Deloitte.

Scottish Allocation Formula

GMS workload model

26 August 2016

Contents

Contents 2

Executive summary 2

1 Introduction 4

2 Methodology 5

2.1 Econometric model estimation 6
2.2 Econometric model application 9
2.3 Allocation weights 9
2.4 Health inequalities and unmet need 9

3 Data 11

3.1 Data summary 11
3.2 Representativeness of the data sample 13
3.3 Descriptive analysis 1

4 Results 16

4.1 Model selection 16
4.2 Model results 17
4.3 Allocation weights: comparison with 2004 SAF formula 20
4.4 Differences between relative weights: SAF 2004 vs. 2016 review 21

5 Appendix 25

5.1 Correlation analysis of the additional need indicators 26
5.2 Dispersion tests 26
5.3 Model coefficients 27
5.4 Read codes vs. consultation weights 28

6 References 29

Important Notice from Deloitte

This final report (the "Final Report") has been prepared by Deloitte LLP ("Deloitte") for the Scottish Government in accordance with the contract with them dated 20 December 2015 ("the Contract"), as extended by email dated 01 August 2016, and on the basis of the scope and limitations set out below.

The Final Report has been prepared solely for the purposes of reviewing and assessing the Scottish Allocation Formula for the Scottish Government, as set out in the Contract. It should not be used for any other purpose or in any other context, and Deloitte accepts no responsibility for its use in either regard, including its use by the Scottish Government for decision making or reporting to third parties.

The Final Report is provided exclusively for the Scottish Government's use under the terms of the Contract, however it may be made available to TAGRA, the Technical Advisory Group on Resource Allocations in Scotland, solely for the purpose of evaluating the workload update of the Scottish Allocation Formula. No party other than the Scottish Government, including TAGRA, is entitled to rely on the Final Report for any purpose whatsoever and Deloitte accepts no responsibility or liability or duty of care to any party other than the Scottish Government in respect of the Final Report or any of its contents. If TAGRA choose to rely on the Final Report, they do so at their own risk and without recourse to Deloitte.

The information contained in the Final Report has been obtained from the Scottish Government and third party sources that are clearly referenced in the appropriate sections of the Final Report. Deloitte has neither sought to corroborate this information nor to review its overall reasonableness. Further, any results from the analysis contained in the Final Report are reliant on the information available at the time of writing the Final Report and should not be relied upon in subsequent periods.

Any decision to invest, conduct business, enter or exit the markets considered in the Final Report should be made solely on independent advice and no information in the Final Report should be relied upon in any way by any third party. This Final Report and its contents do not constitute financial or other professional advice, and specific advice should be sought about your specific circumstances. In particular, the Final Report does not constitute a recommendation or endorsement by Deloitte to invest or participate in, exit, or otherwise use any of the markets or companies referred to in it. To the fullest extent possible, both Deloitte and the Scottish Government disclaim any liability arising out of the use (or non-use) of the Final Report and its contents, including any action or decision taken as a result of such use (or non-use).

Executive summary

The Scottish Allocation Formula (SAF) is a weighted capitation formula used to allocate the General Medical Services (GMS) budget to general practices across Scotland. The budget allocated by the SAF is known as the Global Sum and covers the largest part of the payment to general practices.

The SAF was first used in 2004 and aims to compensate practices for their workload and differences in unit costs. The workload dimension of SAF is a function of the list size and the population relative need. The latter is a function of population age, gender, and morbidity and lifestyle circumstances factors (deprivation, standardised mortality ratio, long-term illness ratio, etc.). The unit cost dimension aims to compensate practices for costs associated with variation in staff costs across regions and provision of GMS in rural and remote areas.

This report provides an update of the workload dimension of the SAF (the unit cost dimension is out of the scope of this report). In particular, a series of econometric models have been estimated in order to quantify relative need and expected workload across practices. Compared to the 2004 SAF, the formula currently used to determine allocation weights, the variables that capture differences in relative need have been reviewed and the model coefficients have been refreshed using more recent data. Most importantly, the methodology used to estimate relative need has also been revised. The 2004 SAF estimates the impact of age-gender and additional need in isolation. This review estimates their impact simultaneously, within the same multivariate model. Given the correlation between these variables, a multivariate approach is considered more appropriate.

The results of the analysis suggest that

- A large share of the significant variation in workload across practices can be explained by the model. Model forecast accuracy at the practice level is around more than 95%.

- The age and gender demographic profile of the registered population is a significant driver of GMS workload. The older the population, the greater the workload. For instance, patients 75 of age or older have up to three times higher need compared to young patients (10 years old or younger).

- There are additional factors that drive workload, beyond age and gender. It was found that workload is positively related to morbidity and life circumstances factors such as long term sick and unemployed indicators, deprivation deciles and limiting long term illness ratios. For instance, patients in the least deprived decile have an average utilisation of GMS that is 14% lower than patients in the most deprived decile, all other things being constant. When taking into account other additional need factors such as long term sick and unemployed and limiting long term illness, patients in the most deprived decile have up to 25% higher need.

- There are significant differences in the predicted weights between the 2004 SAF and this review. This is primarily due to the difference in the estimated impact of morbidity and life circumstances factors on workload. The model presented in this report finds a stronger impact of deprivation on GMS utilisation compared to the 2004 SAF. There are also significant differences in the estimated relationship between utilisation and age-gender. The 2004 SAF places more weight on younger populations compared to the model presented in this report. The difference in age-gender weights can be partly explained by the treatment of zero-consultation patients. The 2004 SAF has excluded zero-consultation patients from the analysis, which seems to have led to biased estimates of the workload-age relationship.

The results of this work need to be seen in the following context:

- The purpose of the formula is not to determine the total GMS budget but rather an allocation of the budget proportionally to each practice workload.

- Practice workload is measured by the number of read codes and by the number of consultations by patient. However, workload could also be measured by file opening times per read code or consultation. If the variation in file opening times is different than that of read codes or consultations, then the model may over/under-estimate relative need. Future data collection could also seek to capture such information.

- The allocation weights presented do not take into account any potential health inequalities associated with supply issues or patient behaviour. Health inequalities related to patient behaviour (e.g. specific types of patients do not present themselves to health providers, which could lead to unmet healthcare need) or geographical shortage of GPs are, by and large, beyond the control of existing practices and therefore could not be significantly addressed through the workload model. Addressing these sources of health inequalities requires a separate analysis and potentially allocation mechanism.

1 Introduction

The Scottish Allocation Formula (SAF) has been developed in order to allocate the core general medical services (GMS) budget in proportion to the population's relative need and to compensate general practices for their workload and unavoidable differences in unit costs.

The budget allocated by the SAF is known as the Global Sum, which constitutes the largest component of the general practice payment (c.55% of total GMS budget). Other key payment components are related to Quality and Outcomes Framework targets, enhanced services, premises and seniority.

The Global Sum formula has two dimensions: (i) a workload and (ii) a unit cost dimension. The workload dimension allocates the budget on the basis of each practice's expected workload, which depends on the list size and the corresponding population's relative need. The latter is a function of the population's age, gender, and morbidity and lifestyle circumstances factors. The unit cost dimension compensates practices for unavoidable costs associated with the staff costs using the Market Forces Factor and the provision of GMS in rural and remote areas. This report focuses on the workload dimension.

The SAF was first used in 2004 and has been regularly updated since then to take into account changes in practices' list size and demographic composition. This report presents a model update. In particular, the variables that capture differences in relative need have been reviewed and the model coefficients have been refreshed using more recent data. Furthermore, the methodology used to estimate relative need has also been revised. The 2004 SAF estimates the impact of age-gender and additional need in isolation. This review estimates their impact simultaneously, within the same multivariate model. Given the correlation between these variables, a multivariate approach would be more appropriate.

A series of econometrics models have been estimated with the aim to quantify the impact of age, gender and morbidity and life circumstances on practice workload. The estimation of the models is carried out using patient level data obtained from the Practice Team Information (PTI) for a sample of the Scottish population covering 5.4%[1] of the registered population. The sample has been determined by data availability on GMS utilisation (read codes and consultations). After the econometric model was estimated, it was applied on the total registered population to calculate the expected workload and the associated allocation weights for all Scottish general practices.

The analysis presented in this report seeks to estimate relative need for GMS and not total population need. Equally, the purpose of the formula is not to determine the total GMS budget but rather a fair allocation of the budget across general practices.

The remainder of this report is organised as follows:

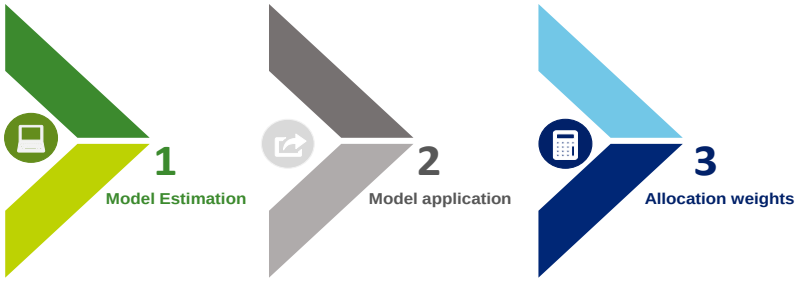
- Section 2 sets out the methodology;
- Section 3 discusses the data used in the analysis together with some descriptive analysis; and
- Section 4 presents the results of the analysis.

[1] SAF 2004 was last updated using the PTI sample covering 2011/12 data from 57 practices covering 5.4% of the registered patient population.

2 Methodology

Relative need and allocation weights associated with the GMS workload dimension of the Global Sum formula are estimated in three steps.

Figure 1: Over-arching methodology



1. **Econometric model estimation.** A series of econometric models has been estimated using patient level data with the aim to quantify the relationship between GMS utilisation and patient demographics and morbidity and life circumstances (MLC) factors.

2. **Econometric model application.** The econometric model has been applied to the Scottish population in order to determine the expected utilisation and relative need for each practice given the practice's demographic and MLC distribution.

