Summary. The paper focuses on the Hospital and Community Health Service (HCHS) component of the current formula for the target allocation of annual “weighted capitation” funding of England’s 304 Primary Care Trusts — a formula that responds to political pressure to tackle continuing health inequalities. The history of allocation formulae for the National Health Service is briefly reviewed, before exposing the composition and provenance of the HCHS component. Under the heading “Logical and statistical deficiencies” we consider different frameworks, model mis-specification, allowance for age, ins & outs of supply, unmet need & wrong signs, and replicability & robustness — questions relevant to trust in the current formula based on the regression of age-standardized current utilization on a multiplicity of socio-economic variables. We conclude that it might be better to put future resources into developing direct, rather than proxy, measurement of health needs.

Keywords: Direct measurement, Health need, Primary Care Trusts, Proxy variables, Weighted capitation funding.

Address for correspondence: 25 Hawtrey Drive, Ruislip, HA4 8QW, UK. E-mail: mervyn at stats.ucl.ac.uk

“O what a tangled web we weave...” (Sir Walter Scott)

1 Introduction

1.1 Socio-economic factors and health inequality

In “Tackling Health Inequalities: a Programme for Action”, the Department of Health (DH) argued the case for “integrating health inequalities into the mainstream of service delivery, with a focus on disadvantaged areas and groups” (Department of Health, 2003a, p.3). They supported this, and other proposals affecting a wide range of public sector activity, by a dramatic and influential graph reproduced here as Figure 1. DH’s title for Figure 1 is “The Widening Mortality Gap between the Social Classes”. The graph cannot bear the weight of this interpretation. It uses an inappropriate reference population, it ignores the huge shift in the proportion of the population in the Unskilled category between the 1930s and the 1990s, and it takes no account of the changes in the major causes of death over that period. These and other deficiencies are elucidated in Appendix A.

Whether differences in health outcomes are increasing or decreasing, the DH’s aim is to provide “equal health care for equal need”. If we ignore the difficulty with this prescription—that it neglects the question of how to judge the priorities of widely different forms of health care—the problem remains of how to identify need and to quantify the cost of provision. The current target allocation of funds is not based on directly measured need but on more easily accessible measures, such as the Standardized Mortality Ratio (SMR) and other small area demographic, socio-economic and health statistics, as proxies for need which are then combined in complex formulae. The theoretical challenge presented by the work examined in this paper is whether these formulae are based on the science of health need measurement or whether they are simply misleading like Figure 1. Is it possible that bad science is being used to cloak essentially political decisions?

1Research Report No. 242, Department of Statistical Science, University College London. Date: April 2004
1.7 Figure 1 shows that the gap in mortality between professional and unskilled manual men—social classes I and V—since 1930/32 has increased almost two and a half times.

Figure 1: The Widening Mortality Gap Between the Social Classes

1.8 The Government’s aim is to reduce health inequalities by tackling the wider determinants of health inequalities, such as poverty, poor educational outcomes, worklessness, poor housing, homelessness, and the problems of disadvantaged neighbourhoods. This approach is supported by a national public service agreement (PSA) target:

- by 2010 to reduce inequalities in health outcomes by 10 per cent as measured by infant mortality and life expectancy at birth

1.9 The PSA target is underpinned by two more detailed objectives:

- starting with children under one year, by 2010 to reduce by at least 10 per cent the gap in mortality between routine and manual groups and the population as a whole
- starting with local authorities, by 2010 to reduce by at least 10 per cent the gap between the fifth of areas with the lowest life expectancy at birth and the population as a whole

1.10 This Programme for Action is based on the commitments made by all Government departments in the CCR. Chapter 3 sets out the progress of existing programmes and initiatives reflected in the CCR summary.

1.11 This programme also sets out:

- proposals for delivering change across the four themes (chapter 4)
- the changes required in the roles and responsibilities of the key players to help deliver the national target and impact on the underlying causes of health inequalities (chapter 5)
- how progress will be measured and monitored (chapter 6)

Source: Office for National Statistics, Decennial Supplements, analysis by DH Statistics Division

1.2 Allocation of funds to Primary Care Trusts in 2003/4

Since 2001/2, 304 Primary Care Trusts (PCTs) have been responsible for the delivery of health care in their areas. In 2003/4, DH distributed 45 billion pounds to these PCTs. The amount received by an individual PCT was influenced by a target allocation based on a weighted capitation formula (Department of Health, 2003b). Figure 2 plots the 2003-4 target allocation against a complex estimate of the population for which a PCT is responsible. The per capita figure varied from £675 for Wokingham in Surrey to £1326 for Tower Hamlets in London’s East End (figures based on Department of Health, 2003c, Tables 4.2 and 5.2). Figure 2 sheds light on the relative influence on target allocation of differences in population size and the other differences between PCTs that are allowed or encouraged to influence it. The line from the origin through Wokingham appears to be a sort of lower boundary to the points for PCTs with population up to 200,000, along which allocation is roughly proportional to population with a constant per capita figure.

1.3 The weighted capitation formula for Hospital and Community Health Services

Three-quarters (34 billions) of the weighted capitation formula (WCF) costs is for Hospital and Community Health Services (HCHS). This paper will focus on two indices that largely determine the HCHS component of WCF and that are intended to make due allowance for inequalities other than age-profile differences. These indices are \( I_1 \) for acute and maternity services and \( I_2 \) for mental health services. But first a little history from the last three decades!
to respond to regional inequalities was the ex-cathedral “Crossman formula” that defined allocation as:

allowance for growth, plus an allowance for scandals” (Department of Health, 1989a). The first attempt
allocation to geographical areas was based on the crude principle of “What you got last year, plus an
The NHS took over the management of almost all health services in 1948. For over two decades,

2 Historical formulae

The NHS took over the management of almost all health services in 1948. For over two decades, allocation to geographical areas was based on the crude principle of “What you got last year, plus an allowance for growth, plus an allowance for scandals” (Department of Health, 1989a). The first attempt to respond to regional inequalities was the ex-cathedral “Crossman formula” that defined allocation as:

\[ A \propto \text{Population} + \text{Bed-years}/2 + \text{Cases}/4 \]  

(1)

(The constant of proportionality matters for health service managers but not here.) Used from 1971 to 1976, this formula was criticised for paying inadequate attention to regional differences. The Resource Allocation Working Party (RAWP) was set up with orders to devise a “pattern of distribution responsive objectively, equitably and efficiently to relative need”. The RAWP formula, in substantial use from 1976 to 1988, was essentially:

\[ A \propto \sum_c \text{Population}_c \times \text{SMR}_c \]  

(2)

where \( \text{Population}_c \) is the population weighted by bed usage categorised by age and sex for condition \( c \), and \( \text{SMR}_c \) is the regional Standardized Mortality Ratio for condition \( c \). The \( \text{SMR}_c \)s in (2) gave it a responsiveness to differential death rates that led to a massive transfer of resources from five Regional Health Authorities (RHAs) in the South East to RHAs in the North with higher \( \text{SMR}_c \)s. Dissatisfaction with the formula built up until 1985 when DH decided that some “fine tuning” was needed. This resulted in the RoR (Review of RAWP) formula. As Sheldon and Carr-Hill (1992) tell it in their excellent review, the RoR team devised a formula based on regression analyses of the dependence of the variation between wards within RHAs of the “total acute non-psychiatric, non-maternity deaths and discharge rates” on explanatory variables that included SMR75 (the standardized mortality ratio for death before 75) to the power of 0.44 and a measure of service availability. DH found this fine tuning altogether too empirical (Department of Health, 1989b) and preferred the formula:

\[ A \propto (\text{Population weighted by national resource utilisation in age bands}) \times \sqrt{\text{SMR75}} \]  

(3)
Although the Sheldon and Carr-Hill (1992) critique and others (Carr-Hill, 1990; Sheldon et al, 1993) reflected widespread dissatisfaction with (3), it was nevertheless used from 1988 until 1996, returning to the South some of the resources that had been moved to the North by the RAWP displacement of the Crossman formula.

