially cancer, heart disease, broken bone, diabetes and stroke. Individuals who regard themselves as being in poor health have more than $6,000 more in annual health expenditure than an average person. This supports the use of self-assessed health, available in most data sets, as a measure of innate health. Individuals with a heart disease have about $2,200 in health expenditure every year and those with breast or prostate cancer have about $1,800 in annual expenditure. Unhealthy lifestyle also increases fixed effects, in addition to the existing chronic condition. The morbidly obese have an extra $286–$480 on top of any cost of chronic conditions. Smoking also increases cost by $340. Being underweight and alcohol abstinence are associated with higher fixed effects, perhaps reflecting consequences of chronic conditions. Private health insurance coverage tends to increase fixed effects. This result is consistent with insuree's moral hazard leading to overutilisation of health services. Highest income is also associated with larger fixed effects, which may capture taste in investment in health or health technology among the high income individuals.

[Table 4 here]

The doctor variables have small impact on the fixed effects. Those with a family GP tend to have lower fixed effects, consistent with the hypothesis that a relationship with doctors avoids duplications of diagnostics and tests, but this effect is not precisely estimated despite the large sample size. GP fees also do not predict fixed effects, contradicting the conjecture that high fixed effects are due to high charging doctors.

[9] Our results contrast with Newhouse et al (1989) which finds that about 20-30% of total variation is explained by individual level fixed effects, whereas in our model fixed effects explain about 56% of total variation. As explained below, socio-demographic factors, utilisation, pharmacy, and health conditions explain about half of the total FE in our data, while time invariant survey factors explain about 15% of the FE.

The second column reports the results for (6). Some covariates have opposite effects on time-varying expenditure and fixed effects. The time-varying component of total health expenditure is higher for females, those who speak a foreign language (not English) at home, have a health card, have low education and income and have skin cancer. As these characteristics are negatively associated with fixed effects, individuals with these characteristics are likely to be type III who are vulnerable to large swings in their health expenditure due to adverse health shocks. Older individuals, those in poor health, obese individuals, and those with diabetes have both high fixed effects and time-varying expenditure, suggesting that these groups of individuals are likely type I. On the other hand, sufferers of breast/prostate cancer, other cancer, heart disease, depression and broken bones have high fixed effects but lower time-varying expenditure suggesting that they are more likely to be type II having ongoing treatments.

The GP variables suggest that for each $1 increase in the average out of pocket charged by a family GP, time-varying expenditure is reduced by $0.80. This may be explained by frequent users tending to search for a GP with low or no co-payments.

Our study is unusual in having multiple years of administrative data to merge with survey data gathered in multiple years. It is useful for assessing how well survey information gathered in one year improves the predictive power of the diagnoses and pharmacy information in the same and different years. Since much of the information gathered in a survey varies only slowly from year to year (e.g., education, location, chronic conditions) it is of use to see whether its contribution to total predictiveness is especially large in the same or subsequent year in which the information is gathered. Table 5 presents various R-squares for various subsets of our total sample. Each row in the table corresponds to a different year in which the survey was conducted, while each column corresponds to a different year of the outcome variable. The first notable feature is that the 2008 data is more predictable than either 2007 or 2009, which we believe is due to the lower CV and skewness of spending in that year. The predictiveness of the survey information for a given year can be detected by how much higher the R-square is relative to the average for that sample year. Hence we see that the 2007 survey responses resulted in an above average predictiveness for 2008 than the average (0.649 versus 0.648) while for the 2008 survey the 2009 value is higher than the average. (0.649 versus 0.633). While detectable, the gain from using survey from the same year is modest, suggesting

that most of the value of adding survey information is in the time invariant factors rather than time changing information. This has important implications on how survey information may be useful as further modifiers for plan payment formulas and provider profiling in the context of risk adjustment models, suggesting that long- rather than short- duration information from surveys is more important.

Table 6 examines the predictive power of different types of information, both with and without individual level fixed effects using the adjusted R-square measure. The models without FE illustrate that models containing diagnoses (RCC), pharmacy (RX), and utilisation measures each improve on the modest predictive power of a simple model with just age and gender, which are available in nearly all administrative data and widely used in basic payment formulas. Adding the survey and year dummies increases the predictive power only slightly. The next three rows show that using the pharmacy alone explains 15% of the variation while, utilisation measures alone or survey information alone explain only 10% to 11%. The bottom three rows show that the survey information does about as well predicting spending in years one or two years before or after the survey as in the year of the survey itself. Finally the model with fixed effects shows that all of the observed measures, including time varying information, adds at most an additional 2% to predictions made using person-level fixed effects: 56% of health spending variation is across individuals, and only an additional 2% explained by time varying diagnoses, pharmacy and utilisation measures.

5 Conclusion

This paper exploits the linkage between a very large survey (n=262,293) of the over 45 population and a comprehensive set of health administrative records, to predict total health care expenditure for each individual. In doing so, we estimate the size of health subsidies received by an individual and add this to his/her private health expenditure. The subsidies account for more than 65% of total health expenditure.

We adopt a prospective approach, using health expenditure in the next year as the outcome variable. Considering several commonly used econometric models of health expenditure in the literature, we find that a linear model produces the best in-sample, as well as out-of-sample, fit of mean expenditure, conditional on predictors, compared to Gamma and log models, which tend to give too much weight to the small number of individuals with extremely high expenditures.

For the over 45 population, by isolating the expenditure that is time-invariant or intrinsic to an individual (fixed-effects), we find that those with high fixed-effects tend to be old, sick and engage in unhealthy lifestyles (having a smoking history and/or being obese). We do not find evidence that fixed effects are driven by a relationship with a general practitioner nor by fee-setting behaviour. In about 7% of cases we find high time-varying expenditure and high fixed-effects suggesting further deterioration in health. Together with the 14% who are likely to require ongoing treatments, these cases are expected to rely heavily on the hospital sector. Meanwhile, in another 14% of cases we find positive time-varying expenditure and low fixed-effects which may reflect vulnerability to an adverse health shock. They are likely to rely heavily on out-of-hospital services. Socio-demographic factors that predict high time-varying expenditure include low income, low education, poor health and being obese.

At a time when many are looking to predictive models to help inform providers and payers, or even to use for payment innovations (Ash and Ellis, 2012), the results have implications even to the US or Europe, since they suggest that much of the additional predictive power of survey information is from time-invariant or slow moving variables. Occasional rather than annual surveys may add most of the modest incremental predictive power from this source, and help predict subsequent years nearly as well as the survey year.

References

Ash, A.S., Zhao, Y., Ellis R.P., et al. 2001. Finding future high-cost cases: Comparing prior cost versus diagnosis-based methods. *Health Services Research* 26(6) Part II December 194-206.

Ash, A.S., Ellis R.P. 2012. Risk-Adjusted Payment and Performance Assessment for Primary Care. *Medical Care*. In press http://www.ncbi.nlm.nih.gov/pubmed/22525609

Australian Institute of Health and Welfare. 2010. Health system expenditure on disease and injury in Australia, 2004-05. *Health and welfare expenditure series no. 36*. Cat. no. HSE 87. Canberra: AIHW.