3. **Allocation weights.** A weighted list size is derived for each practice by multiplying each practice's relative need profile, derived in the previous step, by the corresponding registered population. Practice allocation weights are then calculated by dividing each practice's weighted list by the sum of the weighted lists across all practices.

Each of these steps is discussed in detail in the remainder of this section.

The overall framework used in this review is similar to the one used in the 2004 SAF, however, there are three main differences between the two approaches:

1. The 2016 review uses more up-to-date patient information from 2012/13 to estimate the relative need;

2. The 2004 SAF classifies the population into eight age groups whereas this review uses more detailed age groupings (18 groups); and

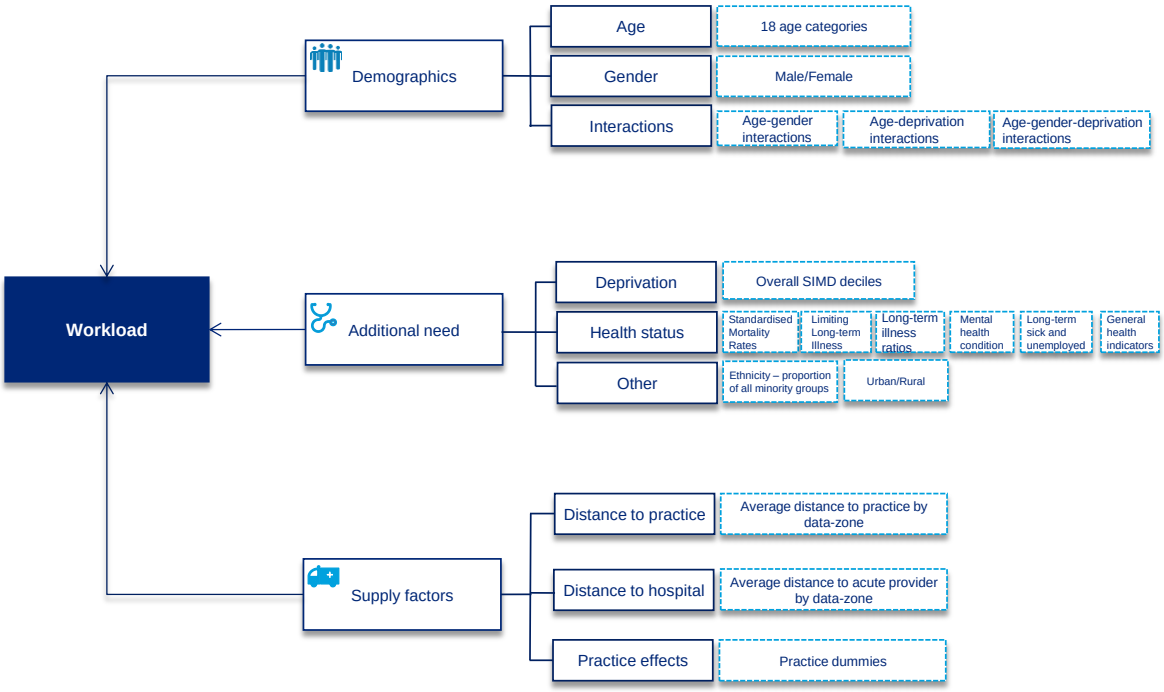
3. The 2004 SAF modelled the impact of gender and age separately from the effect of additional need measured by the MLC indicators, whereas this review quantifies the impact of patient demographics and MLC simultaneously, within the same model. The simultaneous approach is designed to measure the marginal impact of the MLC variables and is preferred given the collinearity between the age-gender utilisation profiles and MLC variables.[2]

[2] This approach is similar to the 2016/17 NHS England primary care allocation methodology. NHS England 2016: *Technical Guide to Allocation Formulae and Pace of Change.* Online available at*:* https://www.england.nhs.uk/wp-content/uploads/2016/04/1-allctins-16-17-tech-guid-formulae.pdf.

2.1 Econometric model estimation

The econometric model estimates the impact of patient demographics and MLC on the utilisation of GMS, which, in principle, is equivalent to the practice workload. The general specification of the econometric model used is illustrated in Figure 2.

Figure 2: Model specification



Source: Deloitte analysis

Following recommendations from a panel of experts[3], workload has been approximated by the total number of read codes[4] associated with each patient visit to the practice. While the total number of consultations per patient, another potential proxy for GMS utilisation, is also available, it was recommended that read codes could capture more accurately the complexity of the patient and/or intensity of GP's workload. As a sensitivity check, the econometric analysis in this study has used both read codes and consultations.[5]

Workload depends on whether the consultation is delivered at the practice or at the patient's home. In order to take this into account, the panel of experts recommended adopting the approach followed in the 2004 SAF whereby read codes associated with home visits are multiplied by three to reflect the increased amount of time necessary to see a patient at home. This adjustment is consistent with estimates of unit costs per patient for primary health care computed in previous studies[6].

GMS workload is modelled as a function of three types of indicators (detailed description of all the data is provided in Section 3.1 and Appendix):

[3] This includes primary care researchers and representatives of TAGRA, the Technical Advisory Group on Resource Allocations in Scotland, which is responsible for overseeing and maintaining the development of the Scottish Allocation Formula (SAF) for GMS.
[4] Scottish general practices record clinical information (symptoms, diagnoses or other activities) using read codes, which are the recommended national standard coding system. Each consultation generates at least one read code. GPs can record an unlimited number of read codes per consultation.
[5] The Appendix compares the practice allocation weights generated by models based on read codes to weights generated by models based on consultations. The correlation between the two sets of weights is 0.99, suggesting that the choice between read codes and consultations has an insignificant impact on the estimated practice weights.
[6] See the study on unit costs for primary health care published by Brilleman et al. (2014) / Journal of Health Economics 35, 109-122.

- **Demographics**. The practices' registered population's age and gender is expected to explain most of the variability in the utilisation of GMS, with elderly patients having relatively higher need for healthcare services.

- **Additional need.** Utilisation is expected to vary across populations within the same age-gender group due to differences in MLC. Thus, additional need indicators, for instance, deprivation and limiting long-term illness, have also been considered in the model.[7]

- **Supply-side factors.** The observed workload is a function of both demand and supply factors. Although the focus of the analysis is on the determinants of relative need, which is driven by demand-side factors, supply-side factors have been included in the model to control for confounding supply effects.[8] This is particularly important if the supply of GMS is restricted in some parts of the country. If, for example, there is an under-supply of services in deprived areas, the estimated relationship between utilisation and deprivation could be biased if the degree of supply of services is not taken into account. Two sets of supply-side factors were considered in the econometric analysis.

    - **Access.** The average distance to the practice could control for the travelling time that may impact utilisation[9]. All other things being constant, practices that serve a dispersed population might have a lower utilisation of GMS. Distance to the nearest acute provider could be considered a proxy for the availability of alternative settings of care. The restricted supply of community and acute services may lead to a higher utilisation of GMS.[10]
    - **Practice dummies.** Practice dummies can, in principle, control for access but also other unobserved differences in the supply of GMS, i.e. type of services provided, waiting times, and practices' productivity. As practice dummies also account for supply differences between practices, including additional access variables in model specifications would not significantly improve to the explanatory power of the models. The practice dummies could also control for any possible supply-side effects associated with historical over- and under-funding of practices: number of consultations for a practice that has been under-funded may be relative low, all other things being equal, because there are not enough resources to service the local population. In contrast, consultation rates for an over-funded practice could be higher than average because of "over-supply" of GPs and low waiting times.

Although the access variables were initially considered in the analysis, they were eventually disregarded due to potential endogeneity issues, i.e. the areas that have populations with relatively low demand for GMS might have a lower number of practices, and therefore, the accessibility may be lower than average. In other words, accessibility might affect workload but workload might also have an impact on access. This endogeneity might lead to biased model estimates.

Model estimation

As the workload proxy is measured as count data, i.e. only non-negative integer values are observed, the econometric analysis is based on generalised linear models.[11] The model is estimated using negative binomial regressions for three main reasons.

[7] MLC variables are highly correlated with each other (see Appendix), which could make it difficult to isolate their individual impact in regression analysis or assess which model specification is a better description of the underlying relationship. Several alternative model specifications have been estimated with the aim to determine the model specification that best fits the data.
[8] If demand and supply factors are correlated then failing to control for the latter could lead to biased estimates of the demand effects and allocation weights.
[9] SAF 2004 takes into account patients' accessibility to GMS practices in the rurality and remoteness adjustment.
[10] Also, the complexity of the patients' cases may be higher if patients substitute community and acute services for GMS.
[11] See Box 1 for more details.

- **Ordinary Least Squares (OLS).** While standard regression models could be used to estimate the relationship between the GMS workload and the population characteristics, this approach is not ideal. For instance, an OLS regression model could predict negative values for the GMS workload, which is theoretically implausible.

- **Poisson regression.** Poison regressions are designed to estimate models with count data. However, as read codes are over-dispersed (the variance of the workload distribution is higher than the mean), negative binomial regressions are considered to be more appropriate.[12]

- **Empirical testing.** The three approaches have been empirically evaluated and it was found that the negative binomial provides a better in-sample and out-of-sample fit than the OLS and Poisson regressions.

All the model specifications presented in Section 4.1 are estimated using a log-linear functional form associated with the negative binomial distribution described by equation (2) in Box 1.

Box 1: Overview of count regression models considered

Poisson models

The most common technique used to model count data (or nonnegative integer values) is a Poisson regression. The Poisson distribution describes the number of occurrences or realised outcomes that occur in a given period of time, with the associated mean given by the average number of outcomes per period. The parameterisation of the Poisson model is usually given by the exponential mean of occurrences where the conditional mean depends on the linear combination of a set of variables. The set of regressors included in Poisson regression equations is selected in a similar manner as in linear regression models:

$$E[y_i|X_i] = \exp(\boldsymbol{x_i'\beta}) = \exp(\beta_1 + \beta_2 x_{2i} + \cdots + \beta_k x_{ki}) \quad (1)$$

Model (1) is usually estimated as a log-linear specification, as part of the family of generalised linear models, available in many popular statistical software programs. Thus, (1) can be re-written as:

$$lnE[y_i|x_i] = \boldsymbol{x_i'\beta} = \beta_1 + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} \quad (2)$$

In such a model specification, the coefficient estimates β_i represent the expected change in the log of the mean of the outcome variable per unit change in predictor x. If *y* is equal to zero, the estimation of the log-linear function (2) can still be performed as only *the conditional mean* must always be positive, however, the actual outcome can be zero.