In 1993, a University of York team that included Carr-Hill and Sheldon took up the search for better formulae, to improve sensitivity to geographical variations in the need for health care. Rather than produce a single formula, the York team recommended the use of separate formulae to cover acute services and mental health services, to be supplemented by further measures of need for other components of Hospital and Community Health Services. They also recommended that “consideration should be given to initiating a national cohort study as an alternative means of measuring the link between social circumstances and health care needs of individuals” (Carr-Hill et al, 1994; “York” for short; p.5). As part of their study, the York team fitted a multi-level model regressing the logarithm of utilisation, indirectly standardised by age and sex, on a selection of logged socio-economic and health-related variables. The units of analysis were over 5000 electoral wards (single or amalgamated) at level 1 and 186 District Health Authorities (DHAs) at level 2. For acute services, this gave an analogue of (3) with $\sqrt{\text{SMR}_{75}}$ replaced by

$$
(\text{SMR}_{75})^{0.162}(x_1^t)^{0.253}(x_2^t)^{0.076}(x_3^t)^{0.044}(x_4^t)^{0.029}
$$

where

- $x_1^t$ = Standardized illness ratio for those aged 0–74 in households only
- $x_2^t$ = Proportion of pensionable age living alone
- $x_3^t$ = Proportion of dependants in single carer households
- $x_4^t$ = Proportion of economically-active unemployed.

Formula (4) and similar ones for mental health and other services were used for new Health Authorities until 1998 and for their Primary Care Group replacements until 2002. The case for replacing the York team formulae was started in 1998 and growing political concerns were expressed as a need for a “fairer formula for the new NHS” (Department of Health, 1989a). The reformulation was carried out by a team led by a health economist in the University of Glasgow. Their report was entitled “Allocation of Resources to English Areas” (Sutton et al, 2002; “AREA” for short). It will guide English PCT funding until 2006. The remit for the study included:

(i) “investigating the possibility of constructing a single index of need at PCT level”

(ii) “investigating alternative methodologies to replace the age-related components of the existing formulae”

(iii) “proposing a methodology to adjust for unmet need”

(iv) “ensuring that the proposed need indices were robust at PCT level”

(our italics, AREA p.2).
3 The current Health and Community Health Services formula

3.1 Age-profile and inequality adjustments

The WCF document (Department of Health, 2003b) gives a clear account of how the HCHS component has been constructed. Allocation to PCTs is now guided by

\[ A \propto C_{\text{age}} \times I \times I' \times I'' \] (5)

where \( C_{\text{age}} \) is the estimate of the cost of providing services based on national averages of cost in seven age bands. Specifically for PCT \( i \)

\[ C_{\text{age},i} = \sum a n_a c_a N \] (6)

where \( n_a \) is the population in age-band \( a \) and \( c_a N \) the corresponding national per capita cost of HCHS services. The indices \( I' \) and \( I'' \) adjust for regional cost differences between PCTs and will not concern us here. Index \( I \) adjusts for differences between PCTs in the explanatory factors selected by the AREA team. From the worked arithmetic examples, we deduce that \( I \) has the algebraic form

\[ I = 0.8553 \times I'_1 + 0.1447 \times I'_2 \] (7)

where \( \{I'_t, t = 1, 2\} \) are the indices of inequality for acute and maternity care and for mental health care given by

\[ I'_t = I_t / \sum_j P_j I_{ij} \] (8)

for PCT \( i \), where \( P_j \) is the proportion of the national population for which PCT \( j \) is responsible. The weights 0.8553 and 0.1447 in (7) are the proportions in which the total HCHS expenditure for 1998-2001 was divided between the two activities.

The values of \( \{I_{1i}, I_{2i}\} \) are of the linearly constructed indices

\[ I_1 = -0.152 + 0.008x_1 + 0.013x_2 + 0.070x_3 + 0.026x_4 + 0.108x_5 + 0.103x_6 + 0.225x_{39} + 0.548x_{40} + 0.375x_{41} \] (9)

\[ I_2 = 0.385 + 0.358x_7 + 0.338x_8 + 0.034x_9 + 0.636x_{42} \] (10)

where

\begin{align*}
    x_1 &= \text{ID2000, deprivation index for education} \\
    x_2 &= \text{Low birthweight babies per 1000} \\
    x_3 &= \text{SMR 75, standardized mortality ratio under 75} \\
    x_4 &= \text{Proportion over 75 living alone} \\
    x_5 &= \text{Standardized birth ratio} \\
    x_6 &= \text{ID2000, deprivation index for income} \\
    x_7 &= \text{CMF65, comparative mortality factor under 65} \\
    x_8 &= \text{Proportion over 60 claiming income support} \\
    x_9 &= \text{ID2000, deprivation index for housing} \\
    x_{39} &= \text{Nervous system morbidity index} \\
    x_{40} &= \text{Circulatory system morbidity index} \\
    x_{41} &= \text{Musculo-skeletal morbidity index} \\
    x_{42} &= \text{Psycho-social morbidity index}.
\end{align*}

and \( x_1, \ldots, x_9 \) have been standardized to have a population-weighted average of zero and s.d. equal to the coefficient of variation of the unstandardized variable. The indices \( I_1 \) and \( I_2 \) are taken from the
AREA team report, complete with already estimated regression coefficients (AREA, Table 8.4.4, p. 142 and Table 8.4.2, p. 140). The \( x \)-variables have to be calculated for every PCT to determine the target allocations. The break in their numbering is because they were drawn from a pool of 80 or so variables — 46 so-called need variables \( x_1, \ldots, x_{46} \) (including eight morbidity indices specially constructed from socio-economic variables) and 34 so-called supply variables \( x_{47}, \ldots, x_{80} \).

Department of Health (2003b) tells us that \( I_1 \) and \( I_2 \) are portions of linear regressions on larger numbers of variables. For \( I_1 \), in addition to the need variables in (9) the explanatory variables included the following eight supply variables:

\[
\begin{align*}
  x_{47} &= \text{Mean waiting time} \\
  x_{48} &= \text{Distance to general practice} \\
  x_{49} &= \text{Distance to hospital} \\
  x_{50} &= \text{Proportion of outpatients seen under 13 weeks} \\
  x_{51} &= \text{Residential/nursing homes} \\
  x_{52} &= \text{Access to private providers} \\
  x_{53} &= \text{Number of hospital beds} \\
  x_{54} &= \text{Distance to maternity hospital}
\end{align*}
\]

and two more need variables (intriguingly labelled “other”)

\[
\begin{align*}
  x_{10} &= \text{Proportion of ethnic minorities} \\
  x_{11} &= \text{ID2000, deprivation index for employment}.
\end{align*}
\]

For \( I_2 \), in addition to those in (10) the full regression formula included \( x_1 \) (a “need” variable in \( I_1 \) but now reclassified as “other”), \( x_{10} \) (again classified as “other”) and the supply variable

\[
x_{55} = \text{Average distance to a mental health hospital}.
\]

A need variable classified or reclassified as “other” makes no direct contribution to the relevant index. Instead, it joins the supply variables to form a class of silent variables that play a hidden role in determining PCT allocations. The role of silent variables and the “other”-ness mystery will be explained in Sections 4.2 and 5.5.

### 3.2 Relative importance of the age adjustment and the need inequality adjustment

Figure 3 reveals the relative influence of age-profile and the inequality index \( I \) on the HCHS component of WCF. Age-weighted population is the conceptual population that the PCT would have if the national population were divided in the proportions of \( \{C_{age}\} \): the age-weighted index is then the ratio of age-weighted population to actual PCT population. It is widely supposed that variations in age-profile between PCTs have a greater influence on allocation than the inequality index, perhaps because the national per capita cost \( c_{a,N} \) depends so strikingly on age \( a \) — from £225 for a child between 5 and 15, to £2357 for the over-85s (Department of Health, 2003b). Figure 3a shows how the 2003-4 HCHS allocations would have varied with population if those allocations had been simply determined by the allowance for age-profile in the formula \( A \propto C_{age} \). There is little deviation from a constant per capita allocation. That is no longer the case with the additional adjustment for inequality in the formula \( A \propto C_{age} \times I \). Figure 3c gives an alternative (symmetrical) view of the relative influence of the two adjustments, revealing that they are strongly negatively correlated (\( r = -0.34 \) and the \( 2 \times 2 \chi^2 \) is 29, \( P < 0.0001 \)) — mainly because ageing populations do not qualify for the largest upward adjustments.
Figure 3: Comparison of the influence of age-profile and inequality index on the national estimates of PCT costs of HCHS services. The ordinate scale in (a) and (b) is chosen so that the total notional allocation in those Figures is the 34 billions of Section 1.3. PCT codes: ED is East Devon, TH is Tower Hamlets and W is Wokingham. Numbers in the quadrants of (c) are counts of PCTs.

by the inequality index \( I \). Rural East Devon (ED), with 28% over-65s compared with 16% for England as a whole, clearly illustrates the latter feature.

