Social Science & Medicine 361 (2024) 117346

Contents lists available at ScienceDirect

Social Science & Medicine

journal homepage: www.elsevier.com/locate/socscimed

Adjusting the risk-adjustment: Accounting for variation between organisations in the responsiveness of their expenditure to need

Sean Urwin [a,*], Laura Anselmi [a], Emmanouil Mentzakis [b], Yiu-Shing Lau [a], Matt Sutton [a]

[a] *Health, Organisation, Policy and Economics, School of Health Sciences, University of Manchester, United Kingdom*
[b] *Department of Economics, City, University of London, London, United Kingdom*

ARTICLE INFO

Handling editor: Joanna Coast

Keywords:
Risk adjustment
Resource allocation
Weighted capitation
Need
Cost

ABSTRACT

There is concern that basing healthcare budgets on risk adjustment estimates derived from historical utilisation data may reinforce patterns of unmet need. We propose a method to avoid this, based on a measure of how closely local health organisations align resources to the needs of their populations. We refer to this measure as the 'responsiveness of expenditure to need' and estimate it using national person-level data on use of acute hospital and secondary mental health services in England. We find large variation in responsiveness in both services and show that higher expenditure responsiveness in mental health is associated with fewer suicides. We then re-estimate the national risk-adjustment model removing the data from the organisations with the lowest expenditure responsiveness to need. As expected, higher need individuals are estimated to have higher expenditure needs when less responsive organisations are removed from the estimation of the risk-adjustment. Removal of organisations with below-average responsiveness results in the neediest deciles of individuals having an extra £163 (7%) annual need for acute hospital care and an additional £79 (27%) annual need for mental health services. The application of this approach to risk adjustment would result in more resources being directed towards organisations serving higher-need populations.

1. Introduction

Many publicly-funded healthcare systems throughout the world allocate resources based on risk adjusted formulae derived from information on historical healthcare use. Examples include the four countries of the United Kingdom, Canada, Australia and New Zealand (Rice and Smith, 2001).

Risk adjustment models are generally used to predict healthcare budgets or insurance premiums based on the observable characteristics of individuals, applied by governments and social insurers in numerous healthcare contexts worldwide (Juhnke et al., 2016). Unlike for health insurance purposes, the aim of risk adjustment models used for resource allocation purposes in publicly-funded healthcare systems is to produce a prediction of healthcare costs that ensures an equitable distribution of resources according to healthcare needs. The objective is to create a level playing field based on population need, rather than existing access to care (Rice and Smith, 2001; Smith, 2006). This is primarily done by including variables that account for differences in supply across regions and small areas and setting their values to the national average at the prediction stage (Anselmi et al., 2022; Gravelle et al., 2003). However, there are concerns that these formulae solidify patterns of unmet need because they are based on historical healthcare use (Santana et al., 2021).

There is a large literature on the application of risk adjustment methods that covers a variety of settings, but mainly relates to insurance markets (Juhnke et al., 2016; McGuire and van Kleef, 2018). The aim and methods used for risk adjustment depend on the context (McGuire and van Kleef, 2018). Recent literature discusses issues surrounding variable selection (McGuire et al., 2021), solutions to address incomplete information (Rose et al., 2017), the use of machine learning methods to improve the risk prediction accuracy (Rose, 2016) and the identification of undercompensated groups (Zink and Rose, 2021). Other related work has considered place-based adjustment factors to account for regional differences in diagnostic intensity (Finkelstein et al., 2017) and the inclusion of socio-economic variables (Schokkaert et al., 2018). One focus of this literature, and most related to our study, are attempts to account for differences in healthcare use between groups. For example, Zink and Rose (2020) develop 'fair regression' methods to account for underprediction by an insurer and therefore undercompensation for certain groups.

* Corresponding author. Williamson Building, 176 Oxford Road, University of Manchester, Manchester, M13 9QQ, United Kingdom.
E-mail address: sean.urwin@manchester.ac.uk (S. Urwin).

https://doi.org/10.1016/j.socscimed.2024.117346
Received 3 April 2024; Received in revised form 11 August 2024; Accepted 13 September 2024
Available online 14 September 2024

Unmet needs as a concept is multifaceted and difficult to define (Santana et al., 2021; Smith and Connolly, 2020). It can broadly be described as when "patients experience a health shortfall that could be treated in a cost-effective and appropriate manner with the available technology, but is not met by healthcare supply" (Santana et al., 2021). There is limited evidence that quantifies the unmet healthcare needs of populations across all health and social care sectors. Among the available evidence, it is suggested that there exists substantial variation in the degree of unmet needs across high-income settings (OECD, 2021). This variation may arise because of inadequate health planning or prioritisation, a higher concentration of individuals in some areas with lower demand, or limited service delivery in high need areas and to high need populations.

The existence of unmet needs in the context of a formula predicted on administrative records from past healthcare service use means that certain groups may not be allocated adequate resources that are reflective of their needs. Unmet needs may bias the formulae used in England in two ways. First, from the omission of need measures that are not considered predictors of costed healthcare service use. If appropriate measures of need are not included in the formulae because they are not observed, predictions of healthcare need will not reflect the full healthcare needs of populations. The second cause of bias would be through an underestimation of how much additional cost a need measure generates. For example, per-person healthcare expenditure on older age groups living alone may be smaller, on average, compared to younger age groups and those not living alone, even after accounting for need-related characteristics. A possible explanation could be differences in the accessibility of healthcare services, if older age groups and those living alone face barriers in obtaining more healthcare. Therefore, predictions based on historical utilisation may reinforce these spending patterns used to allocate healthcare budgets. The effectiveness of healthcare resource allocations in reducing existing health inequalities may be diminished if they reflect patterns of unmet needs and result in widening health inequalities (Barr et al., 2014).

The funding formulae used in England aim to allocate healthcare resources according to need (NHS England, 2023). The approach to allocation and the formulae used have evolved over time (Smith, 2008) one of which, is through the introduction of person-based data and information in line with most risk adjustment models used in settings with insurance markets (Dixon et al., 2011). Despite this improvement, there is a recognition that the formulae may not fully capture differences in need and the current approach to accounting for unmet needs, alongside health inequalities, which involved the distribution of a proportion of the budget according to the standardised mortality ratio (NHS England, 2018), and more recently avoidable mortality (NHS England, 2024). For 2018 to 2022, this proportion is 10% of the general and acute budget. This approach was intended to be a temporary method for the NHS until a more data driven approach can be developed (NHS England, 2018).

In this paper, we propose a way of approaching risk adjustment that can take account of unmet needs. We focus on the second cause of unmet needs, generated by an underestimation of how much additional cost a needs indicator generates. Unmet need is likely to manifest in areas where utilisation is less responsive to variations in the need of populations. We refer to this as the 'responsiveness of expenditure to need'. Once identified, the less responsive areas can be removed from estimation of the formula to provide coefficients that generate more responsive needs-weightings. The resulting adjustment ensures that the formula needs weights are calibrated excluding data from the care systems that are less successful at aligning their resources with need among their higher need populations.

The concept of estimating resource allocation formulae on a subselection of regions and applying the estimated needs weights nationally has been explored by Sutton and Lock (2000) with aggregate data from Scotland. They identified the most progressive regions based on a Kakwani index of progressivity which captures the degree to which healthcare use rises in relation to need across areas within regions. A similar concept was later proposed, but this time based on selecting areas depending on their performance on nationally-monitored performance indicators (Morris et al., 2010).