OLS regressions

An alternative to generalised linear models is OLS regressions that estimate the log of *y* on **x**. However, if *y* can be zero, then, usually, *y* needs to be transformed (e.g. using ln(*y*+0.5)) to be able to take the log of positive values. A guide of the alternative model specifications is provided in Mullahy (2001) and Cameron and Trivedi (2005).

Negative binomial models and test for overdispersion

Very often, the Poisson model is inadequate as it restricts the conditional mean of *y* to equal its conditional variance (a feature of the data called equidispersion). If, however, this assumption does not hold, then the Poisson standard errors are wrong. For instance, if *y* is overdispersed, with the variance being higher than the mean, then a common and more general model that can be used is the negative binomial model. Overdispersion tests of the workload measures based on the PTI data used in this study confirm the presence of overdispersion. Therefore, a negative binomial specification is appropriate for estimating the workload model specifications considered.

Alternative count regression models

Other model specifications for count data include two-part and mixture models. These specifications treat the processes of zero outcomes differently from the non-zero counts. Two-part count models, also known as hurdle models, include a count equation for positive outcomes (Poisson, negative binomial, etc.) and a binomial equation (known as hurdle component) for the zero observations. Similarly, zero-inflated models can handle data sets with a large number of zero outcomes by combining a point mass at zero with a count distribution (Poisson, negative binomial, geometric, etc.). Cameron and Trivedi (2005) provide an in-depth overview of the count regression models presented here.

2.2 Econometric model application

The econometric models specified in section 2.1 are estimated using a sample of the Scottish population registered with a sample of 56 practices due to data availability.[13] In order to estimate the expected workload for primary care services across all Scottish practices:

a. The econometric model has been applied to the entire registered Scottish population. The practice dummies have been *sterilised* so that only the demand-side factors are taken into account in the computation of allocation weights;[14]

b. Expected workloads at the patient level have been aggregated at the practice level; and

c. Relative need by age-gender profiles at the practice level have been derived by dividing each practice's expected workload to the overall workload across all practices. Thus, the relative need estimates effectively reflect the practices' age, gender and MLC distributions.

2.3 Allocation weights

Allocations weights should reflect the practices' relative workload, which is a function of list size (practice population) and patient complexity/average need. In order to take both of these into account, a weighted list size is computed by adjusting list sizes by each corresponding practice's average relative need, estimated in the previous step. Allocation weights are then defined as the ratio between a practice's weighted list over the total (sum of all practices) weighted list.

2.4 Health inequalities and unmet need

In a recent paper, McLean et al. (2015) show that Standardised Mortality Ratios (SMR) in deprived areas tend to be higher than in affluent areas, which could be interpreted as evidence of health inequalities.

Health inequalities could be associated with factors not amenable to the provision of health care services (life-style choices) and/or unmet need. The latter could be the result of one or more factors.

- **Relative under-funding of existing practices.** For example, if practices that serve deprived populations or complex patients are not compensated adequately for their workload, they would not have sufficient resources to meet local population need.

- **Patient behaviour.** Unmet need and health inequalities might be the result of patient behaviour, e.g. specific types of patients who do not present themselves to health providers.

- **Under-supply.** Under-supply of primary care services in a region might lead to longer waiting times and/or poorer access, e.g. longer travelling times to a practice, and subsequently under-utilisation of health care services.

Health inequalities associated with existing practices' funding should be, in principle, addressed through the workload allocation formula. Insofar as the allocation formula sufficiently captures the impact of demographics and MLC factors on GMS utilisation, the allocation of funds to existing practices would be proportional to their workload, facilitating a fair budget allocation. However, McLean et al. (2015) provided some evidence which shows that total practice funding in Scotland is disconnected from population relative need as proxied by consultation rates, deprivation and multi-morbidities.

In this report, apart from age-gender profiles, workload is modelled as a function of deprivation and several other socio-economic and morbidity factors, including SMR, limiting long term illness and mental health conditions that might be associated with higher need. In other words, the additional need associated with deprivation or morbidity is explicitly incorporated into the allocation weights, and

[13] Data on GMS workload measures are available only for the sample of 56 practices tracked by PTI, representing 5.6% of all Scottish practices. The explanatory variables included in all the model specifications tested, however, are available for the entire Scottish population, thus allowing the application of the models to the wider population.
[14] The estimated coefficients of the practice effects are replaced with the average of all the practice effects in the computation of relative practice weights.

practices with more deprived or complex patients receive more funds, keeping age and gender constant. As shown in section 4.4, the formula presented in this report predicts significantly more funds for deprived areas compared to the 2004 SAF.

Dealing with the other two sources of health inequalities would potentially require a different type of allocation mechanism. Patient behaviour or geographical shortage of GPs are, by and large, beyond the control of existing practices and therefore could not be significantly addressed through the GMS contract. These sources of health inequalities could be potentially dealt with by Health Boards through specific interventions such as incentivising GPs to open practices in deprived areas. Furthermore, health inequalities may be associated with other parts of the system, e.g. acute or community setting of care, however, an evidence-base on the sources of health inequalities is currently lacking. A health inequality budget could be allocated to Health Boards who could potentially use local knowledge and discretion to deal with unmet need more effectively. In essence, this is the mechanism proposed by NHSE, which allocated a health inequalities budget to CCGs through the CCG allocation formula, and not directly to general practices or other health care providers.

3 Data

This section discusses the data used in the analysis and presents a series of summary statistics. In particular, the following topics are discussed:

- **Data summary.** This provides a discussion of the variables considered in the analysis, both the workload proxies and the explanatory variables included in the econometric models.
- **Representativeness.** The sample data have been compared to the general population in order to assess its representativeness in terms of age, gender, deprivation and urban/rural mix.
- **Descriptive analysis.** A series of descriptive statistics have been generated to sense check the data and gain a high-level understanding of the underlying relationships.

3.1 Data summary

The 56 practices within the sample represent 5.6% of all Scottish general practices and 5.4% of the Scottish registered population. The sample includes 289,144 patients with at least one read code/consultation and 74,849 inferred patients with zero consultations.[15]

- **Workload**. The workload data have been obtained from the Practice Team Information (PTI) and refer to the year 2012/13.[16]
- **Demographics.** The gender and age of the patients have also been taken from the PTI. Patients' age was categorised into 5-year age bands, from 0-4 to 85+ years old.
- **Additional need.** The MLC variables come from the 2011 Census and ISD Scotland. These variables are measured at the data zone[17] level and have been attributed to patients on the basis of their data zone code.

[15] The Appendix discusses the data used in this study in more detail.
[16] These data were the most recent GMS workload measures available. Workload proxies are also available by three types of visit: patient's home, surgery and out of office hours.
[17] Data zones represent small-area statistical geographies that have between 500 and 1,000 household residents.

Table 1: Data summary

Variable type	Variable	Description	Source
Workload proxy	Number of consultations	The sample contains the records of 289,144 patients who visited 56 practices during 2012/13. For each patient, the sample contains the number of read codes and the number of consultations that occurred during the year.	PTI 2012/13 records
	Number of read codes		
Demographics	Age	Patients' age and gender, with age grouped in 18 age categories across 5-year age bands from 0 – 4 to 85+ years old.	PTI records and GP practice list sizes
	Gender		
Additional need	Scottish Index of Multiple Deprivation (SIMD) deciles	Overall Scottish Index of Multiple deprivation provides a relative ranking of the Scottish data zones based on a weighted average of seven deprivation measures: income, employment, heath, education, skills and training, housing, geographic access and crime. The relative overall SIMD ranking for each data zone is assigned to a decile with 1 representing the most deprived 10% of data zones, while decile 10 contains the least deprived 10% of the data zones.	PTI records and Scottish Neighbourhood Statistics
	Standardised mortality ratios (SMR)	Covers all causes for people under 70 years old.	Scotland's 2011 Census
	Mental health condition	Covers all conditions related to mental health of residents within each data zone.	
	Long-term limiting illness ratio (LLTI)	Proportion of residents within each data zone whose day-to-day activities were limited by a long-term illness.	
	Bad general health	General health self-assessment of residents within each data zone, from very bad (a score of 5) to very good (a score of 1).	
	Long-term sick and unemployed	Proportion of residents within each data zone that are economically inactive due to a long-term sickness or disability.	
	Ethnicity	Proportion of all minority ethnic groups by data zone.	
	Location: urban vs. rural	Each data zone is categorised as being located in an urban or rural area by ISD Scotland.	ISD Scotland

3.2 Representativeness of the data sample

The distribution of the key variables within the PTI sample has been compared with the distribution of the Scottish population to assess its representativeness. The representativeness has been assessed in terms of four key variables.

- **Age** (see Figure 3). The age distribution of the patients covered by the PTI sample is similar to the distribution of the Scottish population, however, the PTI sample tends to over-represent the 40-44 age group and to under-represent the 20-24 group.

- **Gender** (see Figure 4). The female population is slightly under-represented in the sample (50.5% vs. 51.4% in the population).

- **Deprivation** (see Figure 5). There are some differences in the distribution across deprivation levels. The sample under-represents the population in the most deprived decile (SIMD =1) and over-represents the population in the fourth, fifth and tenth deciles.

- **Urban vs. rural classification**[18] (see Figure 6). The sample slightly under-represents the urban population (80.9% vs. 82.7% in the population).

Overall, the population appears to be well represented by the sample. Although the sample covers only a subset of the registered Scottish population, it is large enough (it contains more than 363,000 observations) and has sufficient variation in order for the model to provide reasonably accurate estimates.