4 Derivation of the inequality indices \( I_1 \) and \( I_2 \)

4.1 Units of analysis and choice of dependent variable

The data used by the AREA team to construct the indices are not available at the level of individuals but can be aggregated to make analytical units out of electoral wards, general practices or even Primary Care Trusts themselves. The units chosen for fitting the regression models were unamalgamated electoral wards — 8410 for \( I_1 \) and 7982 for \( I_2 \).

For the choice of dependent variable, the AREA team used age-standardized utilization i.e. relevant expenditure on individuals in the ward adjusted for the ward’s age-profile: “In general, direct standardization was the method of choice, but an indirectly standardized model was sometimes preferred on the basis of statistical tests” (AREA, p.91). So \( I_3 \) came out of modelling the indirectly standardized

\[
Y_1 = \frac{C_1}{\sum_n n_a c_{1aN}}
\]  

(11)
but \( I_2 \) came from the directly standardized

\[
Y_2 = \frac{\sum a n_a c_{2a}}{C_{2N}}. \tag{12}
\]

Here \( C_1 \) is the estimate of the ward’s crude utilization (cost) for the acute and maternity services component of HCHS, \( n_a \) is now the ward population in age-band \( a \), \( c_{1aN} \) is the corresponding per capita national cost estimate for age-band \( a \), \( n_aN \) is the national population in age-band \( a \) out of 19 bands (not seven as in (6)), \( c_{2a} \) is the per capita estimate of the ward’s cost of mental health services for age-band \( a \), and \( C_{2N} \) is the estimate of the total national cost of mental health services.

Had the AREA team chosen direct standardization for acute and maternity services, DH would now be using

\[
I_1 = \text{constant} + 0.010x_{11} + 0.014x_2 + 0.072x_{12} + 0.040x_4 + 0.079x_5 + 0.078x_6
+ 0.268x_{39} + 0.468x_{40} + 0.356x_{41} + 0.162x_{43}. \tag{13}
\]

instead of (9). As well as the changes in the coefficients, \( x_{12} \) (CMF75) would replace \( x_3 \) (SMR75), and we would get an extra variable \( x_{43} \) (the respiratory morbidity index). Furthermore, the constant term would have to be calculated for a different linear function of the national means of the silent variables. The AREA report does not say what effect these changes would have on PCT allocations or how the effect would be defended.

4.2 The choice of explanatory variables

Appendix D of the AREA report lists the 38 “need” variables excluding the morbidity indices and the 34 “supply” variables. This multiplicity of highly correlated explanatory variables was reduced “on the basis of their updateability and plausibility and the correlation between them” and then

“The choice of supply and need variables to include...was by forward and backward stepwise selection from the shortlist of variables....Stepwise regression methods indicated which variables were important determinants of utilization. However, particularly in the backward stepwise models, many of the included variables were highly correlated and consequently variables had unexpected signs. Model specification was improved by considering the correlations between the included variables, [their] importance in explaining residual variation and RESET tests for misspecification. It is from this model that we tested whether particular variables with unexpected signs could be reliably interpreted as evidence of unmet need. An incorrect specification caused by omitting relevant variables or using the wrong functional form can give misleading or biased coefficients...To assess the model specification we use the RESET (Regression Specification Error Test) test (Ramsey, [1970]).” (AREA, §5.3)

In addition to \( x_{47}, ..., x_{80} \), the AREA team introduced two more supply factors: 95 Health Authorities (HAs) and an unspecified number of providers (mental health hospitals). Wards are subdivisions of HAs and may make use of more than one mental health hospital. In fitting the regression that generated \( I_1 \), an additive dummy variable was used for each HA. For \( I_2 \), things were more complex — the extra variables were the proportions \( \{x_q|q = \{81, 82, ..., 80 + Q\}\} \) of a ward’s activity delivered by the unspecified number \( Q \) of providers. When multiplied by their coefficients, the HA dummies and the provider variables are taken to represent components of supply and so join the already large class of silent variables. The HA dummies act like any other variable in linear regressions. Their numbering, not required in this account, would be from \( 81 + Q \) to \( 175 + Q \).
From Tables 8.4.4 and 8.4.2 (AREA) we can extract the estimated regression coefficients for all variables except the HA dummies and the provider variables. For a ward in HA \( h \), equation (9) expands to include an intercept \( \hat{\alpha}_h \) and 10 silent variables:

\[
\hat{Y}_1 = \hat{\alpha}_h + 0.008 x_1 + 0.013 x_2 + 0.070 x_3 + 0.026 x_4 + 0.108 x_5 + 0.103 x_6 + 0.225 x_39 + 0.548 x_{40} + 0.375 x_{41} - 0.013 x_{10} - 0.158 x_{11} - 0.101 x_{47} - 0.047 x_{48} - 0.021 x_{49} + 0.160 x_{50} - 0.003 x_{51} - 0.034 x_{52} + 0.013 x_{53} - 0.023 x_{54}
\]

where the \( t \)-values are shown over the corresponding coefficients. The least-squares fitting was weighted by ward populations and gave \( R^2 = 0.76 \) and \( \text{RESET} = 0.45 \) with \( P = 0.72 \) (see Appendix B for explanation of the RESET test).

For mental health services the expansion of (10) is

\[
\hat{Y}_2 = \sum_{q=81}^{80+Q} \beta_q x_q + 0.358 x_7 + 0.338 x_8 + 0.034 x_9 + 0.636 x_{42} - 0.046 x_1 - 0.034 x_{10} - 0.072 x_{55}
\]

with \( R^2 = 0.36 \) and \( \text{RESET} = 47 \) (\( P < 0.0001 \)). The \( t \)-values in (14) and (15) were calculated using heteroscedasticity-consistent estimates (White, 1980; see Appendix B) of the standard errors of the associated coefficients. The constants \(-0.152\) in (9) and \(0.385\) in (10) were determined from (14) and (15) by giving all the silent variables, including the HA and provider variables, their national averages over wards.

The “otherness” of \( x_{10} \) or \( x_{11} \) in (14) or (15) is now explicable. Although highly statistically significant according to their \( t \)-values, their coefficients are negative — which is the wrong sign for a “need” variable (see Section 5.5). The AREA team therefore reclassified them as silent variables so that their PCT values in the allocation formulae are replaced by their national averages.

Now that we know what has been done to get the current PCT allocation formula, we can ask whether it should have been done and whether the use of the resulting formula may be less acceptable than something less complicated.

## 5 Logical and statistical deficiencies

### 5.1 Framework or Procrustean bed?

The AREA report opened with a clear statement of the prior case for the approach which was then developed over 187 pages. It outlined the general framework in which the DH remit would be accommodated:

"The allocation of resources for health care across geographical areas in the NHS is based on the principle that individuals in equal need should have equal access to care, irrespective of where they live. To implement the principle it is necessary to measure need for health care in different areas. But those allocating resources do not have sufficient information to measure need directly. The basic assumption underlying the resource allocation procedure is that use of health services is determined by patient need and by supply. Statistical modelling of the relationship between utilisation, socio-economic variables (including
measures of health) and supply factors can identify which socio-economic variables are indicators of need because of their effect on utilisation. Information on the level of such socio-economic variables across areas can then be used to allocate resources in a way that reflects need.” (AREA, §1.1)

In stark contrast are these criticisms voiced in recent years:

(i) “...simultaneous modelling techniques cannot deal adequately with a system where demand, utilization and supply are so inextricably intertwined...The use of the formulae gives too much importance to variables which only account for some of the variation...The fundamentally political nature of the construction of an allocation policy and the choice of indicators of need must be acknowledged.” (Sheldon and Carr-Hill, 1992)

