Rejecting an empirical ‘person-based’ formula for funding CCGs in favour of the farming analogue of one-year-ahead extrapolation

by

Mervyn Stone

November 2012

What need the bridge much longer than the flood?
The fairest grant is the necessity.
Look what will serve is fit
Much Ado About Nothing, Act 1, Scene 1

Contents

1. Doctors in the waiting room
2. Predictand (‘dependent variable’)
3. Variables for the recommended formula
4. Construction of the formula
5. Performance assessment and its empirical basis
6. Structural assumptions and the ‘curse of dimensionality’
7. ‘Small would be beautiful’ for covariates
8. Individual level adjustments of all sorts and sizes
9. Discrepancies between the recommended formula and its provenance
10. Elevating the formula to practice level
11. What’s in the freezer?
12. Additional difficulties of the PBRA approach
13. Back to the drawing board
14. Bloggers
Footnote
References

1. Doctors in the waiting room

A potentially important paper[1] appeared in the 2011 British Medical Journal (BMJ). It is important because the Department of Health DoH) may take it as later justification for some version of a person-based resource allocation (PBRA) funding formula for GP-practice commissioning within individual Clinical Commissioning Groups (CCGs) for 2013/14 and beyond. The allocations to practices are due to be announced by the New Year and are likely to provoke a lively response by the private sector of NHS that GPs occupy.

The research was funded by DoH. Three of the authors were members of DoH’s own Advisory Committee on Resource Allocation during the period of the analysis fully reported in reference [5]—but this updating paper assures us that they were ‘independent of the sponsoring body’ DoH. My own ‘conflict of interest’ is that years ago I was asked (and probably paid) by DoH to comment on and rank six tenders for the research contract for this line of research.

The tender of the PBRA research team—and the one proposed by the even-more independent health economist, Professor Sheena Asthana—were the two that held most promise. I had few reservations about lending strong support to the Asthana tender, which promised to break new ground, but there was a question as to whether the other would be able to make a clean break with the sad history[2,3,6] of the national PCT funding formula, given that four of those responsible for (and even happy to defend) that history were members the PBRA team and appear to have been responsible for the 'nitty-gritty' of the statistical analysis (experts who deserve to be recognized for their contributions as 'old hands', rather than as 'the usual suspects' of a DoH source.) The lead author of the present summary account[1] of novel and part-successful research is Dr Jennifer Dixon, Director of the Nuffield Trust, who is not a statistician and claims only to have contributed 'ideas for analysis' as project leader.

The novel and successful part is that the recommended formula for allocation to a practice was constructed at the level of millions of individuals—and is largely determined by the age/sex and morbidity profile of the practice register, variables that facilitate any value judgement of whether or not the formula 'looks right' . The failing part of the research is its inability to dispense with an econometric approach based on inherited and unjustifiable assumptions.

The BMJ paper is the focus of the following critique of the PBRA approach, which tries to lighten some of the technicalities—just as Drs Dixon and Bardsley have managed to do in an easily readable Nuffield Trust research summary[1a].

2. The predictand ('dependent variable')

The paper has just one paragraph on this component of the PBRA study. It is reproduced here since, as will become clear, the work-load and ease of acquisition of the data needed to determine the predicted are major factors in considering extrapolation as a feasible alternative to prediction:

> *Constructing the dependent variable—costs per individual in 2007-8*
>
> *The costs for 2007-2008 of NHS inpatient and outpatient care (excluding costs for services outlined above) for each unique individual registered with a general practice on 1 April 2007 was estimated by costing every inpatient admission or outpatient attendance that they had incurred. Because the interest was in modelling the costs of commissioning care, we sought wherever possible to use the prices paid by commissioners. We therefore used national tariff prices (Health Resource Group version 3.5) or, where this was absent, the average national reference cost to cost an individual's inpatient and outpatient activity. We could cost 96% of outpatient attendances in 2007-8 with the national tariff, and the rest were costed using national reference costs or speciality average cost.*[1]

This novel study started with individual cost but the objective was the successful prediction of practice cost. There is no mention in this paragraph of any difficulty the PBRA team may have had in assembling the data for their subsequent calculation of actual practice cost (the sum of individual costs over the practice list)—comparable to the only partly resolved problems of assembling data for the prediction of practice cost. An alternative to prediction from troublesome data and with a questionable formula would be the extrapolation of a last-year total of justifiable and monitored costs of individuals in the practice list.

3. The variables for the recommended formula

The person-based resource allocation (PBRA) formula in Table 2[1] was conceived and devised (for the then PCTs) as if it were an operational formula for the notional funding of an individual. It was a prediction of cost for 2007-08 based on data about that individual for the previous two years, for 321 explanatory variables selected from a very much larger number. It is not often that a model that fits 321 parameters is described as 'parsimonious', but statisticians will be able to relax when they see that the 'number of observations' for the least-squares fit was more than 5 million.

There are 160 factors (159 'need' factors) specific to an individual. The first six 'need' covariates are the so-called 'attributed need' variables, either drawn from the small ONS-defined area in which the individual lives or that depend on data for the individual's GP practice. The three supply covariates are the variables for specific NHS services to the practice and the locality in 2005-07. Covariates are variables selected because they have been found to have statistically significant 'covariation' (correlation) with the predictand:

Factors
f1 ∼ Age/sex factor (38 categories: 19 five-year bands and 2 sexes!)
f2 ∼ PCT factor (PCT identity is an unlisted supply factor in Table 2[1])
f3-f159 ∼ Diagnoses (157 whether-or-not factors for ICD-10 categories—157 according to Table 2[1] but 150 in bmj.com's 'web extra' Table B)
f160 ∼ Private/NHS (whether-or-not privately-funded inpatient treatment)
Need covariates
n1 ∼ Area proportion of persons in social rented housing
n2 ∼ Area proportion of persons claiming any disability allowance
n3 ∼ Area proportion of persons aged 16-74 with no educational qualifications
n4 ∼ Area proportion of 'mature city professionals'
n5 ∼ Area proportion of students
n6 ∼ Area asthma prevalence
Supply covariates
s1 ∼ A numerical measure of the quality [quantity?] of primary and secondary stroke care divided by [surprise, surprise!] the *weighted population*[2,4]
s2 ∼ Area measure of access to magnetic resonance imaging
s3 ∼ Catchment population of the hospital trust with the most inpatient admissions from the individual's GP-practice.