[18] This classification follows the Scottish Government's 2 fold Urban Rural Classification where rural areas represent settlements of less than 3,000 people and urban areas represent settlements of 3,000 or more people.

Figure 3: Sample representativeness: Age

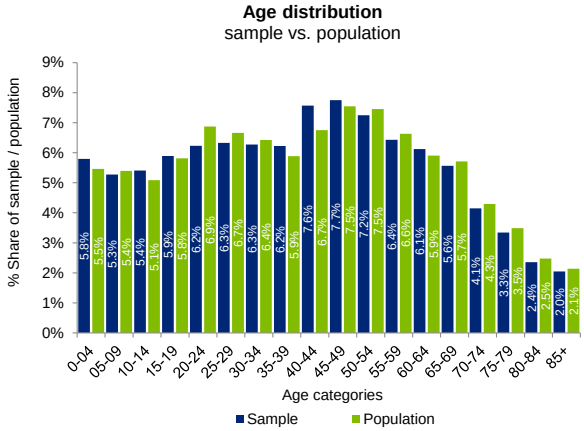


Figure 4: Sample representativeness: Gender

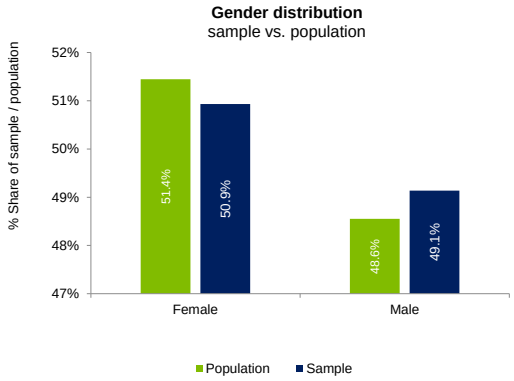


Figure 5: Sample representativeness: Deprivation

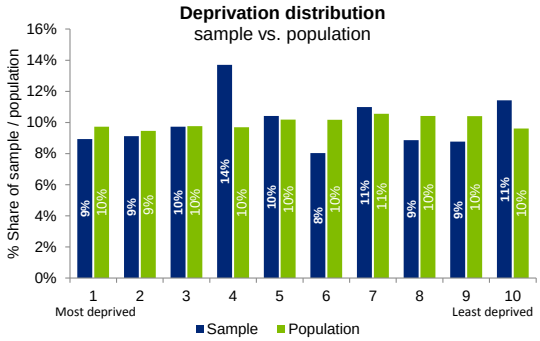
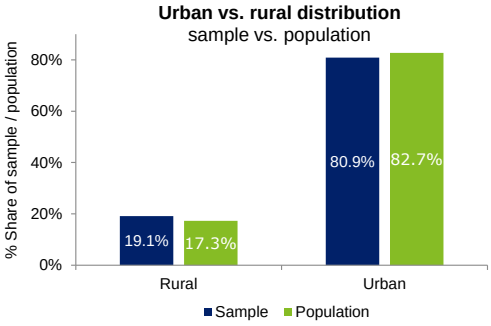


Figure 6: Sample representativeness: Urban/rural



Source: ISD Scotland; PTI, Deloitte calculations

3.3 Descriptive analysis

This section presents some descriptive statistics of the sample data used in the econometric analysis. The main features of the data are summarised below.

- **Workload distribution** (Figure 7). The majority of the patients (63%) had between 0 and 10 read codes in 2012/13, while a significant portion of registered patients (20%) had zero read codes. Consultation rates exhibit a similar pattern with 70% of patients having between 1 and 10 consultations and only c.9% with more than 15 consultations.[19]

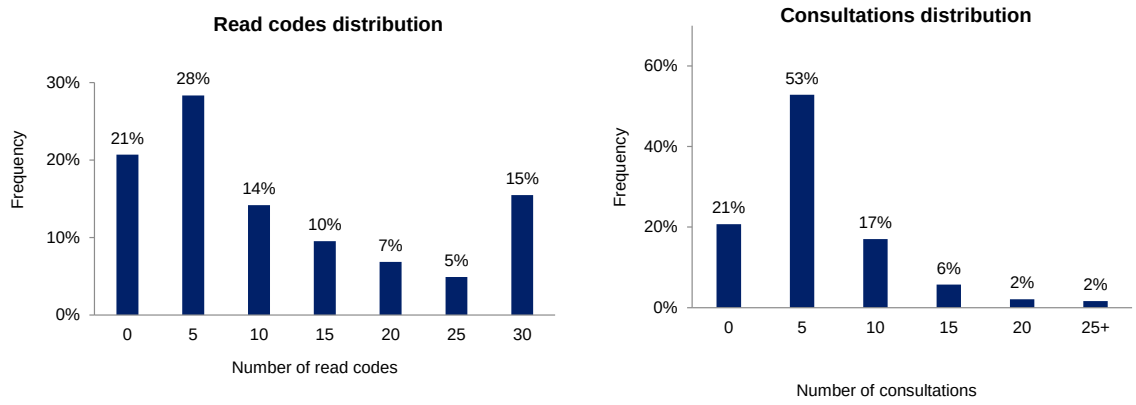
- **Age and gender** (Figure 8). There are significant differences in the GMS utilisation by age. Patients older than 75 years old have the largest utilisation with an average of 31 read codes and eight consultations. Patients between five and 14 years old have the lowest utilisation with an average of three read codes and less than two consultations during 2012/13. Female patients between 15 and 54 years old have a higher utilisation than male patients whereas the opposite is true for patients older than 65 years.

- **Deprivation** (Figure 9). Average utilisation is fairly equally spread across deprivation deciles, with slightly higher averages across deciles 1 to 5 relative to deciles 5 to 10.

[19] In the econometric analysis, patients with read codes higher than 145 (the top 0.1% of the distribution), representing 383 observation, have been removed from the sample as these outliers could skew the results of the analysis.

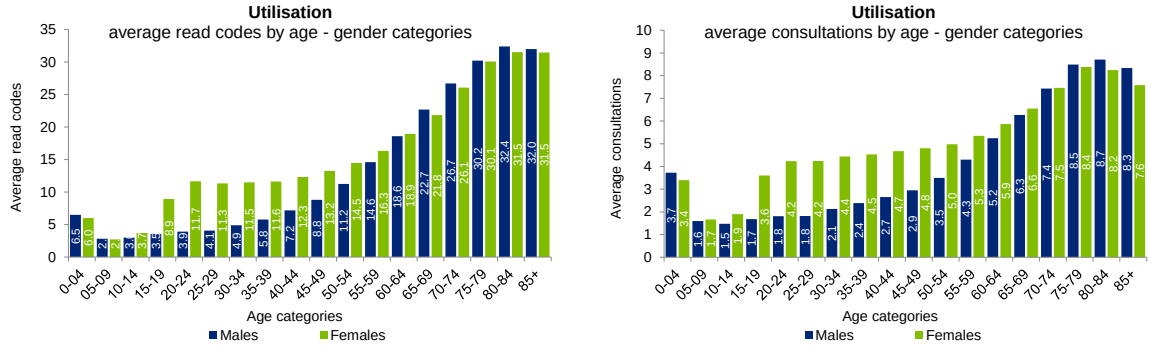
- **Read codes vs. consultations** (Figure 10). The two workload measures are highly correlated with a higher number of consultations associated with a high number of read codes. The correlation coefficient between these two measures is 0.84.

Figure 7: Workload distribution: consultations and read codes



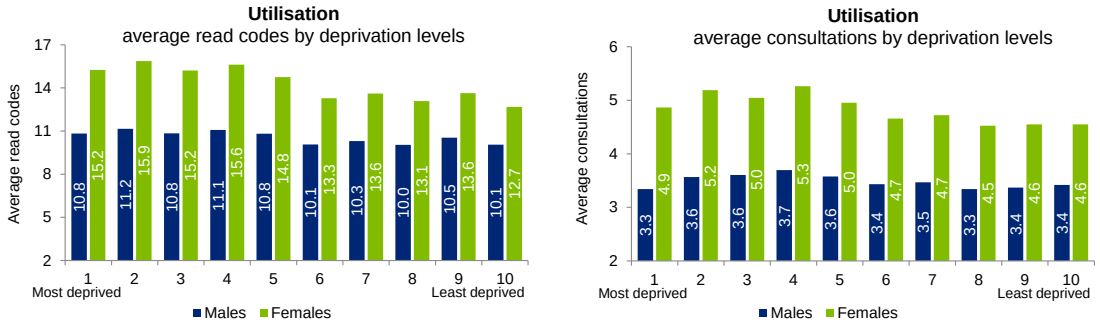
Source: PTI, Deloitte calculations

Figure 8: Utilisation by age-gender categories



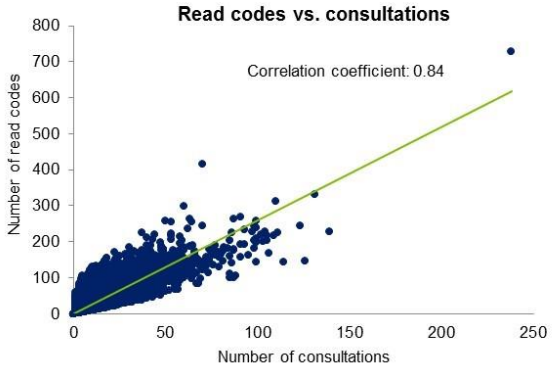
Source: PTI, Deloitte analysis

Figure 9: Utilisation by deprivation deciles



Source: PTI, Deloitte analysis

Figure 10: Workload proxies: read codes vs. consultations



Source: PTI, Deloitte analysis

4 Results

This section discusses the results of the econometric analysis together with the estimated allocation weights. In particular, the following modelling aspects and results are set out:

- Model selection process.
- Model coefficient estimates.
- Comparison with the 2004 SAF weights.

