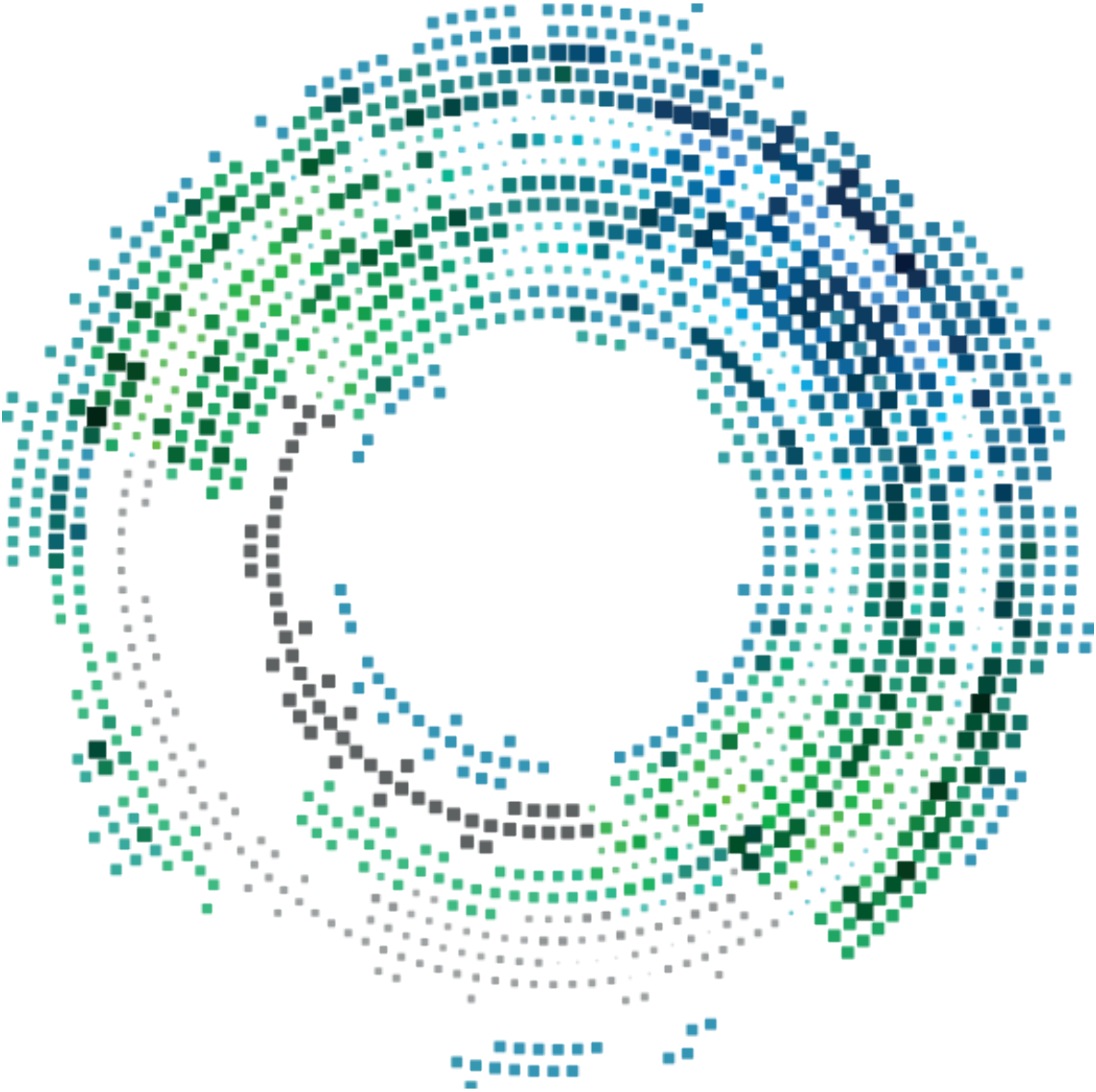
Deloitte.



Scottish Allocation Formula – General Medical Services

Unit cost formula review

8 November 2016

Contents

Important Notice from Deloitte 2

Executive summary 3

1 Introduction 5

2 Sources of geographical variation in costs 7

3 Rurality and remoteness review 9

4 MFF review 13

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 14 April 2016 ("the Contract"), as extended by email dated 01 August 2016, and on the basis of the scope and limitations set out below, and on the basis of the scope and limitations set out below.