Table 2[1] classifies the dummy (whether-or-not) variable f160 as a need variable rather than a factor. Factors do not have units of measurement, covariates do, but the table does not give them. The order of listing here has exchanged f3 and f4-f159 to simplify later explanation. The 160 factors required over 300 dummy variables, whose values made up the coded part of the record for each of 5 million individuals; the records were completed by the straightforwardly numerical values of the 9 covariates.

4. Construction of the formula

The approach to statistical modelling of this paper is one that DoH's Advisory Committee for Resource Allocation (ACRA) has countenanced for years. What ACRA has not acknowledged is that a formula does not become evidence-based just because, in its trial-and-error selection, every conceivable and possibly influential factor is somehow 'taken into account'—especially when the taking-into-account is largely unprincipled. That said, the PBRA approach incorporates sensible statistical method when it uses random selection of individuals for analysis and then validation of a formula by a variant of cross-validation. (The application of the latter technique is somewhat flawed at practice level since it appears that 10% of individuals were in both the construction and validation samples, thereby probably inflating R^2-values a little.)

The way the above factors and covariates contribute to the initial fitting of the Table 2^1 formula can be expressed in different ways. The one with most symmetry starts the expression with the average of the 5m values of the predictand (the estimated individual cost) in the construction sample. This overall average is then subject to a *jointly-determined* plus-or-minus adjustment (adj.) for each of the factors and covariates. So the cost-prediction formula for an individual that first comes out of the formula-fitting procedure ('single-stage ordinary least-squares regression') can be written as the sum of 15 terms that exhibits the pragmatic assumption of *additivity*:

Average of 5m costs + *Age/sex adj.* + *PCT adj.* + . . . + *Catchment adj.*

In this representation, the individual adjustments for factors other than diagnoses (the first dot!) have a common structure, in which the paper's invocation of empiricism/pragmatism excuses the absence of theoretical justification. Taking *Age/sex adj.* for example, the least-squares fitting of the whole formula gives it an optimizing weight for each age/sex category. In the fitting, there is no restriction on the assigned number and each individual gets the value for his/her category. The 5m values are averaged and an individual's age adjustment is simply the deviation of the value from the average. The adjustment for diagnoses differs from this in that each of the 157 diagnoses gets its own assigned weight (provided all diagnoses are represented in the 5m construction sample)—and the adjustment is the deviation of a diagnosis score (the sum of the weights for the individual's diagnoses) from its average over the 5m sample. Note that the diagnosis score will be zero for the majority of individuals, who will have had the good fortune to escape any diagnosis.

The covariate adjustment terms are taken, again pragmatically, to be the *product* of a least-squares assigned coefficient and the deviation of the individual's covariate value from its average in the construction sample. Take, for example, the first need covariate n1—*Area proportion of people in social rented housing* (*SRH* for short). The adjustment is the product of a coefficient (the 0.35 in Table 2^1) and the plus-or-minus deviation of the individual's *SRH* from the average of *SRH* in the 5m construction sample:

SRH adj. $=$ *Coefficient from fitting the extended model* × *Deviation*

If the definition of n1 is that of the 2009 PBRA Team study[5], n1 is indeed a proportion with a national 2001 Census average of 0.18, standard deviation of 0.15 and range of next-to-zero to next-to-unity. An individual living in an area whose SHR is half-way up the top 5% (2.5% is 2 standard deviations above average) would get a contribution of $0.35 \times 2 \times 0.18 = 0.13$ to his/her model 5 formula (13p if the formula is in pounds).

The recommended PBRA formula for cost prediction is equivalent to a *truncation* of the above-displayed full expression—a truncation that 'freezes' at zero the adjustments of the three supply covariates together with that of the PCT factor, by replacing these variables by their national average for every individual. The truncation/freezing is made on the grounds that supply variables and supply factors:

'*should not influence the target allocations for practices ... the point of the formula is to distribute available resources on the basis of health needs as far as possible, not influenced by any variations in the supply of services that the individual has access to'.*

The previous PBRA Team report[5] made the case for freezing with:

By freezing supply variables at the same level for all individuals, differences in local supply conditions are removed from the estimated expenditure which then varies across individuals only with variables deemed to reflect need.

The expectation here is that the truncation would make a level playing field. The underlying assumption is that what is left of the formula can be plausibly labelled 'need' and serve as a 'fair shares' formula for resource allocation for individuals and hence, by aggregation, for practices. This, I suggest, may be no more than econometric pretension.