4.1 Model selection

A number of alternative model specifications have been estimated in order to identify the best model that describes the underlying relationship between GMS utilisation and patients' demographics and MLC indicators. The best model is assessed on the basis of three criteria:

- **In-sample fit**. This is measured by the Bayesian Information Criterion (BIC). The smaller the BIC, the better the fit.[20]
- **Out-of-sample fit**. The models' out-of-sample prediction accuracy has been computed by splitting the sample into a training sample, covering 70% of the observations, and a test sample covering the remaining 30% of the observations. The training sample was used to estimate the models, while the test sample was used to assess the prediction accuracy out-of-sample. The prediction accuracy is measured by the Mean Absolute Percentage Error (MAPE). Greater weight is placed on the in-sample fit as it is calculated using all data as opposed to out-of-sample fit, which is calculated on 30% of the data.
- **Expected signs of the coefficients**. The sign of the coefficients of the covariates of interest should be consistent with prior expectations. For instance, SMR is expected to lead to higher need and workload. If the effect of SMR is negative then SMR is excluded from the model.

Table 3 presents the in-sample and out-of-sample fit of a sample of alternative model specifications that have been tested (model coefficients are shown in the Appendix):

- Model 1 and 2 are the simplest models. Both of them include age-gender interaction variables, whereas Model 2 also controls for practice-specific effects.[21] A comparison between these two models suggests that the inclusion of practice dummies increases the in-sample and out-of-sample fit considerably indicating that there are significant unobserved practice-specific effects, potentially associated with supply-side factors.
- Model 3 includes the rurality indicator, which improves the fit of the model slightly, as compared to model 2.
- Models 4 to 9 augment Model 3 by including MLC indicators. These models have similar explanatory and predictive power and improve upon the in- and out-of-sample fit of Model 3, indicating that these MLC variables have an impact on workload.

[20] BIC consists of two components: (1) the log-likelihood of the model and (2) the sample size and the number of parameters included in the model. The former measures the model's goodness of fit and the latter penalises the model's goodness of fit for the number of parameters included. BIC tends to select the best model more often than other measures of fit (for instance, AIC).
[21] The age-gender interaction allows the impact of age to be different between male and female patients.

- Some of the variables in Models 4 to 9 are statistically insignificant. Model 10 includes multiple MLC indicators but only statistically significant variables are retained. This model has the best fit and all variables included have the expected sign - see the next section.

- Finally, in Model 11 deprivation is interacted with age and gender, entertaining the possibility that the impact of deprivation differs across age-gender groups. This hypothesis is not supported by the econometric results as the model's fit decreases considerably compared to the other specifications

Overall, the GMS utilisation appears to be a function of age-gender and a number of MLC variables. Model 10 has the best in-sample fit and one of the best out-of-sample fit, and on balance, it is the preferred model. However, there are several alternative models that have a good fit and could be used for the determination of the allocation weights. In the remainder of the report, for illustrative purposes, Model 10 is used as the baseline model.

Table 2: Model selection

Model	Model Specification	BIC	MAPE
1	Age-gender	1,684,985	14.72
2	Age-gender + Practice effects	1,681,615	4.87
3	Model 2 + Rurality effect	1,681,408	4.88
4	Model 3 + SIMD + Standardised mortality rate (SMR)	1,680,626	4.92
5	Model 4 + Limiting long-term illness ratio	1,680,510	4.91
6	Model 5 + Bad general health	1,680,521	4.91
7	Model 6 + Mental health condition	1,680,532	4.90
8	Model 7 + Long-term sick and unemployed	1,680,541	4.90
9	Model 8 + Ethnicity - proportion of all ethnic minorities	1,680,539	4.89
10	Model 3 + SIMD + Limiting long-term illness ratio + Long-term sick and unemployed + Ethnicity - proportion of all ethnic minorities	1,680,508	4.89
11	Age-deprivation-Gender + Rurality effect + Practice effects + Limiting long-term illness ratio + Long-term sick and unemployed + Ethnicity - proportion of all ethnic minorities	1,683,285	4.63

Notes: The models have been estimated using patient level data from the 56 practices tracked by PTI; all models have been estimated using negative binomial regressions and the number of read codes associated with each patient's visit as dependent variable; BIC represents the Bayesian Information Criterion and MAPE represents the mean absolute percentage error of the out-of-sample experiment;

Source: Deloitte analysis.

4.2 Model results

Table 3 reports the coefficient estimates of Model 10:

- **Elasticities.** All the coefficients are interpreted as elasticities. The age-gender interactions, the rurality effect and deprivation variables are recorded as indicator/dummy variables and their interpretation is as follows: female patients aged 0-4 years old, for instance, have 54% fewer read codes than female patients aged 45-49 (base category), all other things being equal. The morbidity indicators (long-term limiting illness ratio and long-term sick and unemployed) and the proportion of all ethnic minorities are continuous (numerical) variables and their coefficients measure the percent change in GMS workload for one unit change in the MLC variable, all else constant.

- **Statistical significance.** All the coefficients are statistically significant at the 1% level. The only exceptions are deprivation – deciles 2 and 3, which are statistically insignificant and male patients aged 50-54 years old, which is significant at the 10% level.

- **Age impact.** Workload generally increases with age across both genders. Table 3 shows that the coefficient estimates for female patients aged between 75 and 85+ years old and males between 60 and 85+ years old have a larger average GMS utilisation (the coefficients are larger than 1) relative to the reference age category 45 to 49 years old across both genders.

- **Deprivation.** The results show that there is a positive relationship between workload and deprivation. Patients in the least deprived decile have 14% fewer read codes than the most deprived patients (base category), all other things being constant. However, deprivation is highly correlated with LLTI (see correlation matrix in Appendix 5.1) which is also included in the model. The combined effect of deprivation and LLTI would imply significantly more funds for deprived areas than those indicated by the deprivation effect alone.[22] The positive relationship between workload and deprivation was not evident in the average utilisation by deprivation decile presented in the descriptive analysis section (Figure 5). After accounting for age-gender interactions and MLC variables, the deprivation impact is more accurately measured. These results highlight the importance of analysing the underlying relationships within a multivariate framework.

- **Rural effect.** The coefficient of rural/urban indicator suggests that patients living in rural areas have, on average, 9% less utilisation of GMS relative to patients in urban areas, all other things being equal.[23]

- **MLC variables.** Areas with relatively higher morbidity indicators (e.g. higher limiting long-term illness ratios) have higher GMS utilisation. The estimated impact on workload of ethnic minorities is negative, implying that areas with a relatively larger share of ethnic minority residents tend to have relatively lower GMS utilisation rates. In the application stage the ethnicity variable could be sterilised so that its estimated negative impact in not taken into account in allocation weights.[24]

[22] Model 4 in Appendix 5.3, which excludes LLTI, indicates that the difference in utilisation between the least and most deprived areas could be up to 25%.

[23] The negative coefficient of rurality may reflect the lower access to GMS in rural areas or potential omitted factors. Although this should be controlled for in the model, its impact could be sterilised in the application stage if it is believed that providing less funds to rural areas could intensify any existing access challenges.

[24] The negative impact of ethnicity may be because minor ethnic groups have lower tendency to visit a GP for a given level of need (http://www.scotpho.org.uk/downloads/scotphoreports/scotpho160621-hospital-admissions-by-ethnic-group-v1.pdf) or under-supply of GMS.

Table 3: Model coefficient estimates (Model 10)

Age-Gender	% change in workload relative to 45-49 age group	p-value
Female 00-04	-0.54	0.00
Female 05-09	-0.79	0.00
Female 10-14	-0.72	0.00
Female 15-19	-0.32	0.00
Female 20-24	-0.12	0.00
Female 25-29	-0.14	0.00
Female 30-34	-0.12	0.00
Female 35-39	-0.11	0.00
Female 40-44	-0.07	0.00
Female 45-49	0.00	
Female 50-54	0.10	0.00
Female 55-59	0.25	0.00
Female 60-64	0.44	0.00
Female 65-69	0.65	0.00
Female 70-74	0.97	0.00
Female 75-79	1.25	0.00
Female 80-84	1.34	0.00
Female 85+	1.34	0.00
Male 00-04	-0.26	0.00
Male 05-09	-0.68	0.00
Male 10-14	-0.65	0.00
Male 15-19	-0.59	0.00
Male 20-24	-0.55	0.00
Male 25-29	-0.53	0.00
Male 30-34	-0.44	0.00
Male 35-39	-0.35	0.07
Male 40-44	-0.18	0.00
Male 45-49	0.00	
Male 50-54	0.28	0.00
Male 55-59	0.65	0.00
Male 60-64	1.13	0.00
Male 65-69	1.62	0.00
Male 70-74	2.01	0.00
Male 75-79	2.46	0.00
Male 80-84	2.60	0.00
Male 85+	2.63	0.00

SIMD Deciles	% change in workload relative to the reference category	p-value
1 (Most deprived)	**0.00**	
2	0.00	0.84
3	-0.01	0.39
4	-0.03	0.04
5	-0.05	0.00
6	-0.08	0.00
7	-0.08	0.00
8	-0.08	0.00
9	-0.10	0.00
10 (Least deprived)	-0.14	0.00
Rurality effect		
Urban	0.09	0.00
Rural	**0.00**	

MLC indicators	% change in workload for every unit increase in each indicator	p-value
Limiting long-term illness ratio	0.27	0.00
Long-term sick and unemployed	0.94	0.01
Ethnicity - proportion of all ethnic minorities	-0.37	0.00

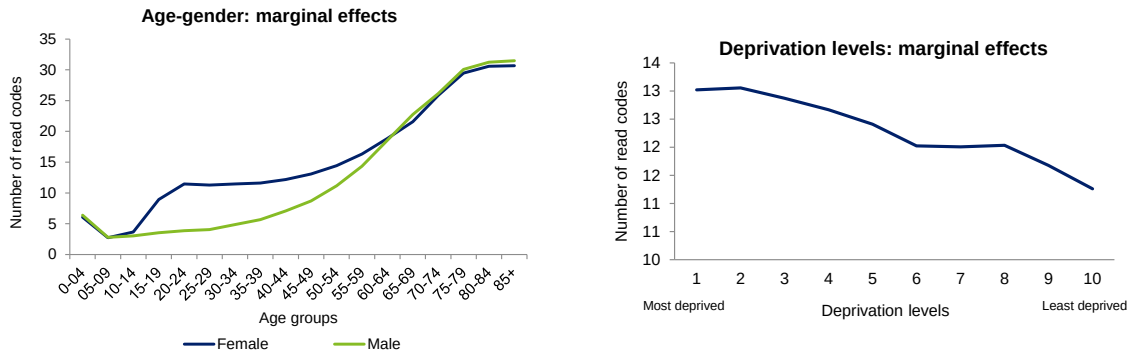
Notes: Reference categories are highlighted in bold

Source: Deloitte analysis

Figure 11 depicts the model predicted average utilisation by age-gender and deprivation.[25] These plots reflect the coefficient estimates discussed above and present the model predicted workload by the three key variables (age, gender and deprivation).