Building on the idea of Sutton and Lock (2000), we first produce a measure at a local health organisation level that captures how responsive a local health system is to the needs of their populations derived from person-level data. We explore the correlation of responsiveness with health outcomes at a health system level, to address the concern that the responsiveness of expenditure to need in highly responsive areas may reflect poor performance in other dimensions and less favourable health outcomes. We then use our measure of the responsiveness of expenditure to need to identify areas that we should exclude when calculating the risk adjustment formula and show the impact this has on the predictions of service need for all areas nationally. As empirical examples, we produce a measure of responsiveness for two types of healthcare services in England, General and Acute services and secondary Mental Health services. These services represent the largest share of the budget for services commissioned locally within the NHS, and are allocated based on regression models estimated on person-level data, therefore demonstrating two important applications of the approach.

2. Resource allocation formulae in England

Healthcare resources in England were allocated to local health organisations, referred to as Clinical Commissioning Groups (CCGs), based on weighted capitation formulae. There were 211 CCGs on their creation in 2013 which reduced to 106 by 2021, following a series of mergers. After July 2022, CCGs were replaced by 42 Integrated Care Boards (ICBs) and future allocations will be made to these organisations. Weighted capitation formulae have a long history in England and been applied by the NHS to allocate healthcare resources since 1976 (Smith, 2008). There are separate formulae for each funding stream which includes general and acute, maternity, mental health, primary care and prescribing (NHS England, 2023). Each formula is used to generate need weights which are combined with population and other factors to determine target budget shares.

The approach used in England to derive the need weights is referred to as 'utilisation based', as opposed to one using disease prevalence (Vallejo-Torres et al., 2009). Since 2011 a person-based formula has been used by the NHS which exploited individual level data, including all individuals registered to a GP practice (Dixon et al., 2011). Since then this approach is used where person-level data are available, including for the formulae for general and acute (Chaplin et al., 2016; NHS England, 2022) and mental health services (Anselmi et al., 2020).

The formula is estimated by modelling person-level expenditure as a function of need, supply and local care system indicators using ordinary least squares regression (Anselmi et al., 2020; Chaplin et al., 2016; Dixon et al., 2011):

$$y_{ijt} = \alpha_1 + \alpha_2 x_{2ij,t-1} + \alpha_3 x_{3ijk,t-1} + \gamma . S_{ij,t-1} + u_j + \varepsilon_{ij} \quad (1)$$

where y_{ijt} represents costed service use for individual i in CCG j in time t, $x_{ij,t-1}$ represents the individual needs variable, $x_{3ijk,t-1}$ represents attributed area-level need variables measured at a lower level of geography than CCGs, S_{ij} represents the supply variables, u_j the CCG indicators, α_1 is the constant and ε_{ij} is the error term. Variables included in each model vary across funding streams. Generally, needs variables measured at an individual level, $x_{2ij,t-1}$, include demographic characteristics such as age, gender, ethnicity and those more closely linked to health such as diagnosis codes. Needs variables measured at an area level, $x_{3ijk,t-1}$, are assigned to an individual based on their registered GP practice or Lower-layer Super Output Area (LSOA). LSOAs are a geographic area of approximately 1000–3000 people. Supply variables tend to measure the capacity of the health system to provide care (beds, waiting times) and geographical characteristics such as distances to nearest healthcare providers. Indicators for each CCG are included to capture any remaining unmeasured supply differences between areas that are not measured by other variables. We measure service use in the financial year, time t, while all included need and supply variables are measured before the start of the financial year, at t-1. Individuals who do not use healthcare services in the financial year of analysis are included with zero utilisation and all their need and supply variables. Diagnostic markers based on diagnoses recorded on any hospital utilisation in the previous two years are included for all persons.

After estimating the coefficients, the values of the supply and local care system variables are set to the national average values before predictions are calculated. Therefore, predictions of service need ($\widetilde{y}_{ijt}$) reflect variations in need rather than supply. Individual level need predictions are derived by:

$$\widetilde{y}_{ijt} = \widetilde{\alpha}_1 + \widehat{\alpha}_2 x_{2ij,t-1} + \widehat{\alpha}_3 x_{3ijk,t-1} \tag{2}$$

Where $\widetilde{\alpha}_1$ is the adjusted constant term and $\widehat{\alpha}_2$ and $\widehat{\alpha}_3$ are the needs coefficients. The individual level need predictions can then be aggregated to give weighted capitation budgets for purchasing organisations, or any sub-grouping which may be helpful and identifiable within the data, such as local areas.

Formulae are refreshed regularly. For example, the General and Acute formula was refreshed in 2011/12, 2015/16 and then for 2022/23. In the intervening years, the allocations based on these formulae are updated for changes in population size and age-gender composition only.

3. Data and methods

3.1. Data

We produce our measure of local responsiveness of expenditure to need for two healthcare services representing the funding streams that absorb most of the spending commissioned locally – general and acute hospital services and mental health services. For both sectors the main datasets analysed are at a person-level and have been previously used by the NHS for resource allocation purposes. The approach that has been used by the NHS to estimate each of the two formulae, the costing approach, and variables used for the estimation of each formula are described in Chaplin et al. (2016) and NHS England (2022) for general and acute hospital services and in Anselmi et al. (2020) for mental health services.

The general and acute formula is estimated on the whole population of 58 million individuals. It includes the 2018/19 Secondary Uses Dataset (SUS) which contains information from admitted patient care (APC), outpatient attendances (OP) and accident and emergency (AE) attendances for each service user. Costs were assigned to activity using national tariffs and reference costs to obtain the per-person total healthcare cost. We replicate the model of the general and acute formula and use all explanatory variables applied to the general and acute formula which includes but is not limited to: age; gender; ethnicity; morbidity indicators (measured using 152 indicators for groups of ICD-10 codes (Chaplin et al., 2016)); household type; deprivation (measured using components of the Indices of Deprivation (Ministry of HousingCommunities & Local Government, 2019) and area level characteristics such as the proportion of those 16–74 years old who have never worked. Supply variables include waiting times, distance to the nearest hospital assigned to each individual and measured at LSOA level as well as the inclusion of indicators for each CCG. Chaplin et al. (2016) and NHS England (2022) provide further details on the rationale for model choice within the general and acute formula.

The mental health formula is estimated on the whole adult population of 43 million who are at least 20 years old. We apply costs to mental healthcare use in 2015/16 using a combination of activity data from: The Mental Health and Learning Disabilities Data Set (MHLDDS); the Mental Health Services Data Set (MHSDS); Secondary Uses Service (SUS) data-set; and the Improving Access to Psychological Therapies (IAPT) data-set. To obtain the total per-person mental healthcare cost, a combination of reference and unit costs was applied to secondary care inpatient stays based on intensity and security, outpatient contacts based on the professional and on IAPT contact based on an average cost per appointment. We use all explanatory variables applied to the mental health formula which include: age; gender; ethnicity; past physical health diagnosis with a severe mental illness; the proportion of the population residing in a given LSOA and receiving out of work benefits; whether an individual is registered to a student GP practice; and the prevalence of severe mental illness at the individual's registered GP practice. Supply variables include distance to the closest mental health trust headquarters and binary indicators for each CCG. Anselmi et al. (2020) provide further details on the rationale for model choice within the mental health formula.