(ii) “...the search for an empirically based resource allocation formula of high precision in the name of promotion of equity is largely fruitless...Attention should increasingly be focused on how resources are used — the effectiveness and appropriateness of health service interventions...weighting in the dark [is] an industry which is abstracted from research into, and the management and delivery of, health services.” (Sheldon et al, 1993)

(iii) “We have become besotted with the production of ever more refined empirically based formulas.” (Sheldon, 1997)

(iv) “...a professional consensus should be developed regarding the major methodological issues in this field. If this cannot be provided, it implies either an arrogance among statisticians or the absence of a robust methodological core to the work of the profession, neither of which I believe to be the case...I wonder whether there is a role for the [Royal Statistical] Society in encouraging and endorsing transparency in this area.” (M. Derbyshire in discussion of Smith et al, 2001)

(v) “My third concern is with the existence of the large amounts of unexplained variation in the author’s models.” (H. Goldstein in Smith et al, 2001)

(vi) “...however sophisticated the methods, estimating needs weights is essentially contested, that is to say a concept whose application is inherently a matter of dispute.” (Bevan in Smith et al, 2001)

AREA’s general framework could well have been used eight years earlier by the York team but, despite their common approach, the resulting target allocations differ substantially. The changes contribute to the gap between actual and target allocations, which is so large that, at the DH-proposed rate of implementation, it would take over 20 years to achieve the targets based on the AREA formula (Hacking, 2003). Such differences in outcome are to be expected if statistical techniques are used without a compelling logic to guide them. The AREA team in particular did not respond to expressions of logical concern by students of the English scene over decades. This inattention is highlighted by the fact that there is only one item in common between the AREA team’s 49 references and the York team’s 51 references.

There would be no need at all for regression formulae based on demographic and socio-economic variables (“proxy” variables for short) if we could estimate directly the appropriate utilization for the health care needs of each PCT, for example, by a combination of routine records for met needs and sample surveys for unmet needs. But this approach has been thought to be either infeasible or too costly. The AREA report simply says that “those allocating resources do not have sufficient information to measure need directly” (§1.1) but makes no judgement whether or not such information is unobtainable in principle.
The simplest alternative to proxy formulae would be to use crude utilization estimates such as $C_1$ in (11), but this would merely preserve the status quo with little responsiveness to real differences in health care needs. However, the case has not been made that modelling crude utilization with demographic and selected socio-economic variables will provide sufficiently accurate estimates of appropriate utilization. The logical gap might be filled if it were possible to measure directly appropriate utilization for a sample of individuals large enough to validate the proxy formula. This is presumably what the York team meant by:

"...if an empirical basis is sought for identifying the determinants of utilization, the only method likely to yield significantly more robust and credible results than the present study is the use of long term cohort studies of individuals, which can relate the individual’s needs profile to their actual use of health service resources.”(York, §7.18)

But there could still be large residual errors even after such validation. Both teams seemed happy with large unexplained residuals and the large adjustments to existing allocations that such residuals would generate if the recommended adjustments were fully implemented. The York team was explicit on this point in reference to formula (4):

"...we are not necessarily seeking to develop a model which offers the best fit to the data. This could be achieved simply by indiscriminately adding further socio-economic variables. Instead, we have tried to identify a model that explains that part of the variation in utilization caused only by legitimate needs.” (York, §6.11)

Both teams may have “tried” — but is there any sense in which they can claim to have succeeded when the criterion of success must be fairness in the funding allocations? When the formula is not a good fit — and that is what transpired in both studies — the way is open for serious and undetected biases in an inadequate formula. The quoted criticisms (i) and (v) make this point implicitly. The claim in the AREA report that “Statistical modelling...can identify which socio-economic variables are indicators of need because of their effect on utilisation” (our italics) might be acceptable in the context of a scientific investigation using exploratory data analysis to get clues that would guide further research. But what PCTs want is measurement that is as accurate as possible — modest inaccuracy could lose them millions of pounds in their target allocation.

What is to be made of the participation in the York team of two of the critics quoted above — especially in the report’s claim to have found a “convincing model” on the basis of “sound statistical techniques” (York, §6.64 and §7.12)? The following sections explain some of the reasons why we are unconvinced. They reinforce the widespread critical concern, even if one or two of them may look like arranging deckchairs on the Titanic to anyone who takes a quickly rejectionist line.

5.2 Model mis-specification

The term “specification” has a reassuring ring to it — things that are built “to specification” are things we can live with without question. But, in any major statistical study, reassurance has to be backed up by evidence if what we have constructed comes at the end of a long search for suitable building materials. The crucial influence of search in model building has been studied by Leamer (1978) and Hjorth (1994). Working in a less critical framework, Godfrey (1988) offers numerous significance tests for model specification through which, it is claimed, reassurance can be achieved. None of the significance tests used by the two teams are anything but conditional i.e. they are not concerned to weigh up the influence on the outcome of the test of choices en route to any particular formula. To
accept the results of these tests as a reason for confidence in any formula puts a heavy responsibility on
the subjective judgements of the earlier search process.

The other reason for questioning the support offered by such tests is that, when we are nowhere near
explaining most of the variation in the data, non-significance in some specification test can easily be
consistent with the formula being rejectable on some other fruitful branch of the search tree. In other
words, the specification test may be down a cul-de-sac. The following artificial example illustrates how
this can easily work out for the RESET test.

With three explanatory variables, 1000 independent observations were generated for the model
\[ U = 10 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + e \quad \text{with} \quad x_1, x_2, x_3 \text{and} \ e \ \text{all random Uniform}(-1, 1) \text{and} \ \beta_1 = \beta_2 = 1. \]

If, presented with these data, an inspired search had led us to the true model, we would have got the
estimates \( \hat{\beta}_1 = 1.06, \hat{\beta}_2 = 1.03 \) with \( t \)-values 19 and 30 and an \( R^2 \) of 0.57. But fitting the uninspired
formula \( \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \) to the same data gives significance only to the coefficient of \( x_3 \) with
\( t = 27 \) and \( R^2 = 0.415 \), thereby selecting the model \( \beta_0 + \beta_3 x_3 \) which has an \( R^2 \) of 0.414. Applying
RESET to test the “specification” of the simpler but less truthful formula gives a non-significant \( F \) with
\( P = 0.24 \).

One would need a prior conviction that models must be linear and free of interaction between the
explanatory variables to say that such counterexamples are irrelevant. The AREA team placed less
emphasis than the York team did on formal statistical criteria. Interest in the \( t \)-values for the
inclusion/exclusion of explanatory variables, or in RESET to reveal misspecification, or in \( R^2 \) for
overall satisfaction with a model was almost perfunctory — willing to ignore high significance levels
e.g. for the RESET of the fitting of formula (15). Moreover these objective tests had to defer to
subjective assessments — such as those concerning the coefficients that turned out to have the “wrong
signs” e.g. those for \( x_{10} \) (Proportion of ethnic minorities) and \( x_{11} \) (ID2000 deprivation index for
employment) in (14) or (15).

The AREA report did not acknowledge that low \( R^2 \)-values are associated with massive adjustments of
(standardized) utilization to get to what the need index calls for in the allocation formula. These
adjustments must be assumed to express only random error where they are not adjustment for HA
dummies and supply variables. Low \( R^2 \)’s leave potential room for much better formulae than those
recommended. Acceptance of poorly fitting models defies the Popperian thesis that we should never
accept a hypothesis unless the only alternatives can be positively excluded.

We have already extracted the key \( R^2 \)-values: 0.76 for \( I_1 \) and 0.36 for \( I_2 \). It is important to note that
these already unimpressive values do not reflect the influence of need variables alone. They include the
contributions to the fitting from all the supply variables as well from the levels of the 95 HAs for \( I_1 \) and
\( Q \) providers for \( I_2 \) — but the \( R^2 \)-values have been presented without comment as if the HAs and supply
variables were not relevant.