5. Performance assessment and its empirical basis

The paper suggests that the PBRA team arrived at a favourable assessment of the recommended predictor at practice level, largely on the basis of the value 0.7735 in Table 2[1] of the square of a cross-validated performance assessment statistic R—namely, the common-or-garden correlation coefficient between the predicted total costs of a 10% random sample of 797 practices and the actual (predictand-based) total costs for the individuals registered with these practices. The paper says that considerations of *statistical robustness, face validity, parsimony and freedom from perverse incentives* were involved in deciding which of the many models analysed to recommend, but its conclusion section mentions only the performance of their preferred model—one of *the best models tested*—that it :

'*... performed in line with best international standards in predicting future health costs, predicting upwards of 77% of next year's hospital costs per practice.'*

What the paper does not highlight is the financially significant variation in the formula that would be recommended for any of these nearly equal-performance best models—that would all satisfy some health economist's judgement of their statistical robustness, face validity, parsimony and freedom from perverse incentives. In effect, the PBRA team are presenting the 77% (and the performance features associated with it) as the main case for their approach and are wisely acknowledging the absence of supporting theory when their broad approach is described as empirical:

> *'For over 20 years, the English NHS has adopted an* ***empirical*** *approach to developing formulas to distribute NHS resources between geographical areas, in which a statistical model is developed to predict the costs of care for an area based on the needs of the local population, after adjusting for any variations in supply. We adopted the same principles . . . '.*

Empiricism can take pride of place when its consequences are correctly validated, as they can be seen to have been in the evaluations listed in Table 1[1]. An empirical predictor can be accepted for use when it has unexpected success, in which (a) the phenomenon of misleading over-fitting by a complex *ad hoc* formula can be excluded and (b) predictions are so close to predictand values as to leave little room for any improvement of the model to exhibit appreciably better predictions. When all that happens to come together (a rare occurrence, because most successful formulae in science have a modicum of theoretical inspiration if not support), the fitted model can be said to be a satisfactory approximation to either truth or optimality.

How far does the PBRA formula come to satisfying these criteria, bearing in mind that, in the present context, the formula is to be used as a financial instrument in a financially competitive NHS? The size of the construction sample (in the millions) was probably enough to satisfy criterion (a) despite the very large battery of expressions (models) that the millions of records were being asked select from in order to maximise predictive performance. But whether or not criterion (b) is satisfied must be a matter of subjective judgement. The paper does give the reader a fairly good idea of what an R^2 of the order of 77% means for practices with a list size over 500—in terms of the gaps between cost predictions (either truncated and full) and actual costs in the construction sample. Table 4[1] gives percentages of a random sample of practices for four gap levels defined by the ratio 'distance from full prediction' (DfF)—the undisplayed ratio of the paper:

$$DfF = \frac{Actual\ practice\ cost\ of\ commissioning}{Cost\ prediction\ for\ the\ practice\ from\ the\ full\ formula}$$

Since the denominator is the least-squares fit of the numerator, the 'benchmark' value of DfF is unity. The most significant finding was that, for 32% of the 797 practices in the 10% random sample, DfF was more than 10% away from unity.

By cutting two divisions out of the definition of the paper's 'distance from target allocation' (DfT) and inverting it, we get:

$$DfT = \frac{Actual\ practice\ cost\ of\ commissioning}{Need\ prediction\ for\ the\ practice\ from\ the\ frozen\ formula}$$

This is the definition in reference [5], whose Table 10.5 provides the percentages in the table below and summarized in the paper as:

> '*61 % of practices (were) more than 5% above or below their target, and, of these, 35% were more than 10% and 14% were more than 20% above or below their target.*'

The 'of these' in this is best deleted to avoid ambiguity. GP practices may want to know more about the 14% of their number for which the gap between prediction and target was more than 20%. Would such a gap be considered acceptable by GPs on the wrong side of the benchmark or by their CCG managers (on more impartial grounds)? This table facilitates comparison of the two separately presented performance measures:

Distributions of deviation from benchmark

Measure	0-5%	5-10%	over 10%		over 20%
DfF	38	30	32	100%	?
DfT	38.9	26.5	34.6	100%	14.0

For DfT, neither data source allows a symmetrical table (centred on the unity benchmark) to be constructed. For DfF, the source percentages that are divided are almost perfectly symmetrical. Table 10.9[5] gives the average DfT as 1.8 (80% above benchmark) for the 24 practices with a list size between 500 and 1000, reflecting either high cost or low prediction—and an average of 1.04 (4% above unity) for the 509 between 1000 and 2000. These two list bands make up 6.5% of practices, which leaves open the possibility that for most practices the two measures might be much the same. To decide that question, the paper would have had to give us a paired comparison such as the values of the ratio

$$\frac{DfF}{DfT} = \frac{Need\ prediction\ for\ the\ practice\ from\ the\ frozen\ formula}{Cost\ prediction\ for\ the\ practice\ from\ the\ full\ formula}$$

that cuts out the actual cost and therefore the 'random' component of cost unexplained by prediction. As it is, we are left in the dark about how this potentially informative ratio varies from practice to practice.

6. Structural assumptions of the model and the 'curse of dimensionality'

'The more it changes, the more it's the same thing' may be a Gallic cynicism too far. Nevertheless, most statisticians will see the contribution of econometric practice and theory to this paper (analysed in the next two sections) as an 'abuse of regression'—just as they did in the discussion of the response by Galbraith *et al* [3] to a critical DoH report. The abuse of most consequence lies in bland acceptance of the assumptions of additivity and product structure of the adjustments. The additivity issue was addressed with commendable clarity and transparency on page 11 of the PBRA Team's 2009 report[5]:

'A cruder but more ***practical*** *approach to addressing the "curse of dimensionality" is to assume that risk factors* [predicting variables] ***operate independently*** *and to quantify the* ***independent*** *marginal contribution to the capitation payment* [predicted cost] *of each additional needs factor. That is, an individual's risk factors are simply* ***added up*** *to create a capitation payment, and no interactions between risk factors—or only a small number of interactions—are considered. This has been the basis of many practical systems of person-based capitation formulae developed since the mid-1990s. The additivity assumption makes it* ***easier*** *to use multiple regression methods to quantify the effects of potential risk factors by reducing the number of coefficients to be estimated.'* (my emphases)

Here we can see how the PBRA's 'practical' approach is making a deliberate choice of an interaction-free model (equivalent to the assumption of additivity) and thereby ignoring knowledge accumulated by statisticians over at least eight decades of the hazards of face-value acceptance of jointly estimated coefficients when there are strong correlations between variables. As soon as coefficients are jointly estimated to optimise the predictive performance of a formula that expresses their joint effect, the contributions are no longer 'independent'. The strong correlations between variables mean that prediction is a communal activity and that the adjustment term of a covariate cannot be claimed to express an 'influence' specific to that variable.