[25] These are also called marginal effects and have been computed by setting the values of all other variables at the sample mean values.

Figure 11: Marginal effects: predictive workload by age-gender profiles and deprivation levels



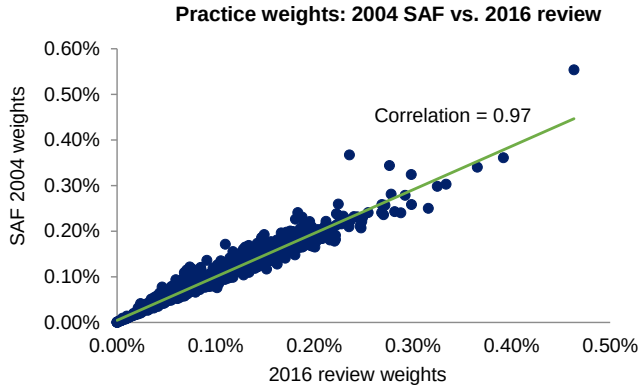
Source: Deloitte analysis

4.3 Allocation weights: comparison with 2004 SAF formula

Figure 12 compares the practice weights from the 2004 SAF[26] with the weights estimated in this report (2016 review) using Model 10. Although the correlation between these two sets of weights is high (0.97), there are significant differences across several practices.[27] This is clearly presented in Table 4, which shows that the difference in weights between the two formulae is more that 10% for c.35% of all Scottish practices. Only for c.39% of the practices the difference in weights is less than 5%. The sources of the differentials between the 2004 SAF and 2016 review weights are explored in the next section.

Should the allocation weights be derived using the results from Model 10 discussed above, about 35% of all the Scottish practices would see a relatively large change - of more than 10% - in their weighting. Thus, there is the potential for the Scottish Government to consider introducing a pace of change policy in order to gradually move towards the target allocations based on the updated SAF analysis. A similar approach has been implemented by NHS England in allocating the 2016/17 to 2020/21 national budget to Clinical Commissioning Group areas[28].

Figure 12: Practice allocation weights: 2004 SAF vs. 2016 review



Source: Deloitte analysis

[26] The unit cost adjustment has not been included in the 2004 SAF weights in order to make a like-for-like comparison with the weights generated in this report.
[27] Both formulae have been applied to the same dataset, therefore the difference cannot be explained by differences in the application of the formulae.
[28] Study available online at: https://www.england.nhs.uk/wp-content/uploads/2016/04/1-allctins-16-17-tech-guid-formulae.pdf

Table 4: Differences in practice weights between the 2016 review and SAF 2004

Absolute percentage difference between relative weights: 2016 review vs. SAF 2004	Number of practices
0% - 5%	39.3%
5% - 10%	27.2%
10% - 20%	25.7%
20%+	7.8%

Source: Deloitte analysis

4.4 Differences between relative weights: SAF 2004 vs. 2016 review

There are two types of factors that might explain the difference in the weights presented in the previous section: (1) differences in the underlying data used to estimate the impact of the need drivers on utilisation and (2) differences in methodology. Differences in the methodology may lead to differences in the estimated impact of age-gender, MLC and rurality on workload. In particular, there are four hypotheses that have been investigated:

1. **Age-gender utilisation profiles (underlying data).** The impact of age-gender in the 2004 SAF formula has been estimated using the 2011/12 PTI sample, whereas the 2016 review uses the 2012/13 PTI sample. If the age-gender utilisation profiles differ between these two datasets, for instance, because of sample variability, the corresponding estimated coefficients could be different.

2. **Age-gender coefficients.** The estimated impact of age-gender on workload might differ across the two formulae due to different methodologies applied.

3. **MLC coefficients.** The estimated impact of additional need variables may also be different between the 2004 SAF and 2016 review. SAF 2004 uses a single composite indicator, the Arbuthnott index[29], for each practice to capture the impact of the MLC factors on workload, estimated separately from the impact of age-gender interactions. The 2016 review uses a set of separate MLC factors (deprivation deciles and morbidity indicators) corresponding to each patient data zone estimated within the same multivariate regression.

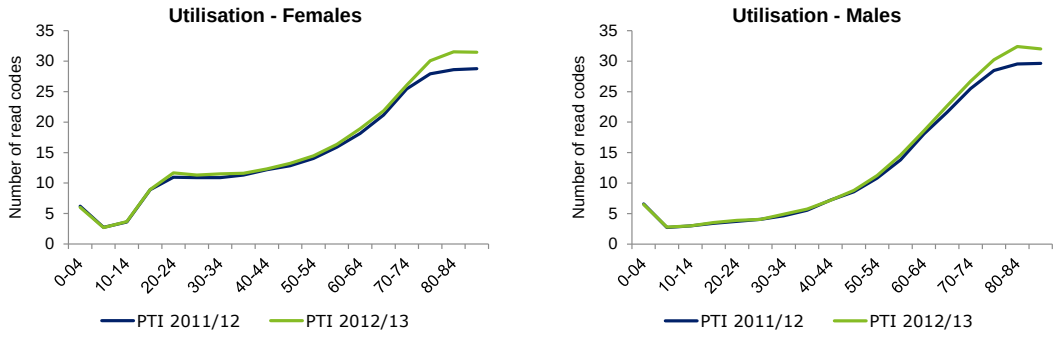
4. **Rural/urban adjustment.** The 2004 SAF workload model does not include a rural/urban indicator.

Age-gender utilisation profiles

Figure 13 compares the average GMS utilisation by age-gender between the 2011/12 and 2012/13 datasets. Average utilisations for both samples mirror closely each other. There is a small difference in utilisation at the tail end of the age distribution, which is quite small to explain the differences between the two formulae (especially given that a relatively small proportion of the population is situated at this part of the distribution).

[29] The Arbuthnott index is calculated as a weighted average of four metrics: (1) the standardised mortality rate for people under the age of 65; (2) the unemployment rate; (3) the proportion of elderly people claiming income support; (4) households with two or more indicators of deprivation.

Figure 13: Unconditional utilisation averages by age-gender: PTI 2011/12 vs PTI 2012/13

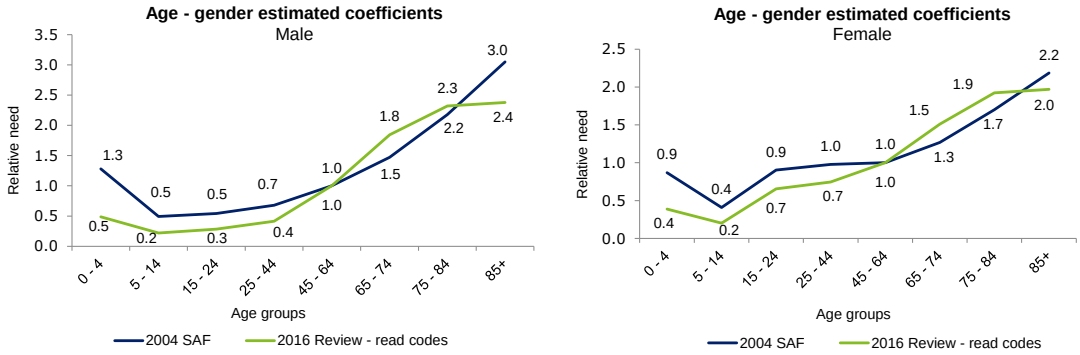


Source: PTI 2011/12 and 2012/13

Age-gender coefficients

Figure 14 shows the age-gender estimated coefficients used by the two formulae. The difference between the two sets of coefficients is significant for several age categories. The 2004 SAF places more weight on younger patients compared to the 2016 review. Both models include age-gender variables; MLC indicators are excluded. Therefore, the difference between these two sets of results is driven by the differences in the methodological approaches used in SAF 2004 vs the 2016 update. The main difference between SAF 2004 and the 2016 update is the inclusion of patients with zero consultations in the 2016 sample.

Figure 14: SAF 2004 weights vs. 2016 age-gender coefficients



Source: Deloitte analysis

The 2004 SAF estimated the age-gender impact on workload using data for patients who had at least one consultation whereas the 2016 review uses the entire registered population within the sample including zero-consultation patients. In order to investigate the impact of the exclusion of zero consultation patients on the allocation weights, the 2016 review model has been estimated using data only for patients with at least one consultation.[30]

The results of this analysis are presented in Table 5. The second column shows the difference between the weights generated by the two formulae with zero-consultation patients being included in the estimation sample of the 2016 review model. The third column shows the same comparison but the zero-consultation patients have been excluded from the 2016 review sample. The results suggest that part of the difference in the age-gender coefficients between the two formulae can be explained by the fact that 2004 SAF has excluded the zero-consultation patients from the analysis.

[30] In order to investigate the impact of zero-consultation patients, both the 2004 SAF and 2016 review weights used in this analysis use only age-gender as need drivers (MLC and other factors are excluded from the models).

