Estimating Health Care Expenditure in Scotland: A review of modelling and research design for an ageing population

Claudia Geue, Paula Lorgelly, Andrew Briggs
Section of Public Health and Health Policy, University of Glasgow, Tel: +44 (0)141 330 3294
E-Mail: c.geue.1@research.gla.ac.uk

Abstract

Background: Population ageing and its impact on public expenditure is a major concern for developed countries. Compared to the rest of the UK, population ageing in Scotland will be even more pronounced due to differences in the demographic pattern. It is a general perception that older people use more health care services as health deteriorates with increasing age; however, a number of recent studies have shown that proximity to death may be a much stronger indicator for health care expenditure than age alone. There is an ongoing debate in the literature regarding the most appropriate modelling approach; robust estimation is important, especially if the analysis is to inform budgeting.

Aim: This paper seeks to review and summarise methods of estimating health care expenditure when an individual's age and time to death is controlled for in order to inform the design of a subsequent study, particularly in terms of available data linkage options.

Methods: A literature review was undertaken, summarising the main contributions to the research field. Strengths and weaknesses of different approaches are outlined and discussed. Current data linkage methods for Scottish health data are explored and alternatives with respect to our particular research question proposed.

Results: A mix of methodologies has been applied in different national studies, leading to varied results. This paper has highlighted issues in terms of sample selection, econometric modelling and data requirement.

Discussion: Many issues remain unanswered including the potential for bias by focussing on decedents only; issues of endogeneity in the modelling approach and the ability to project future expenditure.

1. Introduction

Population ageing is a major concern for developed countries in terms of the amount of public funding that will need to be made available in the forthcoming decades to pay for health care, long-term care and pensions. Population ageing is generally caused by two factors: decreasing mortality rates, as people live longer, and decreasing birth rates. It is usually described as an increase in the proportion of people over the age of 65 in a population (Payne et al. 2007). The general perception is that older people use more health care as health deteriorates with increasing age (Payne et al. 2007).

With the largest population cohorts approaching the age of 65, it has previously been estimated that costs for public services such as health care (HC), long term care (LTC) and pensions will increase. One study has suggested that population ageing will increase age-related social expenditure from under 19% of GDP in 2000 to almost 26% of GDP by 2050 (Dang et al). It is estimated that expenditure on HC and LTC will be responsible for approximately half of this increase.

Compared to the rest of the UK, population ageing in Scotland will be even more pronounced due to differences in the demographic pattern. The Scottish population is expected to rise over the next two decades and peak in 2031 followed by a decline thereafter. This will result in an even higher proportion of the very old in the population (GRO for Scotland, 2007).

The first research to establish the relationship between population ageing and HC expenditure was undertaken by Abel-Smith and Titmuss in 1956. When forecasting future expenditure needs they assumed that demographic changes would be the only factor influencing the cost of the National Health Service and other factors, such as incidence and type of disease and type and quality of treatment would remain unchanged (Abel-Smith and Titmuss, 1956). The authors estimated future HC expenditure combining population projections from the Registrar-General with Census data on the proportion of the population in hospital by sex and age.

Over the last two decades though there has been a growth in literature that takes into account factors other than population ageing *per se* to explain and estimate future HC spending. Fuchs, one of the earlier examples of this, found that cross-sectional differences in HC expenditures by age overestimate the changes that would result from an ageing population (Fuchs 1984). Using Medicare data he showed that HC spending is more a function of time to death (TTD) than it is of age and that the reason why expenditure

increases with age in cross-sectional data is the increasing proportion of people near death. Grouping people as survivors or decedents Fuchs showed that after adjusting for age-sex differences in the survival group, most of the age-related increase in expenditure is eliminated.

A number of more recent studies that considered proximity to death (PTD) in addition to age and how this affects HC expenditure have used more advanced econometric modelling techniques and there now seems to be a general agreement in the literature as well as amongst policy makers that age alone does not drive HC expenditure. The association between age and HC costs seems to reflect a possibly stronger relationship between PTD and HC expenditure which seems to be a much better predictor of acute HC costs than population ageing *per se*. A conclusion that can be drawn from these findings is that the inclusion of the additional factor "time to death" or "proximity to death" in resource allocation formulae may provide a more robust estimate for future HC expenditure. If the remaining time to death is not accounted for, these future costs could be overestimated.

This paper will proceed by outlining general modelling techniques of health care expenditure and how these have been applied in different national studies. This will be followed by a presentation of research questions and an exploration of data linkage methods to answer these questions.

2. Modelling Techniques for HC Expenditure Estimation

Different methodological approaches have been used to explain the relationship between age, time to death and expenditure. Some studies have used descriptive methods to compare HC expenditure for survivors and decedents (McGrail et al. 2000; Busse et al. 2002; Polder et al. 2006). More advanced methods were applied by other authors, who used regression analysis, which allows for the estimation of several independent factors simultaneously (Zweifel et al. 1999, 2004; Seshamani and Gray 2004a, 2004b; Hakkinen et al. 2008). However, as expected, different methods of data collection, sample selection, coverage of HC costs and model estimation have lead to varying results. It is important to consider all these issues so to identify the most appropriate approach, otherwise we reduce our ability to accurately inform budgeting.