Consider for explanatory purposes the fitting of the additive model based on the first two factors in my listing:

Average of 5m costs + Age/sex adj. + PCT adj.

There are 38×152=5776 joint categories of age/sex and PCT. With a construction sample as large as 5 million, there is a good number of individuals in each category—more than enough to give good enough predictions for registration-weighted aggregation over categories for each practice. The Nuffield paper gives no indication that any comparison was made of the performance of this additive model with the 'category model' of the sort that might be used in, for example, Sweden, that would assign a prediction number (such as the category average) for each of the 5776 categories. The key difference is that the additive model makes stronger assumptions: for example, that the difference between the predictions for the young and the old should be the same for every PCT, whereas the Swedish model allows that difference to depend on ('interact with') the PCT.

My uninformed guess is that the R^2-values would differ—and go in favour of the Swedish 'formula' as a better predictor—making a case for reflection before adding the remaining factors and covariates to the additive model. But the new individual-morbidity-based approach needs the diagnoses to be added to show their predictive value, just as they do in model 3 of Table 1[1]:

Average of 5m costs + Age/sex adj. + PCT adj. + Diagnoses adj.

The problem with this superficially modest extension of the 2-factor model, is that the number of categories increases to 906,832. And with only(!) 5 million individuals in the sample used for fitting the formula, many of these categories are almost certainly empty and the corresponding Swedish model would have been unable to make a prediction for any individual who turns up in the future in one of the empty categories. The PBRA team escaped the 'curse of dimensionality' only by recourse to an econometric discipline and practice that purports to 'take into account' an almost unlimited number of 'explanatory' variables.

7. 'Small would be beautiful' for covariates

Without the specialized simplicity of their product (coefficient×deviation) form, there could be no theoretical objection to a covariate adjustment being *added*, if what was being added for a particular category of factors were allowed to be a general (flexibly adaptive) function of the size of the deviation—with a separate function for each category to achieve the optimal adjustment. The specialized product form can be defended against the charge of trial-and-error empiricism only by supposing that:

(a) deviations are *small* enough for the general function for each category to be replaced by what mathematicians call a 'first-order approximation' (making the adjustment *proportional* to the deviation),

(b) the coefficients are the *same* for all categories.

Would either assumption be considered plausible in the PBRA context? And the same considerations apply with even greater force to the simultaneous inclusion of all 9 covariates, since the assumption of additivity requires supposition (a) to be made for the adaptive function of all 9 covariates that would be optimal for prediction of cost.

In this analysis, 'small' has to mean small compared with the range of values in which curvature (non-linearity) of the optimal adaptive functions would inhibit a first-order approximation and violate the proportionality assumption of the adjustment. Seven of the variables are defined for the 'area' in which the individual resides, defined as an ONS census 'lower layer super output' area (with an average population of 1500). Assessing whether or not the variation of these variables is 'small' in the sense just specified has to be a matter of subjective judgement. For one of the area proportion covariates, n1, we have seen the variation is large enough for values near 0% and values near 100% to be well represented in the construction sample. For such variables, curvature might be expressed as a sizeable discrepancy between the value of the adjustment at the 50% value and what the optimal adjustment would be, if we knew it. If the 50% value were close to the median of the 5 million well-spread values, the discrepancy might affect an appreciable proportion of them, and any defence of what has been done needs a theory that would dictate adequate linearity over the corresponding range.

8. Individual level adjustments of all sorts and sizes

There is information in Table 2[1] which the paper does not uncover. It is in the untabulated *t*-values (ratios of coefficient to standard error) for factor f160 and the 9 covariates that range from 3 to 24. They are all significantly different from zero—enough, as the PBRA Team paper[5] explains, to win a place in the formula. The MRI covariate s2, for example, has a t-value of 4.8—the ratio 7781.37/1625.73 of its Table 2[1] coefficient to its standard error. The value 4.8 has a *very* significant P-value, well below the 0.01 level of a Table 2 footnote—in other words, there must be some lawful (not necessarily causal) relationship between s2 and cost. The question here is how much of the *terra incognita* represented by $1 - R^2$ does this 'explain'?—to be specific, how much has it contributed to the R^2-value of 0.1272 (13%) in Table 1[1] ?

No assumptions are needed to answer that question if you use least-squares to fit the formula to 5 million costs. The elementary algebra in the *Discounting t-values* section of reference [3] gives the answer. The final addition of the s2 adjustment covers only 1/227,000th of the terra incognita that remains to be searched for better models when the formula has been fitted using all the other variables. Its contribution to the 0.1272 is a mere 0.000004, which does not even shake the 4^{th} decimal place of R^2! Another way of expressing the smallness of the contribution is to compare the predictions for individuals from the full formula with the predictions you would get if you left s2 out and made a least-squares prediction with the rest of the formula. The paper does not give any individual values for comparison, but the 4.8 tells us, via unquestionable least-squares machinery, that the sum of the squares of the differences between them would be only 3/100,000th of the sum of the squares of the deviations of the full formula from the national average. Admittedly, these contributions would look less pitiful if you did not have to square things in order to interpret the given results (squaring makes 'small' look even smaller). But the overall message would be much the same—namely, what is the scientific case for the inclusion in the model of such a weakly explanatory covariate?

There is no inconsistency, only paradox, in the conjunction of very high statistical significance and very low explanatory power. The paradox comes about because a very weak 'signal' (fortuitously discovered at individual level by an empirical procedure) becomes detectable in the 'noise' when you have 5 million observations containing the same very weak relationship. In its nicely informative 'minus' game, Table 3[1] omitted the age and sex factors from three models including the recommended model 5, but we are not told what omitting all three supply covariates from that model and refitting would give. What Table 3[1] finds, on the other side of the coin as it were, is that when the bundle of f3 and all 9 covariates were *added* to the basic model, cross-validated R^2 increased only slightly from 0.1227 to 0.1229—an increase that represents the combined explanatory power of the whole bundle.