The Final Report has been prepared solely for the purposes of reviewing the methodology underpinning the unit cost dimension of the Scottish Allocation Formula for General Medical Services, 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.

The Final Report is provided exclusively for Scottish Government's use under the terms of the Contract. No party other than the Scottish Government 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.

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.

All copyright and other proprietary rights in the Final Report remain the property of Deloitte LLP and any rights not expressly granted in these terms or in the Contract are reserved.

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 and has been developed with the aim to allocate the primary care budget in proportion to practice expected workload and account for unavoidable differences in the costs of providing General Medical Services (GMS) in different parts of the country.

The budget allocated by the SAF is known as the Global Sum, which constitutes the largest component of the general practice payment. The allocation 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 unit cost adjustment compensates practices on the basis of their degree of rurality and remoteness (R/R) and the Market Forces Factor (MFF).

The R/R adjustment recognises that rural and remote practices might have higher than average unit costs due to unavoidably small list sizes and the presence of economies of scale in the provision of GMS. The MFF seeks to compensate practices for unavoidable staff costs due to regional differences in pay rates.

The objective of this document is to review the unit cost dimension of the SAF and provide recommendations on how to potentially improve the methodology underpinning the unit cost adjustment (the workload dimension and other practice payments are out of the scope of this review). The results of the review suggest that the current methodology has some substantial weaknesses which might over- or under-estimate the unavoidable costs practices actually face.

1. **Double-counting.** The model used to estimate the R/R adjustment models practice income, which reflects costs, fees and allowances, as a function of three R/R proxies measuring the degree of rurality and population sparsity. However, costs may vary across practices because of other variables, for example, local population need partly determined by age and deprivation, MFF and the type of services provided. Given these factors are not controlled for, the R/R variables may capture their effect, which is also accounted for in other parts of the formula, potentially leading to double-counting.

2. **Rurality/remoteness - Income variable.** The variable analysed in the R/R model reflects practice fees and allowances, which are based on the previous payment system and therefore the model estimates may reflect both the underlying R/R cost factors but also the historical payment mechanism. If higher expenses of rural practices are influenced by previous payment systems, then *practices* should not be compensated for this proportion of costs.

3. **Rurality/remoteness - Explanatory variables.** The model assumes that the more rural or remote a practice is, the higher the costs, which may not be a valid assumption. For instance, there might be practices in rural areas that have significant list sizes and therefore do not require an allowance for unavoidable smallness.

4. **MFF.** The current MFF adjustment is applied at the Health Board level and hence it does not take into account possible differences in staff costs across practices within Health Boards: Health Boards cover large geographical areas with, on average, 70 practices, which are unlikely to face the same labour market conditions. Furthermore, it assumes that an overall wage index, based on all occupation groups, may adequately reflect regional pay differentials for doctors and nurses.

This report provides a number of recommendations on how to potentially improve the current unit cost formula.

- **Unavoidable smallness.** The costs of unavoidable smallness could be estimated by the cost-scale relationship, which could be used to infer the economies of scale and the excess costs a small list size imposes on rural and remote practices. This is similar to the approach used by NHS England (NHSE) in the 2016/17 allocation formula.

- **MFF.** Three alternative approaches are proposed. The simplest one is based on wage data recorded by occupation and local authority, which could provide more granular information to estimate and benchmark staff costs. The more sophisticated approach relies on practice specific data and regression analysis.

The improvements proposed in the analysis are intended to provide more accurate estimates of the costs related to practices' locations and geographical variations in employee costs.