45 and Up Study Collaborators, 2008. Cohort profile: The 45 and Up study. *International Journal of Epidemiology* 37: 941 – 947.

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

Einav, L., Finkelstein, A., Ryan, S. and Cullen, M. 2011. Selection on moral hazard in health Insurance. SIEPR Discussion Paper No. 10-028. Stanford Institute for Economic Policy Research.

Hausman, J.A. and Taylor, W.E. 1981. Panel data and unobservable individual effects. *Econometrica* 49: 1377–1398.

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

Newhouse, J.P., Manning, W.G., Keeler, E.M. and Sloss, E.M. 1989. Adjusting capitation rates using objective health measures, and prior utilization. *Health Care Financing Review* 10 (3): 41–54.

Zhao, Y., Ash, A.S., Ellis, R.P., et al. 2005. Predicting Pharmacy Costs and Other Medical Costs Using Diagnoses and Drug Claims *Medical Care* 43(1): 34-43.

Table 1: Summary statistics of total health expenditure ($'000)

	2007	2008	2009
% zero	0.027	0.026	0.027
Mean	4.057	4.681	5.009
Std. deviation	8.614	9.619	11.79
Coef. of Variation	2.123	2.054	2.353
Skewness	9.688	8.096	11.10
5^{th} percentile	0.060	0.073	0.201
25^{th} percentile	0.565	0.661	0.652
Median	1.703	1.950	1.877
75^{th} percentile	4.200	4.832	4.700
95th percentile	15.24	17.86	19.71
N	262,293	262,293	262,293

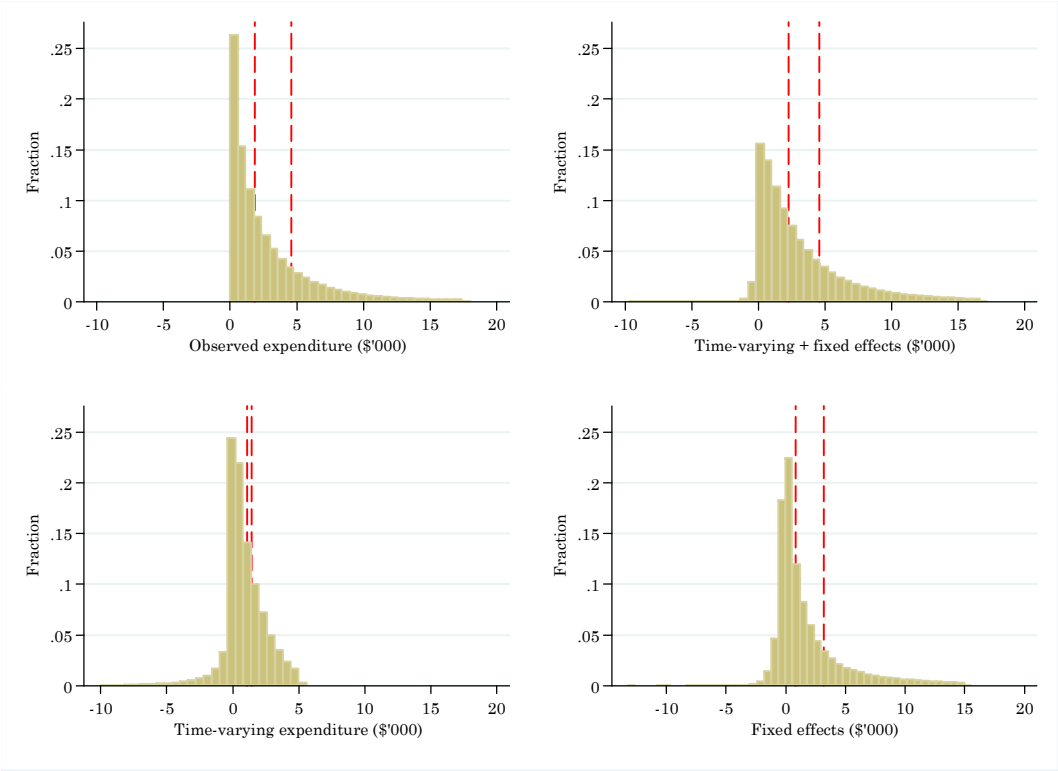
Note: Sample weighting is used which reflects the 45+ NSW population by region due to oversampling in regional areas. Values in constant $2009.

Table 2: Summary statistics of survey-based and GP variables

Demographic	Mean	**Demographic**	Mean	**Economic**	Mean	**Self-reported chronic conditions**	Mean	**Lifestyle**	Mean
Age: 45-19	0.137	Foreign born	0.121	Income: <$20k	0.189	SAH: excellent	0.147	BMI: underweight	0.013
Age: 50-54	0.163	Foreign language	0.278	Income: $20k-<$30k	0.087	SAH: very good	0.352	BMI: normal	0.344
Age: 55-59	0.170	Skin: very fair	0.157	Income: $30k-<$40k	0.074	SAH: good	0.326	BMI: overweight	0.363
Age: 60-64	0.146	Skin: fair	0.543	Income: $40k-<$50k	0.070	SAH: fair	0.118	BMI: obese (I)	0.146
Age: 65-69	0.119	Skin: light olive	0.249	Income: $50k-<$70k	0.102	SAH: poor	0.021	BMI: obese (II)	0.042
Age: 70-74	0.087	Skin: dark olive	0.016	Income: >=$70k	0.254	SAH: missing	0.037	BMI: obese (III)	0.017
Age: 75-79	0.065	Skin: brown	0.024	Income: missing	0.223	High blood pressure	0.356	BMI: missing	0.075
Age: 80+	0.113	Skin: black	0.001	Full time	0.352	Skin cancer	0.277	Ever smoke	0.422
Male	0.467	Skin: missing	0.010	Part time/other work	0.143	Breast/prostate cancer	0.059	Alc: no drink	0.350
Married	0.684	**Region**		Fully retired	0.364	Other cancer	0.062	Alc: low/moderate	0.490
Never married	0.066	Remote	0.094	Disabled	0.039	Heart disease	0.120	Alc: risky (>2/day)	0.137
Widowed	0.090	Outer region	0.079	Not in Labour Force	0.100	Stroke	0.031	Alcohol: missing	0.023
Divorced	0.073	Inner region	0.156	PHI with extra	0.515	Diabetes	0.091	**GP**	
Separated	0.028	Major city	0.671	PHI without extra	0.142	Asthma	0.024	Family GP	0.279
Unknown	0.006			No PHI	0.343	Depression	0.126	Mean OOP family GP	1.754
Partner	0.053			Health card	0.281	Broken bone	0.117		(5.706)
High school	0.132							Mean OOP other GP	3.315
Certificate	0.314								(6.820)
Trade/diploma	0.310								
University	0.244								