3.2. Responsiveness measure

Our responsiveness measure is computed as a needs-based slope index of inequality that requires cost weighted healthcare use, a composite needs measure and the ranking of individuals by need. We construct the measure in four stages.

First, we calculate a composite measure of predicted service need. We obtain this need measure from the fitted need values of the healthcare use formula provided in equation (2) which uses the variables detailed for general and acute and mental health respectively for the whole population nationally.

Second, we rank all individuals by their predicted need nationally. We use the national rank to achieve comparability. This is obtained by dividing each individual's national rank by the total population size. This fractional rank ranges from zero, which represents the individual with the lowest estimated need, to one, which represents the individual with the highest predicted need nationally.

Third, we calculate for each individual i in CCG j the gap between their expenditure and their predicted need.

Finally, we estimate a linear regression for each CCG to relate these gaps (denoted y^*_{ij}) between expenditure (y_{ij}) and estimated service need ($\widetilde{y}_{ij}$) to the individual's rank in the national distribution of need. The slopes of the regression lines (denoted β_j) are the needs-based slope index of inequality:

$$y^*_{ij} = y_{ij} - \widetilde{y}_{ij} = \rho_j + \beta_j * rank_i \tag{3}$$

For each CCG, the slope indicates how the gap between expenditure and need would differ between the individual with the highest need and the individual with the lowest need nationally, if the expenditure patterns in this CCG were replicated nationally. The constant (ρ_j) indicates how much the individual with the lowest need (where rank is 0) is under or over funded. The sum of the constant and the slope ($\rho_j + \beta_j$) gives the value of the expenditure minus service need gap for the individual with the highest need where the rank is 1. This approach is consistent with the application of an absolute measure of health inequality to the fairness gap, as proposed by Fleurbaey and Schokkaert (2009).

Our responsiveness measure assumes that all CCGs have individuals present at all points of the national need distribution. To provide evidence on this assumption, we graphically present the distribution of the number and percentage of individuals from each CCG across 1000 quantiles of the national need distribution.

3.3. Interpretation of the responsiveness measure

Fig. 1 presents a graphical example of a negative and positive responsiveness value, for example for CCGs at the 5th and 95th percentile of the national distribution by responsiveness. We demonstrate this with two positive slopes - an expenditure and a predicted need

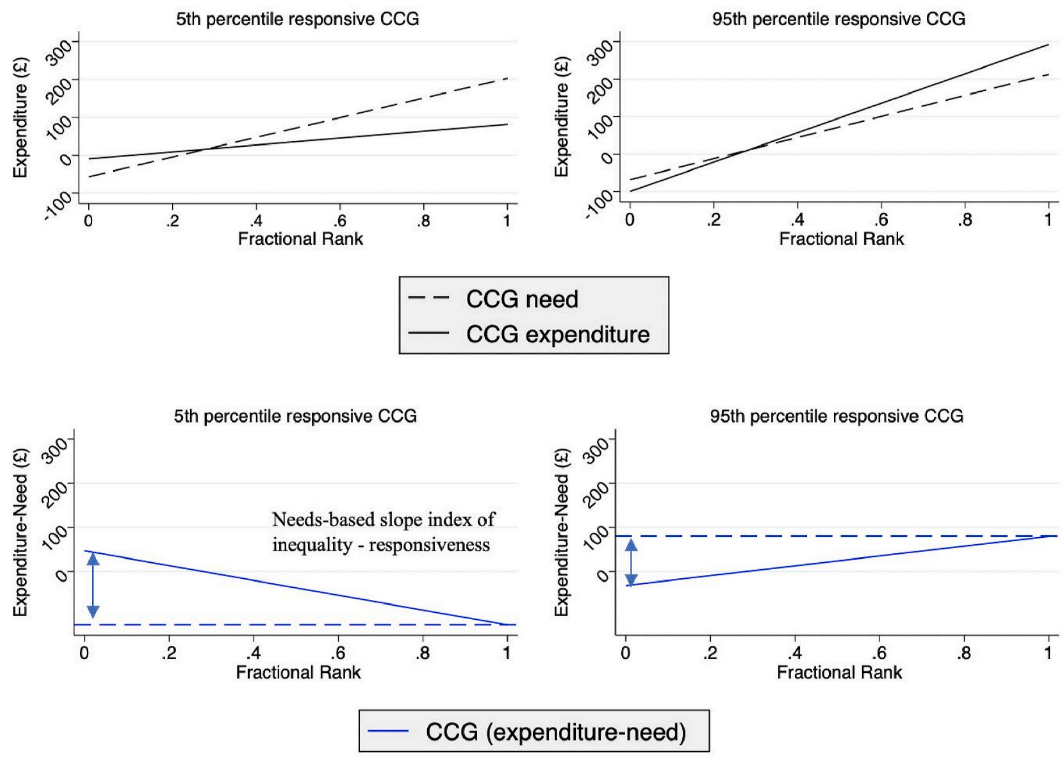

Fig. 1. Examples of Low (left panels) and High (right panels) Responsiveness and relative Needs Based Slope Index of Inequality.
Note: The needs ranking runs from lowest to highest need. In the bottom panels the gap between the blue solid and dashed line is responsiveness – the needs-based slope index of inequality. The intercept of the solid blue line is the difference in the constant terms for the CCG expenditure and CCG need lines.

slope. A CCG has a negative responsiveness value if the difference in expenditure between the lowest and highest need individual is smaller than the equivalent difference based on predicted need. This is represented in the upper left-hand side graph of Fig. 1 where the expenditure slope is flatter than the predicted need slope. A CCG has a positive responsiveness value if the difference in expenditure between the lowest and highest need individual is greater than the equivalent difference based on predicted need. This is represented by the upper right-hand side graph of Fig. 1 where the expenditure slope is steeper than the predicted need slope.

The two graphs in the bottom panel of Fig. 1 show the gradient of the gaps in the expenditure and predicted need slopes denoted by β_j from equation (3). The intercepts of these slopes denoted ρ_j from equation (3), indicate the gap between expenditure and predicted need for the person with the lowest level of need. The double arrowed line in the bottom two graphs indicates the calculation of the gap between expenditure and predicted need for each individual. If actual expenditure increases less than predicted need over the need distribution in the lower left-hand side graph, i.e. where $\beta < 0$ from equation (3), then the CCG has a responsiveness value less than zero. The reverse applies for CCGs with a responsiveness value above zero.

3.4. Association of responsiveness with outcomes

We test whether responsiveness is positively or negatively related to CCG level health outcomes. We do this to address a possible concern that responsiveness in highly responsive areas may reflect less favourable health outcomes. We estimate the relationship between responsiveness and two CCG outcome measures obtained from the Office of National Statistics (ONS). Our outcome measures are the age-standardised avoidable mortality rate per 100,000 population (ONS, 2024) and the age-standardised suicide rate per 100,000 population (ONS, 2019). We chose avoidable mortality because this includes deaths from causes that are considered treatable or preventable by healthcare services such as general and acute hospital care. We chose suicide rates as a nationally-available population health outcome relevant for mental health services.