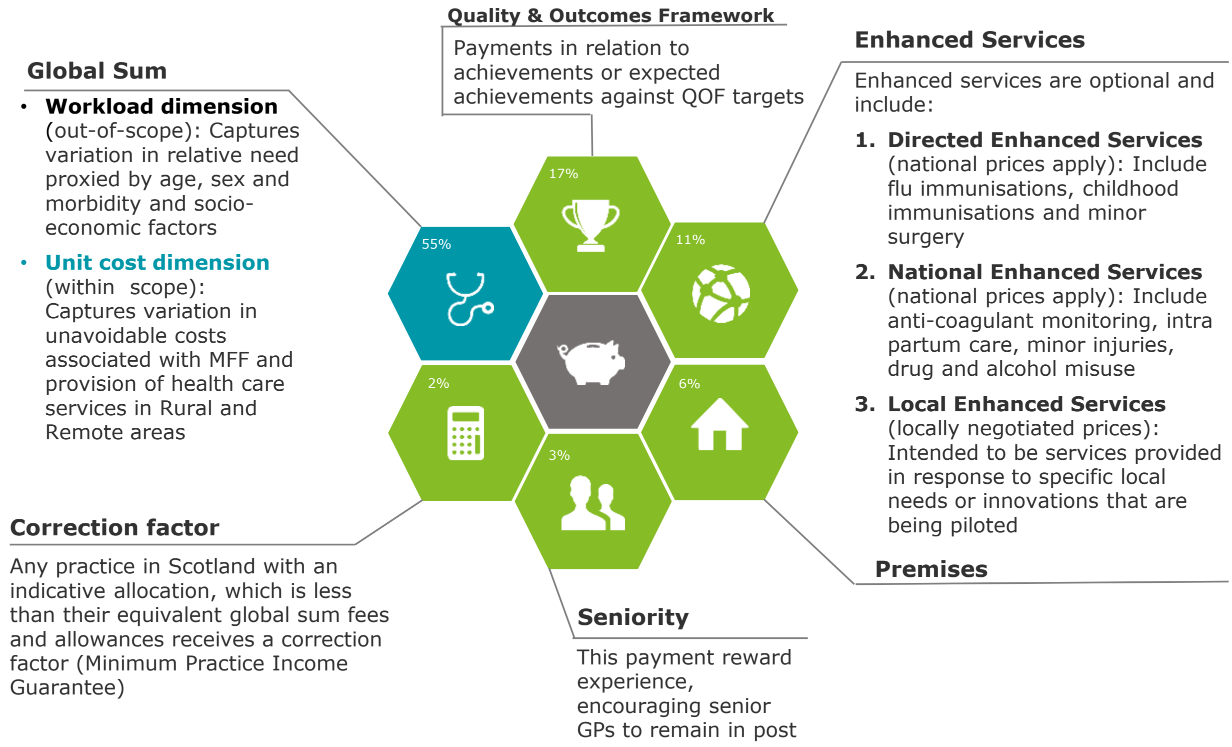
1 Introduction

The budget allocated by the Scottish Allocation Formula (SAF) is known as the Global Sum. The Global Sum is the largest component of practice income and intended to fund the core general medical services (GMS) required by local populations.

1.1 Background

Figure 1 describes the key components of the general practice funding. The Global Sum accounts for c.55% of the total GMS funding and has two dimensions: (i) a workload and (ii) a unit cost dimension. The workload dimension is designed to account for differences in population relative need and practice workload. The unit cost dimension aims to capture cost differences in the provision of GMS associated with the location of a practice.

Figure 1: How practices are funded



Quality & Outcomes Framework

Payments in relation to achievements or expected achievements against QOF targets

Global Sum

- **Workload dimension** (out-of-scope): Captures variation in relative need proxied by age, sex and morbidity and socio-economic factors
- **Unit cost dimension** (within scope): Captures variation in unavoidable costs associated with MFF and provision of health care services in Rural and Remote areas

Enhanced Services

Enhanced services are optional and include:

1. **Directed Enhanced Services** (national prices apply): Include flu immunisations, childhood immunisations and minor surgery
2. **National Enhanced Services** (national prices apply): Include anti-coagulant monitoring, intra partum care, minor injuries, drug and alcohol misuse
3. **Local Enhanced Services** (locally negotiated prices): Intended to be services provided in response to specific local needs or innovations that are being piloted

Correction factor

Any practice in Scotland with an indicative allocation, which is less than their equivalent global sum fees and allowances receives a correction factor (Minimum Practice Income Guarantee)

Premises

Seniority

This payment reward experience, encouraging senior GPs to remain in post

Source: http://www.nhshistory.net/gppay.pdf; *Invitation to Quote (Ref: SG/RES/2015/051): Review of the Scottish Allocation Formula: Modelling the Variation in Unit Costs of Providing Primary Care.*

The unit cost dimension makes two general types of adjustments: (i) a rurality/remoteness (R/R) and (ii) a Market Forces Factor (MFF) adjustment. The R/R adjustment primarily aims to compensate practices for unavoidable smallness, as rural/remote practices may have higher than average unit costs due to unavoidably small list sizes and the presence of economies of scale in the provision of GMS. The MFF seeks to compensate practices for unavoidable staff costs due to regional differences in pay rates.