2.1 Modelling Techniques – Theory

In the first step of building a model, individuals with no activity (no HC utilisation) are usually distinguished from individuals with activity (HC utilisation), and the probability of using HC services in a defined time period (usually a year) is estimated. This regularly leaves a high proportion of people who do not access HC services and therefore not incur any costs. The high number of zero cost observations is an issue that needs to be addressed in the second step of the model, as including zero cost observations leads to a highly skewed distribution of the cost variable. Left-censoring of the sample and excluding zero cost observations on the other hand could lead to biased and overestimated results as ill people may be over-represented in the model.

Different time periods have been used to analyse at which time point from death individuals may become more expensive to the HC system (Zweifel et al. 1999; Seshamani and Gray 2004; Hakkinen et al. 2008), while different groups of users and non-users of HC services have also been compared in these studies.

In general there are four groups of individuals that are of potential interest for the analysis of HC expenditures in a defined period of time; survivors and decedents and HC users and non-users, as described below:

	HC utilisation	**No HC utilisation**
Decedents	Decedents who utilised HC services	Decedents who did not utilise HC services
Survivors	Survivors who utilised HC services	Survivors who did not utilise HC services

The following sections summarise, in chronological order, the main contributions in this research area and compare the different approaches.

2.2 Modelling Techniques – The "Old Swiss Way"

In their first study, Peter Zweifel and colleagues used longitudinal data from a Swiss sickness fund and examined the effect that age has on HC expenditure after controlling for TTD. The data are available for a period of 10 years, based on two data sets of HC expenditure records of deceased individuals (Zweifel et al. 1999). Zweifel et al estimated costs for decedents in their last two (and five) years of life separately for individuals aged 65 and over and for individuals under the age of 65. Decedents who did not access HC in the last two years prior to death (zero cost observations) are excluded from the second step of their estimation. Survivors with or without cost observations were also excluded from the analysis. As data comes from sickness funds, HC costs included in the model extend beyond hospital costs and also include other expenditure, e.g. for primary care. The authors employed a Heckman sample selection model, estimating a decedent's probability of HC utilisation in the first step using probit regression. They calculated the inverse Mills ratio *λ.* Positive HC expenditures (log transformed to mitigate skewness of the data) were estimated using OLS in the second step. The second step included the same regressors as the first step of the model and *λ* was included to account for potential selection bias, which may have been caused by excluding zero cost observations.

Model specification:

First part: not presented in original paper

Second part:

$$\ln(HCE) = \beta_0 + \beta_1 A + \beta_2 A^2 + \beta_3 SEXF + \beta_4 (A * SEXF) + \beta_5 INS + \beta_6 \lambda + \sum_{q=1}^{7} \gamma_q Q_q + \sum_{t=1982}^{1992} \delta_t * Y_t + \varepsilon$$

A: calendar age in years, *SEXF*: gender dummy (1=female), *INS*: dummy (1=extra insurance), *λ*: inverse of the Mill's ratio (capturing impact of potential sample selection issues), Q_q: dummy indicating quarter before death, Y_t: year dummy

Zweifel et al acknowledge that excluding deceased individuals, who did not incur costs in their last 2 years of life, may bias their results, as sick individuals may be over-represented and also incur higher costs. They argued that the bias would only be relevant if the inverse Mills ratio λ turns out to be significant in the second step. If this sample selection test is insignificant though, Zweifel et al suggested that excluding zero cost observations does not cause bias. Even if there was bias, the authors argued this would be corrected by including λ in the second part of the estimation.

Their findings revealed that for deceased individuals aged 65 and over, in their last two years of life, HC expenditure does not depend on age when controlling for closeness to death (Zweifel et al. 1999).

2.3 Modelling Techniques – The "English Way"

Seshamani and Gray criticise the model selection of Zweifel et al, and used longitudinal hospital data from Oxfordshire to replicate the Swiss model (ZFM model) (Seshamani and Gray 2004a). Their data include people, who were aged 65 and over in 1970 and died during the follow-up period until 1999. The first part of their paper replicated the ZFM model as closely as possible, but with differences in the cost variable. Seshamani and Gray have used hospital costs whereas Zweifel et al have used all HC costs that are covered by the sickness fund. In their replication they found that neither age nor PTD had a significant effect on hospital costs. Seshamani and Gray attribute weaknesses in the ZFM model to model selection. Their main criticism is that multicollinearity is present when using a Heckman model without an 'exclusion restriction' (a regressor that is significant in the selection part of the model but not in the second part of the model which estimated the level of costs). Seshamani and Gray further argued that a sample selection model is only necessary if the selection process is unobserved and can therefore not be modelled separately. All zero cost observations in the ZFM model are observed. The model therefore treats zero cost observations as missing data and not as zero costs. Consequently the model may not inform actual budgeting for HC (Seshamani and Gray 2004a).