The previous section has given reason for scepticism about whether the full (non-frozen) formula has adequately explored the unknown territory revealed by the calculations of this section. How confident can we be that there is no other imaginative and inspired way of combining variables that would double the 0.1272 and be 'validated' with a cross-validated practice level R^2 impressively larger than 0.77—and deliver a frozen formula with very different practice predictions. What should be recognised by DoH's independent advisors on the CCG formula is that the empirical validation of prediction performance at practice level is not a validation in any scientific sense of the model at individual level, whose empirically-determined shape dictates the practice allocation formula. Any nonsense in the formula at the construction level (individuals) is simply amplified and elevated to practice level.

9. Discrepancies between the formula recommended in November 2011 and the PBRA research[5] of 2008-9 on which it is based

All but the last of the nine coefficients for the recommended formula of Table 2[1] (described in the text as ‘parsimonious model 5’) are exact reproductions of the coefficients column for the ‘new parsimonious model’ in Table 7.3 of the PBRA Report[5]—down to the final digit of the six-significant-figure coefficient 7781.37 for s2. The discrepancy between these two columns is that, by 2011, the last two coefficients of the ‘new parsimonious model’ (145.43 for *practice list size* and 3.17 for *residential home beds accessibility score*) had been replaced by a single coefficient (-0.0000354 for *catchment population of hospital trust that supplied practice with largest number of inpatient admissions (2006-7)*).

At first sight, it may be imagined that two different models have been fitted and that we have come across a reassuring robustness of the coefficients for the nine covariates that were in common. It would be difficult to find a statistician who would find reassurance in such a coincidence. It is beyond any reasonable statistical possibility that the exchange of variables would have left all the common coefficients completely unchanged (to the 2nd decimal place). What is likely is that at least one of the two columns is erroneous—but which? If only one, suspicion will fall on Table 7.3, where Dr Rex Galbraith has drawn my attention to another un-reassuring coincidence—the duplication of the same pair of numbers, (3.17, 3.94), for a needs covariate and a supply covariate! If the Table 7.3 results are thereby thrown in doubt, the question is then whether we can trust the duplication of the financially influential coefficients in Table 2.

There is an additional discrepancy that adds to doubts about both tables. My last section made inferential use of values of the test statistic t calculated as coefficient/standard error in Table 2 but given as the last column of Table 7.3—a ratio that spills out of the least-squares computation of the coefficients. Here are the nine pairs of t-values that are clearly different: (4.8,4.4), (5.8,6.2), (5.1,5.2), (-3.5,-3.1), (-11.1,-10.3), (-24.1,-31.1),(3.8,3.9), (-4.4,-4.7),(4.8,5.3). Why the difference?

10. Elevating the formula to practice level—and higher

Before commenting on the section *Data lags and changes to practice populations*[1] concerning later years (2010-11 and beyond), there are some critical residual questions about the prediction for 2007-08.

For 2007-08:

Unlike the age/sex, diagnosis and private/NHS factors, the PCT factor and covariates in the full (non-frozen) formula have the same value for individuals on the same practice list (with relatively few exceptions). So when their adjustments are aggregated to practice level, the balance of influence between factors and covariates is drastically changed and the covariates speak with a relatively strong united voice, as Table 1[1] shows. The factor adjustments for age/sex and diagnoses have a relatively weaker (though still strong) voice at practice level due to the partial cancellation of positive and negative individual deviations, whereas for the covariates the practice level adjustments encapsulate all the variation at individual level and their $0.1229 - 0.1223$ contribution to cross-validated R^2 at individual level is inflated to $0.7735 - 0.6084$ at practice level. It is therefore not surprising that the covariates are no longer pitifully weak contributors to the performance of the formula at individual level but begin to play a significant role in the full formula and therefore in its recommended truncation.

On April 1 2007, there were 8238 practices with a list-size over 500. That is more than the 6781 'middle layer super output areas' and nearly as many as the 8414 electoral wards that the AREA (DoH's RARP 26) and CARAN (DoH's RARP 30) teams, respectively, used as units of analysis to constuct basic formulas for 10 years' allocations. The 8238 would be enough to do some (not here recommended!) formula selection at practice level. Did the PBRA team care to refit their selected model to the aggregated explanatory variables at practice level? If they did, why are the results not reported? If not, why not? The aggregation would have attenuated the probably very large influence of outlier variables and costs of the sort that gave DfT averages of 1.8 for the smallest practices and a probably statistically significant 1.04 for the 509 next-smallest practices—and the re-estimated coefficients might show financially significant differences for practices from the coefficients of the recommended formula.

When the research[5] was done on which this BMJ paper is based, there were no CCGs. As collections of practices, CCGs constitute an even higher level of aggregation of the formula funding for individuals—equivalent, of course, to aggregation over the practices that constitute them. It would be pertinent to know what the frozen and non-frozen formulas look like when the aggregation is carried to that higher level, and what their performances would have been in pseudo-prediction of the 2007-08 costs. There would now be a CCG-level cross-validated R^2 and analogues of the practice-level DfF and DfT measures. My guess is that the new R^2 would be found to exceed 0.77 by a good margin—which would be good news for the empiricist defence of the formula. However, because CCGs appear to be substantially coterminous (for practices) with a former PCT, the PCT adjustment might even dominate the age/sex adjustment—and, when the PCT adjustment is zeroed by the freezing, the full formula's DfF and the truncated formula's DfT would probably differ much more than at practice-level. The question would then arise as to whether the supposedly 'fair shares' formula for practices can be seriously defended at CCG level—and the need to justify the differential effects of freezing for CCGs would come centre-stage!