Sheldon et al (1993) described such empirical allocation formulae as “weighting in the dark”. Over a
decade later, we still lack anything to lighten the darkness. If the scepticism already expressed is
justified, the AREA team’s rich selection of explanatory variables in (14) may get us no nearer the
allowance for inequality that direct measurement of health needs might be able to justify than does the
York team’s selection of the handful of proxy variables in (4). The York report gives no measure of
goodness of the multilevel fit of (4) equivalent to the \( R^2 \) of 0.54 that is generated by least squares-
regression on the logarithms of the five variables. For the latter (see (18) below), the York team’s claim
that “The \( R^2 \) statistic indicates that the chosen acute sector needs variables account for 54% of the
variance in age/sex adjusted utilization...” may be as misleading, by neglecting the contribution of the
14 RHA dummies, as the AREA team’s similar neglect or misinterpretation of the influence on \( R^2 \) of its
much larger number of HA dummies and supply variables. For instance, the latter were omitted from
the models for the last two columns of Table 5.9.1 (in which the Yes labels should have been No) and
for these models \( R^2 \) was observed to be above 0.5. Given that there are big differences in utilization
between HAs, the textual comment that “...the needs variables still explain over half the variation in
utilization when there are no HA effects [sic]” should not be interpreted as evidence that these needs variables “capture overall need” whatever that means. For a valid inference along these lines, we would use only the HA dummies and supply variables before getting the additional contribution of needs variables.

A different problem is that the choice of units (wards) for fitting models does not correspond to the recognisable NHS beneficiaries — neither individuals nor the PCTs that receive and spend the allocated funds. Even an equation that managed to describe the dependence of standardized utilization at ward level may not apply at the level of individuals (the ecological fallacy) or at PCT level (if, for example, there were economies or diseconomies of scale).

5.3 Allowance for age

The York and AREA teams used a two-step approach, first adjusting for age and then for other sources of inequality. For the latter they recommended models that would describe the relationship between potential explanatory variables and standardized utilization rather than utilization itself. In the York report we read:

“Levels of morbidity, and therefore of the need for health care, clearly vary with an individual’s age and sex. ...In practice...the difference between direct and indirect standardization was found to be negligible. We therefore used indirect standardization throughout in order to be consistent with standardized mortality ratios, which are indirectly standardized. ...Standardization is intended to obviate the need to include any age or sex variables in our model of demand. In practice, however, standardization might not fully remove the effect of demography on demand, so — once a preferred model was identified — a large number of demographic variables were added to the chosen model to check that no residual impact of age or sex remained unexplained.” (York, §4.4 & §4.7)

The “intended” here appears to be politically motivated — standardization, claimed without documentation to accommodate all the “effect of demography”, also accommodates the “reluctance to open up the issue of the share of the NHS cake between age groups” (York, §3.14).

The finding of negligible difference between direct and indirect standardization referred, it seems, to the estimates of the coefficients of (it should be noted) already selected variables. This conditional robustness is shown with the model fitted in the AREA report’s Table 8.4.1, where the only major discrepancy appears to be in the values of the RESET statistic ($P = 0.002$ for direct but $P = 0.15$ for indirect). If we accept that, at least for estimating models with data like that used for $I_1$, direct and indirect are practically equivalent, we have here an interesting paradox: a method that sets out to model actual per capita cost by a formula like (21) in Appendix C that involves the ward’s age-profile is equivalent to one that models a fixed-weights weighted combination of the age-specific per capita costs by a formula like (22) that does not involve the age-profile at all! Appendix C shows why such first-order agreement of coefficients may be expected and why the observed discrepancy in RESET should not surprise us. Such agreement is almost a mathematical tautology — it does not mean that such standardizations are bringing us any closer to the goal of measuring true need.

The AREA team followed its remit by “investigating alternative methodologies to replace the age-related components” of the York team formula i.e. alternatives to the two-step procedure. But the only alternative on which they reported (Table 5.7.1) was to fit separate regressions in 19 age-bands to $C_1$ (without maternity) for a single choice of explanatory variables (the same in all age-bands) — a choice, moreover, that had been made in a quite separate analysis of an age-standardized utilization. Finding that “many of the age group models are misspecified”, the team was unable to recommend such
an alternative. This is a particularly weak argument for preferring the two-step procedure rather than searching for a better one-step procedure.

The AREA team also followed its remit by investigating the possibility of replacing $I_1$ and $I_2$ by a single index derived from an aggregate model for acute and maternity and mental health, but found the results for an aggregated model were “less promising and the overall explanatory power...reduced once insignificant variables were dropped” (AREA, §8.4.4). For use rather than derivation, however, DH has adopted the single index $I$ given by (7). We could write $C_{age} I$ as

$$
(C_{1, age} + C_{2, age})(0.8553I_1^* + 0.1447I_2^*)
$$

where $C_{1, age} = \sum_a n_a c_{1aN}$, $C_{2, age} = \sum_a n_a c_{2aN}$ and $c_{1aN}$, $c_{2aN}$ are the acute and maternity and mental health components of national per capita utilization, respectively. The inconsistency here in applying the separately derived indices to the combined utilization would be removed in the alternative formula

$$
C_{1, age} I_1^* + C_{2, age} I_2^*
$$

which could differ appreciably from (16). The difference between (16) and (17) as a fraction of (17) can be shown to be $(F_1 - 0.8553)(I_1^* - I_2^*)/(0.8553I_1^* + 0.1447I_2^*)$ where $F_1 = C_{1, age}/(C_{1, age} + C_{2, age})$. There is yet another pressing alternative to (16) or (17). Since the formula for $I$ was derived by fitting ward costs, the alternative is simply to add the ward formula costs $(\Sigma_a n_{aw} c_{aN})I_w$ for the wards in a PCT, where $n_{aw}$ is the ward population in age-band $a$ of ward $w$ and $I_w$ is now the inequality index for just the ward. It is easily established that this would add $\Sigma_a n_a c_{aN} p_{aw} (I_w - I)$ in which $p_{aw} = n_{aw}/n_a$ is the proportion of the PCT population of age-band $a$ that are in ward $w$. There may well be age-bands in some PCTs for which $d_a$ would have made a significant positive difference to the allocation.

Would DH be able to defend the difference to any PCT that would have benefitted if either of these alternatives had been adopted?

### 5.4 The “ins and outs” of supply

The methods used by both teams for their final estimations do not reflect the complexity of the arguments that led up to them, in which both teams employed the subtle concept of endogeneity without any precise and general definition. Appendix E shows how the concept can be best explained via the complementary concept of exogeneity.

The York team prepared the ground for its multilevel model ((25) in Appendix D) thus:

“In developing a resource allocation formula, we wish to correct for variations in supply between areas. Effectively, this means assuming that all supply in an area is at some national average level appropriate to the level of needs found in an area. In calculating a measure of relative need, therefore, the variation in utilization due to variation in supply variables should be considered only to the extent that supply reflects variations in legitimate need for health care. The requirement is to develop a measure of “normative utilization”: the utilization that would obtain in an area if the response to its needs was at the national average level.” (York, §5.17)

“The analytic task is to find that part of the supply effect which is attributable to factors unrelated to needs indicators, and to remove that part of the supply effect from the model.” (York, §5.18)

“This is achieved by undertaking an ordinary least squares [OLS] regression of utilization on the needs variables identified... The coefficient on each needs variable will then capture its direct and indirect effect on utilization.” (York, §5.21)
The general idea in these quotations is not affected by the York team’s ultimate preference for multilevel modelling rather than any form of OLS. The team accepted that Appendix D’s (25) could have been written with an additional linear term in supply variables as, in a sense, a more truthful representation of how the logarithm of (standardized) utilization might be modelled. But the omission of that term does not make the smaller model necessarily untruthful — if the additional term were normally distributed with expectation a linear combination of the need variables in (25). The coefficients in that expectation would be the biases in OLS that the endogeneity of a missing variable is conventionally abused for — but for the York team they represent the adjustments to \( \beta_{\text{need}} \) that might then represent the “normative utilization” of the second quotation.

However, we see the York team’s logic as faltering in its neglect of any analogous adjustment for the endogeneity in the DHA levels of utilization — whose correlation with the need variables would be likely to express “variations in legitimate need”. The neglect comes from estimating the need coefficients from the within-DHA variation.