Seshamani and Gray proposed an updated model, which included zero cost observations of decedents and used an extended two-part model introducing additional variables (diagnosis, source of admission, place of discharge and marital status) for the second part. Their final criticism of the ZFM model relates to the fact that each quarterly cost observation was treated as an independent observation not allowing for correlation between observations of the same patient when using cross-sectional analysis. Seshamani and Gray also suggested clustering by patient to derive more accurate standard errors. Results of their updated model showed that age as well as TTD had a significant effect on hospital costs.

An updated version of their first paper uses random effects regression in the second step of the estimation to take account of the longitudinal nature of the data (Seshamani and Gray 2004b)

Model specification:

First part:

$$\Pr(HCE > 0) = \alpha + \beta_1 A + \beta_2 A^2 + \beta_3 S + \beta_4 A * S + \sum_{q=2}^{24} \gamma_q Yr_q + \sum_{t=1971}^{1999} \delta_t Y_t + \sum_{c=2}^{5} \chi_c C_c + \sum_{s=2}^{5} \zeta_s Soc_s$$

A: patient age, *S*: patient sex*, Yr:* number of years remaining until death*, Y:* calendar year*, C:* cause of death*, Soc:* social class

Second part:

$$\ln(HCE) = \alpha + \beta_1 A + \beta_2 A^2 + \beta_3 S + \beta_4 A * S + \sum_{q=2}^{20} \gamma_q Yr_q + \sum_{t=1971}^{1999} \delta_t Y_t + \sum_{c=2}^{5} \chi_c C_c + \sum_{a=2}^{7} \phi_a Adm_a$$

$$+ \sum_{d=2}^{7} \pi_d Dis_d + \sum_{m=2}^{4} \mu_m M_m + \sum \delta_x Dx_x + \sum \varsigma_s Soc_s$$

As above with the addition of, Adm: source of admission*, Dis:* place of discharge*, M:* marital status*, Dx:* hospital episode diagnosis

2.4. Modelling Techniques – The "New Swiss Way"

In reply to the methodological critique of their first model, Zweifel et al updated their analysis, addressing some of the issues that were raised by Seshamani and Gray and Salas and Raftery (Seshamani and Gray 2004a; Salas and Raftery 2001). Addressing the issue of multicollinearity between λ and the exogenous regressors the authors agree that the second part of a Heckman selection model is only identified when λ is a non-linear function of the regressors (Zweifel at al, 2004), which was not the case in their first analysis in 1999.

Salas and Raftery had argued that TTD may be endogenous as it could be influenced by previous and current HC expenditure. This means that weak exogeneity requires PTD in a given month not to be affected by expenditure in that same month and strong exogeneity requires PTD not to be influenced by expenditure in any previous periods either. To reduce the problem of endogeneity, Zweifel have adopted a quasi-instrumental variable (IV) approach that was suggested by Salas and Raftery. Zweifel et al added lagged HC expenditures to the OLS part of their model (cost observation in one year; TTD as a single explanatory variable measured from the end of that year). The IV approach is only valid though if the lagged variables are not correlated with the error term. Zweifel et al suggest that the error term could contain an element that is time-invariant and cannot be observed or measured. In this case the orthogonality condition would be violated, leaving the authors to

point out that their original model which found PTD rather than age as the main predictor for HC expenditure must be robust to any error that is caused by endogeneity.

The authors introduced an updated two part model without the inverse Mill's ratio and included zero costs and survivors.

Model specification:

First part:

$$p(HCE_i > 0) = \beta_0 + \beta_1 A + \beta_2 \frac{A_i^2}{1000} + \beta_3 S_i + \beta_4 (A_i * S_i) + \sum_{k=1}^{6} \beta_{5k} W_{ik} + \beta_6 D_i + \beta_7 (D * A_i) + \beta_8 TTD_i + \varepsilon_i$$

Second part:

$$HCE_i \mid HCE_i > 0 = \beta_0 + \beta_1 A + \beta_2 \frac{A_i^2}{1000} + \beta_3 S_i + \beta_4 (A_i * S_i) + \sum_{k=1}^{6} \beta_{5k} W_{ik} + \beta_6 D_i + \beta_7 (D * A_i) + \beta_8 TTD_i + \varphi_i$$

as before with: W = {*REG, ACC, HOSP, ODI, DED, EI*}, REG = regional dummy, *ACC, HOSP, OSI* and *DED* = dummy variables for supplementary insurance and optional high deductibles*, EI* = average HC expenditure level in the individual's community, *D* = decedents-survivor dummy

Expected costs are calculated by multiplying an individual's probability of incurring costs with the conditional amount of expenditure. The authors did not use a log transformation of the cost variable as this introduces difficulties (if heteroskedasticity is present) when re-transforming the variable to the original scale.