The QOF payment, the second largest component of the practice payment, aims to incentivise practices to improve health outcomes. Payments associated with Enhanced Services seek to compensate practices for the provision of non-core services such as minor surgery, flu immunisation, and other types of GMS that might be required in response to local needs or innovations that are piloted. Other payments include Premises (c.6%), Seniority (c. 3%) and Minimum Practice Income Guarantee (c.2%).

1.2 Scope of this report

The purpose of this report is to review the unit cost dimension of the SAF and provide recommendations on how to potentially improve the current approach. The workload dimension and other practice payments are out of the scope of this review. More specifically, the review focuses on the following areas:

1. **Sources of variation.** Understand the degree to which the current formula is designed, in principle, to capture all key sources of geographical differences in costs;

2. **Methodology.** Evaluate the degree to which the formula is expected to provide reasonable estimates; and

3. **Possible improvements.** Provide recommendations to improve the current framework.

The remainder of this report is organised as follows:

- Section 2 identifies the key sources of geographical variation in general practice costs;
- Section 3 reviews the R/R adjustment used in the unit cost formula; and
- Section 4 reviews the MFF adjustment.

2 Sources of geographical variation in costs

Table 1 identifies seven key sources of geographical variation in the costs of providing GMS across Scotland including case-mix, type of services provided and unavoidable smallness. Another factor that might be associated with geography is travelling costs associated with home and nursery visits, however, these are expected to be relatively small.[1]

Factors related to rurality and remoteness might have a positive, negative or ambiguous impact on practice costs. For example, costs related to the nature of services provided and unavoidable smallness might be higher, premises and access costs could be lower while the case-mix and staff costs could have an ambiguous impact. Furthermore, most of the sources associated with geographical variation in costs are captured in other parts of the payment system, as listed in Table 1. The exceptions are factors associated with unavoidable smallness, access and alternative settings of care. This has important implications as the overall payment mechanism may double-count R/R.

Table 1: Geographical variation in practice costs

Source	Description	Impact on costs	Incorporated through other parts of the payment system?
Case-mix	R/R areas typically have older populations which have greater need for health care services (see 2004 SAF). On the other hand, patient need and complexity associated with morbidity and life circumstances factors, such as deprivation, limiting long-term illness ratio, may be lower in rural and remote areas.	+/-	SAF (workload component)
Nature of services	R/R practices may provide additional and/or more complex services than urban practices due to the limited availability of alternative settings of care.	+	Enhanced Services
Unavoidable smallness	R/R areas may have unavoidably small list sizes and if there are economies of scale in the provision of GMS, they would have higher costs than urban practices, all other things being equal.	+	No
Staff costs	Staff salaries may vary across Scotland due to geographical differences in costs of living and area attractiveness. Lower cost of living in rural areas would translate into lower staff costs whereas potentially lower location attractiveness	+/-	MFF

[1] Petrol costs might be higher in rural areas as GPs are likely to have to travel longer distances. However, travelling time might be higher in urban areas due to congestion.

Source	Description	Impact on costs	Incorporated through other parts of the payment system?
	might require practices to offer higher pay rates to attract staff.		
Premises	Rent and premise values are typically lower in R/R areas which would suggest lower costs of GMS provision, keeping everything else constant.	-	Premises
Access	Practices in R/R areas cover more dispersed populations, which require longer travel times to access GMS. If utilisation of GMS is negatively correlated with travel times, then rural/remote areas will have lower workload per registered patient, all other things being equal.	-	No
Alternative settings of care	Due to access constraints to alternative settings of care (community and/or acute) associated with longer travelling times, patients in R/R areas might use GMS more frequently than patients in urban areas.	+	No

3 Rurality and remoteness review

This section reviews the R/R part of the unit cost component of the SAF providing:

- A brief description of the 2004 SAF approach;
- Possible limitations of the current approach;
- Recommendations to potentially improve the R/R methodology; and
- A simulation aiming to illustrate the scale of unavoidable smallness.

3.1 2004 SAF approach

The R/R adjustment in the current formula is based on a regression model that aims to explain differences in costs across practices, approximated by practice fees and allowances, as a function of three R/R proxies:

- Population density (number of hectares per resident);
- Population scarcity (proportion of people living in settlements of less than 500 people); and
- The proportion of practice list attracting road mileage payments.