The AREA team rejected the York team’s logic:

“Supply variables may be correlated with past values of the need variables because of the operation of explicit resource allocation formulae. ...Hence the error term will be correlated with the supply variables, leading to biased estimates of the coefficients on the supply variables in the utilization equation.” (AREA, §3.2.1)

and then explored ways in which the problem of supply might be resolved, before adopting an approach that simply uses the HA terms in (14) and the \( \{x_q\} \) in (15) to countenance least-squares as a valid estimation technique. For (14), there was a brief justification (AREA, §3.2.4):

“...a better method of allowing for endogeneity is to incorporate HA fixed effects into the utilisation model. In brief the argument is that past resource allocation procedures have allocated resources to HAs based on HA level need variables. Hence when HA dummies are included in the utilisation equations in addition to the potential need variables and supply variables, they pick up unobserved need and other factors affecting utilisation and so produce coefficients on the other variables which are not contaminated by endogeneity.”

The AREA team argued that “dropping the supply variables at the final stage is not appropriate since the relationship between supply and need is less likely to yield useful information about need than the relationship between utilization and need with supply held constant.” As already noted, the supply variables are retained for the estimation of the need coefficients in (14) and (15) and then, to determine the additive constants in \( I_1 \) or \( I_2 \), are replaced by their national averages. In this, the AREA team did not concede that there may be expression of response to legitimate need in \( \{\alpha_h\} \) that cannot be accounted for in the need variables selected for the allocation formula — which may now be punishing enterprising PCTs.

These are murky waters and we will not try to adjudicate between the two approaches — but focus instead on questions that challenge the framework in which this particular disagreement has arisen.

### 5.5 Unmet need and “wrong” signs

An estimated coefficient has a “wrong” sign when parties agree that it goes against or even violates expectation. The problem arose in the York team’s Table 6.1 for the variable BLACK* (Proportion not in black ethnic groups). This “wrong” sign was (exceptionally) positive because the variable had been defined negatively to avoid logarithms of zero. It suggested that
“areas with higher proportions of black residents exhibit lower utilization than expected. The York team and its advisors interpreted this finding as reflecting supply rather than need, perhaps because wards with large ethnic minority populations tend to be close to acute hospitals. The BLACK variable was therefore not considered as a needs variable.” (York, §6.3)

No link was made with the concept of “unmet need” and the offending variable was not introduced again by the York team.

The AREA team was, however, instructed to propose a methodology to adjust for unmet need. The historian (Department of Health, 1989a) did not record how the team’s advisors thought one could tell there is “unmet” need when the data are only indirectly related to need and one sets out to determine that relationship empirically. Nevertheless, the team was by its own account (§3.4) well-prepared to handle “wrong” signs. Subtle arguments for relating such signs to “unmet” need are widely distributed throughout the report, concentrating on the ethnic proportion variable $x_{10}$. The arguments were used to justify re-classifying $x_{10}$ as a silent (“other”) variable and thereby make an implicit positive adjustment of the final formula in favour of PCTs with higher ethnic proportions.

For AREA, the wrong signs emerged in an ultimately rejected model — the “basic model” for $I_1$ of AREA’s Table 8.4.4. For the basic model, both the ethnic proportion $x_{10}$ and the unemployment index $x_{11}$ had seriously negative coefficients with lower fitted standardized utilization for higher values of $x_{10}$ and $x_{11}$. Concentrating on $x_{10}$, three quite different non-exhaustive explanations are:

(i) We can believe the model and conclude that ethnic minorities are healthier.
(ii) We can disbelieve the model and conclude that ethnic minorities have unmet need which has not already been accounted for in the allowance that the formula makes for supply or accessibility of services (as the supply variables in the formula are designed to do).
(iii) We can believe that there is no unmet need but that an influential morbidity measure has been omitted from the formula — a measure that is negatively correlated with $x_{10}$ which acts as a proxy for it.

To test hypothesis (iii), AREA (§4) used interview data from individuals in the household samples of the Health Survey for England. AREA reports has a variety of regression analyses of the (largely self-reported) health status on sets of explanatory variables that included ethnic status. AREA declined (§2.10) to use the available objective data on the grounds of substantial sample reduction and associated bias, while acknowledging that self-reporting quality may vary with ethnic group and socio-economic status. A footnote to each table of results explains that the regressions were also done with dummy variables for age and sex (known to be the major influences on health status) but the results were not indicated.

From all that, the indices $\{x_{39}, \ldots, x_{46}\}$ for individual morbidity were constructed as logistic functions of linear combinations of socio-economic variables including ethnic status, standardized for age and sex (Appendix 7, Department of Health, 2003b). In the words of the report itself:

“We argued in section 3.4 that the effect of the introduction of additional morbidity indices on the coefficients of socio-economic variables which appear to have the “wrong” sign would provide evidence about whether the negative coefficients on variables like the ethnic minority proportion and employment deprivation indicate unmet need. There are two possible explanations for the negative coefficients...The first is that there is unmet need: individuals in ethnic minorities or who are employment-deprived receive less care given their morbidity. The second is that individuals in these groups are treated in the same way as the general population and the general morbidity variables, like SMR or SIR, included in these models do not fully reflect their better health. To distinguish between these explanations we require additional measures of morbidity. If the coefficients on the ethnic

16
minority and employment deprivation variables remain negative...this strengthens
the
evidence for unmet need.” (AREA, §4.1)

In the light of the findings of continued “wrong signs” for $x_{10}$ and $x_{11}$ in (11) and (12) when morbidity
indices are included, AREA rejects hypotheses (i) and (iii) and concludes that there is evidence of
unmet need. The case here is vulnerable in that it assumes that the formula as a whole, in which the
contentious negative coefficients play a role, is a fairly close approximation to some underlying reality.

Both $x_{10}$ and $x_{11}$ are strongly age-related. Given the ambient low $R^2$-values, should the possibility be
excluded that the negative coefficients are simply artefacts, stemming from a reluctance of the
utilization-influential age variable to comply with the Procrustean bed intended to accommodate it by
indirect age-standardization of utilization coupled to an interaction-free additive linear model?
Moreover, if the coefficients of $x_{10}$ and $x_{11}$ are wrong for the measurement of need (appropriate
utilization) then the idea, contained in hypothesis (ii), that the supply variables can fully account for
unmet need is undermined — and the coefficients of $x_1$ to $x_9$ may likewise be unsuitable for measuring
true need.

5.6 Replicability and robustness

Although the financial consequences of the repeated formula changes over the last three decades will
have been closely monitored by those directly affected (whether they were gainers or losers), they have
received little attention in the statistical thinking provoked by those changes. In particular, the literature
shows little interest in the scientific question of replicability. This question is typically diluted as one of
circumscribed “robustness”, interpreted as a test of whether the formula that emerges from some
particular approach does not change drastically when subjected to variations within the framework of
that approach — such as the addition of variables, their transformation to a logarithmic scale or some
other change that leaves much of the framework unaffected.
Each research team that is sequentially engaged is somehow able to develop its own confidence in the
changes that it makes to the immediately preceding formula. An onlooker who sees that this confidence
is usually dependent on subjective considerations may think that what is lacking is a robustness test of
the sort that would split the data-base randomly and give the two halves to separate teams for
completely independent constructions of a prediction formula. Keynes (1940) put the challenge of such
replicability in a particularly cogent fashion:

“It will be remembered that the seventy translators of the Septuagint were shut up in
seventy separate rooms with the Hebrew text and brought out with them, when they
emerged, seventy identical translations. Would the same miracle be vouchsafed if seventy
multiple correlators were shut up with the same statistical material?”

But even the Keynesian challenge may not be very useful unless the criterion for a successful
replication is clearly specified. The York team almost suggests that we do not even need a criterion:

“Because of the high level of correlation between many of the variables, the variables
selected were to some extent arbitrary...and not too much meaning should be attached to the
actual variables selected...the variables selected are proxies for unmeasurable social factors
which could be equally successfully captured by other variables, so the precise variable
selected is less important than the social factor it partially captures.” (York, §8.4.1)

So the York team might not object to the replacement of its formula (4) by the completely different
AREA team’s counterpart $I_1$ (if reworked to exclude maternity). Arguably the rival formula might have
the same “explanatory power” and “capture” the “same needs characteristics” — but how we would ever know that is not explained. For individual PCTs, the question of replicability is a very practical one — would the allocation of funds be substantially different if one formula were chosen rather than the other? The AREA team did not answer that question. Instead it explored a limited form of replicability that bypasses the issue of selection of variables — more a re-application than a replication.