Including survivors in the estimation poses the problem of unknown TTD. Zweifel et al therefore assume that this must naturally be greater than the maximum TTD of deceased individuals (42 months in their study) and set TTD for survivors to 43 month. They restrict age to be between 30 and 95 years (Zweifel et al 2004).

2.5 Modelling Techniques – The "Finnish Way"

A more recent study undertaken by Hakkinen et al in Finland has used individual level linked data representing a 40% sample of the Finnish population aged 65 and older in 1997, followed up until 2002 (Hakkinen et al. 2008). The authors used linked data available from a number of sources that includes survivors as well as decedents. To prevent multicollinearity and endogeneity, costs are studied in one year and TTD is measured from the end of that

year as a single explanatory variable (lagged). This also allowed inclusion of patients that survived until the end of the follow-up period. Survivors only seem to be included for descriptive analyses, but not for regression analysis. The authors decomposed costs into several components: LTC and non-LTC and estimated eight different models (see Figure 1), with the first one estimating the likelihood of being an LTC user. Subsequent models were estimated separately for LTC users and non-LTC users. For non-LTC users a two-part model was then used to estimate the probability of using any of the defined HC services and the second part estimated HC costs incurred. Hakkinen et al compared predictions for expenditure with and without TTD included. The authors concluded that total expenditure for both LTC and HC rises with age, although this association is weaker when controlling for TTD. They also projected future costs and predicted a 13% lower projection for total expenditure (HC and LTC) when including TTD compared to a model that does not include TTD.

2.6 Modelling Techniques – The "Czech Way"

The most recent study that has been undertaken in this area comes from the Czech Republic (Pavlokova 2009). Pavlokova used data from the largest Czech health care insurance provider. The author studied the last 12 quarters before death and all ages were included (in 18 five-year age bands). Most of the Czech population is registered with a GP and as they are paid by capitation there are no zero cost observations for individuals. This resulted in the author's selection of a one-part model, since there is no sample selection present and almost all individuals have a minimum cost observation (Pavlokova 2009). The author estimated two models, GLM (using a random sample) and OLS without log transformation to avoid the problem of re-transformation, which is claimed to be justifiable in large datasets. Survivors and decedents are included in the regression, but it is not obvious, how TTD for survivors has been handled.

Model specification:

$$HCE_i = \alpha + \beta sex_i + \sum \chi_k agecategory_{i,k} + \sum_{l=1}^{12} \delta_l quartertodeath_{i,l} + \varepsilon_i$$

In both estimations, Pavlokova found the effect of TTD to be more pronounced than the effect of age in terms of prediction HC expenditure (Pavlokova 2009).

2.7 Modelling Techniques – The "Scottish Way" – so far…

One other study has looked at PTD and health care utilisation in Scotland and it was undertaken by Graham and Normand (2001). It represents a preliminary approach to exploring the relationship between age, PTD and expenditure using Scottish data. The researchers used linked data from the Scottish Morbidity Records (SMR01) and the General Register Office for Scotland for 1998 and 1999. Their data included decedents and survivors. The dataset for decedents included information on people, who had died in 1999 and the 12 months preceding their death. The dataset for survivors included people, who did not have a death record, but had at least one episode of acute hospital care in 1999 and information for the last 6 months of 1998. Graham and Normand assumed that the rest of the population did not utilise HC during that period. It is not clear from their report whether decedents, who did not utilise HC, are included in the dataset (Graham and Normand, 2001). Work from this report has not been published and the methods employed are not detailed in their final report to the CSO.

Graham and Normand found that acute HC costs for survivors increase with increasing age, whereas costs for decedents decrease with increasing age. The authors also analysed variations in these patterns by different socio-economic groups and found significant differences in costs incurred by survivors or decedents between deprivation categories and by age group.

2.8 Summary of modelling techniques

Table 1 summarises the mix of methods that have been applied in different national studies to explain the relationship between age, proximity to death and expenditure.

As the frequency and extent to which health care data (administrative data) is collected differs from country to country, any analyses that can be undertaken may be restricted by the data available and also by the possibilities to link health data records to other administrative datasets or survey data. One further issue that seems to evolve from the literature review is that model selection depends on the health care system and reimbursement mechanisms in place as well as on the type of health care costs studied. If cost data from health care insurers, which usually include costs for GP visits etc in addition to hospital costs, are analysed, the first step of the model does reflect a patient's choice to a greater extent than the first step of a model that analyses acute inpatient hospital episodes only, where a patient's choice or decision to access health care is less obvious. Also, if a GP acts as a

gatekeeper for access to secondary or specialist care a patient's choice is not clearly reflected. There are therefore issues that should be considered in terms of sample selection and whether this is observed or unobserved.