The model estimates indicate that the more rural or remote a practice is the higher the costs. These estimates are used to compute a R/R index, which is applied on the weighted list size predicted by the workload model. The resulting weighted list seeks to account for differences in practices workload as well as unavoidable costs associated with R/R.

3.2 Limitations

There are three key limitations in the current approach.

1. **Income variable.** Because income captures both the underlying practice costs but also fees and allowances determined by the previous payment system, the model could reflect both the underlying R/R factors but also the historical payment mechanism. If higher expenses of rural practices are influenced by previous payment systems, then these costs should not be compensated by the Global Sum formula.

2. **Double-counting.** The model includes only the R/R proxies listed above. If these proxies are correlated with other factors, the estimated impact of R/R on costs would also capture the effect of omitted factors, which could bias the estimates. Most importantly, if omitted factors underlie other parts of the payment system, then the overall practice payment would double-count the impact of R/R. Examples of double-counting include *patient case-mix captured by the workload SAF or staff costs accounted for in the MFF adjustment.* Given the discussion in the previous section, double counting of R/R effects is likely and could be non-trivial.

3. **R/R variables.** The model assumes a linear relationship between excess unavoidable costs and the degree of R/R, which may not necessarily be valid. For instance, there might be practices in rural areas that have significant list sizes and therefore do not require an allowance for unavoidable small scale.

3.3 Recommendations

As discussed in Section 2, there are three geographical sources of potential cost variation across practices that are not captured by other parts of the practice payment system:

- Access to healthcare services;
- Alternative settings of care; and
- Unavoidable smallness.

The first two sources are related to workload: the workload per registered patient could be higher in rural and remote areas because of lack of alternative settings of care. On the other hand, access issues in these areas could lead to lower workload. In contrast, unavoidable smallness is related to the unit cost of GMS provision: serving a patient in rural or remote areas could be more expensive than in urban areas because of small list sizes. The proposed approach effectively suggests analysing these two sets of factors separately.

Access and alternative settings of care

Access and alternative settings of care could be estimated and accounted for in the workload component of the SAF. This could be implemented by including proxy variables, such as population density, or dummy variables in the workload model like in the 2016 model review.[2]

Unavoidable smallness

Unavoidable smallness could be estimated in two steps as depicted in Figure 2.[3]

1. **Cost-scale relationship.** Estimate the relationship between costs and list size whilst controlling for confounding effects.[4] The list size coefficient would measure the degree of economies of scale. If there are economies of scale, the unit cost of GMS provision should decline as the list size increases.[5]

2. **Unavoidable smallness.** Use the cost-scale relationship to compute the excess costs associated with small scale. The latter would be the difference between the unit cost of a small practice, estimated by the model, and the unit cost of a benchmark practice (again estimated by the model). The benchmark practice effectively reflects the list size and unit cost a practice would have assuming that it is not located in a rural or remote area. The benchmark practice could be defined as the “average” list size practice.[6]

[2] The rural dummy was found to have a negative and statistically significant effect on workload, indicating that the access factor dominates potential alternative of settings of care effects. Incorporating this effect in the overall formula, however, could potentially intensify the access issues in rural areas and therefore could be sterilised in the application of the model.

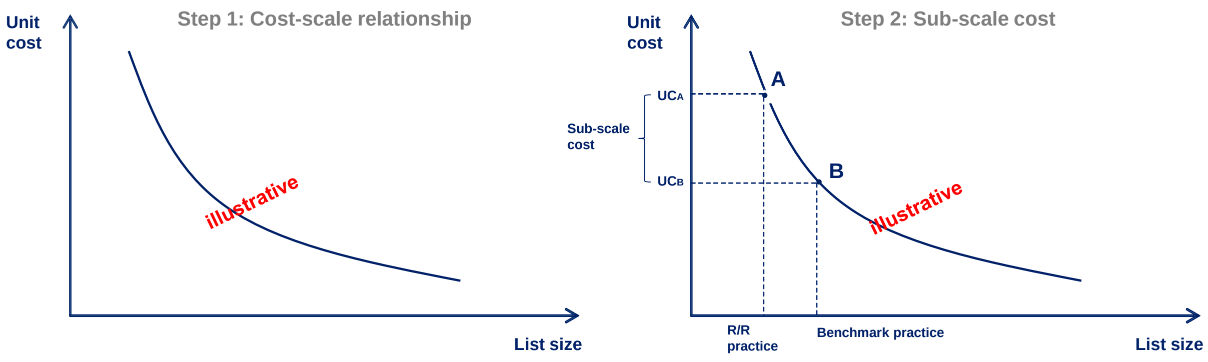
[3] This approach is similar to the method used by NHSE to estimate the unavoidable cost of smallness for acute providers in the 2016/17 CCG allocation formula. See ACRA (2015) 36: Costs of unavoidable smallness due to remoteness and Deloitte (2006): Adjusting the General Medical Services Allocation Formula for the unavoidable effects of geographically-dispersed populations on practice sizes and locations. A report for NHS Employees.