The AREA team replicated the York team’s OLS fitting of the logarithms of the five explanatory variables in (4), a fitting that ignored the 186 District Health Authorities but had dummies for the 14 Regional Health Authorities. That fitting had changed (4) into

\[
(SMR_{75})^{0.130} (x_1')^{0.202} (x_2')^{0.132} (x_3')^{0.043} (x_4')^{0.050}.
\]

For 2000-1 data — and presumably putting sex back into the indirect standardisation of the fitted utilization and excluding maternity to match what the York team had done — the replication gave:

\[
(SMR_{75})^{0.210} (x_1')^{0.491} (x_2')^{-0.102} (x_3')^{0.045} (x_4')^{-0.116}.
\]

There are striking disparities between the earlier and the later exponents — the positive exponent of \(x_4'\) in (18) with a \(t\)-value of 5 becomes negative in (19) with a \(t\)-value of 13! For the least-squares fitting of (19), \(R^2 = 0.55\), RESET = 2.7 with \(P = 0.05\), and an unspecified homoscedasticity test statistic had \(P = 0.0002\). The AREA team claimed that these \(P\)-values and the disparities in the exponents suggest that the York team formulae “may lead to allocations failing to reflect current levels of need”. But one explanation for the difference may have been overlooked. The York team’s OLS used dummy variables for the 14 then extant RHAs but the AREA team “did not include fixed effects for Health Authorities” in their re-application of the York team model. If the term “Health Authorities” here embraces the former RHAs, it is to be expected that the exponents should differ to the extent that differences between the aggregate (logged) RHA utilizations are and have been residually correlated with the average of the (logged) explanatory variables for their constituent wards.

6 Conclusion

Current and past allocations of HCHS funds have been based on the use of proxies for need such as age, SMR, proportion of elderly living alone, and level of unemployment. However, the York and AREA reports provide no objective validation that their methods can measure actual need for health care. More specifically, there is no justification of the claim that consideration of supply effects can convert a formula to predict “standardized utilization” into one to predict appropriate standardized utilization. The York report (§2.21) made only passing reference to the possibility of direct measurement, and the AREA team ruled it out completely in the general framework that we have already quoted in Section 5. It seems that the York team thought direct measurement might provide some justification for the proxy approach (York, §7.18). But it is at least conceivable that a harder push in that direction would allow the proxy variable approach to become a historical curiosity.

Achieving that goal might have to await change in the way that government departments such as DH administer the first-stage contractual arrangements of projects that call for external expertise. The case for such a change and an inexpensive proposal for effecting it — relaxation of the restrictive contractual arrangements in the early stages of any enterprise that now inhibit the free expression of honest opinion — are outlined in Stone (2005). The first stage here would be to motivate a large number of independently situated individuals and small groups — in PCTs and universities — to come up with their best thoughts about the problem. That would include ideas about how to pursue things beyond the first stage!

Should we be deterred by the potentially high cost of finding, developing and testing a nationally acceptable way of directly measuring the aggregate health needs of individual PCTs? The acquisition of
the necessary data — and the hard thinking and value judgements needed to guide and exploit it — would not come easily. But just one percent of the PCT budget for one year would give £450M. Perhaps we should ask ourselves how confident we can be that the current waste from the continued use of proxy formulae is not well in excess of that one percent.

Acknowledgements

We thank DH’s Stuart Perry for guiding us to the WCF document, Gwyn Bevan for broader references and advice, John Hacking and referees for their critical interest in an earlier version, Professor James Duguid for his warm encouragement, and Rex Galbraith for assisting with the graphics.

Appendix A: Indices of mortality and the measurement of need

Whether restricted to under-75s or to particular causes of death, SMRs have strongly influenced funding since 1976. The indirectly standardized SMR75 and the directly standardized CMF65 are doing this in the WCF indices. An SMR is simply the ratio of the total number of deaths in the population of interest to the total number there would have been if its death-rates in different age-bands had been those of some standardizing population. The associated CMF is simply the corrective reciprocal of the SMR that exchanges the roles of population of interest and standardizing population. Fox (1978) has pointed out that it is CMF not SMR that is “strictly the correct measure to use if the mortality rates for [subgroups] are to be compared among themselves as well as with those for the standard population”. An SMR uses

“...the age distribution of the population studied...In cases where this differs from either the standard population age distribution or the age distributions of other populations with which comparisons are made, the SMR may be misleading.”

If there are problems with these indices in scientific study, how do they square up as measures of need? Any claim that SMRs iron out differences caused by differences in age structure is only partially true for the scientific understanding of mortality and not at all true when it is a question of measuring need. Practical allocation should be based on the relevant multidimensional data and, as far as possible, on openly acknowledged and broadly agreed value judgements (the kind of thing that the National Institute for Clinical Excellence deals with). We cannot rely on half-understood indices to do our decision-thinking for us.

The improprieties of Fig.1 go well beyond such gentle considerations:

(i) implying a currently “widening mortality gap” when the data are more than 10 years old,

(ii) ignoring the massive change in the proportion of the population in the Unskilled category between the 1930s and the 1990s,

(iii) ignoring the changes in the causes of mortality over that period with fewer deaths from class-blind infectious diseases and more from diseases related to life styles and

(iv) using an ancient standardizing population. (Why is it thought necessary to go via a third population — and stir up the ditty-line “That night we went to Birmingham by way of Beachy Head”? A third party can give a ratio of SMRs different from unity even when the two categories being compared have identical age-specific death-rates. The categories ought to be directly comparable by some tailored index and, if interest lies in how that index changes with time, another graph would do the job.)
Appendix B: RESET and White’s standard errors

Reset Equation Specification Error Test (RESET). The null hypothesis $H_0$ is $y = X\beta + e$ where $X$ is $n \times p$ of rank $p$ and $e$ is $N(0, \sigma^2 I)$. Consider an alternative hypothesis $H_1$ with $y = X\beta + A\gamma + e$ where $A$ is $n \times q$ of rank $q$. If $RSS_0$ and $RSS_1$ are the sums of squared residuals for OLS fitting of $H_0$ and $H_1$, then $F = [(RSS_0 - RSS_1)/(n-p)]/[RSS_1/(n-p-q)]$ has an $F_{(q,n-p-q)}$ distribution under $H_0$. Both $RSS_0$ and $RSS_1$ are functions of the OLS residuals $r_1, ..., r_n$, which are distributed independently of the OLS estimator $\hat{\beta}$ under $H_0$. It follows that the F-distribution also applies when $RSS(\hat{\beta})$ is substituted for $RSS_1$, where $RSS(\hat{\beta})$ is the RSS for OLS fitting of the formula $X\beta + A(\hat{\beta})\gamma$ where $A$ is now allowed to depend on $\hat{\beta}$. The RESET test statistic simply uses this $F$-statistic to test $H_0$ with $A(\hat{\beta}) = (\hat{y}^2, \hat{y}^3, \hat{y}^4)$ where $\hat{y} = X\hat{\beta}$ and $\hat{y}^T = (\hat{y}_1, ..., \hat{y}_n)^T$.

White’s standard errors. If $H_0$ is heteroscedastic (i.e. if the variances of the $e_i$ are not all equal) the standard errors of $\hat{\beta}_1, \hat{\beta}_2, ...$ are the square roots of the diagonal elements of

$$
(X^T X)^{-1} X^T \text{diag}(\sigma_1^2, ..., \sigma_n^2) X(X^T X)^{-1}
$$

(20)

The standard errors are estimated by replacing $\sigma_i^2$ by $r_i^2$ in (20), $i = 1, ..., n$. Since $E(r_i^2) = (1 - P_{ii})^2 \sigma_i^2$ where $P = X(X^T X)^{-1} X$, the estimates may be biased. Since trace $P = p$, the average of the $P_{ii}$ is $p/n$ and, if $p/n \ll 1$, the biases may be small. By embedding this formulation in a sequence of increasingly informative cases, White (1980) was able to establish asymptotic consistency.