3. Issues for consideration

The following details a number of issues that should be taken into consideration when attempting to analyse the relationship between age, proximity to death and HC expenditure.

1. Should survivors as well as decedents be included in the analysis?
2. Should zero cost observations be included?
3. Which age groups should be studied?
4. Should cross-sectional or longitudinal data be used?
5. Which period before death should be studied?
6. Which modelling technique should be applied?

3.1 Should survivors as well as decedents be included in the analysis?

As raised by Hakkinen et al, the effect of age on HC expenditure may be different for survivors and decedents. Another issue for consideration is, whether there would be selection bias if only decedents were included. The problem that arises when including survivors not only in the descriptive part of the analysis, but also in the regression part is the unknown TTD. To predict TTD for survivors, the application of propensity scores or a matched data set could be explored. Zweifel et al have coded TTD as being equal for all survivors and being greater (one month) than the maximum TTD studied for decedents (Zweifel et al 2004). This seems to be a very crude way of imputing missing values for TTD for survivors and assumes a constant TTD across all survivors regardless of any other characteristics. In other studies though, it is not obvious how TTD for survivors was imputed (Pavlokova 2009).

3.2 Should zero cost observations be included?

As highlighted by Zweifel et al, excluding decedents, who do not incur any costs, could lead to biased estimates. If the data set only includes people, who were ill and therefore likely to incur higher costs, then any results that would inform actual budgeting for HC expenditure could be misleading. To correctly inform budgeting procedures it appears important to

include zero-cost observations and so to prevent selection bias. If the sample selection process is observed rather than unobserved, excluding zero costs results in missing data rather than a reflection of the underlying selection process.

In order to prevent over-representation of a particular group of the population it should be considered to include survivors as well as zero cost observations.

3.3 Which age groups should be studied?

Most of the previous research has studied individuals at the age of 65 or over. The proportion of individuals aged 65 and over that do not use HC is likely to be very small, so that concentrating the analysis on this age group will potentially provide fewer zero-cost observations. Also, it avoids the problem of right censoring of the data as there are fewer patients with unknown TTD (Seshamani and Gray 2004a). As population ageing is the focus of this research and an ageing population is usually described as the proportion of the population over the age of 65 it seems plausible only to include elderly people. But since the research interest also focuses on HC expenditure, factors that may influence HC expenditure simultaneously and that affect individuals below the age of 65 should not be ignored. Not including individuals at younger ages would not consider the problem of premature deaths. This is interesting, especially in Scotland, with life expectancy for males being as low as 54 years in one area of Glasgow (WHO, 2007). Inclusion of younger age groups may add significantly to explaining variations in HC expenditure in relation to age and PTD. Graham and Normand had analysed their data by socio-economics groups and found significant differences in costs incurred by survivors or decedents between deprivation categories and by age group (Graham and Normand, 2001). There is evidence that “poorer people” may cost more in terms of HC (Cookson and Laudicella, 2009).

3.4 Should cross-sectional or longitudinal data be used?

The advantage of using longitudinal data is that a number of baseline characteristics can be analysed such as co-morbidities that existed when people entered the study. Furthermore, longitudinal or panel data allows to control for unobserved heterogeneity in patients, and multiple observations per patient can be treated as dependent observations. Analysing costs for decedents only, requires a relatively long period of follow up when individuals enter the data at 65 years or over. In order to allow for a sufficient number of deceased individuals the follow-up period is about 25-30 years. One drawback of using this approach though is the lack of a “survivor group”. Another problem when using longitudinal data is the potential lack

of survey data that provides baseline characteristics and covers the same time period and that can be linked to administrative HC utilisation data. Cross-sectional data is usually more readily available for analysis.

3.5 Which period before death should be studied?

Depending on data availability, different times away from death have been looked at in terms of when individuals start to become more expensive. Seshamani and Gray have established that TTD becomes a significant predictor for HC expenditure as far as 15 years away from death (Seshamani and Gray 2004b). Shorter periods have been studied by Zweifel et al, Hakkinen et al and Graham and Normand (Zweifel et al. 1999; Hakkinen et al. 2008; Graham and Normand 2001).

3.6 Which modelling technique should be applied?

Conventionally, a two-part model is used to estimate the relationship between age, proximity to death and HC expenditure, with the first part (probit or logit) predicting the probability of using HC services in a defined period of time. Hakkinen et al have used a slightly different approach by distinguishing LTC users from non-users in their first part and then estimated different types of HC costs separately in subsequent models (Hakkinen et al, 2008). The second, linear part of the model then usually estimates costs for those individuals, who incur positive expenditure, using either a log transformation or GLM to account for skewness of the cost data.

Reviewing the literature, it would appear that model selection should be determined by the health care system and by the nature of cost data analysed. Cost data that comes from health care insurers and includes costs for a number of different HC services that have differing restrictions in terms of access (GP visits, specialist visits, hospital care, and expenditure for pharmaceuticals), may be very different from cost data that is analysed for acute inpatient care in hospitals.