Table 5: Impact of zero-consultation patients

Absolute percentage difference between relative weights: 2016 review vs. SAF 2004	Number of practices - Including zero consultation patients	Number of practices - Excluding zero consultation patients
0% - 5%	68.2%	77.2%
5% - 10%	22.8%	18.5%
10% - 20%	7.3%	3.0%
20%+	1.7%	1.2%

Source: Deloitte analysis

MLC coefficients

Table 6 investigates the impact of MLC variables on the difference in allocation weights predicted by the two formulae. The second column reflects weights estimated using both age-gender and MLC variables whereas the third column reflects weights that are based only on age-gender models. The inclusion of the MLC variables in the allocation models leads to a significant increase in difference between the 2004 SAF and 2016 review weights suggesting that differences in the MLC adjustment between the two formulae has a significant effect on allocation weights.

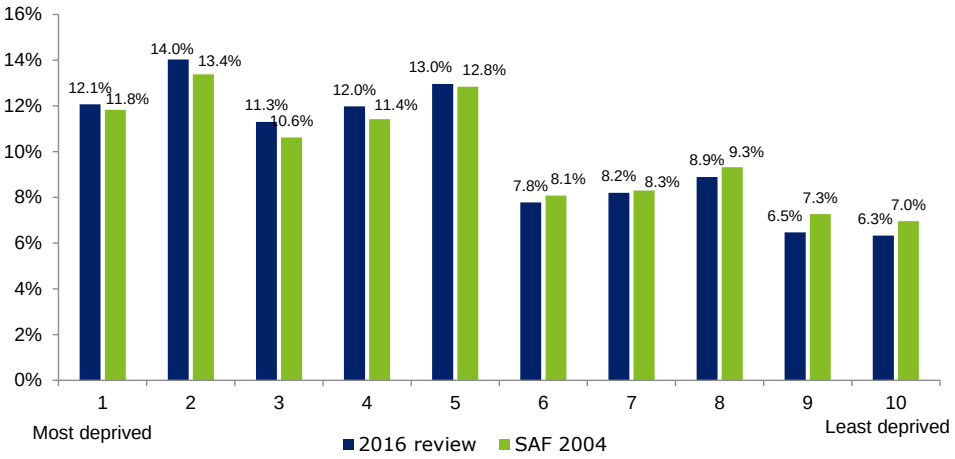
Table 6: Differences in relative weights due to the inclusion of MLC indicators

Absolute percentage difference between relative weights: 2016 review vs. SAF 2004	Number of practices - Including MLC effects	Number of practices - Excluding MLC effects
0% - 5%	39.3%	68.2%
5% - 10%	27.2%	22.8%
10% - 20%	25.7%	7.3%
20%+	7.8%	1.7%

Source: Deloitte analysis

Figure 15 shows the allocation of funds predicted by the 2004 SAF and 2016 review by deprivation deciles[31]. The 2016 review allocates more funds to the more deprived areas compared to the 2004 SAF. For instance, under the 2016 review formula, 37.4% of the GMS budget is allocated in the three most deprived deciles, which is 1.6% more than under the 2004 SAF.

Figure 15: Allocation of GMS budget by deprivation decile



Source: Deloitte analysis

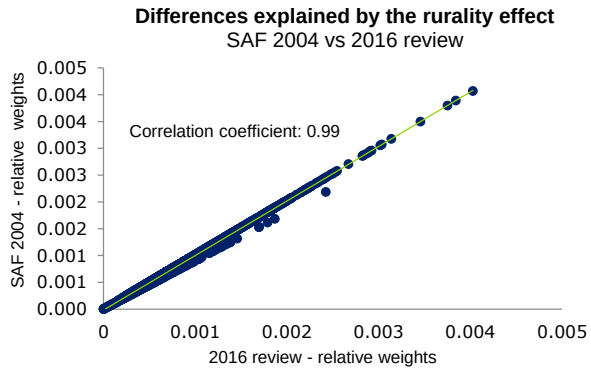
Rural/urban adjustment

The impact of the rurality was tested by comparing the results generated by a 2016 model with and without the rurality adjustment. The weights implied by these two models are compared in Figure 16.

[31] Practices are assigned to SIMD deciles based on their post code. The allocation weights for the 2016 review presented in Figure 15 are derived using Model 10.

The difference between them is not statistically significant suggesting that the difference between the 2004 SAF and 2016 review cannot be explained by the rurality adjustment.

Figure 16: Impact of the rurality adjustment on practice relative weights



Source: Deloitte analysis

Overall, the evidence presented in this section suggests that the main difference between the two formulae is the MLC adjustment. The two formulae also assume a different impact of age-gender on workload. This difference is partly explained by the exclusion of the zero-consultation patients from the 2004 SAF model.

5 Appendix

Table 7 presents summary statistics related to the distribution of the registered population and utilisation of GMS by age-gender.

Table 7: Sample summary statistics

	Age	Total number of patients	Patients with zero consultations	Average consultations	Average read codes*	Share of patients with zero consultations
Female	**0-04**	10,302	925	3.4	6.0	9.0%
	05-09	9,392	3,348	1.7	2.7	35.6%
	10-14	9,668	3,307	1.9	3.7	34.2%
	15-19	10,705	1,547	3.6	8.9	14.5%
	20-24	11,569	1,270	4.2	11.7	11.0%
	25-29	11,650	1,486	4.2	11.3	12.8%
	30-34	11,406	1,412	4.4	11.5	12.4%
	35-39	11,351	1,523	4.5	11.6	13.4%
	40-44	13,854	1,934	4.7	12.3	14.0%
	45-49	13,908	1,943	4.8	13.2	14.0%
	50-54	13,202	1,909	5.0	14.5	14.5%
	55-59	11,813	1,592	5.3	16.3	13.5%
	60-64	11,217	1,355	5.9	18.9	12.1%
	65-69	10,407	898	6.6	21.8	8.6%
	70-74	8,029	588	7.5	26.1	7.3%
	75-79	6,666	471	8.4	30.1	7.1%
	80-84	5,078	364	8.2	31.5	7.2%
	85+	5,043	329	7.6	31.5	6.5%
Male	**0-04**	10,832	906	3.7	6.5	8.4%
	05-09	9,790	3,529	1.6	2.8	36.0%
	10-14	10,008	3,963	1.5	3.0	39.6%
	15-19	10,751	4,036	1.7	3.5	37.5%
	20-24	11,118	4,208	1.8	3.9	37.8%
	25-29	11,378	4,629	1.8	4.1	40.7%
	30-34	11,435	4,312	2.1	4.9	37.7%
	35-39	11,296	4,014	2.4	5.8	35.5%
	40-44	13,689	4,532	2.7	7.2	33.1%
	45-49	14,287	4,481	2.9	8.8	31.4%
	50-54	13,169	3,455	3.5	11.2	26.2%
	55-59	11,582	2,479	4.3	14.6	21.4%
	60-64	11,059	1,632	5.2	18.6	14.8%
	65-69	9,855	1,049	6.3	22.7	10.6%
	70-74	7,071	561	7.4	26.7	7.9%
	75-79	5,497	407	8.5	30.2	7.4%
	80-84	3,501	277	8.7	32.4	7.9%
	85+	2,415	178	8.3	32.0	7.4%
All patients		**363,993**	**74,849**			**20.6%**

*Read codes adjusted for home visits (multiplied by 3).

Source: PTI data set, GP workforce list sizes, Deloitte analysis

5.1 Correlation analysis of the additional need indicators

Table 8 shows the correlation coefficients between all the MLC variables considered in the analysis.

Table 8: Correlation matrix between the additional need factors tested

	Limiting long-term illness ratio	All causes SMR 70 and under	BGH - bad or very bad	Unemployed and sick	Mental health condition	Deprivation deciles
Limiting long-term illness ratio	1.00	0.71	0.93	0.03	0.77	-0.85
All causes SMR 70 and under	0.71	1.00	0.70	0.05	0.65	-0.67
BGH - bad or very bad	0.93	0.70	1.00	0.05	0.76	-0.81
Unemployed and sick	0.03	0.05	0.05	1.00	-0.13	-0.05
Mental health condition	0.77	0.65	0.76	-0.13	1.00	-0.67
Deprivation deciles	-0.85	-0.67	-0.81	-0.05	-0.67	1.00

Source: Deloitte analysis

5.2 Dispersion tests

The dispersion test used to examine the presence of over-dispersion in the data is the LM test proposed by Cameron and Trivedi (1990). This tests the null hypothesis that a Poisson model is appropriate to model the workload data against the alternative of over-dispersion and/or under-dispersion.

The null hypothesis is consistent with the standard Poisson model assumption that requires the mean of the data distribution to equal its variance:

$$Var(y) = E(y) = \mu \ .$$

The null hypothesis is assessed against an alternative form for the variance of the data distribution, consistent with the negative binomial models:

$$Var(y) = (1 + \alpha) * \mu$$

Baseline model	Test statistic	P-value	Sample dispersion estimate
Consultations	117.07	0.00	4.79
Read codes	175.73	0.00	15.28

Source: Deloitte analysis

The dispersion tests for both consultations and read codes show that the sample dispersion parameter is higher than 1, thus rejecting the null hypothesis of equidispersion and suggesting that a negative binomial model is more appropriate than a Poisson parametrisation.

5.3 Model coefficients

This section sets out the coefficients of the models discussed in section 4.1. In order to interpret the coefficient estimates, they need to be exponentiated (e^{β}). This would change the interpretation by considering the *multiplicative* changes in workload for each unit change in the explanatory variables included (as in Section 4.2). For example, the impact of living in an urban area as opposed to a rural area is estimated in Model 10 to be 1.088. Thus, the GMS workload associated with patients living in urban areas are on average 8.8% higher than patients living in rural areas, everything else being held constant. Similarly, for every unit increase in limiting long-term illness ratios, patients have an increase in their GMS workload by 26% on average, all else constant.