We use outcome measures for periods that are broadly contemporaneous with our responsiveness measure as well as for periods that come after the responsiveness measure. We use avoidable mortality as an outcome for General and Acute responsiveness in the calendar years 2018 and 2019 and for Mental health in the calendar years 2015 and 2016. We use the suicide rate for mental health for two periods that cover the calendar years 2014 to 2016 and 2016 to 2018.

We use ordinary least squares regression to estimate the relationship between outcomes and responsiveness:

$$\boldsymbol{y_j} = \gamma + \pi^*\beta_j + \varnothing \boldsymbol{X_j} + u_j \tag{4}$$

Where β_j is the measure of responsiveness for CCG j, π is coefficient of interest on the responsiveness value, $\boldsymbol{y_j}$ is the CCG outcome, $\boldsymbol{X_j}$ is the vector of covariates, γ and u_j are the constant and error terms respectively.

We adjust for the characteristics of populations and health systems that may confound the relationship between responsiveness and each outcome. For example, an older, rural and deprived area may find it more difficult to align their resources to the needs of their population compared to a younger, urban and affluent area. Similarly, areas with better leadership, closer working relationships and are closer to their allocated share of the healthcare budget may find it easier to align their resources with need. These factors may also be related to avoidable mortality and thus should be accounted for. The choice of these covariates was informed by engagement with CCGs' leaders and by discussions with members of the Technical Advisory Group (TAG) and Advisory committee of Resource Allocation (ACRA).

Specifically, the population characteristic measures include: the proportion of the CCG population aged over 65 years old; the proportion of the CCG population residing in an urban area; and the average index

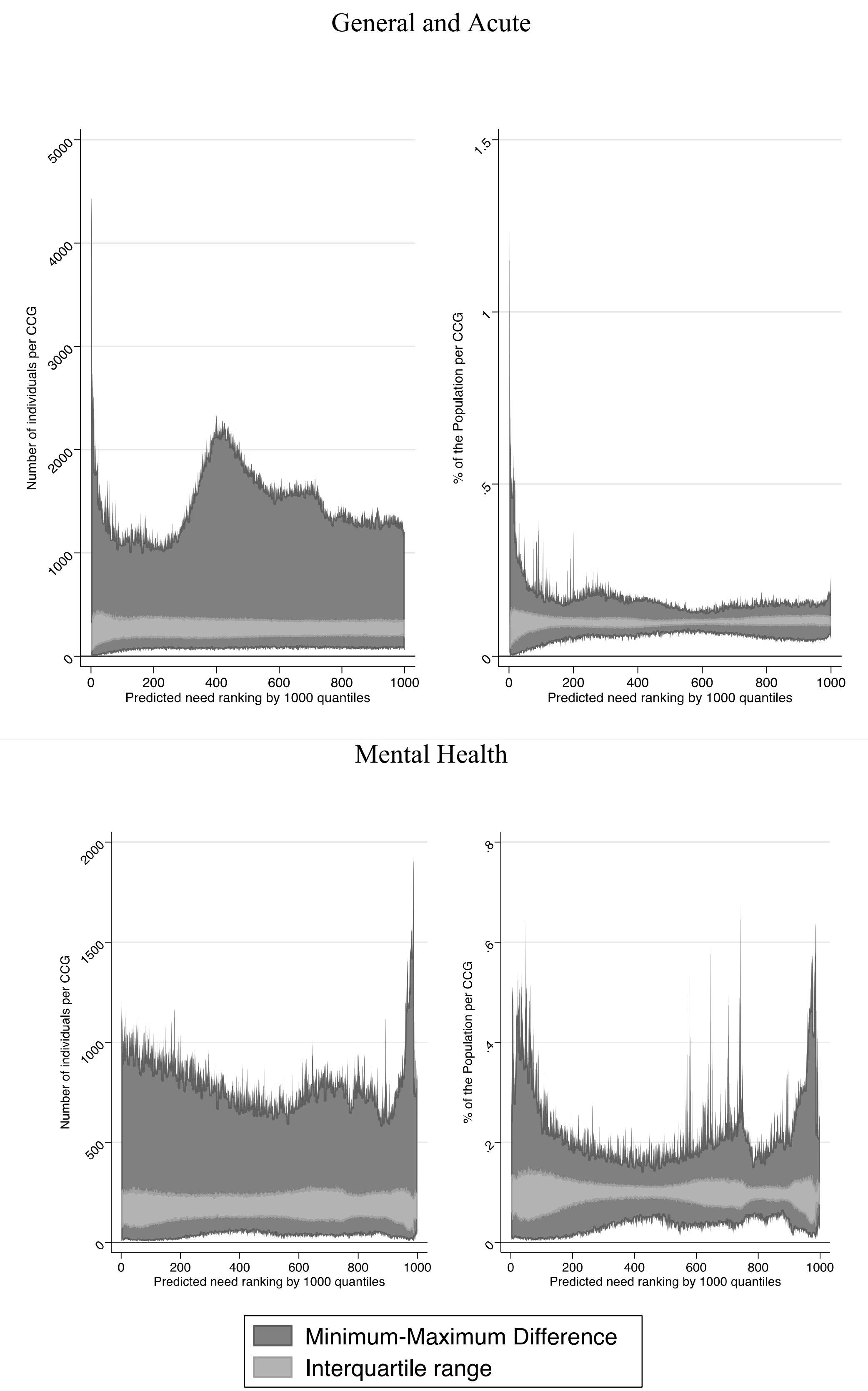

Fig. 2. Distribution of individuals by CCGs across 1000 quantiles of national service need.

of multiple deprivation rank for each CCG coded into deciles. The health system characteristics include: the quality of CCG leadership (measured on a one to four scale where one is the highest level of quality); the effectiveness of local working relationships measured as a percentage (where a higher percentage is higher effectiveness); and the distance each CCG's budget is from their target allocation as a percentage. A measure specific to general and acute analysis is the number of type one accident and emergency departments within a 20-km radius. A measure specific to mental health analysis is the proportion of a CCG's population in the lowest national mental health decile as measure by the Small Area Mental Health Index (SAMHI). Additional details on the measures we use for regression analysis are provided in appendix Table A1.

3.5. Impact of removing the least responsive areas

We re-estimate the coefficients on the variables included in the General and Acute and in the Mental Health models using data on individuals living in the more responsive CCGs. We produce five different coefficient sets based on the removal of individuals from the CCGs with the lowest level of responsiveness nationally. These are outlier (two general and acute and three mental health) CCGs, the lowest 5%, 10%, 25% responsive CCGs and all CCGs with a responsiveness value below zero. The latter removal, in other words, restricts the formula to all CCGs where, between the lowest and highest need individual, the expenditure increase is larger than predicted need increase. For each coefficient set we obtain predicted need values for all individuals with supply variable values set to the national average as derived in equation two.