The following section outlines the data that is available in Scotland and how analyses could be undertaken, given the data sets available.

4. Scottish data availability and linkage

The Information Services Division Scotland (ISD) hold information on every acute inpatient episode in Scotland (SMR01 data), reaching back to 1981. The SMR01 datasets are routinely linked to the Registrar General's death records (GRO) as well as to a number of other national data (cancer registry, Scottish Health Survey).
The Community Health Indicator (CHI) is a unique NHS identifier for every patient in Scotland, who is registered with a General Practitioner or in receipt of a screening service. It includes more than 99% of the Scottish population (Fischbacher et al. 2007).

Linkage of medical records in Scotland goes back to the late 1960s and is undertaken by ISD using probability matching (Fischbacher et al. 2007). Using the method of probability matching allows for imperfections of the data compared to an exact matching approach, which could miss up to 15% of true links. Linkage is usually undertaken on a number of core items: surname, initial, year, month and day of birth with a discrepancy rate of up to 3% in pairs of records belonging to the same individual. As datasets tend to be very large, probability matching is usually undertaken on a subset of pairs of records involved in a linkage procedure. Subsets are those, which share a minimum level of identifying information (sorting files into blocks or pockets). Probability matching is done by assigning probability weights (every time an item of identification is the same in two records, the probability that these two records belong to the same person is increased and vice versa) (The Scottish Medical Record Linkage System 1993).

Specifically to this project, what data and data linkage approach would be required to study the role of PTD and age in Scotland?

4.1 SMR01 and GRO

The SMR01 – GRO linked data would provide information on hospital episodes (positive costs) for individuals with a death record. However, it does not provide information on individuals with a death record, who did not utilise HC services. It gives information on survivors, who incurred costs, but not on survivors, who did not incur costs (those not in the GRO). The appropriate linkage approach should be determined by the research question. Depending on the groups of individuals that we want to study there might be different data linkage methods that should be considered. Graham and Normand analysed costs for decedents and survivors, who had incurred positive expenditure and therefore the widely used method of linking GRO data to SMR01 data was used (Graham and Normand, 2001).

Other studies that have used the same direction of linking GRO data to SMR01 data were mainly interested in the mortality of a subset of the population with a certain specified disease (Stewart et al, 2002).

A population based approach that also includes individuals without (positive) HC expenditure would require a different method of linking data sets, but would have the advantage of including the whole Scottish population and consequently provide more generalizable results for the relationship between HC expenditure, age and PTD. The following explores alternative methods to recover groups of individuals that have been "lost" in current data linkage approaches.

4.2 Recovering zero cost observations for decedents

One approach to retrieving decedents with zero costs that could be explored is an alternative linkage of SMR01 and GRO records. There are potential implications on the number of observations that are left in the final dataset depending on the direction the matching is done. If death records are matched to SMR01 data, the "master dataset" is based on hospital episodes (=positive cost observations) all zero cost observations for deceased individuals will be lost. An alternative method to link SMR01 and GRO records is to use GRO data as the "master dataset" (dataset that is used as the baseline data and additional records are added) and then merge in all hospital records. This would provide individual records for decedents, who did not access HC in a defined period before their death.

Even though the method above would recover observations for decedents without HC utilisation, this would still not consider the whole population as surviving non-users are still missing from the data. Using the CHI dataset, which includes 99% of the Scottish population, could solve this issue and provide data on non-users of hospital services in both, the decedents and survivor group. Figure 2 shows how linkage could be undertaken to obtain a dataset that initially includes the whole population and which by a process of elimination enables us to identify survivors and decedents with and without HC utilisation.

The two groups of individuals, who did not utilise HC services and therefore could not be analysed using the existing SMR01-GRO link, could also be retrieved by using linked survey data (like MIDSPAN and the Scottish Health Survey) that follows individuals over time and either records their death independent of prior hospitalisation or provides information on the survivor group without HC utilisation. This analysis would however be based on a small

sample of the total population (Renfrew/Paisley) and may not consider important variations between different areas in Scotland.

It would also be worthwhile exploring the suitability of the Scottish Longitudinal Survey (SLS), which would provide a random sample of the Scottish population. This would technically represent a subsample of the CHI dataset. A number of administrative datasets are linked with the SLS, such as the 1991 and 2001 census. SMR01 data can be linked to the SLS sample members, if requested. The SLS also offers information on other vital events, such as deaths and marriages. This dataset would provide information on all four groups of individuals.

5. Issues for HESG discussion

- Logistics of handling a large dataset; 5 million people live in Scotland, will we need a random sample?

- Issues of model convergence (sample size and GLM?)