For 2010-11 and beyond:

The same individuals were used for both the predictions and the costs underlying the percentages reported in Table 4[1]—namely, a 10% sample of the 50 million on GP lists on April 1 2007. The calculations were all made within the carefully assembled data-base. But when the PBRA team had to tackle the problem of predicting the practice costs for 2010-11, the calculation had to be for lists with many new individuals, for whom there were need and supply variables that could not be feasibly ascertained—and variables that would in any case, with the lapse of three years, have fallen outside their sell-by-date for those individuals who were still on the same list as in 2007. The researchers felt no need to get new data for a fresh least-squares estimation of weights for factors and of coefficients for covariates (the paper used the term 'weight' for the product of coefficient and covariate deviation). What they needed was a pragmatic solution to the problem of the missing variables for new individuals or out-of-date variables for the others.

If I have understood the careful explanation in the sub-section *Change to the practice registered population*, this is how the solution might have been found for prediction of a practice cost for 2010-11 based on the recommended model 5. The database of explanatory variables was extended to include 2007-8. For each practice and for each of the years 2005-6, 2006-7, 2007-8, the individuals in the practice list on April 1 of the year were divided into the 38 age/sex bands—and, for each band, the average was calculated of the frozen formula for the individuals in the band. These averages were found to vary 'hardly at all' and were combined (in some way) into a single average for the age/sex band. The prediction for 2010-11 would then be the sum of the products of these averages with the corresponding numbers of individuals in the April 1 2010 list.

11. What's in the freezer?

To give freezing a fair wind, this section will for a while suspend any disbelief in the model: the devils of disbelief are in the boring detail (the assumptions of additivity and product form of adjustments). As already noted, the strong correlations between variables mean that prediction is a communal activity and that the adjustment term of a covariate cannot be regarded as a ring-fenced contribution, specific to that variable and insulated from the swirl of connected activities that is a least-squares fitting. This observation is enough to undermine any claim that a level playing field is being created by cutting out the supply variables. It is a claim that cannot have empirical defence because there is no independent, value-free measure of the unevenness of the field (of justifiable costs). So there has to be something 'true' about the 'freezing' of supply variables, in the sense that the truncated formula has to be a better predictor of the justifiable cost of true need for health care (once any true, i.e. causal not just correlational, influence of supply variables has been taken out of it).

We are given the value of R^2 for the full formula but, because we do not know yet what the justifiable cost is, we have no measure of how well the frozen formula predicts the justifiable cost. All we have is our faith in econometrics-based assurances that its coefficients have not been frost-bitten, so to speak, by the joint fitting of all the coefficients in the full formula. Evidence for such assurances is not provided in this paper. If the PBRA formula is implemented, how will DoH meet the objections of a CCG maintaining that some coefficient does not match their experience of how the need for health care depends on the variation of the corresponding variable between its practices?

I cannot believe that the PBRA Team would have kept the freezing game going unless it thought it was changing the level of the playing field. There must have been confidence that the supply variables have a true-in-some-sense influence on cost and one that is worth ‘freezing’. I hope that, when DoH publishes the Exposition Books that will show, almost to the last penny, how many millions each practice will get for 2013-14, they will publish the formula calculation for a random sample of individuals showing the joint and individual contribution (preferably as adjustments!) of the supply factor and supply covariates. There is a paradox in the freezer! It is how can the size of those adjustments (if they are appreciable) be reconciled with the above findings of the analysis of both t and R^2 values that, in the comparison of models at individual level with and without the supply covariates, their influence appears to be almost non-existent?

To resolve the paradox, it is necessary to go once more into the assumption-free theory of the least-squares fitting machinery. The paper does not mention (and reference [5] does not dwell on) the thorny question of high, even very high, correlations between explanatory variables. The long-standing view has been that this has been a problem facing researchers at the formula construction level when selection is needed to get a (superficially) parsimonious model. Once the fitted full formula is in hand, it seems that correlations do not feature when its frozen offspring is recommended for implementation by DoH. To resolve the paradox, we have to consider how correlations affect the least-squares fitting of an additive expression when some variables in the expression are highly correlated, as we may suppose them to be in the full (non-frozen) formula. There appears to be no specific evidence for this supposition in either of the two PBRA papers[1,5]. But we can invoke the intimately related AREA and CARAN studies (DoH reports RARP 26 & 30 respectively) which are concerned about high correlations and openly discuss the rejection of plausibly explanatory variables because they happen to be highly correlated with other more plausible variables.

The least-squares machinery dictates that, when one of two highly correlated explanatory variables is omitted, the remaining one can conceal the omission by taking over and substantially reproducing the same prediction on its own (in the scenario when the machinery does not bring other variables in to fill the gap). This is what is probably happening when, for example, the MRI covariate s2 is omitted and the predictions do not perceptibly change (as shown in the *Individual adjustments of all sorts and sizes* section). Here, I am assuming that the Exposition Books will show at least a financially significant adjustment for s2 in the full formula, at the very least more significant than a last penny, when aggregated to practice level prior to freezing (if I am wrong about that for all three supply covariates, the case can move to then certainly dominant supply factor PCT for which the following discussion would have to be reworded). The hypothesized second variable (that comes to fill the gap left by s2) need not be one of the other variables in the full formula—the least-squares machinery is so powerful that it will automatically find the additive combination of those variables that does the job. The substitution phenomenon is quite symmetrical—s2 is sharing the prediction load with the second variable (or combination of variables) whatever the strength of its causal influence on cost.