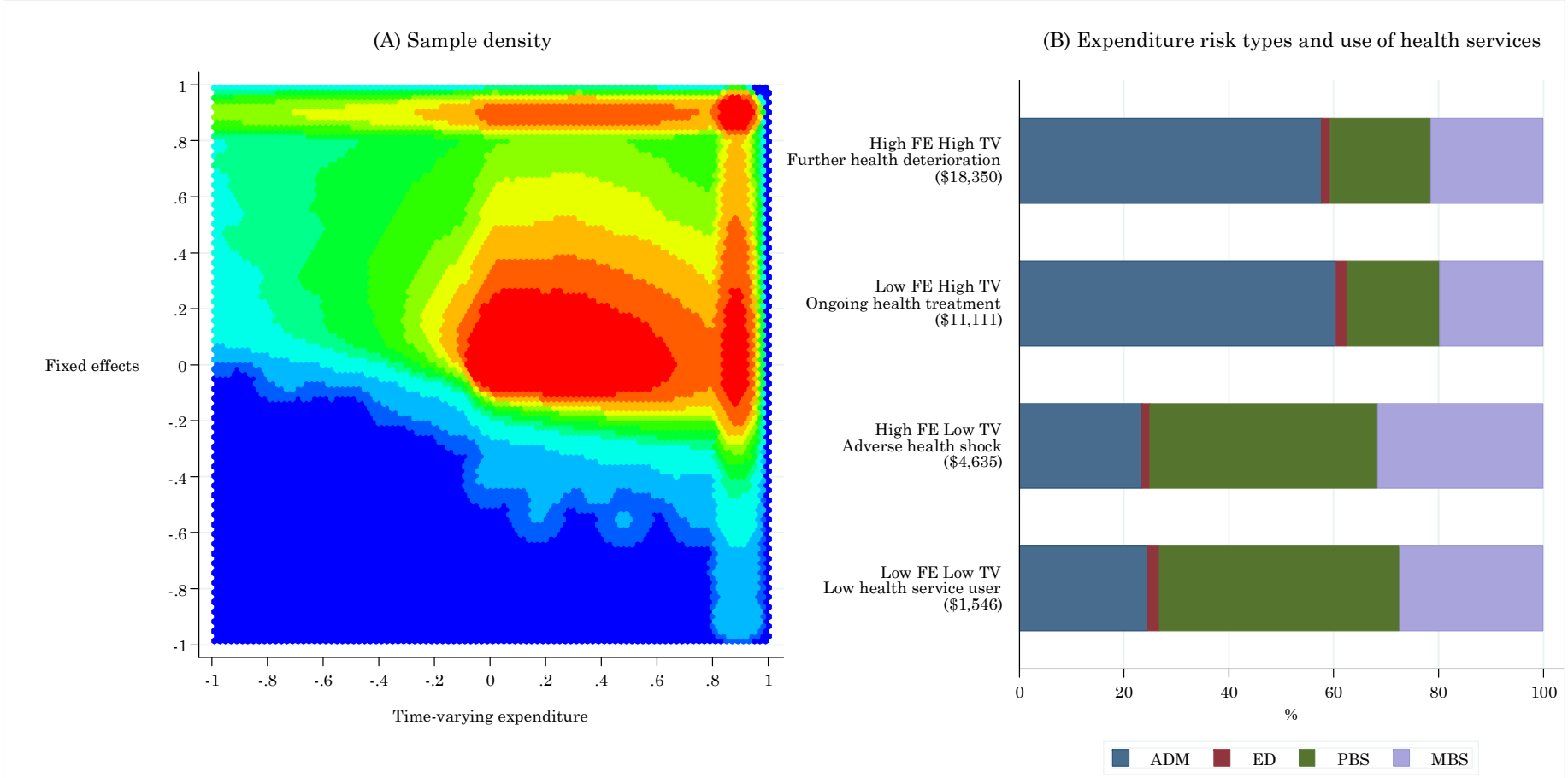
Note: sample weighting is used which reflects the 45+ NSW population by region due to oversampling in regional areas. Figures are sample proportions, except mean OOP family GP and other GP which are continuous variables. Standard deviations are provided in parentheses. SAH stands for self-assessed health. Where a variable contains missing values these values are flagged by a dummy variable which is included in the regressions.

Figure 1: Distribution of observed and predicted health expenditure



Note: distributions are truncated at the 95th percentile. The two vertical lines mark the median and the mean (which is always bigger than the median). Fixed effects estimation is equivalent to estimating a linear model by OLS with a dummy variable for each individual. The estimated individual fixed effects are restricted to sum to zero or equivalently are parameterised as deviations from their overall mean which appears as the estimated intercept in the fixed effects estimation. This intercept is included with the fixed effects plotted in the bottom-right quadrant rather than the time-varying component in the bottom-left quadrant.

Figure 2: Sample density, expenditure risks and use of health care services



Note: in panel (A), the axes represent health expenditures scaled between -1 and 1. Blue region indicates no observation. Red region indicates the highest density of observations. Approximately, 'High FE and High TV' is indicated by the region where the y-axis value is above 0.4 and the x-axis value is above 0.6. In panel (B), the mean total expenditure for each type is reported in parentheses. Further health deterioration comprises 6.6% of the sample, ongoing treatment is 13.6%, adverse health shock is 13.6% and low health service user is 66.2%.

Table 3: Fixed effect regression results ($'000)

	Coefficient		Coefficient		Coefficient
Chronic conditions (from claims)		Forearm & hand disorders and injuries	-0.3160**	Stroke	-0.2804
Infections	-0.7490***	Hemorrhagic conditions	-1.3390**	Post-stroke paralysis	-1.7790***
Solid tumors	0.8787***	Anemia	-0.8747***	Sequelae of cerebrovascular events	-0.2429
Blood and lymph neoplasm	0.9977**	Disorders of immunity	0.4199	Cerebro-vascular impairment	-1.9284***
Carcinoma in situ	1.1594***	Cognitive disorders	-1.6290***	Peripheral atherosclerosis	-0.2069
Benign/uncertain neoplasm	-0.0534	Drug abuse	-1.4541**	Other peripheral-vascular conditions	-0.7292***
Other neoplasms	-0.3551***	Alcohol abuse	0.1987	Thrombosis/phlebitis	-0.4976**
Diabetes by co-morbidity level	0.1196	Tobacco use	1.6642***	Lung intervention and complications	-24.7549***
Type i diabetes	-1.9508***	Personality disorders	-4.9541***	Lung infection	-1.0124***
Malnutrition	-4.9553***	Other mental conditions	-0.1818	Lung congestion and effusion	-4.1343***
Hyperlipidemia and lipidoses	-0.2876*	Psychoses	-1.8708***	Lung fibrosis	-1.2857***
Endocrine conditions	-0.2143	Eating disorders	5.9107*	Other lung conditions	-1.1122***
Excess weight	-0.8810***	Mood and anxiety disorders	-1.1454***	COPD and asthma	-0.3793*
Other nutritional and metabolic conditions	-1.0057***	Chromosomal and developmental disorders	-2.1460*	Diabetic/other retinopathy	0.4414
Liver intervention and complications	-8.6832***	Neurological trauma	-2.1494***	Blindness	-2.9524*
Liver failure	0.2787	Paralysis and coma	0.5837	Eye infection and inflammation	-0.7457
Billary and gallbladder conditions	-1.2441***	Seizure disorders	-0.8031*	Eye intervention and complications	-0.4115
Hepatitis	-1.0305**	Degenerative neurological conditions	0.464	Other eye conditions	-0.5778***
Alcoholic liver, cirrhosis, and infarct	-0.0374	Myoneural conditions	-3.3034	Significant ENT disorders	-0.4675
Peptic ulcer and related conditions	-1.0311***	Other neurological conditions	-0.7921***	Hearing impairment	-0.1096
Other gastrointestinal conditions	-0.2598***	Headache	-0.0834	Other ENT disorders	-0.2616*
Pancreatic disorders	-1.3404***	Respiratory arrest	-5.8156***	Urinary system intervention and complications	11.2432***
Inflammatory bowel disease	-0.1185	Cardiac arrest	-6.9949***	Chronic kidney disease and failure	-0.8159***
Knee disorders and injuries	-1.3363***	Cardiovascular intervention and complications	-3.1802***	Bladder and other urinary conditions	-0.1114
Hip disorders and injuries	-2.3461***	Coronary artery disease	-0.4686***	Nephritis	3.0746***
Back disorders and injuries	-0.4237**	Congestive heart failure	0.0696	Urinary system infection	-0.4555**
Other musculoskeletal conditions	-1.1196***	Heart valve and pericardial conditions	-0.8013***	Female genital conditions	-0.4636***
Musculoskeletal infection	-2.6152***	Congenital heart conditions	2.6986***	Male genital conditions	-0.6157***
Inflammatory musculoskeletal conditions	-0.4461	Cardiac arrhythmias	-0.7535***	Completed/terminated pregnancy	0.6496
Lower leg & foot disorders and injuries	-0.032	Other heart conditions	-1.5969***	Other pregnancy	4.8321
Shoulder & upper arm disorders and injuries	-0.4362*	Hypertension	-0.1881*	Uncompleted pregnancy	-1.8893