To assess the impact of removing CCGs from the pool of individuals used to estimate the needs weights, we calculate: (i) the change in the average level of predicted need in each need decile compared to the baseline formula where the need deciles are calculated using the baseline formula and (ii) the fractional national needs gradient from equation three regressed on each new set of need predictions. In other words, the difference in predicted need for each new set of need predictions using the ranking of the lowest and highest need individual from the baseline formula. We use an individual's position in the need ranking from the baseline formula to calculate the predicted need deciles and ranking as individuals may change their position in national rankings based on different prediction sets. This ensures we are comparing the same individuals within each decile or for each rank when we calculate predictions based on a subset of CCGs. More specifically, the estimated needs gradient in (ii) is derived from five separate OLS regressions using each set of new need predictions ($\widetilde{y}_i$) derived after progressively removing the least responsive CCGs and the baseline fractional ranking from the original formula (in equation three):

$$\widetilde{y}_i = \eta + \delta * rank_i \tag{5}$$

Where δ represents the difference between the lowest and highest ranked individual according to need and η is the constant term.

4. Results

4.1. Descriptive statistics of responsiveness and its components

Ranking individuals in the national needs distribution could result in some CCGs only having individuals in certain parts of the distribution. Fig. 2 presents the interquartile range of the number and percentage of individuals per CCG for every one-thousandth of the general and acute and mental health needs distributions (approximately 40,000 individuals for mental health). The interquartile ranges of the number of individuals from each CCG for each of the thousandths of the national mental health need distribution range from 5.0 to 10.0% of a CCG's population at the 25th percentile to 10.7–13.9% of a CCG's population at the 75th percentile. For general and acute need deciles, the equivalent figures are for 6.1–9.4% of a CCG's population at the 25th percentile and 10.5–12.5% of a CCG's population at the 75th percentile. Therefore, every CCG has a meaningful number and proportion of its individuals at each portion of the national need distribution.

Table 1 (and Fig. A1 in the appendix) presents the distribution of responsiveness for general and acute and mental health, respectively. General and Acute responsiveness has a wider range of values than the responsiveness measure for mental health. The mean value of general and acute responsiveness is £-17 (SD: £181) across 192 CCGs with the lowest value at £-453 and the maximum value at £558. In other words, between the lowest and highest individual in the needs ranking, expenditure is £17 lower than the predicted need value on average across all CCGs. The equivalent mean value for mental health responsiveness is £-40 (SD: £88) across 211 CCGs.

Table 1 also shows statistics on the expenditure and need slopes. Although the average responsiveness values are similar for general and acute compared to mental health, the gap in expenditure between lower and higher need individuals are far larger for the former. For example, the difference between the lowest and highest need individual is £1769 higher in terms of expenditure for general and acute relative to £216 higher in mental health expenditure. We present summary statistics for the constant values that give the expenditure and predicted need for an individual with the lowest need in appendix Table A2.

4.2. Association of responsiveness with CCG outcomes

The mean and standard deviation of the outcomes and covariates we use for regression analysis are presented in appendix Table A3. The mean avoidable mortality rate is 226.34 in 2019 and 227.27 in 2016 per 100,000 population for the General and Acute and Mental health CCG set, respectively. The suicide rate is 9.81 for 2016 to 2018 and 10.14 for 2014 to 2016 per 100,000 population for the Mental health CCG set.

We present the association of responsiveness with two CCG outcome measures in Table 2. Higher responsiveness in general and acute hospital expenditure is not related to avoidable mortality. It is correlated

Table 1
Descriptive statistics of CCG responsiveness to need.

	Mean (SD)	Percentile						
		Min	10th	25th	50th	75th	90th	Max
General and Acute (192 CCGs):								
Responsiveness to need	−17 (181)	−453	−224	−156	−34	75	227	558
Expenditure slope	1769 (221)	1319	1509	1594	1753	1908	2033	2671
Predicted need slope	1785 (172)	1441	1591	1657	1778.5	1884	1997	2421
Mental Health (211 CCGs):								
Responsiveness to need	−40 (88)	−235	−151	−107	−41	14	76	216
Expenditure slope	216 (93)	43	105	142	208	275	347	483
Predicted need slope	256 (27)	209	228	236	251	270	292	363

Note: *The responsiveness measures are derived from equation (3). The reported expenditure slope values are the* β_j^{exp} *from a regression of:* $y_{ij} = \rho_j^{exp} + \beta_j^{\,exp} * rank_i$ *for all* $j = 1, \ldots, J$ *where j denotes CCG and i denotes the individual. The predicted need slope values are the* β_j^{need} *from a regression of:* $\widetilde{y}_{ij} = \rho_j^{need} + \beta_j^{\,need} * rank_i$

Table 2
The association between responsiveness and CCG outcomes measured per 100,000 population.

	General and Acute		Mental Health			
	Avoidable mortality		Avoidable mortality		Suicide rate	
	2018	2019	2015	2016	2014–16	2016–18
Responsiveness (£)	0.296	−1.345	−0.540	1.330	−0.496***	−0.447***
	(1.234)	(1.420)	(2.420)	(2.220)	(0.150)	(0.160)
% Population over 65 years old	−1.082	−0.093	0.617	0.045	0.203***	0.207***
	(0.684)	(0.787)	(0.712)	(0.653)	(0.044)	(0.047)
IMD decile	13.606***	10.463***	12.324***	12.768***	0.238***	0.347***
	(0.786)	(0.905)	(0.994)	(0.911)	(0.062)	(0.066)
% Population in urban areas	0.499***	0.419**	0.296*	0.266*	−0.005	0.009
	(0.142)	(0.163)	(0.156)	(0.143)	(0.010)	(0.010)
Quality of CCG leadership	−0.520	2.507	7.062**	8.150**	0.302	−0.013
	(2.909)	(3.347)	(3.462)	(3.175)	(0.215)	(0.228)
Effectiveness of local working relationships (%)	0.281	0.430	−0.039	−0.304	−0.019	−0.032
	(0.291)	(0.335)	(0.317)	(0.291)	(0.020)	(0.021)
Distance from target (%)	−0.690	0.692	−1.467***	−1.330***	0.045*	0.006
	(0.647)	(0.745)	(0.369)	(0.339)	(0.023)	(0.024)
Number of Type 1 A&Es within 20 km	−6.079***	−4.848***	–	–	–	–
	(0.714)	(0.821)	–	–	–	–
Mental health severityˆ	–	–	0.597***	0.612***	0.008	0.008
	–	–	(0.144)	(0.132)	(0.009)	(0.009)
Constant	188.924***	157.414***	101.119***	127.776***	5.614***	5.824***
	(22.662)	(27.102)	(30.086)	(27.591)	(1.866)	(1.985)
Number of CCGs	189	189	211	211	211	211

Note: Higher deciles of IMD represent the least deprived areas. The lowest decile of the SAMHI measure represents the worst mental health. Higher values of quality of CCG leadership and effectiveness of local working relationships represent favourable outcomes. ˆMental health severity is measures as the % of a CCG in lowest SAMHI decile. The SAMHI is derived from five measures measured at lower super output area level: NHS-Mental health-related hospital attendances, Prescribing data – Antidepressants, QOF - depression, and DWP - Incapacity benefit and Employment support allowance for mental illness. Standard errors in parentheses. *p < 0.1; **p < 0.05; ***p < 0.01.

positively by 0.296 ($p > 0.1$) in 2018 and negatively by 1.345 ($p > 0.1$) in 2019 per 100,000 population. Higher responsiveness in mental health expenditure is not related to avoidable mortality but is strongly related to a lower suicide rate by 0.496 ($p < 0.01$) in 2014 to 2016 and 0.447 ($p < 0.01$) in 2016 to 2018 per 100,000 population. Therefore, there is evidence that areas with higher mental health responsiveness have more favourable suicide rates.