- Effect of time lags: If using decedents and survivors, we need to be careful that the period selected for the cost observation is not actually a survivor's TTD. Therefore a lag between cost observation and observation of hospitalisation needs to be applied- what is the ideal lag?

- How does frequency at which costs are measured or available (quarterly, yearly) affect results?

- Would there be selection bias if only decedents were included?

- How should any potential endogeneity between costs and TTD be addressed?

- Ultimately the analysis will be undertaken using long-term care expenditure, for which only aggregated data seems to be available to date for Scotland, how can we consider long term care costs?

References

Abel-Smith B, Titmuss R (1956). The cost of the National Health Service in England and Wales. Cambridge: Cambridge University Press

Busse R, Krauth, C and Schwartz, F W (2002). "Use of acute hospital beds does not increase as the population ages: results from a seven year cohort study in Germany." J Epidemiol Community Health **56**(4): 289-93.

Cookson R, Laudicella M (2009). “Do the poor still cost more? The relationship between small area income deprivation and length of stay for elective hip replacement in the English NHS from 2001/2 to 2006/7.” HEDG Working Paper 09/07. April 2009

Dang T, Antolin P, Oxley H (2001). Fiscal Implications of Ageing: Projections of age-related spending. Economics department working papers no 305, ECO/WKP(2001)31. Paris: Organisation for Economic Co-operation and Development.

Fischbacher C M, Bhopal, R, Povey, C, Steiner, M, Chalmers, J, Mueller, G, Jamieson, J and Knowles, D (2007). "Record linked retrospective cohort study of 4.6 million people exploring ethnic variations in disease: myocardial infarction in South Asians." BMC Public Health **7**: 142.

Fuchs V R (1984). ""Though much is taken": reflections on aging, health, and medical care." Milbank Mem Fund Q Health Soc **62**(2): 143-66.

General Register Office for Scotland. (2007). “Projected Population of Scotland (2006-based).” http://www.gro-scotland.gov.uk/files1/stats/projected-population-of-scotland-2006-based/projected-population-of-scotland-2006-based.pdf

Graham B., Normand C. (2001). “Proximity to death and acute health care utilisation in Scotland”. Final Report. Chief Scientist Office http://www.isdscotland.org/isd/servlet/FileBuffer?namedFile=ptdfinalreport.pdf&pContentDispositionType=inline

Hakkinen U, Martikainen, P, Noro, A, Nihtila, E and Peltola, M (2008). "Aging, health expenditure, proximity to death, and income in Finland." Health Econ Policy Law **3**(Pt 2): 165-95.

Kendrick S. W. and Clarke JA. (1993). The Scottish Medical Record Linkage System. Health Bulletin (Edinburgh). 51, 72-79. http://www.isdscotland.org/isd/files/The%20Scottish%20Record%20Linkage%20System.doc

McGrail K, Green, B, Barer, M L, Evans, R G, Hertzman, C and Normand, C (2000). "Age, costs of acute and long-term care and proximity to death: evidence for 1987-88 and 1994-95 in British Columbia." Age Ageing **29**(3): 249-53.

Pavlokova K. (2009). “Time to death and health expenditure of the Czech health care system”. IES Working Paper 5/2009. IES FSV. Charles University.

Payne G, Laporte, A, Deber, R and Coyte, P C (2007). "Counting backward to health care's future: using time-to-death modeling to identify changes in end-of-life morbidity and the impact of aging on health care expenditures." Milbank Q **85**(2): 213-57.

Polder J J, Barendregt, J J and van Oers, H (2006). "Health care costs in the last year of life--the Dutch experience." Soc Sci Med **63**(7): 1720-31.

Salas C and Raftery, J P (2001). "Econometric issues in testing the age neutrality of health care expenditure." Health Econ **10**(7): 669-71.

Seshamani M and Gray, A (2004a). "Ageing and health-care expenditure: the red herring argument revisited." Health Econ **13**(4): 303-14.

Seshamani M and Gray, A M (2004b). "A longitudinal study of the effects of age and time to death on hospital costs." J Health Econ **23**(2): 217-35.

Stewart S, McIntyre K, Capewell S, McMurray JJV. (2002). “Heart Failure in a Cold Climate. Seasonal Variation in Heart Failure-Related Morbidity and Mortality.” JACC **39**(5): 760-6.

WHO report 2007. Health Inequality, Inequity And Social Determinants of Health. http://www.who.int/social_determinants/resources/interim_statement/csdh_interim_statment_inequity_07.pdf

Zweifel P, Felder, S and Meiers, M (1999). "Ageing of population and health care expenditure: a red herring?" Health Econ **8**(6): 485-96.

Zweifel P, Felder S, Werblow A. (2004). “Population Ageing and Health Care Expenditure: New Evidence on the ‘Red Herring’”. The Geneva Papers on Risk and Insurance **29**(4): 652-666.