[4] Control variables could include practice and GP characteristics (e.g. location, GP age and gender), MFF, and any factors that could capture population relative need, e.g. population age, gender, and morbidity and life circumstances factors.

[5.] In the example shown in figure 2 no dis-economies of scale are assumed, thus unit cost continues to decline with an increase in list size even for large practices.

[6] The unavoidable smallness adjustment should be applied to staff costs as other practice costs are captured by different parts of the formula.

Figure 2: Estimating unavoidable smallness



Is a small list size unavoidable?

Excess costs associated with small list size are not expected to be unavoidable for all practices. A practice may have a small list size due to factors that are within the management/GPs control, such as the type of services provided and the quality of service. Furthermore, a scheme that compensates all practices for small scale could introduce perverse incentives, for example rewarding GPs for running small rather than optimal size practices. The practices that are unavoidably small could be identified by looking at the number of practices within an area and/or the distance to the nearest second practice.

Key challenge

A large estimation sample might be required to sufficiently control for the effect of other factors in the cost-scale regression. For instance, relative need might be difficult to control for at practice level as average age or deprivation by practice might not vary enough.

3.4 Simulation

The simulation presented in this section aims to illustrate the potential excess costs associated with an unavoidably small practice list size and relies on four assumptions.

1. **Elasticity of costs with respect to list size.** Three alternative elasticity assumptions have been considered: (i) -0.10; (ii) -0.20, and (iii) -0.3. For example, an elasticity of -0.1 implies that a 10% increase in list size reduces cost per patient by 1% whereas an elasticity of -0.3 assumes that a 10% increase in list size reduces cost per patient by 3%.[7]

2. **Cost per consultation.** The average cost of consultation is assumed to be £45, although the results are insensitive to this assumption.

3. **Small practice.** A small practice is defined as one with 1,500 registered patients which corresponds to the bottom decile of the list size distribution in Scotland.[8]

4. **Benchmark practice.** The benchmark practice is defined as a practice with a median list size of about 5000 patients.[9]

The results of the simulation are presented in Figure 3 and suggest that the burden of unavoidable smallness ranges between 11% and 30%.[10] If a small practice is assumed to have a list size of 1,000

[7] The -0.1 and -0.2 scale effect assumptions are consistent with the economies of scale found in the provision of acute health services (ACRA (2015) 36). The -0.3 value is consistent with the results from Deloitte (2006) which estimated the relationship between general practice costs (premises and staff) and list size.

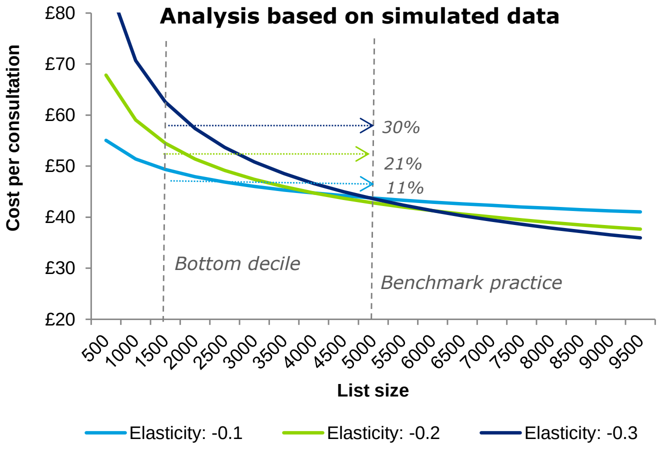
[8] tp://www.isdscotland.org/Health-Topics/General-Practice/

[9] The average list size is 5,820, slightly higher than the median, the top decile is c.10,000 and the maximum c.32,000.