4.3. The impact of removing the least responsive areas

Fig. 3 shows baseline predicted need using all CCGs in the formula. Predicted need is £2415.15 on average per person within the highest need decile and £90.41 within the lowest need decile for general and acute services. For mental health services, average need is £-0.90 in the lowest need decile and £291.35 in the highest need decile.

Fig. 3 (and appendix Table A4) also shows differences from the baseline service need prediction, at each need decile, after progressively removing less responsive CCGs. Removal of outlier CCGs, results in changes to predicted need deciles within the range of £4 for general and acute and the range of £1 for mental health. Removal of all CCGs that have a negative value of responsiveness of general and acute expenditure, or, in other words, where an increase in expenditure is lower than an equivalent increase in predicted need, results in the lowest predicted need decile at £89.79 lower and the highest predicted need decile at £163.48 higher relative to the baseline formula. Therefore, the gap between the highest and lowest need deciles increases by £253.27 per person, on average, relative to the baseline general and acute formula.

Table 3 shows how the national predicted needs gradient changes from using sub-samples of CCGs based on responsiveness. Each value in Table 3 is derived from the coefficient on the fractional ranking coefficient from equation (5). The gap between the lowest and highest need individual is £1743.98 and £248.43 for the general and acute and mental health formula, respectively. However, removing all CCGs with a negative responsiveness value from estimation of the formula produces an equivalent gap that is 10.77% and 44.27% higher for the general and acute and mental health formula, respectively.

5. Discussion

There are concerns that the allocation of healthcare resources based on variations in historical utilisation may reflect and reinforce patterns of unmet need (Santana et al., 2021). However, there is little evidence, in the context of publicly-funded healthcare systems on how to address this issue. We offer a solution to these concerns through building on the idea highlighted by Sutton and Lock (2000) to estimate the formula on areas that are least affected by unmet need, or more responsive to need. Our measure of the responsiveness of expenditure to need indicates which local health systems devote more resources to their needier populations. Therefore, resource allocation formulae estimated using coefficients from more responsive areas ensures each region's share of the health care budget are based on re-calibrated weights that devote more resources to needier populations.

We find substantial variation in responsiveness across health organisations in England demonstrating the scope for calibrating the weights from resource allocation formulae on more responsive areas. Removal of outlier health organisations with the lowest responsiveness values and those with a negative responsiveness value from the estimation of the general and acute formula results in the top decile of the need distribution with a higher predicted need value relative to the baseline formula, on average, at £3.00 and £163.48, respectively. For predictions with the mental health formula the equivalent values for the neediest decile are £0.35 for the removal of outlier health organisations and £78.89 for the removal of all negatively responsive health organisations. The removal of more health organisations based on responsiveness results in a progressive increase in the difference between the lowest and higheest need individuals.

Our measure of responsiveness builds on the idea of Sutton and Lock (2000) to select sub-groups of local health organisations to estimate resource allocation formulae. However, in contrast to Sutton and Lock (2000) we are able to use person-level data and capture the progressivity of spending over need using an adaptation of the slope index of inequality rather than the Kakwani index of progressivity. We suggest removing the least responsive areas rather than retaining the most

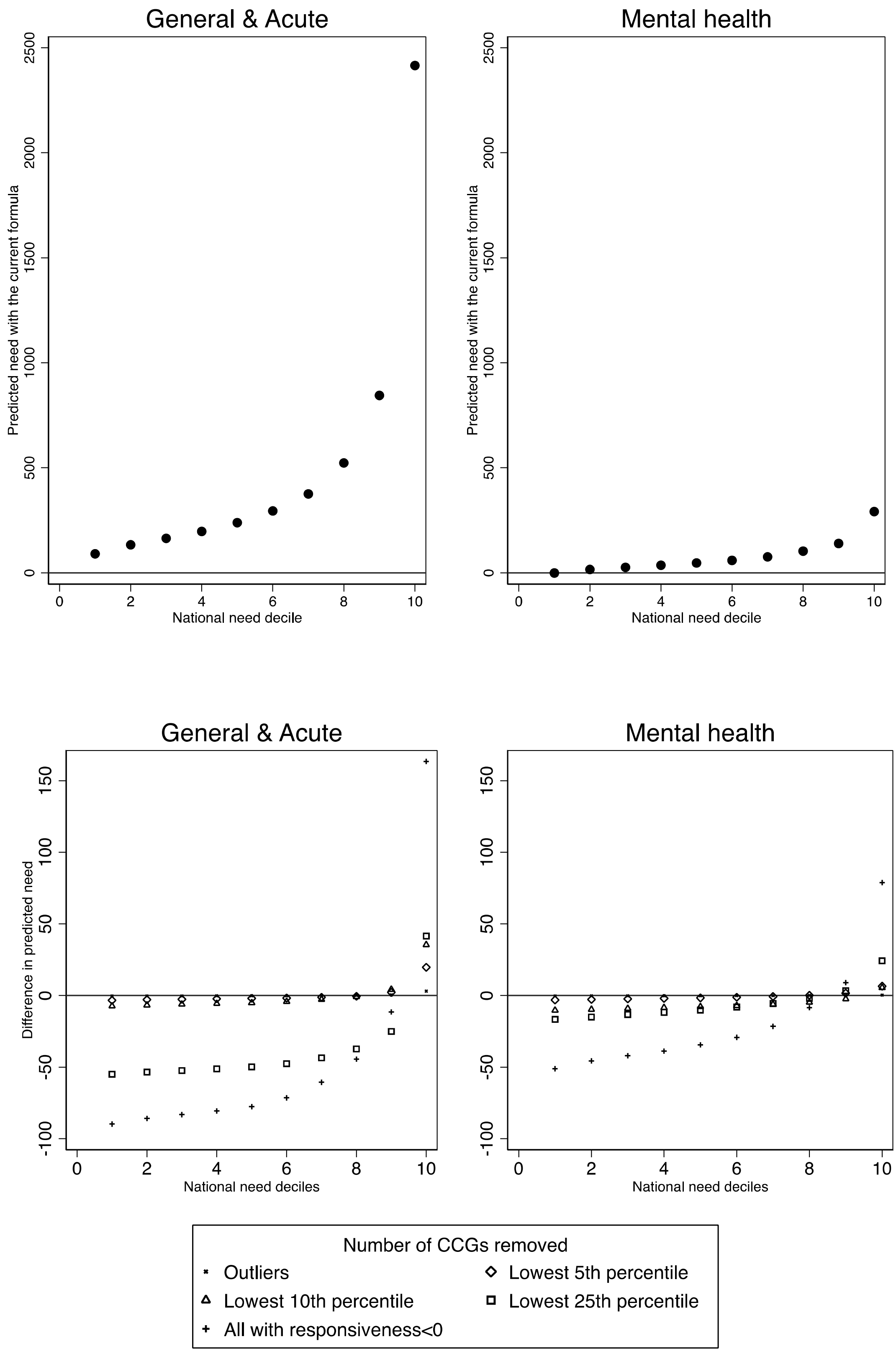

Fig. 3. Need predictions by national need decile.

responsive to recalibrate the coefficients of the risk adjustment model. The choice of using a slope index of inequality is in line with other metrics used to evaluate inequality performance in the English health system (Asaria et al., 2016; Cookson et al., 2018). We describe our measure as the responsiveness of expenditure to need to avoid confusion with the WHO broader definition of health system responsiveness which describes "how well the health system meets the legitimate expectations of the population for the non-health enhancing aspects of the health system" (Mirzoev and Kane, 2017).