That is the explanation that resolves the paradox, but it is a faith in the model (as a true measure of need) that allows some to rely on another bit of least squares theory to dictate the inclusion of s2. For s2, believers in the model can say that without s2 the model is (technically) *biased* but that when it is included you have a true measure of the expectation of cost, giving the actual cost when purely random error is added. With faith, you can be reassured that there is a 95% confidence interval, 7781.37 plus-or-minus 1.96× 1625.73, i.e. 4,600 to 11,000, for the true value of the coefficient of s2, and that the freezing of the corresponding adjustment is the central estimate of what has to be frozen to level the cost playing field. Such faith ignores the *ad hoc* and theoretically-unsupported way in which the individuals' formula was eventually arrived at.

12. Additional difficulties of the PBRA approach

The paper expresses two difficulties with commendable clarity:

'*Data lags. In developing the model, the data used on individual level needs were based on data at least 18 months old because of the lags in the production of cleaned hospital episode statistics at national level .... Notwithstanding improved data collection efforts, such lags will* ***always*** *be a feature of resource allocation methods.'* (my emphasis)

'*Changes to the practice registered population. Perhaps more significant was that during the lag period there might be significant changes to a practice's registered population ....*'

As we saw in the last-but-one section, the study developed a '*pragmatic solution*' to these challenges, which was to calculate the *per capita* prediction in each age-sex group of each practice (where '*the needs and supply variables in the formula hardly varied at all*' over the construction period)—before using the age-sex breakdown of the most recent registration numbers in the aggregate prediction of practice cost. An ingenious solution—but one that does not overcome a third difficulty, namely, the questionable '*accuracy of population size registered with general practices*'. (The phenomenon of list inflation is known to vary appreciably from area to area.)

Another unresolved difficulty concerns the exclusion of expenditure on mental health from the scope of the PBRA formula—a legacy of the separate treatment of mental health and acute services in the PCT-funding formula, for which there was seemingly good reason:

> '*the data available to construct a formula are different and the factors influencing need for mental healthcare were thought to be sufficiently different from those for more general acute care.*'[1]

The paper does not explain why, with mental health excluded from 'spend' i.e. from the predictand, it was nevertheless a contributor to predictions as one of the 152 ICD10 categories—according to page 42 of the PBRA Team[5] report. Previous formula builders obliged to include mental health swallowed hard and incorporated the dubious mental health formula more as an *ad hoc* filler of an administrative gap than as a trustworthy financial instrument. The present rapid growth of dementia numbers means that this difficulty is one that should no longer be ignored, but it may call more for lateral thinking than clever exercise of pragmatism.

13. Back to the drawing board

My written and oral evidence to the House of Commons Health Committee 2006 inquiry[2] into NHS deficits had been analytically critical of the then resource allocation formula. With the strong support of two other witnesses, it may have influenced the committee to recommend that

> '*consideration be given to basing the formula on actual need rather than proxies of need*'

a recommendation that might just have prompted the Nuffield Trust to investigate direct measurement. The oral evidence had gone out on a limb by suggesting, rather incoherently,

> '*a sampling of GP-registered patients in a pilot study (using) a small fraction of the money . . . wasted by this formula . . . to investigate, using trained nurses, sampling patients—some of whom will not have had any costs on the Health Service* ***in the previous year*** *. . . A trained nurse would be able (to estimate) as near as one could . . . the real health need of a patient. . . . But that would bring up the question of value judgements.*' (my emphasis)

The unspoken idea was that, instead of prediction with a nonsense formula (unlikely to be a good measure of justifiable health care need) we should simply extrapolate from a continually updated direct assessment of the already justified (historical!) cost of health care to the next year's *per capita* justifiable need of each PCT. If the idea could be developed, piloted and judged feasible, on the basis of data for stratified random samples of individuals on practice-lists, it could be implemented England-wide with enough precision to fix the allocation to individual CCGs. And if government had the nerve, GP-practices could even be left like ferrets in their CCG sack (an average of 40 or so) to divide the allocation, perhaps with the aid of a 'fair shares' formula of their own choosing.

The reluctance of health economists to put historical costs into a formula is an understandable response to the evidence that such costs have often been subject to political manipulation and influences unrelated to healthcare needs. But it is surely time that DoH conceded that the decades-long program to construct a capitation formula alternative has failed. The PBRA Team report[5] in which the 2007-08 cost data were used to construct the predictand is dated 14 October 2009. Perhaps the rejection of such 'historical' costs should be reconsidered, even if they are biased in some CCGs. The breadth and depth of the PBRA analysis suggests that the 2007-08 data may have been ready for use some time in 2009. With better organization, this particular data-lag might be shortened—enough to provide a CCG-level covariate for refining the England-wide sample estimates (use of a lawfully biased covariate to improve an honest estimate is a well-understood statistical tool).

There is an analogy with farm economics that should encourage lateral thinking. A one-year-ahead extrapolation of justifid cost would carry less risk than the loan that a farmer gets from an obliging bank to see him/her through the year, where the loan is based on the farmer's accounts for previous years. If 'private-sector' farmers can do it with a high level of risk, why not 'private-sector' GPs with a much lower one? Almost at a stroke, the above four difficulties in the prediction-by-formula approach would be banished. The long and inescapable delays in the acquisition of data needed to update or implement the formula would be replaced by a lag of at most one year in getting the accountant-certified 'historical' costs. The difficulty of population changes would largely vanish because the 'population' of GP-registered individuals making any significant contribution to costs is automatically built into the construction of the predictand, while the 'population' of GP-registered individuals who are non-users of the NHS in the collection year may have a small contribution that could be estimated with adequate precision. The mental health difficulty would vanish, assuming that historical costing for usage of mental health services could be made as easy as it is for acute care.