Note: *, **, and *** indicate statistical significance at 5%, 1% and 0.01% respectively. This result corresponds to equation (3). The sample size is 786,879.

Table 3: Fixed effect regression results ($'000) (continued)

	Coefficient		Coefficient		Coefficient
Severe burns	4.7988**	NRTIs	1.2021	Potassium-sparing diuretics	-0.4252
Skin ulcers	-0.8052**	Hepatitis treatments	2.4653	Aldosterone antagonists	-0.2988
Other skin conditions	-0.2333*	Miscellaneous antivirals	-25.5440***	Pulmonary hypertension drugs	12.0767***
Head injury	-0.7159*	Amebicides	2.848	Thiazide diuretics	0.1735
Traumatic amputation	2.4907	Anthelmintics	0.3966	Vasodilators	1.2066***
Other injuries	-0.5936***	Antimalarials	-0.6125	Vasopressors	-0.214
Poisoning	-0.7605**	Leprostatics	-0.436	Alzheimer's/age related dementia	1.6913***
Symptoms	0.0081	Anticoagulants (warfarin)	-0.5844***	Antidepressants (SSRI)	-0.2355**
Bone marrow transplant and complications	-21.1070***	Anticoagulants (non-warfarin)	-0.9914***	Antidepressants (non-SSRI)	0.0267
Artificial openings	-4.3890***	Antiplatelet agents	-0.1705*	Antipsychotics	-3.1878***
Amputation status	-4.2822***	Thrombolytics	0.1717	Atypical antipsychotics	-1.2848***
Other v-codes	-1.6694**	Antihemophilic agents	0.5351	ADHD and narcolepsy	0.8259
Other transplant status and complications	-2.8997***	Antineoplastics and chemotherapy adjuncts	-0.4783***	Anxiolytics	0.0067
Chemical and radiation oncology	-0.1514	Hormonal antineoplastics	-0.0202	Sedatives and hypnotics	-0.2455**
Other screening and history	-0.4524***	Colony stimulating factors	-3.9505***	Smoking cessation aids	0.7292***
Post-procedural conditions	-1.6586***	Recombinant human erythropoietins	8.9563***	Antiparkinson agents	0.3867
Implant and device complications	-2.8729***	Miscellaneous iv solutions	-2.8876***	Anticonvulsants (oral)	-0.0375
Other complications	-2.6904***	Vaccines	-0.3434***	Multiple sclerosis agents	3.9884***
Pharmaceuticals (from claims)		Antidotes and drug abuse treatments	17.4216**	Agent for cerebral swelling	0.5537*
Antigout agents	0.4183***	Angiotensin converting enzyme inhibitors	-0.1104	Lou Gehrig's disease (ALS)	3.6576
Headache medication	0.2346	Angiotensin ii inhibitors	0.7654***	Skeletal muscle relaxants (oral)	0.5934
Narcotic analgesics	0.3033***	Antiadrenergic agents, centrally acting	0.6730**	Cholinergic muscle stimulants	0.2636
High-potency narcotic analgesics	-0.4245***	Antiadrenergic agents, peripherally acting	0.7001***	Antipsoriatics	-0.349
Nonsteroidal anti-inflammatory agents	0.1566***	Antianginal agents	0.2984**	Topical acne agents	0.758
Miscellaneous analgesics	0.4312***	Antiarrhythmic agents	-0.4821**	Topical anti-infectives/antifungals/antivirals	-0.3496
Lipid lowering agents (statin)	-0.0099	Antihypertensive combinations	0.4309***	Topical antineoplastics/antiwart	-0.0371
Lipid lowering agents (non-statin)	0.4511***	Beta-adrenergic blocking agents	-0.2504**	Topical steroids/antipruritics	0.0664
Injectable anti-infectives	-1.7763***	Calcium channel blocking agents	0.3384***	Glaucoma agents	0.8988***
Anti-infectives (oral)	0.0819*	Inotropic agents	0.0095	Miscellaneous ophthalmic drugs	0.3957***
Azole antifungals (oral)	-0.9479**	Loop diuretics	0.4571***	Ophthalmic steroids	-0.2512**
Anti-herpetics	-0.0844	Peripheral vasodilators	-2.6446	Ophthalmic anti-infectives	-0.005

Note: *, **, and *** indicate statistical significance at 5%, 1% and 0.01% respectively. This result corresponds to equation (3). The sample size is 786,879.