A possible concern with the application of our responsiveness

Table 3
Predicted needs gradient using responsiveness to select CCG subgroups.

	Baseline prediction	Outlier	Bottom 5%	Bottom 10%	Bottom 25%	Negative
General and Acute national predicted need gradient:						
Removing the least responsive CCGs (% change from baseline)	£1743.98	£1749.85 (0.34%)	£1761.59 (1.00%)	£1779.21 (2.02%)	£1820.32 (4.38%)	£1931.89 (10.77%)
Mental health national predicted need gradient:						
Removing the least responsive CCGs (% change from baseline)	£248.43	£249.42 (0.40%)	£256.78 (3.36%)	£262.21 (5.55%)	£283.93 (14.29%)	£358.40 (44.27%)

measure is that some areas may be too responsive and perform poorly in other dimensions of performance leading to less favourable health outcomes. To address this issue, we have examined whether responsiveness is related to avoidable mortality and suicide rates. The lack of a relationship between responsiveness and avoidable mortality provides some evidence that the less responsive areas we remove from the estimation of the General and Acute formula do not have lower avoidable mortality. A concern with this lack of a relationship between responsiveness and avoidable mortality may reflect a lack of statistical power at this ecological level. However, since we find several factors are significantly related to avoidable mortality such as deprivation, urbanicity and proximity of Type 1 A&E departments, this may be less of a concern. We find some evidence that higher mental health responsiveness is related to lower suicide rates. These results are largely consistent when we allow the time period of the suicide rate to vary. However, we caveat these results as they are based on under 200 observations. Other sectors of the health and social care system, such as primary care, may also influence these outcomes and these outcomes are relatively crude (yet widely used measures).

Our measure of responsiveness is based on coefficients representing relationships between costed utilisation and the recorded measures of need. Unmet need may manifest itself in different ways beyond our exploration here which is based on the need coefficients. For example, with under or over recorded diagnostic information (Finkelstein et al., 2017), where the inputting of diagnostic information represents an avenue for future research in this area. Our work contrasts with the wider risk adjustment literature which has focused on variable selection techniques, incomplete information and compensation within an insurance setting (McGuire et al., 2021; Rose, 2016; Rose et al., 2017; Zink and Rose, 2021). However, different approaches to adjusting for aspects related to unmet needs could be combined to improve the selection of needs variables, correction for incomplete information and the estimation of their effects on service use.

Future work can examine responsiveness over time to identify whether changes in responsiveness relate to changes in area and system level characteristics. This would address the limitation of small observation numbers for these analyses and would also enable a more detailed understanding of the characteristics and possible drivers of what makes a region more or less responsive. In 2022, CCGs in England were reorganised into 42 Integrated Care Systems (or Boards) that cover larger areas than CCGs. Our approach can be replicated to produce a responsiveness measure for these organisations, but can also be applied internationally to other publicly funded healthcare systems. Whilst the geographical units, upon which final proportions of health care budgets are derived from, will have changed to the larger (in terms of population) ICBs, the principles of our methodology would be unchanged whether applied to ICB, CCGs or other smaller geographies.

Our measure of responsiveness is a summary of a linear relationship between an individual's national fractional need ranking and their level of predicted need. We show that individuals from every CCG are present at every point of the national need distribution. However, future work could estimate a non-linear relationship between need ranking predictions.

Our measure of responsiveness can be readily applied in risk adjustment. Excluding less responsive areas from estimation of the formulae increases the relative needs weights for populations identified as being of highest need by the national formula. A major advantage of the proposed approach is that it remains based on patterns of expenditure that are observed in some areas. These are the areas that show a more responsive distribution of expenditure with respect to need than other areas. This adjustment addresses the general concern that the risk adjustment "gearing" is not sufficiently responsive to need. It is also important to note that the omitted least-responsive areas do not necessarily lose out in the calculation of their needs, as the new needs weights are applied to all populations. Decision makers can decide on how many of the least-responsive areas to exclude from the model estimation. This judgement could be based, for example, on consideration on the size of adjustment, the extent of unmet need and its consequences, or some benchmark for responsiveness as an indicator of effective resource targeting.

6. Conclusion

Our study provides an implementable means of accounting for unmet needs in a resource allocation formula. We exploit data that has been previously analysed for informing allocations as well as a range of area and system level characteristics. We show the impact across a range of scenarios of removing different numbers of local health organisations with lower levels of responsiveness. As expected, higher need individuals are estimated to have higher expenditure needs when less responsive local health organisations are removed from the estimation of the risk-adjustment. Whilst we provide an example based on the responsiveness of expenditure to a composite measure of need, the approach could be calculated according to other dimensions, such as deprivation.

Funding and acknowledgements

This project was funded by the National Institute for Health and Care Research (NIHRDH-NIHR130258 Accounting for Unmet Needs in Equitable Resource Allocation).

Laura Anselmi was also funded by the Medical Research Council (MRC) MR/S022554/1.

We are grateful to Ben Barr, Richard Cookson, Chris Bentley, Tarryn Lake, Christodoulos Kypridemos, James Lomas, Miqdad Asaria, NHS England Allocation Team, members of the Technical Advisory Group on Resource Allocation and Advisory Committee on Resource Allocations for continous challenging, feedback and support.

Competing interests

Laura Anselmi is a member of the Technical Advisory Group on Resource Allocations

Matt Sutton is a member of the Advisory Committee on Resource Allocations

CRediT authorship contribution statement

Sean Urwin: Writing – review & editing, Writing – original draft, Software, Formal analysis. **Laura Anselmi:** Conceptualisation, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Writing – review & editing, Writing – original draft. **Emmanouil Mentzakis:** Data curation, Formal analysis. **Yiu-Shing Lau:** Writing – review & editing. **Matt Sutton:** Writing – review & editing, Writing – original draft, Funding acquisition, Conceptualization.

Data availability

The authors do not have permission to share data.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.socscimed.2024.117346.

References

Anselmi, L., Everton, A., Shaw, R., Suzuki, W., Burrows, J., Weir, R., Tatarek-Gintowt, R., Sutton, M., Lorrimer, S., 2020. Estimating local need for mental healthcare to inform fair resource allocation in the NHS in England: cross-sectional analysis of national administrative data linked at person level. Br. J. Psychiatry 216, 338–344. https://doi.org/10.1192/bjp.2019.185.

Anselmi, L., Lau, Y.-S., Sutton, M., Everton, A., Shaw, R., Lorrimer, S., 2022. Use of past care markers in risk-adjustment: accounting for systematic differences across providers. Eur. J. Health Econ. 23, 133–151. https://doi.org/10.1007/s10198-021-01350-9.

Asaria, M., Ali, S., Doran, T., Ferguson, B., Fleetcroft, R., Goddard, M., Goldblatt, P., Laudicella, M., Raine, R., Cookson, R., 2016. How a universal health system reduces inequalities: lessons from England. J. Epidemiol. Community Health 70, 637–643. https://doi.org/10.1136/jech-2015-206742.