[10] The unavoidable cost of smallness in the NHSE acute formula ranges from 0% to 10%, which is lower than the one implied by this simulation. This is primarily driven by the fact that scale variability in general practices in

(corresponding to the bottom 5% of the list size distribution), then the cost of smallness would be between 15% and 38%. This contrasts the current R/R adjustment which ranges between -8% and +734%, and with 25% of the total practices receiving an uplift of greater than 7%. This adjustment seems high considering the results of the simulation and that the unavoidable smallness is expected to be the main driver of the R/R excess costs.

Figure 3: Simulation results



Source: Deloitte analysis based on simulated data

Scotland is higher than in English acute sector (the smallest acute site in English NHS is about 50% smaller than the average acute hospital).

4 MFF review

This section reviews the MFF adjustment incorporated within the unit cost formula providing:

- Some background information around general practice staff;
- A description of the current adjustment;
- The limitations of the current approach;
- Three alternative approaches that could be used to potentially improve the existing methodology; and
- A discussion around whether the MFF adjustment should be applied to Principal GPs.

4.1 Background

General practices have four broad types of staff:

- Principal GPs;
- Salaried GPs;
- Nurses; and
- Admin staff (e.g. receptionists, secretaries, finance managers).

Principal GPs are the owners of practices, typically run as partnerships, whose income is determined by the Global Sum formula and other payments or allowances as well as how efficiently they run the practice. Principal GPs can increase their income by attracting more patients (paid through the workload SAF), reaching QOF targets (QOF payment) and offering additional services (enhanced services payment).

Contrary to other settings of care, there are no explicit pay grades for salaried GPs or nurses whose salaries are determined by market conditions similar to the private sector.

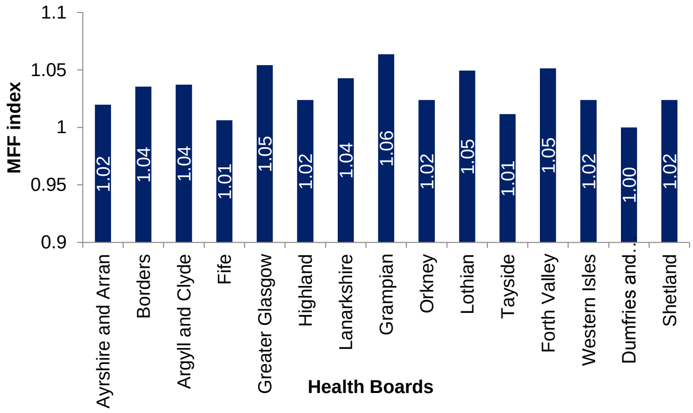
4.2 Current MFF adjustment

Staff costs typically vary by regions because of differences in cost of living, local amenities or general attractiveness of an area. The MFF adjustment recognises that unavoidable differences in staff costs should be taken into account in order for practices to have the necessary resources to meet local population need and facilitate equal access for equal need.[11]

In the 2004 SAF, the MFF is effectively a wage index measuring relative wages across Health Boards. The wage index is based on regional average salaries across all occupations and ranges between 1 and 1.064 (see Figure 4). The regional differences in staff costs are incorporated into the SAF by multiplying the MFF index by the weighted list predicted by the workload formula with the aim to generate a weighted list adjusted by unavoidable costs associated with staff costs.

[11] The two factors determining differences in regional pay rates could have opposite effects: pay rates in rural or remote areas could be lower than urban areas reflecting lower costs of living or they could be higher due to lower area attractiveness.

Figure 4: MFF index in 2004 SAF



Source: Scottish Government

4.3 Limitations

There are three key limitations in the current MFF adjustment.

- **Geographical granularity.** The application of the MFF at the Health Board level does not recognise that there might be differences in staff costs between local areas within Health Boards. Health Boards cover large geographical areas with, on average, 70 practices, which are unlikely to face the same labour market conditions.

- **Wage variable.** Differences in regional staff costs captured by an overall wage index should reflect spatial differences in costs of living and general location attractiveness but also regional variations in skills and occupation or industry make-up of the local economies. Therefore, the overall wage index may not adequately reflect the labour market and the geographical variation in salaries for doctors and nurses.