Table 3: Fixed effect regression results ($'000) (continued)

	Coefficient		Coefficient
Ophthalmic anti-inflammatory agents	-1.1920***	Gallstone solubilizing agents	2.2876
Otic anti-infectives	0.211	Inflammatory bowel disease	0.7367**
Otic steroids with anti-infectives	0.0382	Laxatives	-0.0231
Miscellaneous otic/nasal agents	13.2824***	Ulcer/GERD (PPI)	0.1738***
Adrenal cortical steroids (oral)	0.3828***	Ulcer/GERD (non-PPI)	-0.1221
Androgens and anabolic steroids	0.345	Chronic kidney disease	6.8661***
Contraceptives	0.1295	Female leuprolide (endometriosis)	47.4146***
Endometriosis/fibrosis agents (except leuprolide)	1.3074	Urinary antispasmodics	0.298
Fertility drugs	-1.3872	Antiarthritics	1.8705***
Hormone replacement therapy (female)	0.0898	Immunosuppressive agents	-2.1856***
Hyperprolactinemia treatments	-1.007	Intravenous nutritional products	11.1235***
Octreotide	18.6424***	Other vitamins, minerals, and supplements (injectable)	0.4983***
Osteoporosis treatments	0.1515	Other vitamins, minerals, and supplements (oral)	-0.4891***
Oxytocics	-5.0560**	Other vitamins, minerals, and supplements (other)	-0.3351
Thyroid drugs	0.0676	Cough, cold, allergy mix, antihistamine	1.6591**
Insulin	0.8951***	Cough, cold with narcotics	-0.1916
Oral diabetic agents	0.5325***	Major diagnostic testing	-0.1796
Asthma, COPD (oral)	1.0881*	OTC drugs	-0.4695**
Asthma, COPD (inhaled beta agonist)	0.2479***	Miscellaneous, recognized ATCs	0.4367*
Asthma, COPD (inhaled steroid)	-0.3301**	Ungrouped ATCs	-0.691
Asthma, COPD (inhaled other)	0.2841**	Missing ATC value	0.3097*
Methylxanthines	1.2879**	**Medical services**	
Anticholinergics/antispasmodics	-4.0479	Number of standard GP consultation (item 23)	0.1307***
Antidiarrheals	-0.2061*	Number of initial specialist consultation (item 104)	0.9718***
Antiemetics	-0.0313		
Digestive enzymes	-1.3284**	Constant	3.0472***

Note: *, **, and *** indicate statistical significance at 5%, 1% and 0.01% respectively. This result corresponds to equation (3). The sample size is 786,879.

Table 4: OLS regression of fixed-effects and time-varying expenditure

Dependent variable:	FE ($'000)	TV ($'000)		FE ($'000)	TV ($'000)		FE ($'000)	TV ($'000)
Age 50-54	0.1076	0.0681**	Skin: very fair	0.0301	0.0054	High blood pressure	0.0606	0.3830***
Age 55-59	0.2008***	0.0976***	Skin: light olive	-0.0271	0.0063	Skin cancer	-0.1844***	0.2312***
Age 60-64	0.6325***	0.2328***	Skin: dark olive	-0.0806	-0.0039	Breast/prostate cancer	1.8790***	-0.0043
Age 65-69	0.9635***	0.4912***	Skin: brown	-0.2309*	0.1588***	Other cancer	2.8652***	-0.1987***
Age 70-74	1.2618***	0.6749***	Skin: black	-0.5206	0.2733	Heart disease	2.2500***	-0.1826***
Age 75-79	1.8856***	0.7283***	Skin: miss	0.2136	0.0765	Stroke	1.7396***	-0.2789***
Age 80+	2.4689***	0.4798***	Health card	-0.0726	0.1170***	Diabetes	0.8358***	0.6243***
Male	0.0524	-0.2832***	No PHI	-0.8911***	-0.0535***	Asthma	1.0627	0.2830
Age 50-54 * male	-0.0378	-0.0629	PHI without extra	-0.3390***	-0.0008	Depression	0.4897***	0.0065
Age 55-59 * male	0.2397*	-0.0309	Income: <$20k	-0.2631***	0.2744***	Broken bone	0.9706***	-0.0588***
Age 60-64 * male	0.1756	-0.0114	Income: $20-<30k	-0.2076***	0.1551***	BMI: under	0.5250***	-0.2196***
Age 65-69 * male	0.6681***	-0.0411	Income: $30-<40k	-0.1641***	0.1300***	BMI: over	0.0280	0.0586***
Age 70-74 * male	0.9861***	0.0117	Income: $40-<50k	-0.1145	0.0355	BMI: obese (30-<35)	0.1546***	0.1333***
Age 75-79 * male	1.1136***	0.1162*	Income: $50-<70k	-0.0597	0.0149	BMI: obese (35-<40)	0.2862***	0.1322***
Age 80+ * male	1.1024***	0.1085**	Income missing	0.0846	0.1263***	BMI: obese (40+)	0.4783***	0.2199***
Never married	0.1750**	-0.1371***	Part-time/other work	0.1272**	0.0133	BMI: missing	0.0998	0.0485**
Widowed	0.5378***	-0.2843***	Fully retire	0.4151***	0.1734***	Ever smoke	0.3377***	0.0014
Divorced	0.1687**	-0.0505**	Disabled	3.2765***	-0.0202	Heavy drinking	-0.1808***	0.0227
Separated	0.0607	-0.0501	Not in LF	0.2120***	0.0268	No drink	0.2636***	0.0270*
Unknown	0.1116	-0.0510	Remote	-0.5543***	-0.2470***	Drink: missing	0.1150	0.0593
Partner	0.0005	-0.0824***	Outer region	-0.5315***	-0.2951***	Family GP	-0.0689	-0.0026
Foreign language	-0.4673***	0.1957***	Inner region	-0.3069***	-0.1642***	Mean OOP family GP	-0.0027	-0.0078***
Foreign born	-0.2178***	-0.0005	SAH: very good	0.2775***	0.0569***	Mean OOP other GP	0.0047*	-0.0008
Certificate	0.1001*	-0.1315***	SAH: good	0.9799***	0.1792***	Year = 2008	-0.2791***	-0.8120***
Trade/diploma	0.0456	-0.1387***	SAH: fair	2.7876***	0.2906***	Year = 2009	-0.2565	-2.0000***
University	0.1278*	-0.2043***	SAH: poor	6.1082***	-0.0417			
			SAH: unknown	1.8307***	0.2685***	Constant	0.5261***	1.5595***
						N	226,823	226,823
						R-sq	0.1483	0.0829

Note: *, ** and *** indicate statistical significance at 5%, 1% and 0.01% respectively. FE stands for fixed effects and TV stands for time-varying expenditures. The number of respondents is fewer than the full sample of 262,293 respondents because some respondents were surveyed in 2006.

Table 5: Relation between survey year predictiveness and year of health data

Survey/ Year	2007	2008	2009	All
2007	0.495	0.649	0.613	0.569
2008	0.513	0.641	0.649	0.587
2009	0.539	0.633	0.581	0.580
All	0.515	0.648	0.633	0.585

Note: The table presents R-squares from fixed effects regressions for various subsets of our total sample. Each row in the table corresponds to a different year in which the survey was conducted, while each column corresponds to a different year of the outcome variable.