Barr, B., Bambra, C., Whitehead, M., 2014. The impact of NHS resource allocation policy on health inequalities in England 2001-11: longitudinal ecological study. BMJ 348, g3231. https://doi.org/10.1136/bmj.g3231.

Chaplin, M., Beatson, S., Lau, Y.-S., Davies, C., Smyth, C., Burrows, J., Weir, R., Tatarek-Gintowt, R., 2016. Refreshing the Formulae for CCG Allocations for Allocations to Clinical Commissioning Groups from 2016-17. Report on the Methods and Modelling. NHS England.

Cookson, R., Asaria, M., Ali, S., Shaw, R., Doran, T., Goldblatt, P., 2018. Health equity monitoring for healthcare quality assurance. Soc. Sci. Med. 198, 148–156. https://doi.org/10.1016/j.socscimed.2018.01.004.

Dixon, J., Smith, P., Gravelle, H., Martin, S., Bardsley, M., Rice, N., Georghiou, T., Dusheiko, M., Billings, J., Lorenzo, M.D., Sanderson, C., 2011. A person based formula for allocating commissioning funds to general practices in England: development of a statistical model. BMJ 343, d6608. https://doi.org/10.1136/bmj.d6608.

Finkelstein, A., Gentzkow, M., Hull, P., Williams, H., 2017. Adjusting risk adjustment — accounting for variation in diagnostic intensity. N. Engl. J. Med. 376, 608–610. https://doi.org/10.1056/NEJMp1613238.

Fleurbaey, M., Schokkaert, E., 2009. Unfair inequalities in health and health care. J. Health Econ. 28, 73–90. https://doi.org/10.1016/j.jhealeco.2008.07.016.

Gravelle, H., Sutton, M., Morris, S., Windmeijer, F., Leyland, A., Dibben, C., Muirhead, M., 2003. Modelling supply and demand influences on the use of health care: implications for deriving a needs-based capitation formula. Health Econ. 12, 985–1004. https://doi.org/10.1002/hec.830.

Juhnke, C., Bethge, S., Mühlbacher, A., 2016. A review on methods of risk adjustment and their use in integrated healthcare systems. Int. J. Integrated Care 16, 4. https://doi.org/10.5334/ijic.2500.

Front-matter. In: McGuire, T.G., van Kleef, R.C. (Eds.), 2018. Risk Adjustment, Risk Sharing and Premium Regulation in Health Insurance Markets. Academic Press, pp. i–iii. https://doi.org/10.1016/B978-0-12-811325-7.00021-X.

McGuire, T.G., Zink, A.L., Rose, S., 2021. Improving the performance of risk adjustment systems: constrained regressions, reinsurance, and variable selection. Am. J. Health Econ. 7, 497–521. https://doi.org/10.1086/716199.

Ministry of Housing, Communities & Local Government, 2019. English Indices of Deprivation 2019.

Mirzoev, T., Kane, S., 2017. What is health systems responsiveness? Review of existing knowledge and proposed conceptual framework. BMJ Glob. Health 2, e000486. https://doi.org/10.1136/bmjgh-2017-000486.

Morris, S., Sutton, M., Dixon, P., Wildman, J., Birch, S., Raine, R., Chandola, T., Orr, S., Jit, M., Wolff, J., Atkinson, S., Marmot, M., 2010. Research on the Health Inequalities Elements of the NHS Weighted Capitation Formula.

NHS England, 2024. Technical Guide to Allocation Formulae and Convergence. For 2023/24 and 2024/25 Allocations.

NHS England, 2023. Fair Shares a Guide to NHS Allocations (Allocations Infographics V3 Updated for ICB Resource Allocations 2023/24).

NHS England, 2022. Update of the Formula for General and Acute Hospital Services for 2022/23 Allocations.

NHS England, 2018. Technical Guide to Allocation Formulae and Pace of Change for 2016-17 to 2020-21 Revenue Allocations to Clinical Commissioning Groups and Commissioning Areas.

OECD, 2021. Unmet Needs for Health Care. OECD, Paris. https://doi.org/10.1787/13aff239-en.

ONS, 2024. Avoidable Mortality by Clinical Commissioning Groups in England and Health Boards in Wales.

ONS, 2019. Number of Suicides by Clinical Commissioning Group, 2016 to 2018 Death Registrations.

Rice, N., Smith, P.C., 2001. Capitation and risk adjustment in health care financing: an international progress report. Milbank Q. 79, 81–113. https://doi.org/10.1111/1468-0009.00197.

Rose, S., 2016. A machine learning framework for plan payment risk adjustment. Health Serv. Res. 51, 2358–2374. https://doi.org/10.1111/1475-6773.12464.

Rose, S., Shi, J., McGuire, T.G., Normand, S.-L.T., 2017. Matching and imputation methods for risk adjustment in the health insurance marketplaces. Stat. Biosci. 9, 525–542. https://doi.org/10.1007/s12561-015-9135-7.

Santana, I.R., Mason, A., Gutacker, N., Kasteridis, P., Santos, R., Rice, N., 2021. Need, demand, supply in health care: working definitions, and their implications for defining access. Health Econ. https://doi.org/10.1017/S1744133121000293. Policy Law 1–13.

Schokkaert, E., Guillaume, J., van de Voorde, C., 2018. Chapter 7 - risk adjustment in Belgium: why and how to introduce socioeconomic variables in health plan payment. In: McGuire, T.G., van Kleef, R.C. (Eds.), Risk Adjustment, Risk Sharing and Premium Regulation in Health Insurance Markets. Academic Press, pp. 209–234. https://doi.org/10.1016/B978-0-12-811325-7.00007-5.

Smith, P.C., 2008. Resource allocation and purchasing in the health sector: the English experience. Bull. World Health Organ. 86, 884–888. https://doi.org/10.2471/blt.07.049528.

Smith, P.C., 2006. Formula Funding of Public Services, Routledge Studies in Business Organisations and Networks. Routledge, London.

Smith, S., Connolly, S., 2020. Re-thinking unmet need for health care: introducing a dynamic perspective. Health Econ. Pol. Law 15, 440–457. https://doi.org/10.1017/S1744133119000161.

Sutton, M., Lock, P., 2000. Regional differences in health care delivery: implications for a national resource allocation formula. Health Econ. 9, 547–559. https://doi.org/10.1002/1099-1050(200009)9:6<547::AID-HEC543>3.0.CO;2-E.

Vallejo-Torres, L., Morris, S., Carr-Hill, R., Dixon, P., Law, M., Rice, N., Sutton, M., 2009. Can regional resource shares be based only on prevalence data? An empirical investigation of the proportionality assumption. Soc. Sci. Med. 69, 1634–1642. https://doi.org/10.1016/j.socscimed.2009.09.020.

Zink, A., Rose, S., 2021. Identifying undercompensated groups defined by multiple attributes in risk adjustment. BMJ Health Care Inform 28, e100414. https://doi.org/10.1136/bmjhci-2021-100414.

Zink, A., Rose, S., 2020. Fair regression for health care spending. Biometrics 76, 973–982. https://doi.org/10.1111/biom.13206.