14. Bloggers

The 22 November 2011 paper[1] provoked some lively responses on the BMJ website:

25 November 2011: Healthcare consultant R. P. Jones is pleased
'to see that the simple models are able to explain as much of the cost variation as the most complex versions'
but worried that the effects of weather, air quality and infectious outbreaks
'*may negate the fundamental hypothesis that health care costs are largely due to person-based (socio-economic as well as demographic) factors*'.
Is it possible that a
'*large proportion of the 23% of cost variation not explained by the various person-based models is largely due to the unrecognised role of (changes in) the environment upon cost*?'
Mr Jones suspects that
'*although GPs may be more successful than PCTs at reducing costs the issue of financial risk (and hence the postcode lottery) has not and will not go away...... any more than fluctuations in the weather, etc will go away.*'

27 November 2011: Sittingbourne GP Dr H. J. Beerstecher thinks that the
'*predictive power of the presented models for health expenditure at practice level are impressive*'
and that
'*the approach and selection of variables seems logical and sensible*'
but that it seems illogical to
'*exclude historic healthcare cost, a powerful predictor of individual future healthcare cost, on the basis that its inclusion could either form a perverse incentive for overdiagnosis or supplier induced demand or reflect supply factors instead of need factors . . . past healthcare cost has not formed part of an allocation formula and therefore precedes supplier induced demand and/or overdiagnosis. It would seem more sensible to monitor this after its introduction rather than to dismiss it out of hand.*'
There is selection of and by patients which

‘*determines the health need at practice level. ... Until all patient variables have been accounted for it is premature to discount any predictors of healthcare cost like past expenditure. In the wrong hands, inaccurate budget setting tools will lead to practices being blamed for the cost of healthcare of their registered population. The way forward seems to require a more accurate model to predict the healthcare cost of individual patients which could be used to more accurately predict expenditure at practice level.*’

1 December 2011: With three others, GP-trainee Daniel Peake has
‘*concerns that the Dixon et al use of the national tariff or national reference cost in their equation, which are unproven estimates rather than actual costs*’.
Some University Hospital of North Staffordshire cost data shows a substantial disparity between the two and Dixon et al have
‘*overlooked a major error ... which if not corrected will render their otherwise diligent formula meaningless.*’

14 December 2011: Three of the authors respond to the GP and GP-trainee with commendable but revealing conciseness:
‘*Beerstrecher asks why the historic healthcare costs of individuals were not included in the person-based formulae we developed for allocating resources to general practices for commissioning. Their inclusion would indeed have led to significant improvements in predictive power. However, their use was precluded because of the wish of the purchaser not to create a perverse incentive to increase expenditure in one year order to attract higher funding in future years.*
Peake raises a concern about the accuracy of the national tariff in reflecting the costs actually incurred by hospitals for the care of individuals. The use of actual costs would indeed have given a more accurate picture of cost drivers. However, the purpose of the formula was to model the expenditure needs of GP commissioners and not necessarily the actual costs borne by providers. Therefore the use of the tariff was the correct approach in this study.
So these issues are not shortcomings of our analysis—they reflect the use to which the formula will be put.’

6 January 2012: Dr Beerstecher is censorious about the purchaser-respecting reason he has now been given for the exclusion of historical costs, claiming that
‘*no evidence is presented and no literature is cited that would support excluding powerful predictors from the models. It seems this was based on presumption and not on scientific argument or evidence.*’
But he appears to misinterpret the paper when he finds it shameful that
‘*the authors do not report how much past expenditure improves the predictive power of 12% for individual patients in their models. Formulae that are not based on individual health needs are likely to be too inaccurate for budget setting at practice level if the aim is that practice budgets should move to the target allocation from historical expenditure?*’
The model 2 formula of Table 1[1] achieves a creditable 12% for individuals only because it has the two-years-old diagnosis-score (which could have been expressed as a ‘historical cost’) as one of its two factors.

Footnote

Neither the BMJ paper nor the extensive PBRA Team report[5] on which it is based have made use of the computerized graphical technique[7,8] (used in references [2] and [6]) that displays the sequentially-calculated contributions of factors and covariates to the relative *per capita* allocations to PCTs. The computer software could be redesigned by DoH specialists to show in the present context the sequence-dependent relative contribution of each of the 14 adjustments in the full (non-frozen) formula, including the new supply factor CCG and the three supply covariates. If the supply variables were put at the end of the sequence, the 'frozen' section of the graph would end with the 'target index'[2] that might by then be determining the funding of each of England's more than 8000 practices. For most of us, the obscurity of the formula would be lifted by this insightful graphic.

References

1. Dixon, J., Smith, P., Gravelle, H., Martin, S., Bardsley, M., Rice, N., Georgiou, T., Dusheiko, M., Billings, J., De Lorenzo, M., and Sanderson, C. (2011) A person based formula for allocating commissioning funds to general practices in England: development of a statistical model, *British Medical Journal*, vol. 343, 22 November (freely downloadable as www.bmj.com/content/343 /bmj.d6608).

1a. Bardsley, M. and Dixon, J. (2011) *Person-based Resource Allocation: New approaches to estimating commissioning budgets for GP practices.* London: The Nuffield Trust.

2. Stone, M. (2010) *Failing to Figure: Whitehall's costly neglect of statistical reasoning.* London: Civitas Institute for the Study of Civil Society.

3. Galbraith, Jane and Stone, M. (2011) The abuse of regression in the National Health Service allocation formulae: response to the Department of Health's 2007 'resource allocation research paper' (with discussion), *J. R. Statist. Soc* A, **174**, 517-528.

4. Department of Health (2005) *Resource Allocation: Weighted Capitation Formula.* London: Department of Health.

5. PBRA Team (2009) Developing a person based resource allocation formula for allocations to general practices in England. London: The Nuffield Trust.

6. House of Commons Health Committee (2006) *NHS Deficits, First Report of the Session 2006-07, Vol.II.* London: The Stationery Office Ltd.

7. Stone, M. (2006) Fathoming the PCT funding formula with Excel graphics. Research Report 267, www.ucl.ac.uk/stats/research

8. Stone, M. (2006) Corrections to Eye of Newt and to PCTgrapher.xls. London: Civitas Institute for the Study of Civil Society.