Table 6: Explanatory power of alternative predictors: Adjusted R-square from various OLS models

Included variables	No FE	With FE
[none]	***	0.5631
Age + Gender	0.0466	0.5637
RCC	0.1117	0.5753
RCC + RX	0.1945	0.5769
RCC + RX + Utilisation measures	0.2382	0.5850
RCC + RX + Utilisation measures + Survey	0.2476	0.5850
RCC + RX + Utilisation measures + Survey + Year Dummies	0.2493	0.5876
RX	0.1544	0.5679
Utilisation measures	0.1097	0.5721
Survey	0.1027	0.5631
Survey + year dummies (Using years before survey sample only)	0.1176	***
Survey + year dummies (Using survey year sample only)	0.1274	***
Survey + year dummies (Using years after survey sample only)	0.0908	***

Note: RCC = DxCG related condition categories, RX = DxCG pharmacy groups. Except in the Age + Gender model, age is included as a Survey variable which is constant at the survey year. In the model with FE, Survey variables are effectively dropped because of perfect collinearity with the person dummy variable. The first model with FE, includes only person dummy variables. The sample size (N=786,879) is the same for all rows except for the final three, which use only the years before the survey was conducted (N=209,310), the year in which the survey was actually conducted (N=262,823), and the years after the survey was conducted (N=350,746), respectively.

Appendix

1. Costing hospital admissions

To impute the cost for a hospital episode, we follow the guidelines contained in The Costs of Care Standards 2009/10 released by NSW Health (document GL2011_007). The document can accessed at http://www.health.nsw.gov.au/policies/gl/2011/pdf/GL2011_007.pdf

The services covered by the Standards include acute admitted care, mental health care, sub- and non-acute care, intensive care and emergency department (ED) care. The goal is to cost each hospital separation. In line with this goal, some assumptions have to be made for cases that are not included in the Standards.

Hospital admission costs

Hospital admission costs use the cost weights by AR-DRG provided in The Standards' Appendix 1 and 2. AR-DRG is a patient classification scheme that provides a clinically meaningful way of relating the number and types of patients treated in a hospital to the resources required by the hospital. A cost weight is measure of the average cost of an AR-DRG, compared with the average cost of all AR-DRGs. The average cost of all AR-DRGs is given a cost weight of 1.0. The construction of the cost weights was done in 2006/07 using 2004/05 data by the NSW Department of Health and since then 'escalation factors' applying to the annual average cost have been published to reflect general price inflation in health treatments. After applying the appropriate escalation factor to the AR-DRG cost weights, adjustments are then made for extended length of stay (i.e. length of stay longer than the published 'high trim point' for that AR-DRG), ICU hours and use of ventilation machine. See the Standards for details.

In addition, The Standards group patients into 5 types: (1) psychiatric ward patients; (2) sub- and non-acute care patients (SNAP); (3) patients who were transferred to another facility in the same day or following day; (4) other same day patients; and (5) overnight acute patients and modifications are made to the cost weights. Table A-1 describes the procedure followed The Costs of Care Standards 2009/10 only provides average cost by AR-DRG for all hospitals. To accommodate cost variation across hospitals we use the disaggregation in average cost by peer code available in the previous year's summary publication, Episode Funding Policy 2008/09 – NSW by NSW Health (Document PD2008_063). This document, can be accessed at http://www.health.nsw.gov.au/policies/pd/2008/pdf/PD2008_063.pdf . It reports the average cost by principal referral (A1a or A1b), ungrouped acute (A3), major metropolitan (B1), major non metropolitan (B2) and district (C1). The costs for other hospital groups (C2 to F) are not separately defined and we assume the average cost for all hospitals. An index is then created to reflect the cost variation by hospital type relative to the overall average cost. For example, since the overall average cost was $3,700 in 2008/09 and the average cost for a Principal Referral Group A hospitals was $4,017, the index for Principal Referral Group A hospitals is 1.083.

We adopt the same imputation process for public and private hospitals. The Productivity Commission reports that public and private hospitals have similar average costs, although their composition may vary (Productivity Commission, 2009).

Table A-1:

Type	Freq.	Percent	Costing
Mental health unit (entire stay)	7,464	1.33	Length of stay (LOS) x cost of mental health unit/day ($745 in 2009/10)
SNAP	38,184	6.81	*Geriatric and maintenance care*: respective cost weight x cost per weighted SNAP activity ($11,582 in 2009/10) *Palliative care*: palliative cost weight x cost per weighted SNAP activity. Adjustment for extended LOS (high trim point/high outlier) applies. *Rehabilitation*: rehabilitation cost weight x cost per weighted SNAP activity. Adjustment for high trim point applies.
Transfer	13,717	2.45	Cost rule depends on whether AR-DRG is surgical/non-surgical and the presence of same-day cost weight.
Other, same day	300,150	53.56	Same-day cost weight x hospital average cost e.g. $4,256 in principal referral). Adjustment for high trim point applies.
Overnight cases	200,863	35.84	Overnight cost weight x hospital average cost. Adjustments for high trim point and some care in mental health unit apply.
All			Additional cost for ICU hours ($184 per hour)
All			Additional cost for mechanical ventilation ($95 per hour)
Total	560,378	100	

The following lists further assumptions used when applying the costing rules given in the Standards.

1) The Standards consider cases with maximum length of stay of 365 days. We top-code the extended length of stay at 365 days.
2) For some SNAP classes, the Standards apply low-trim points i.e. excluding cases with length of stay less than a given level. We do not apply this restriction.
3) A few AR-DRGs are indicated in the standards as 'error' (see the Standards for the list of error AR-DRGs) and we regroup them into non-"error" AR-DRGs, which have cost weight information. To do this we use the following imputation rules.
    a) When patients are treated in sub- and non-acute care units and therefore can be classified as SNAP, their costs are constructed according to the SNAP guideline. This affects 29 patients.
    b) For non-SNAP cases, when the major diagnostic group of the patients is known, we group them according to the AR-DRG within that major diagnostic group which has the closest median length of stay as the patient's length of stay. AR-DRGs related to transplants (i.e. those starting with 'A') are excluded. Using median length of stay avoids overestimation due to extreme observations. If several AR-DRGs have the same median, the one with the highest frequency is chosen. Mapping based on length of stay is appealing because cost is greatly affected by length of stay. 886 cases (0.158%) are affected by this alteration.
    c) For non-SNAP cases, when the major diagnostic group of the patients is not known, the median cost by length of stay in a given year is assumed. For long lengths of stay (more than 15 days), bands of length of stay are assumed because of the small number of cases in each year; that is, the median cost for patients with length of stay between specified

day ranges in a given year. Specifically, 6 classes are created: 15-<30 days, 30-<60 days, 60-<90 days, 90-<150 days, 150-<365 days and 365 days. If patients have ICU or ventilator hours, these costs are added to the median cost. 221 cases are affected by this alteration.

4) A few AR-DRGs are indicated in the standards as 'error' (see the Standards for the list of error AR-DRGs) and we regroup them into non-"error.
5) A few AR-DRGs are not listed in the Standards: B82C, C16Z, F42C, F76B, G05C, G10B, G11Z, G47B, G47C, G48C, I01B, I04A, I04B, I05B, I19B, I32C, I79B, J06Z, J07Z, K04B, M62Z, N06A, N06B, N62Z and U63Z. These AR-DRGs are recognised by the National Casemix and Classification Centre (NCCC) at the University of Wollongong which allows us to map them into AR-DRGs that are listed in the Standards. 77 cases (0.014%) are affected by this alteration.
6) In the case of missing AR-DRGs the median cost by length of stay in a given year is assumed. 9 cases are affected.
7) For the purposes of NSW funds allocation, the Standards regard private patients in public hospitals as lower cost to the NSW hospital system than their public patient counterparts, and apply a discount of 9% to private patients because they can be funded by other sources, principally private health insurance. Because we are applying the same cost weights for public and private hospitals, and have no information on how the cost of private patients in public hospitals differs to private hospitals, we do not apply this rule.
8) The Standards increase costs for individuals of Aboriginal or Torres Strait Islander origin by 10%. We do not make this adjustment because Aboriginal or Torres Strait Islander origin status is not disclosed in our data. However we note that Aboriginal or Torres Strait Islander respondents comprise only a very small proportion of the 45 and Up sample, and the adjustment is unlikely to create substantial bias (45 and Up Study Collaborators, 2007).