Table 9: Estimated model specifications for the number of read codes adjusted for home visits

Explanatory variables	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10	Model 11
Female:0-04	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	
Female:05-09	0.449***	0.447***	0.447***	0.45***	0.449***	0.449***	0.449***	0.449***	0.449***	0.449***	
Female:10-14	0.605***	0.601***	0.602***	0.605***	0.605***	0.605***	0.605***	0.605***	0.605***	0.604***	
Female:15-19	1.471***	1.474***	1.476***	1.483***	1.478***	1.478***	1.478***	1.478***	1.478***	1.48***	
Female:20-24	1.904***	1.908***	1.906***	1.9***	1.896***	1.896***	1.896***	1.896***	1.896***	1.896***	
Female:25-29	1.878***	1.881***	1.879***	1.866***	1.863***	1.863***	1.865***	1.865***	1.865***	1.865***	
Female:30-34	1.902***	1.898***	1.898***	1.898***	1.896***	1.896***	1.896***	1.896***	1.896***	1.896***	
Female:35-39	1.916***	1.916***	1.919***	1.923***	1.921***	1.921***	1.921***	1.921***	1.921***	1.921***	
Female:40-44	2.032***	2.012***	2.014***	2.022***	2.018***	2.018***	2.018***	2.018***	2.018***	2.018***	
Female:45-49	2.171***	2.168***	2.171***	2.173***	2.166***	2.166***	2.166***	2.168***	2.166***	2.166***	
Female:50-54	2.382***	2.387***	2.392***	2.399***	2.389***	2.392***	2.389***	2.392***	2.389***	2.389***	
Female:55-59	2.689***	2.689***	2.697***	2.71***	2.702***	2.702***	2.702***	2.702***	2.699***	2.699***	
Female:60-64	3.111***	3.102***	3.108***	3.133***	3.124***	3.124***	3.124***	3.124***	3.121***	3.121***	
Female:65-69	3.55***	3.55***	3.561***	3.589***	3.575***	3.575***	3.575***	3.575***	3.572***	3.572***	
Female:70-74	4.238***	4.259***	4.267***	4.293***	4.28***	4.28***	4.28***	4.28***	4.276***	4.276***	
Female:75-79	4.831***	4.874***	4.879***	4.904***	4.879***	4.884***	4.879***	4.884***	4.879***	4.879***	
Female:80-84	5.008***	5.078***	5.073***	5.094***	5.068***	5.073***	5.073***	5.073***	5.063***	5.063***	
Female:85+	5.008***	5.094***	5.089***	5.104***	5.073***	5.073***	5.078***	5.078***	5.073***	5.073***	
Male:0-04	1.063***	1.059***	1.059***	1.058***	1.058***	1.058***	1.058***	1.058***	1.058***	1.058***	
Male:05-09	0.462***	0.457***	0.457***	0.459***	0.459***	0.459***	0.459***	0.459***	0.459***	0.459***	
Male:10-14	0.499***	0.497***	0.498***	0.501***	0.501***	0.501***	0.501***	0.501***	0.501***	0.501***	
Male:15-19	0.585***	0.583***	0.584***	0.587***	0.585***	0.585***	0.585***	0.585***	0.585***	0.585***	
Male:20-24	0.644***	0.645***	0.644***	0.641***	0.639***	0.639***	0.639***	0.64***	0.64***	0.64***	
Male:25-29	0.674***	0.677***	0.676***	0.67***	0.668***	0.668***	0.668***	0.668***	0.668***	0.668***	
Male:30-34	0.811***	0.811***	0.811***	0.806***	0.804***	0.804***	0.804***	0.804***	0.805***	0.805***	
Male:35-39	0.941***	0.943***	0.942***	0.938***	0.937***	0.937***	0.937***	0.937***	0.937***	0.937***	
Male:40-44	1.185***	1.175***	1.177***	1.175***	1.172***	1.172***	1.172***	1.172***	1.172***	1.172***	
Male:45-49	1.449***	1.438***	1.441***	1.441***	1.438***	1.438***	1.438***	1.438***	1.438***	1.438***	
Male:50-54	1.842***	1.839***	1.846***	1.852***	1.844***	1.846***	1.846***	1.846***	1.844***	1.844*	
Male:55-59	2.356***	2.358***	2.366***	2.38***	2.375***	2.375***	2.375***	2.375***	2.373***	2.373***	
Male:60-64	3.043***	3.04***	3.053***	3.08***	3.071***	3.071***	3.071***	3.071***	3.068***	3.068***	
Male:65-69	3.728***	3.714***	3.728***	3.781***	3.77***	3.77***	3.77***	3.77***	3.766***	3.766***	
Male:70-74	4.289***	4.289***	4.306***	4.345***	4.336***	4.336***	4.336***	4.336***	4.328***	4.328***	
Male:75-79	4.855***	4.923***	4.943***	4.998***	4.983***	4.983***	4.983***	4.983***	4.978***	4.978***	
Male:80-84	5.073***	5.14***	5.155***	5.197***	5.176***	5.176***	5.176***	5.181***	5.176***	5.176***	
Male:85+	5.109***	5.16***	5.176***	5.244***	5.212***	5.217***	5.217***	5.217***	5.212***	5.212***	
SIMD = 1				1.00	1.00	1.00	1.00	1.00	1.00	1.00	
SIMD = 2				0.97**	1.006	1.007	1.007	1.008	1.006	1.003	
SIMD = 3				0.943***	0.994	0.997	0.995	0.996	0.994	0.988	
SIMD = 4				0.914***	0.978	0.981	0.979	0.981	0.98	0.972**	
SIMD = 5				0.882***	0.96***	0.964**	0.962**	0.963**	0.962**	0.953***	
SIMD = 6				0.842***	0.931***	0.935***	0.933***	0.935***	0.933***	0.923***	
SIMD = 7				0.833***	0.93***	0.933***	0.931***	0.933***	0.932***	0.922***	
SIMD = 8				0.826***	0.934***	0.938***	0.937***	0.938***	0.936***	0.924***	
SIMD = 9				0.79***	0.906***	0.91***	0.908***	0.909***	0.908***	0.897***	
SIMD = 10				0.756***	0.874***	0.878***	0.876***	0.879***	0.875***	0.865***	
Rurality effect: rural			1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Rurality effect: urban			1.13***	1.088***	1.083***	1.082***	1.082***	1.082***	1.085***	1.088***	1.085***
SMR - all causes under 70				1.04***	1.014**	1.013	1.015**	1.014**	1.015***		
Limiting long-term illness ratio					1.257***	1.226***	1.24***	1.237***	1.214***	1.266***	1.261***
Bad general health						1.039	1.042	1.039	1.046		
Mental health condition							0.982	0.989	0.995		
Long-term sick and unemployed								1.679	1.852**	1.941**	2.109**
Ethnicity - Proportion of all minor									0.62***	0.633***	0.638***
BIC	**1,684,985**	**1,681,615**	**1,681,408**	**1,680,626**	**1,680,510**	**1,680,521**	**1,680,532**	**1,680,541**	**1,680,539**	**1,680,508**	**1,683,285**

*p<0.1, **p<0.05, ***p<0.01

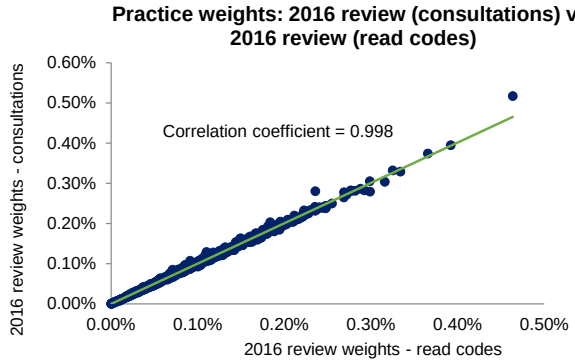
Zero coefficients represent reference categories that are systematically dropped in the estimation associated with categorical variables. Coefficient estimates for practice dummies and age-gender-deprivation interactions in model 11 are not shown.

Source: Deloitte analysis

5.4 Read codes vs. consultation weights

Model 10 was also estimated using the total number of consultations associated with each patient record as a proxy for GMS workload. The results generated by Model 10 based on consultations are very similar to the results generated by the same model specification using read codes instead. Figure 17 shows that the relative weights generated by these two types of models are almost perfectly correlated with a correlation coefficient of almost 1.

Figure 17: Relative weights based on 2016 models using read codes and consultations



Source: Deloitte analysis

6 References

Brilleman S.L, Gravelle H., Hollinghurst S., Purdy S., Salisbury C., and Windmeijer F. (2014). *"Keep it simple? Predicting primary health care costs with clinical morbidity measures".* Journal of Health Economics 35, 109-122.

Cameron A.C. and Trivedi P.K. (2005). "*Microeconometrics: Methods and Applications"*. Cambridge University Press.

Cameron, A.C. and Trivedi, P.K. (1990). *"Regression-based Tests for Overdispersion in the Poisson Model*". Journal of Econometrics, 46, 347–364.

Manning W.G. and Mullahy J. (2001). *"Estimating Log Models: To Transform or Not To Transform?"* Journal of Health Economics, 20(4): 461-494.

McLean G., Guthrie B. and Watt G. (2015). "*General practice funding underpins the persistence of the inverse care law: cross-sectional study in Scotland*". British Journal of General Practice, December 2015.

NHS England and Nuffield Trust (2014). *Meeting need or fuelling demand? Improved access to primary care and supply-induced demand.* http://www.nuffieldtrust.org.uk/sites/files/nuffield/publication/140630_meeting_need_or_fuelling_demand.pdf

Arain M., Nicholl J. and Campbell M. (2013). *Patients' experience and satisfaction with GP led walk-in centers in the UK; a cross sectional study.* BMC Health Services Research 2013 13:142. http://bmchealthservres.biomedcentral.com/articles/10.1186/1472-6963-13-142

NHS England (2016). *Technical Guide to Allocation Formulae and Pace of Change.* https://www.england.nhs.uk/wp-content/uploads/2016/04/1-allctins-16-17-tech-guid-formulae.pdf

Marmot M. (2005). *Social determinants of health inequalities*. Lancet 365: 1099-104 http://www.who.int/social_determinants/strategy/en/Marmot-Social%20determinants%20of%20health%20inqualities.pdf