- **Adjustment.** Because the MFF adjustment is applied on weighted list sizes without a staff weighting, it is effectively assumed that prices of other inputs (such as fixed costs associated with premises) exhibit the same regional variation as staff costs. Furthermore, variation in premises costs should be, in principle, accounted for by the Premises payment, leading to double-counting.

4.4 Recommendations

There are three alternative approaches that could potentially provide more accurate estimates of staff MFF.

Occupation-specific wage index

Instead of using salaries across all occupations, a set of occupations comparable to doctors and nurses could be used to compute the regional wage differentials. The data can be obtained from the Office for National Statistics, Annual Survey of Hours and Earnings (ASHE), which are collected at a local authority level and therefore could also capture variations in pay rates across areas within Health Boards. If the benchmark occupation is the NHS, there will be no geographical variation in staff costs as NHS pay does vary across Scotland.

Main limitation. The main limitation of this approach is that the number of private sector employees within a comparable occupation group and local authority might be relatively small to accurately measure wage differentials. Furthermore, spatial wage differentials might vary because of regional differences in skills, experience and/or industry composition.

Regression analysis using private sector data (NHSE approach)

The English MFF is based on a regression analysis, using data from ASHE, that models regional pay variations in the private sector as a function of regional dummies, age, industry and occupation indicators. The regional dummy variables capture the spatial variation in pay and the remaining variables control for confounding effects.

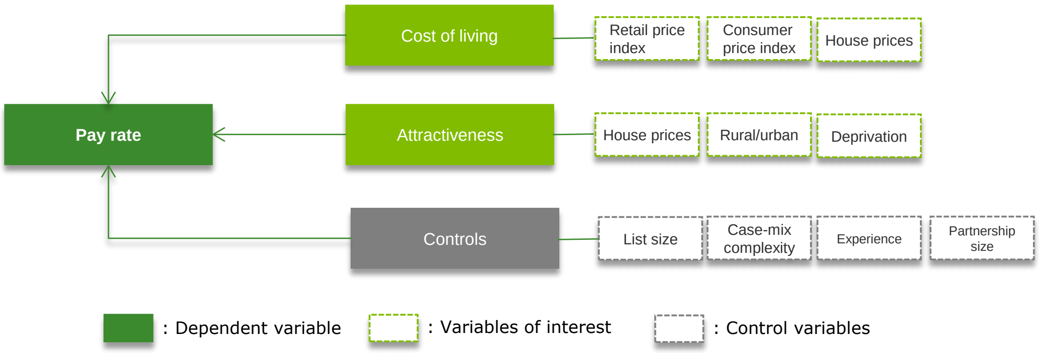
Main limitation. Although this approach could deal with the limitations of the first approach, it assumes that the private sector can sufficiently approximate the labour market for doctors and nurses.

Regression analysis using general practice data

Practice specific data could be used to estimate a model similar to the one presented in Figure 5. Such a model could directly measure the impact of the cost of living, measured by consumer, retail and/or house price indices, and area attractiveness, proxied by deprivation, rurality and house prices, on pay rate. The estimated model could predict the "market" pay rate for a practice given the cost of living and area attractiveness, keeping all other factors constant. Separate models could be estimated for doctors and nurses or by staff grade. The advantage of this approach is that it does not rely on any assumptions around the comparability between different occupation groups.

Main limitation. Pay rates and other practice specific data are not currently available and might be difficult to obtain.

Figure 5: Model specification



4.5 Should the MFF be applied to principal GPs?

The net income of a principal GP is determined by the practice profitability, which compensates the GP for his/her efforts, but also for the cost of living and area attractiveness just as in the case of practice employees.

Elliot et al. (2006)[12] note that the net advantages of setting up a practice in different areas will need to be equalised. This would imply, for instance, that there will be an under-supply of GPs in deprived areas, which could lead to less competition, greater list size and increased profitability. The latter would effectively compensate partner GPs for the low area attractiveness. Elliot et al. provide evidence consistent with this premise.

Variation in the supply of GMS across the country due to regional differences in the cost of living and amenities would imply that patients with the same need might not have equal access to the primary care system. Applying the MFF index on principal GPs income could facilitate a more equal distribution of practices across the country.

[12] Elliot et al. (2006). Adjusting the General Medical Services Allocation Formula to Reflect Recruitment and Retention Difficulties. Health Economics Research Unit 16.