Appendix C: Comparison of indirect and direct standardization

As already noted, the AREA team standardized only by age — both indirectly as in $Y_1$ and directly as in $Y_2$. The indirect standardization modelling is equivalent to fitting the crude per capita utilization

$$
U_1 = C_{1,age} \times \beta^T x
$$

(21)

where $\beta^T x$ is the linear formula that will deliver $I_1$. In the mixed product-linear formula (21), the ward’s age-profile is an explanatory factor of predetermined influence in $C_{1,age}$. The fitting of $Y_2$ that delivers $I_2$ is equivalent to using the formula

$$
U_2 = \omega_{2N} \times \beta^T x
$$

(22)

where $\omega_{2N}$ is a constant (the national per capita cost) and $U_2 = def \sum a p_a N c_{2a}$ is a combination of the ward’s age-specific per capita costs in which the weights are determined by the national age-profile. Note that the ward’s age-profile does not enter (22).

For a general case, let $Y_{ind}$ and $Y_{dir}$ be the indirectly and directly standardized dependent variables for some linear model with wards as units. Then

$$
Y_{ind} = \sum a p_a c_a / \sum a p_a c_{an} = \sum a f_a u_a v_a / \sum a f_a v_a
$$

$$
Y_{dir} = \sum a p_a N c_a / \sum a p_a N c_{an} = \sum a f_a u_a
$$

where $f_a = p_a N c_{an} / \sum a p_a N c_{an}$, $u_a = c_a / c_{an}$, $v_a = p_a / p_{aN}$. Invoking the $X^T y$ component of the classical formula $(X^T X)^{-1} X^T y$, we see that the difference between the two estimates of any coefficient is a linear combination of the differences

$$
\Sigma_a x(Y_{ind} - Y_{dir}) = \Sigma_a \{ x(\sum a f_a u_a v_a - \sum a f_a u_a \sum a f_a v_a) / \sum a f_a v_a \}
$$

(23)
where \( w \) denotes wards and \( x \) represents any one of the explanatory variables in the common model for \( Y_{\text{ind}} \) and \( Y_{\text{dir}} \). For models of type (14) with a “constant” \( \alpha_h \) for each health authority \( h \) and for any regression coefficient \( \alpha_h \), the right-hand side of (23) can be written, with “cov” and “ave” defined with respect to the proportions \( \{ f_a \} \), as

\[
\Sigma_h \Sigma_{\text{weh}} x \, \text{cov}(u, v) / \text{ave}(v).
\]  

(24)

When this \( x \) has been adjusted to have mean zero over the wards of each health authority \( h \), as it can be without affecting the estimation of the regression coefficients, it is likely that (24) will be small compared with the combination of quantities \( \Sigma_w x Y_{\text{dir}} = \Sigma_w x \, \text{ave}(u) \) that determine the regression coefficient for direct standardization. Note that (24) is the sum of covariances (each calculated over the wards in one of the health authorities) between the deviation of \( x \) from its ward average and the ratio \( r = \text{cov}(u, v) / \text{ave}(v) \). So we may expect the indirect and direct coefficient to be not very different. For \( \alpha_h \) however, \( x \) is a dummy variable that cannot now be adjusted to have mean zero over the wards in each health authority. The difference between the two estimates of \( \alpha_h \) is the average of the values of \( r \) in \( h \). In this, any particularity associated with HA \( h \) is not diluted by aggregation over all health authorities, and any sizeable difference in the estimates may affect the way in which RESET uses its quadratic, cubic and quartic powers to pick up such differences (with little effect on the within-HA estimation of the regression coefficients).

Appendix D: The models for within-authority variation

For its final estimation of the factor (4), the York team’s model was

\[
Y = \alpha_{\text{DHA}} + (\beta_{\text{DHA}}^{\text{need}})^T x_{\text{need}} + \epsilon
\]  

(25)

for the logarithm \( Y \) of indirectly standardized utilization, where \( x_{\text{need}} \) is the vector of logged explanatory “needs” variables. In (25), the error \( \epsilon \) is taken to be random \( N(0, \sigma^2) \), and \( (\alpha_{\text{DHA}}, \beta_{\text{DHA}}^{\text{need}}) \) is assumed to be \( N((\alpha, \beta^{\text{mean}}), \Sigma) \). All the unknown parameters are jointly estimated by maximum likelihood. The components of \( \beta^{\text{need}} \) are the recommended need coefficients — an expectation over the infinite population of DHAs of which England has conceptually a random sample. They are all that is required for the product form of (4) — the intercepts \( \{ \alpha_{\text{DHA}} \} \) are on the logged utilization scale and represent multiplicative DHA supply factors that can, it is assumed, be ignored in prescribing the influence of need factors on standardized utilization in the recommended allocation formula. The York team’s model for its least-squares fitting of (4) was

\[
Y = \alpha_{\text{RHA}} + (\beta_{\text{DHA}}^{\text{need}})^T x_{\text{need}} + \epsilon
\]  

(26)

in which the needs coefficients \( \beta^{\text{need}} \) were the same for every DHA and the intercept was taken to be the same for DHAs in the same RHA.

Omitting here the complication of “other” variables, the model for AREA’s (14) for HA \( h \) is

\[
Y_1 = \alpha_h + (\beta^{\text{need}})^T x_{\text{need}} + (\beta^{\text{supply}})^T x_{\text{supply}} + \epsilon.
\]  

(27)

In (27) the variables have not been “logged” and the intercepts \( \{ \alpha_h \} \) therefore represent additive HA supply effects. When the corresponding HA dummies are treated as supply variables, their influence becomes part of the national averaging of supply variables that AREA recommended in order to create a level playing field.
Appendix E: Endogeneity

The York report (§3.17-3.19) recognised that its problem is really a time-series one ("supply and utilization are jointly determined...both affected by the same processes, either at the same time or with lags") but gets its recommended formulae by cross-sectional analysis, as does the AREA study. In a time-series context, Engle et al (1983) defined "weak exogeneity" of a variable with respect to inference about a parameter of interest by its role in

\[ f(Y, x^{\text{supply}}|x^{\text{need}}, \theta) f(x^{\text{need}}|\nu) \]

where \( \nu \) is an (uninteresting) nuisance parameter in the particular parametrisation — “cut” off from \( \theta \) (the parameter of interest because it determines the relationship between \( Y \) and \( (x^{\text{supply}}, x^{\text{need}}) \)). In (28), the marginal distribution of \( x^{\text{need}} \) is informative only about \( \nu \) and it is the conditional distribution of the rest of the data, \( Y \) and \( x^{\text{supply}} \), that is informative about \( \theta \). The components of \( x^{\text{need}} \) can therefore be regarded as predetermined and fixed exogenously. The first factor of (28) can always be further factorized into

\[ f(Y|x^{\text{need}}, x^{\text{supply}}, \theta) f(x^{\text{supply}}|x^{\text{need}}, \theta). \]

If (28) holds, we can say that \( x^{\text{supply}} \) is not exogenous with respect to \( (\theta, \nu) \). But it could still, along with \( x^{\text{need}} \), be exogenous with respect to a different cut \( (\theta_1, (\nu_1, \nu)) \) if the dependencies in (29) were reducible to

\[ f(Y|x^{\text{need}}, x^{\text{supply}}, \theta_1) f(x^{\text{supply}}|x^{\text{need}}, \nu_1) \]

where \( \nu_1 \) is another (uninteresting) nuisance parameter. Whether or not there is endogeneity can be a matter of definition of the parameter of interest.

However, what York had in mind is expressible by putting (29) in the form

\[ f(Y|x^{\text{need}}, x^{\text{supply}}, \theta_1) f(x^{\text{supply}}|x^{\text{need}}, \theta_2). \]

Dropping \( x^{\text{supply}} \) from the first term of (31) for the specialization (25) is an endogeneity-evading ruse that allows \( \beta^{\text{need}} \) to accommodate the coefficient of the missing supply contribution to (25). This requires the strong assumption that the missing supply contribution has an expectation in \( f(x^{\text{supply}}|x^{\text{need}}, \theta_2) \) that is linear in \( x^{\text{need}} \), and more than that.

The AREA treatment of endogeneity also rests on strong unvalidated assumptions. AREA uses the first term of (29) for the specialization (27) with \( Y = Y_1 \), assuming that any supply factor not explicitly covered by \( (\beta^{\text{supply}})^T x^{\text{supply}} \) would contribute the same additive amount for all wards in HA \( h \) and so be absorbable into the intercept \( \alpha_h \).

References

Department of Health (2003a) *Tackling Health Inequalities: A Programme for Action*.
www.doh.gov.uk/healthinequalities/programmeforaction