Nominal costs are deflated to constant 2009AUS$ using the CPI All groups Australia.

We compare the resulting imputed costs with those reported in the AIHW's Australian Hospital Statistics 2007-08 for public hospitals in NSW (Table S12.3). A simple linear regression of the AIHW costs by ARDRG (indexed to 2009) on the costs imputed for NSW using the procedure described above has an R-square of 0.9660 with a slope coefficient of 0.9399. We take the fit and correlation as providing support for the validity of our imputed hospital cost estimates. Our imputed costs have a lower mean and higher standard deviation than the AIHW estimates.

Tables A-2 and A-3 show some examples of cost weights and 45 and UP sample variation for several common AR-DRGs. For same-day AR-DRGs (e.g G46C), there is limited variation and so the NSW and sample cost weights match quite well. For severe AR-DRGs (e.g all AR-DRGs prefixed A) , there is more variation in cost weights, and generally the 45 and Up sample cost weights are larger than the NSW cost weights for the general population. This may be explained by the older population whose treatments may involve extended stays, ICU hours or the use of ventilators more often than younger patients.

Table A-2 AR-DRG Summary Statistics

AR-DRG cost estimate	Number	Mean	Standard. Deviation	Minimum	Maximum
AIHW	610	10,115.82	13939.96	532.476	154,246.9
45 and Up sample based on NSW cost weights	610	9, 539.61	14576.42	387.0055	171,178.6

Table A-3 AR-DRG Sample cost weights

		NSW Cost weight[(a)]	45 and Up admitted patient sample average		
AR-DRG			Cost weight	(Std Dev)	Cost (A$)
B05Z	Carpal tunnel release	0.54	0.54	(0.09)	$1,905
F62A	Heart failure and shock w catastrophic CC	2.52	2.49	(1.02)	$10,665
F62B	Heart failure and shock w/o catastrophic CC	1.15	1.08	(0.69)	$4,137
G46A	Complex gastroscopy w catastrophic or severe CC	2.72	3.03	(2.61)	$11,937
G46B	Complex gastroscopy w/o catastrophic or severe CC	1.44	1.45	(0.84)	$5,191
G46C	Complex gastroscopy, same day	0.43	0.43	(0.02)	$1,485
I03A	Hip revision w catastrophic or severe CC	9.51	9.93	(3.13)	$36,564
I03B	Hip replacement w catastrophic or severe CC or hip revision w/o catastrophic or severe CC	5.35	5.35	(0.86)	$19,479
I03C	Hip replacement w/o catastrophic or severe CC	4.03	4.03	(0.27)	$14,173
I04Z	Knee replacement and reattachment	3.97	4.66	(0.63)	$16,384
K60A	Diabetes w catastrophic or severe CC	2.12	2.18	(1.39)	$8,282
K60B	Diabetes w/o catastrophic or severe CC	0.99	0.88	(0.66)	$3,204
L61Z	Admit for renal dialysis	0.14	0.14	(0.12)	$505
N04Z	Hysterectomy for non-malignancy	1.74	1.74	(0.09)	$6,032
R63Z	Chemotherapy	0.17	0.17	(0.02)	$619

(a) Based on The Costs of Care Standards 2009/10. Cost is in A$2009.

Expenditure disaggregation

Because we do not have cost weights by AR-DRG for private hospital admissions we assume that the public hospital cost weights also apply to private admissions (supported by the Productivity Commission, 2009). However, the public hospital cost weights include medical and pharmaceutical services while private hospital costs do not and private patients are charged for these services with corresponding claims appearing in the MBS and PBS data.

To account for this, we modify the hospital admission costs of private patients, whether in public or private hospitals. Because the admission cost weights reflect the mean expenditure for a given AR-DRG, for each AR-DRG and hospital type, we subtract the mean in-hospital MBS and PBS claimed by private patients during inpatient periods in a given year ($\overline{INHOSPMBS}$ and

$\overline{INHOSPPBS}$), from the admission costs imputed using the public hospital cost weights and add the individual in-hospital MBS and PBS expenditures, ($INHOSPMBS$ and $INHOSPBS$).

We identify in-hospital MBS and PBS expenditures by date matching to a hospital admission (i.e. claims that fall during hospitalisation period). For MBS items we cross check the date matching with a variable in the data that indicates whether the claim is in-hospital or out-of-hospital. MBS claims that coincide with a hospital admission period but are indicated as out-of-hospital by the flag are treated as out-of-hospital and are not assigned to the in-hospital adjustment. MBS claims that are indicated as occurring in-hospital but which do not match an admission in the inpatient data are regarded as "cross-border" cases. Most of the individuals involved reside in border postcodes of NSW and may be admitted to hospitals in a bordering state. They represent 2.33% of the sample. Rather than ignoring the individual recorded expenditure, "cross border" in-hospital MBS claims are classified as admission costs.

In the PBS data, because there is no in-hospital flag we rely only on date matching to assign in-hospital and out-of-hospital pharmaceutical services. The absence of in-hospital indicator also precludes us from identifying cross-border inpatient cases.

Because individuals in the sample may use a mixture of hospitalisation modes in each year we assign to each person, four admission components which we sum to calculate the total in-hospital cost in a year. Ignoring person and year subscripts, we let *a* indicate a public hospital admission and *b* indicate a private hospital admission. We let *j* indicate a public patient admission and *k* indicate a private patient admission. For each person in a given year;

$$HOSPEXP_{aj} = ADM_{aj}$$

$$HOSPEXP_{bj} = ADM_{bj}$$

$$HOSPEXP_{ak} = ADM_{ak} + \overline{INHOSPMBS_{ak}} + INHOSPMBS_{ak} + I\,\overline{NHOSPPBS_{ak}} + INHOSPPBS_{ak}$$

$$HOSPEXP_{bk} = ADM_{bk} + \overline{INHOSPMBS_{bk}} + INHOSPMBS_{bk} + I\,\overline{NHOSPPBS_{bk}} + INHOSPPBS_{bk}$$

The out-of-hospital MBS and PBS expenditures then can be obtained from:

$$OOHMBS = MBS - INHOSPMBS_{ak} - INHOSPMBS_{bk}$$

$$OOHMPBS = PBS - INHOSPPBS_{ak} - INHOSPPBS_{bk}$$

2. Costing emergency department presentations

The costing of emergency visits (presentations) is relatively straightforward compared to the costing of hospital admissions. We follow the Standards and allow cost variation across hospital type by using the average costs by hospital peer group in the 2008/09 report. Apart from hospital type, the cost of ED presentation varies by triage category (more urgent category is more expensive) and whether or not it leads to admission (subsequently admitted cases are more expensive). We use the same imputation process for both private hospital and public hospital emergency presentations. Private hospital emergency presentations comprise less than 2% of the total.