Figure 1: Eight part model (Hakkinen et al, 2008)

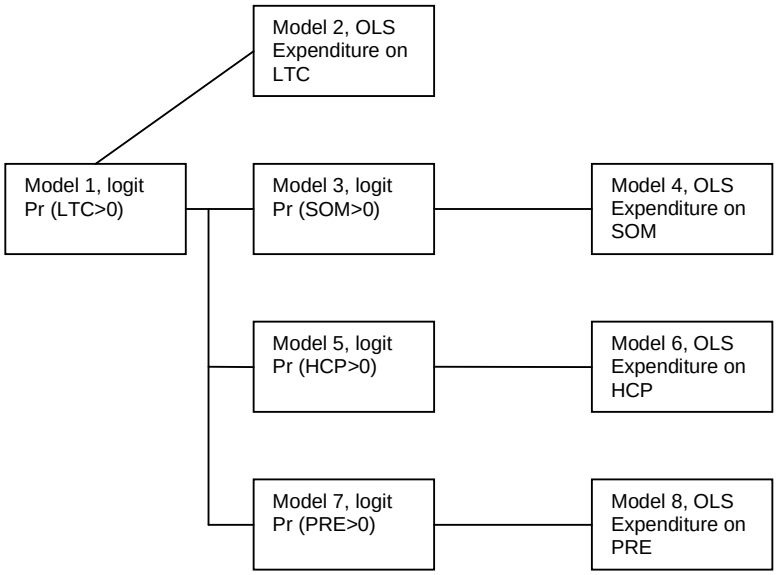

Figure 2: Data coverage from CHI, GRO and SMR01 data sources

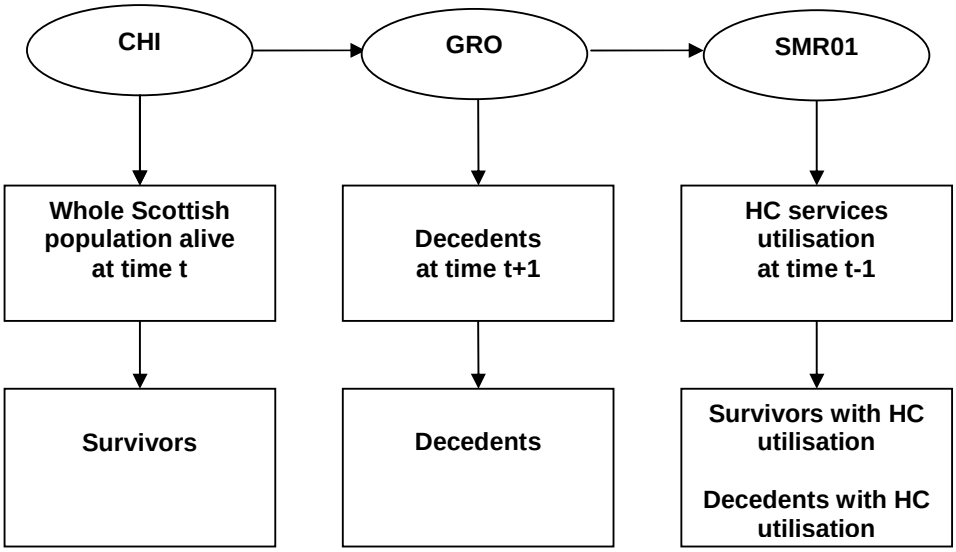

Table 1: Summary of main studies

	Zweifel et al 1999	*Zweifel et al 2004*	*Seshamani and Gray 2004b*	*Hakkinen et al 2008*	*Pavlokova 2009*	*Graham and Normand 2001*
Sample	Decedents with HC utilisation	Decedents and survivors with and without HC utilisation	Decedents with and without HC utilisation	Decedents and survivors with utilisation	Decedents and survivors with HC utilisation (no zero cost observations due to the reimbursement system)	Decedents and survivors with HC utilisation
Costs	All HC costs that are covered by health insurance	All HC costs that are covered by health insurance	Acute inpatient hospital episodes	All HC and LTC services	All costs that are covered by health insurance	Acute inpatient hospital episodes
Data available	10 years	7 years	29 years	5 years	3 years	2 years
Periods away from death	2 (5) years	23 months	15 years	4 years	12 quarters	1 year
Age groups included	All ages and separate analysis for 65+	30-90 years	65+	65+	All ages, 18 five year age bands	All ages? (not specified)
Model used	Cross-sectional analysis, Heckman sample selection model, first part: probit for decedents likelihood to utilise HC, second part: log cost estimation for deceased users	Two-part model, First part estimates probability of positive cost observation, Second part: OLS (no log transformation)	Panel data analysis (random effects) Two-part model, first part: probit for decedent's likelihood of hospital use, second part: GLM (log link, Poisson)	Eight part model- First model estimates probability of using LTC or not. Subsequent models analyse subgroups (Fig. 1)	No selection part GLM (with log link and Gamma), OLS (no log-transformation)	Not specified