3. Annual hospital costs

Total hospital expenditures are the sum across all four admission types and the cost of emergency presentations.

$$TOTHOSPEXP = HOSPEXP_{aj} + HOSPEXP_{ak} + HOSPEXP_{bj} + HOSPEXP_{bk} + ED_a + ED_b$$

4. Annual heath costs

Total expenditure for person in a year is obtained by summing hospital, emergency presentation expenditures and out-of-hospital expenditures.

$$TOTEXP = TOTHOSPEXP + OOHMBS + OOHPBS$$

5. Further notes on statistical method used

In our prospective model, the dependent variable is next year's expenditure. Let y_{it} be health expenditure of individual i=1, ... ,*n* in survey year t=1, ... ,T. Our baseline specification is the following linear model estimated by ordinary least squares (OLS):

(8) $y_{it} = \beta_0 + x_i'\beta_1 + z_{it-1}'\beta_2 + u_{it}$,

where x_i is a vector of time invariant (or slow changing) covariates (e.g. background and socio-economic characteristics, location, general practitioner (GP) characteristics), z_{it} is a vector of time-varying covariates including chronic conditions, various pharmaceuticals and variables capturing medical service use, β_k are conformable parameter vectors to be estimated and u_{it} is the error term.

The standard assumption used in the fixed effects model is that u_{it} can be decomposed into individual-specific effects α_i and a random component ε_{it} as follows:

(9) $u_{it} = \alpha_i + \varepsilon_{it}$.

Under this specification, we are able to consistently estimate β_2, the coefficients of the time-varying parameters, irrespective of any correlation between the covariates included in equation (A1) and time invariant omitted variables. The fixed effects estimator can be thought of as applying OLS to the within transformed model (all variables are expressed in terms of deviations from their sample means calculated over time for each individual) or, equivalently, to a model that includes individual specific constants for all *n* individuals in the data. While providing estimates for β_2, the parameters associated with the time invariant parameters, β_1, will not be identified. However, it is possible to recover estimates for the individual effects. Implementing this procedure, the individual effects are restricted to sum to zero or equivalently are parameterised as deviations from their overall mean which appears as the estimate of β_0 in the fixed effects estimation. Therefore our original linear model with fully flexible fixed effects is written as

(10) $y_{it} = \beta_0 + z_{it-1}'\beta_2 + \alpha_i + \varepsilon_{it}$.

Within estimation proceeds by estimating the following transformed model by OLS:

(11) $(y_{it} - \bar{y}_i + \bar{\bar{y}}) = \beta_0 + \left(z_{it-1} - \bar{z}_i + \bar{\bar{z}}\right)'\beta_2 + (\varepsilon_{it} - \bar{\varepsilon}_i + \bar{\bar{\varepsilon}})$

providing estimates of β_0 and β_2, and estimates of the fixed effects can be recovered using

(12) $\hat{\alpha}_i = \bar{y}_i - \hat{\beta}_0 - \bar{z}_i'\hat{\beta}_2$

where $\hat{\beta}_0$ and $\hat{\beta}_2$ are fixed effects estimates, $\bar{y}_i$ and $\bar{z}_i$ are sample means averaged over the time series observations for each individual and $\bar{\bar{y}}$ and $\bar{\bar{z}}$ are the grand means. Also note that

(13) $\hat{\beta}_0 = \bar{\bar{y}} - \bar{\bar{z}}'\hat{\beta}_2.$

Post estimation, predictions of individual expenditures using these parameter estimates can be constructed as follows:

(14) $\hat{y}_{it} = \hat{\beta}_0 + z_{it-1}'\hat{\beta}_2 + \hat{\alpha}_i$

and will comprise two components; the first we term "time-varying" and the second "fixed effects" expenditure. One natural definition of the time varying component is defined by:

(15) $\tilde{y}_{it}(1) = z_{it-1}'\hat{\beta}_2$

which implies a resulting fixed effects expenditure component defined for each individual as:

(16) $\hat{\alpha}_i + \hat{\beta}_0 = \bar{y}_i - \bar{z}_i'\hat{\beta}_2.$

Alternatively consider the time varying component defined by:

(17) $\tilde{y}_{it}(2) = \left(z_{it-1} - \bar{z}_i\right)'\hat{\beta}_2.$

Here the fixed effect expenditure becomes:

(18) $\hat{\beta}_0 + \hat{\alpha}_i + \bar{z}_i'\hat{\beta}_2 = \hat{\beta}_0 + \bar{y}_i - \hat{\beta}_0 - \bar{z}_i'\hat{\beta}_2 + \bar{z}_i'\hat{\beta}_2 = \bar{y}_i.$

The fixed effect results provide estimates and predictions with minimal assumptions regarding the fixed effects and their relationship with included covariates but they come at the cost of not being able to identify the impact of time invariant variables. In terms of equation (A1) we have sacrificed estimation of β_1. Suppose we add back in these time invariant variables into (A3) to give:

(19) $y_{it} = \beta_0 + x_i'\beta_1 + z_{it-1}'\beta_2 + \alpha_i + \varepsilon_{it}.$

Treating the $\alpha_i$as fixed parameters to be estimated along with β_0 , β_1 and β_2 is not possible because of the perfect collinearity problem. Another possibility is to assume that the α_i are random which yields a random effects specification. The validity of this second specification relies on the tenuous assumption that the random effects are uncorrelated with both x_i and z_{it}. Note that in either of these situations one can still estimate β_2 (and β_0) as before using the within estimator associated with (A4). This suggests a two-step procedure that allows estimation of β_1. Note that the mean outcome averaged over time from (A12) yields:

(20) $\bar{y}_i = \beta_0 + x_i'\beta_1 + \bar{z}_i'\beta_2 + \alpha_i + \bar{\varepsilon}_i$

implying

(21) $(\bar{y}_i - \bar{z}_i'\beta_2) = \beta_0 + x_i'\beta_1 + \alpha_i + \bar{\varepsilon}_i$

and therefore as a second step we replace β_2 by the within estimator and treat $(\alpha_i + \bar{\varepsilon}_i)$ as the disturbance. Operationally this means estimating the following model by OLS:

(22) $(\bar{y}_i - \bar{z}_i'\hat{\beta}_2) = (\hat{\alpha}_i + \hat{\beta}_0) = \beta_0 + x_i'\beta_1 + \omega_i$

which produces consistent estimates of β_1 under an assumption weaker than that associated with random effects, namely that the α_i are uncorrelated with x_i but not necessarily with the z_{it}. More formally, under these assumptions this procedure produces a Hausman and Taylor (1981) estimate of β_1. Equation (15) is what is used in the main text to estimate β_1.